HANDBOOK
of
NUMERICAL ANALYSIS

P. G. CIARLET • Editor

Volume
XIV

Special Volume
Computational Methods for the Atmosphere and the Oceans

R. TEMAM
J. TRIBBIA
Guest Editors

NORTH-HOLLAND

Handbook of Numerical Analysis

General Editor:

P.G. Ciarlet

Laboratoire Jacques-Louis Lions
Université Pierre et Marie Curie
4 Place Jussieu
75005 PARIS, France

and

Department of Mathematics
City University of Hong Kong
Tat Chee Avenue
KOWLOON, Hong Kong

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

North-Holland is an imprint of Elsevier

Volume XIV

Special Volume: Computational Methods for the Atmosphere and the Oceans

Guest Editors:

Roger M. Temam

Indiana University,
Institute for Scientific Computing & Applied Mathematics,
831 E. 3rd Street, Rawles Hall,
Bloomington, IN 47405, USA

Joseph J. Tribbia

National Center for Atmospheric Research,
1850 Table Mesa Drive,
Boulder, CO 80305, USA

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO
North-Holland is an imprint of Elsevier

North-Holland is an imprint of Elsevier
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK
Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands

Copyright © 2009 Elsevier B.V. All rights reserved

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means electronic, mechanical, photocopying, recording or otherwise without the prior written permission of the publisher

Permissions may be sought directly from Elsevier's Science & Technology Rights Department in Oxford, UK: phone (+44) (0) 1865 843830; fax (+44) (0) 1865 853333; email: permissions@elsevier.com. Alternatively you can submit your request online by visiting the Elsevier web site at http://elsevier.com/locate/permissions, and selecting *Obtaining permission to use Elsevier material*

Notice
No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein. Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made.

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

ISBN: 978-0-444-51893-4
ISSN: 1570-8659

For information on all North-Holland publications
visit our website at elsevierdirect.com

Printed and bound in Hungary

09 10 11 12 13 10 9 8 7 6 5 4 3 2 1

Working together to grow
libraries in developing countries
www.elsevier.com | www.bookaid.org | www.sabre.org
ELSEVIER BOOK AID International Sabre Foundation

General Preface

In the early eighties, when Jacques-Louis Lions and I considered the idea of a *Handbook of Numerical Analysis*, we carefully laid out specific objectives, outlined in the following excerpts from the "General Preface" which has appeared at the beginning of each of the volumes published so far:

> During the past decades, giant needs for ever more sophisticated mathematical models and increasingly complex and extensive computer simulations have arisen. In this fashion, two indissociable activities, *mathematical modeling* and *computer simulation*, have gained a major status in all aspects of science, technology and industry.
>
> In order that these two sciences be established on the safest possible grounds, mathematical rigor is indispensable. For this reason, two companion sciences, *Numerical Analysis* and *Scientific Software*, have emerged as essential steps for validating the mathematical models and the computer simulations that are based on them.
>
> *Numerical Analysis* is here understood as the part of *Mathematics* that describes and analyzes all the numerical schemes that are used on computers; its objective consists in obtaining a clear, precise, and faithful, representation of all the "information" contained in a mathematical model; as such, it is the natural extension of more classical tools, such as analytic solutions, special transforms, functional analysis, as well as stability and asymptotic analysis.
>
> The various volumes comprising the *Handbook of Numerical Analysis* will thoroughly cover all the major aspects of Numerical Analysis, by presenting accessible and in-depth surveys, which include the most recent trends.
>
> More precisely, the Handbook will cover the *basic methods of Numerical Analysis*, gathered under the following general headings:
>
> – Solution of Equations in $\mathbb{R}^n$,
> – Finite Difference Methods,
> – Finite Element Methods,
> – Techniques of Scientific Computing.

It will also cover the *numerical solution of actual problems of contemporary interest in Applied Mathematics*, gathered under the following general headings:

- Numerical Methods for Fluids,
- Numerical Methods for Solids.

In retrospect, it can be safely asserted that Volumes I to IX, which were edited by both of us, fulfilled most of these objectives, thanks to the eminence of the authors and the quality of their contributions.

After Jacques-Louis Lions' tragic loss in 2001, it became clear that Volume IX would be the last one of the type published so far, i.e., edited by both of us and devoted to some of the general headings defined above. It was then decided, in consultation with the publisher, that each future volume will instead be devoted to a single "*specific application*" and called for this reason a "*Special Volume*". "*Specific applications*" will include Mathematical Finance, Meteorology, Celestial Mechanics, Computational Chemistry, Living Systems, Electromagnetism, Computational Mathematics etc. It is worth noting that the inclusion of such "specific applications" in the *Handbook of Numerical Analysis* was part of our initial project.

To ensure the continuity of this enterprise, I will continue to act as Editor of each Special Volume, whose conception will be jointly coordinated and supervised by a Guest Editor.

P.G. CIARLET
July 2002

Contents of the Handbook

VOLUME I

FINITE DIFFERENCE METHODS (PART 1)

Introduction, *G.I. Marchuk* 3
Finite Difference Methods for Linear Parabolic Equations, *V. Thomée* 5
Splitting and Alternating Direction Methods, *G.I. Marchuk* 197

SOLUTION OF EQUATIONS IN $\mathbb{R}^n$ (PART 1)

Least Squares Methods, *Å. Björck* 465

VOLUME II

FINITE ELEMENT METHODS (PART 1)

Finite Elements: An Introduction, *J.T. Oden* 3
Basic Error Estimates for Elliptic Problems, *P.G. Ciarlet* 17
Local Behavior in Finite Element Methods, *L.B. Wahlbin* 353
Mixed and Hybrid Methods, *J.E. Roberts, J.-M. Thomas* 523
Eigenvalue Problems, *I. Babuška, J. Osborn* 641
Evolution Problems, *H. Fujita, T. Suzuki* 789

VOLUME III

TECHNIQUES OF SCIENTIFIC COMPUTING (PART 1)

Historical Perspective on Interpolation, Approximation and Quadrature, *C. Brezinski* 3
Padé Approximations, *C. Brezinski, J. van Iseghem* 47
Approximation and Interpolation Theory, *Bl. Sendov, A. Andreev* 223

NUMERICAL METHODS FOR SOLIDS (PART 1)

Numerical Methods for Nonlinear Three-Dimensional Elasticity, *P. Le Tallec* 465

Solution of Equations in $\mathbb{R}^n$ (Part 2)

Numerical Solution of Polynomial Equations, *Bl. Sendov, A. Andreev, N. Kjurkchiev* 625

Volume IV

Finite Element Methods (Part 2)

Origins, Milestones and Directions of the Finite Element Method – A Personal View, *O.C. Zienkiewicz* 3
Automatic Mesh Generation and Finite Element Computation, *P.L. George* 69

Numerical Methods for Solids (Part 2)

Limit Analysis of Collapse States, *E. Christiansen* 193
Numerical Methods for Unilateral Problems in Solid Mechanics, *J. Haslinger, I. Hlaváček, J. Nečas* 313
Mathematical Modeling of Rods, *L. Trabucho, J.M. Viaño* 487

Volume V

Techniques of Scientific Computing (Part 2)

Numerical Path Following, *E.L. Allgower, K. Georg* 3
Spectral Methods, *C. Bernardi, Y. Maday* 209
Numerical Analysis for Nonlinear and Bifurcation Problems, *G. Caloz, J. Rappaz* 487
Wavelets and Fast Numerical Algorithms, *Y. Meyer* 639
Computer Aided Geometric Design, *J.-J. Risler* 715

Volume VI

Numerical Methods for Solids (Part 3)

Iterative Finite Element Solutions in Nonlinear Solid Mechanics, *R.M. Ferencz, T.J.R. Hughes* 3
Numerical Analysis and Simulation of Plasticity, *J.C. Simo* 183

Numerical Methods for Fluids (Part 1)

Navier–Stokes Equations: Theory and Approximation, *M. Marion, R. Temam* 503

VOLUME VII

SOLUTION OF EQUATIONS IN $\mathbb{R}^n$ (PART 3)

Gaussian Elimination for the Solution of Linear Systems of Equations, *G. Meurant* 3

TECHNIQUES OF SCIENTIFIC COMPUTING (PART 3)

The Analysis of Multigrid Methods, *J.H. Bramble, X. Zhang* 173
Wavelet Methods in Numerical Analysis, *A. Cohen* 417
Finite Volume Methods, *R. Eymard, T. Gallouët, R. Herbin* 713

VOLUME VIII

SOLUTION OF EQUATIONS IN $\mathbb{R}^n$ (PART 4)

Computational Methods for Large Eigenvalue Problems, *H.A. van der Vorst* 3

TECHNIQUES OF SCIENTIFIC COMPUTING (PART 4)

Theoretical and Numerical Analysis of Differential–Algebraic Equations, *P.J. Rabier, W.C. Rheinboldt* 183

NUMERICAL METHODS FOR FLUIDS (PART 2)

Mathematical Modeling and Analysis of Viscoelastic Fluids of the Oldroyd Kind, *E. Fernández-Cara, F. Guillén, R.R. Ortega* 543

VOLUME IX

NUMERICAL METHODS FOR FLUIDS (PART 3)

Finite Element Methods for Incompressible Viscous Flow, *R. Glowinski* 3

VOLUME X

SPECIAL VOLUME: COMPUTATIONAL CHEMISTRY

Computational Quantum Chemistry: A Primer, *E. Cancès, M. Defranceschi, W. Kutzelnigg, C. Le Bris, Y. Maday* 3
The Modeling and Simulation of the Liquid Phase, *J. Tomasi, B. Mennucci, P. Laug* 271
An Introduction to First-Principles Simulations of Extended Systems, *F. Finocchi, J. Goniakowski, X. Gonze, C. Pisani* 377

Computational Approaches of Relativistic Models in Quantum Chemistry, *J.P. Desclaux, J. Dolbeault, M.J. Esteban, P. Indelicato, E. Séré* 453
Quantum Monte Carlo Methods for the Solution of the Schrödinger Equation for Molecular Systems, *A. Aspuru-Guzik, W.A. Lester, Jr.* 485
Linear Scaling Methods for the Solution of Schrödinger's Equation, *S. Goedecker* 537
Finite Difference Methods for Ab Initio Electronic Structure and Quantum Transport Calculations of Nanostructures, *J.-L. Fattebert, M. Buongiorno Nardelli* 571
Using Real Space Pseudopotentials for the Electronic Structure Problem, *J.R. Chelikowsky, L. Kronik, I. Vasiliev, M. Jain, Y. Saad* 613
Scalable Multiresolution Algorithms for Classical and Quantum Molecular Dynamics Simulations of Nanosystems, *A. Nakano, T.J. Campbell, R.K. Kalia, S. Kodiyalam, S. Ogata, F. Shimojo, X. Su, P. Vashishta* 639
Simulating Chemical Reactions in Complex Systems, *M.J. Field* 667
Biomolecular Conformations Can Be Identified as Metastable Sets of Molecular Dynamics, *C. Schütte, W. Huisinga* 699
Theory of Intense Laser-Induced Molecular Dissociation: From Simulation to Control, *O. Atabek, R. Lefebvre, T.T. Nguyen-Dang* 745
Numerical Methods for Molecular Time-Dependent Schrödinger Equations – Bridging the Perturbative to Nonperturbative Regime, *A.D. Bandrauk, H.-Z. Lu* 803
Control of Quantum Dynamics: Concepts, Procedures and Future Prospects, *H. Rabitz, G. Turinici, E. Brown* 833

VOLUME XI

SPECIAL VOLUME: FOUNDATIONS OF COMPUTATIONAL MATHEMATICS

On the Foundations of Computational Mathematics, *B.J.C. Baxter, A. Iserles* 3
Geometric Integration and its Applications, *C.J. Budd, M.D. Piggott* 35
Linear Programming and Condition Numbers under the Real Number Computation Model, *D. Cheung, F. Cucker, Y. Ye* 141
Numerical Solution of Polynomial Systems by Homotopy Continuation Methods, *T.Y. Li* 209
Chaos in Finite Difference Scheme, *M. Yamaguti, Y. Maeda* 305
Introduction to Partial Differential Equations and Variational Formulations in Image Processing, *G. Sapiro* 383

VOLUME XII

SPECIAL VOLUME: COMPUTATIONAL MODELS FOR THE HUMAN BODY

Mathematical Modeling and Numerical Simulation of the Cardiovascular System, *A. Quarteroni, L. Formaggia* 3
Computational Methods for Cardiac Electrophysiology, *M.E. Belik, T.P. Usyk, A.D. McCulloch* 129
Mathematical Analysis, Controllability and Numerical Simulation of a Simple Model of Avascular Tumor Growth, *J.I. Díaz, J.I. Tello* 189
Human Models for Crash and Impact Simulation, *E. Haug, H.-Y. Choi, S. Robin, M. Beaugonin* 231
Soft Tissue Modeling for Surgery Simulation, *H. Delingette, N. Ayache* 453
Recovering Displacements and Deformations from 3D Medical Images Using Biomechanical Models, *X. Papademetris, O. Škrinjar, J.S. Duncan* 551
Methods for Modeling and Predicting Mechanical Deformations of the Breast under External Perturbations, *F.S. Azar, D.N. Metaxas, M.D. Schnall* 591

VOLUME XIII

SPECIAL VOLUME: NUMERICAL METHODS IN ELECTROMAGNETICS

Introduction to Electromagnetism, *W. Magnus, W. Schoenmaker* 3
Discretization of Electromagnetic Problems: The "Generalized Finite Differences" Approach, *A. Bossavit* 105
Finite-Difference Time-Domain Methods, *S.C. Hagness, A. Taflove, S.D. Gedney* 199
Discretization of Semiconductor Device Problems (I), *F. Brezzi, L.D. Marini, S. Micheletti, P. Pietra, R. Sacco, S. Wang* 317
Discretization of Semiconductor Device Problems (II), *A.M. Anile, N. Nikiforakis, V. Romano, G. Russo* 443
Modeling and Discretization of Circuit Problems, *M. Günther, U. Feldmann, J. ter Maten* 523
Simulation of EMC Behaviour, *A.J.H. Wachters, W.H.A. Schilders* 661
Solution of Linear Systems, *O. Schenk, H.A. van der Vorst* 755
Reduced-Order Modeling, *Z. Bai, P.M. Dewilde, R.W. Freund* 825

VOLUME XIV

SPECIAL VOLUME: COMPUTATIONAL METHODS FOR THE ATMOSPHERE AND THE OCEANS

Finite-Volume Methods in Meteorology, *Bennert Machenhauer, Eigil Kaas, Peter Hjort Lauritzen* 3
Computational Kernel Algorithms for Fine-Scale, Multiprocess, Longtime Oceanic Simulations, *Alexander F. Shchepetkin, James C. McWilliams* 121
Bifurcation Analysis of Ocean, Atmosphere, and Climate Models, *Eric Simonnet, Henk A. Dijkstra, Michael Ghil* 187
Time-Periodic Flows in Geophysical and Classical Fluid Dynamics, *R. M. Samelson* 231
Momentum Maps for Lattice EPDiff, *Colin J. Cotter, Darryl D. Holm* 247
Numerical Generation of Stochastic Differential Equations in Climate Models, *Brian Ewald, Cécile Penland* 279
Large-eddy Simulations for Geophysical Fluid Dynamics, *Marcel Lesieur, Olivier Metais* 309
Two Examples from Geophysical and Astrophysical Turbulence on Modeling Disparate Scale Interactions, *Pablo Mininni, Annick Pouquet, Peter Sullivan* 339
Data Assimilation for Geophysical Fluids, *Jacques Blum, François-Xavier Le Dimet, I. Michael Navon* 385
Energetic Consistency and Coupling of the Mean and Covariance Dynamics, *Stephen E. Cohn* 443
Boundary Value Problems for the Inviscid Primitive Equations in Limited Domains, *Antoine Rousseau, Roger M. Temam, Joseph J. Tribbia* 481
Some Mathematical Problems in Geophysical Fluid Dynamics, *Madalina Petcu, Roger M. Temam, Mohammed Ziane* 577

Foreword

The development of the modern form of atmospheric sciences is generally traced back to John von Neumann and the meteorological school that he founded in the late 1950s and early 1960s. With such a prestigious background, there are well-established traditions of scientific interactions between specialists of atmospheric and oceans sciences and applied and computational mathematicians. The task is not easy, however, as each of the fields has considerably developed independently: specific problems, specific tools, specific methodologies, and often different languages. Nevertheless, interactions are indispensable. The demands for predictions are numerous, and the interests at task are considerable, from weather predictions for agriculture and many other economical motivations, or evolution of pollution, to large-scale problems like global warming. Phenomena studied by atmosphere and oceans scientists include phenomena for which the physics is not fully known and which need to be modeled or parameterized; such problems do not fall in the objectives of this volume. Other phenomena are well understood and governed by well-known equations. The difficulties then lay in the large number of unknowns to be considered and the large number of equations to be solved. Thus, the size of the problems is and will, for a long time, remain at the limit of available computing power despite the increase of computing power and memory capacities of the computers that we have seen and that we can expect.

Computational mathematicians, applied and numerical analysts, have a different approach. They tend to develop tools that are appropriate for the computational solution of various sorts of problems and equations. It is, therefore, natural to confront these new tools to the formidable problems of atmospheric and oceanic sciences. Mathematicians can also check the well posedness of certain problems and contribute to validate or set aside certain models and certain approaches. So the general purpose of this volume is to bring useful and important geophysical problems to the attention of mathematicians, and to present useful tools developed by mathematicians.

A single volume addressing so important and diverse problems cannot be exhaustive of course, but the articles in this volume address central problems in the area and provide a survey of the frontier of research in these areas.

Two articles in this volume pertain to modeling: the article by B. Machenhauer, E. Kaas, and P. Lauritzen addresses the finite-volume techniques for the atmosphere, and the article by J. McWilliams and A. Shchepetkin is devoted to computational issues in ocean modeling. A series of four articles is devoted to nonlinear methods; the article by

H. A. Dijkstra, M. Ghil, and E. Simonnet studies bifurcation analysis for the atmosphere, the ocean, and the climate; the article by R. M. Samelson relates to time-periodic flows. C. J. Cotter and D. D. Holm introduce the Hamiltonian methodologies, and B. Ewald and C. Penland introduce numerical techniques for integrating stochastic differential equations.

Two articles are then devoted to turbulence: article by P. Mininni, A. Pouquet, and P. Sullivan emphasizes oceanic and astrophysical applications and the article by M. Lesieur and O. Metais addresses geophysical turbulence. Two articles are devoted to data assimilation: the article by J. Blum, F.-X. Le Dimet, and I. M. Navon and the article by S. Cohn. Finally, two articles are devoted to the mathematical analysis of the primitive equations: the article by M. Petcu, R. Temam, and M. Ziane for the viscous case, and the article by A. Rousseau, R. Temam, and J. Tribbia for the nonviscous case.

We hope that this volume will be useful and thank P. G. Ciarlet for his invitation to prepare it and the different teams of Elsevier for its production.

Roger M. Temam and Joseph J. Tribbia
Boulder, 2007

Acknowledgment

For this work, RT acknowledges the support of the NSF grants DMS 0305110 and DMS 0604235, and of the Research Fund of Indiana University.

Finite-Volume Methods in Meteorology

Bennert Machenhauer

Danish Meteorological Institute, Lyngbyvej 100, DK-2100 Copenhagen, DENMARK

Eigil Kaas

University of Copenhagen, Juliane Maries Vej 30, DK-2100 Copenhagen, DENMARK

Peter Hjort Lauritzen

National Center for Atmospheric Research, Boulder, Colorado, P.O. Box 3000, Boulder, CO 80307-3000, USA

Abstract

Recent developments in finite-volume methods provide the basis for new dynamical cores that conserve exactly integral invariants, globally as well as locally, and, especially, for the design of exact mass conserving tracer transport models. The new technologies are reviewed and the perspectives for the future are discussed.

1. Introduction

Finite-volume (FV) methods are numerical methods where the fundamental prognostic variable considered is an integrated quantity over a certain finite-control volume. Thus, instead of grid-point values, finite elements or spectral components, *cell-integrated mean values* are considered. In meteorology, *FV methods* are, therefore, frequently referred to as *cell-integrated methods*. Some FV methods include additional prognostic variables to enhance the numerical accuracy. These variables can be higher order moments or point/face values between the control volumes.

In meteorological applications, so far, the control volumes adopted have generally been the conventional grid cells used in most operational prediction models: i.e., quasi-horizontal regular grid cells in cartesian coordinates on map projections of the sphere or regular grid cells in spherical latitude-longitude coordinates. These grid cells are referred

Computational Methods for the Atmosphere and the Oceans
Special Volume (Roger M. Temam and Joseph J. Tribbia, Guest Editors) of
HANDBOOK OF NUMERICAL ANALYSIS, VOL. XIV
P.G. Ciarlet (Editor)
Copyright © 2009 Elsevier B.V.
All rights reserved
ISSN 1570-8659
DOI 10.1016/S1570-8659(08)00201-9

to as *the Eulerian grid cells*. In the cell-integrated methods, these are complemented by Lagrangian control volumes, which move with the air flow, usually in a quasi-Lagrangian sense, i.e., departing from or arriving at Eulerian grid cells.

Exceptions to the basis of conventional Eulerian grid cells are new operational models based on grids, which are almost uniform on the sphere. Examples are the Massachusetts Institute of Technology general circulation model (ADCROFT, CAMPIN, HILL and MARSHALL [2004]) which is based on the conformal expanded spherical cube, but still has orthogonal coordinates and quadri laterally shaped grid cells, and the German NWP model (MAJEWSKI, LIERMANN, PROHL, RITTER, BUCHHOLD, HANISCH, PAUL, WERGEN and BAUMGARDNER [2002]) that is based on a non-orthogonal icosahedral-hexagonal grid on the sphere. For the sake of simplicity, we shall not go into details with these new grids, which currently is a very active research topic. The same limitation applies to nonuniform grids, such as the one introduced by LI and CHANG (1996). Thus we shall consider only FV methods in conventional grids.

The FV or cell-integrated methods are well suited for the numerical simulation of conservation laws. Before the implementation of FV methods in meteorological modeling, only conservative spatial discretization schemes were developed and used (e.g., ARAKAWA [2000], ARAKAWA and LAMB [1981], BURRIDGE and HASLER [1977], MACHENHAUER [1979], SIMMONS and BURRIDGE [1981]). With these schemes, just the globally integrated discretized time derivative of the invariant quantity in question was zero. Time truncation errors could still cause nonconservation globally. With the introduction of the FV method, the possibility of a conservative full space-time discretization became possible (e.g., MACHENHAUER [1994]). Previously, just global conservation was considered of importance, whereas with the FV methods, local conservation is considered even more important (e.g., MACHENHAUER and OLK [1997]). Conservation laws for mass, total energy, angular momentum, and entropy constitute the fundamental laws for the dynamics and thermodynamics of the atmosphere. Also, potential vorticity is considered a fundamental invariant which should be conserved in an adiabatic friction-free flow. In general, a discretized cell-integrated prognostic equation for a conservative quantity is obtained by integrating the differential flux form of the conservation law in question in space over an Eulerian grid cell and in time over the time-step Δt. The space integration results in an equation stating that the time rate of change of the total quantity in the grid cell is equal to the sum of fluxes through the cell boundaries. The time integration determines the fluxes through the cell boundaries during the time-step. These fluxes are exact if the integration is performed along exact trajectories ending at the boundaries of the regular Eulerian grid cell (also called *the arrival cell*) at time $t + \Delta t$ and originating from the boundaries of an irregular so-called Lagrangian cell (also called *the departure cell*) at time t. With such an exact integration, the integral of the conservative quantity over the arrival cell at time $t + \Delta t$ is equal to the integral over the departure cell at time t, plus changes due to sources and sinks, if any. We shall mainly concentrate on conservation of mass, which is the simplest conservation law, as it has no sources or sinks if precipitation and diffusion of mass is neglected. For this conservation law, called the continuity equation, we shall derive the exact prognostic equation (Eq. (1.8) in Section 1.1). Since exact integrations along exact trajectories will be assumed in the derivation, and since no further approximations are being made, this equation is

referred to as *the exact discretized cell-integrated continuity equation*. It implies exact conservation of mass during a time step, both global conservation, i.e., conservation of the total mass in the entire integration area and local conservation, i.e., conservation of the mass in each individual departure cell. During the derivation of the exact discretized cell-integrated continuity equation, it will be demonstrated that there is equivalence between traditional flux-form FV approaches and newer semi-Lagrangian FV methods. In both formulations, one attempts to approximate the same equation.

The general exact discretized cell-integrated continuity equation describes conservation of mass of "moist air," which is the atmospheric air including all its constituents. Corresponding exact continuity equations for the different constituents in the moist air, for example, water vapor or any chemical constituent, are obtained by simply replacing the density of moist air ρ with the density $\rho_q = q\rho$ of the constituent in question, where q is its specific concentration[1]. In meteorological models, the solution to the continuity equation for moist air is of special importance. The solution determines the flow of air mass, which determines the pressure distribution and thus the dynamics of weather systems, especially the development and decay of weather systems. Spurious mass sources due to local nonconservation of mass might thus influence the simulation of weather systems (Machenhauer and Olk [1997]). The solution determines the flow of all constituents in the moist air since they are transported with the air and thus share trajectories with the air. This is important especially in chemical models as spurious changes in the ratios between linearly correlated (in space) concentrations of reacting chemical constituents are avoided (Lin and Rood [1996]). Thus, in meteorological models, a "correct" simulation of the atmospheric dynamics and all kinds of interactions among constituents depends heavily on the accuracy of the numerical solutions to the continuity equations.

In Section 2, the different mass conserving schemes that have been developed for meteorological applications in two dimensions (2Ds) are described in detail. In the different schemes, different approximations are made in the determination of the trajectories and in the integration along the trajectories over the time-step or in the integration over the departure cell. The approximate schemes presented in Section 2 will be compared with the exact solution. It will be shown that all the different schemes conserve mass globally, simply because they are all constructed so that the mass that leaves a certain face of an Eulerian arrival cell during a time-step is exactly gained in the neighboring cell with which the cell face is shared. This, of course, does not guarantee a high level of accuracy as the global conservation may be obtained even with rather inaccurate local fluxes. However, the accuracy with which the local mass conservation is approximated is a real measure of the accuracy of the local transports of the moist air and its constituents. Section 2 will mainly focus on relatively new schemes, most of which are based on (semi-) Lagrangian approaches. For completeness, a short introduction to the more traditional flux-form schemes is presented as well.

Section 3 provides an overview over the general applicability of FV techniques in meteorology. This section is initiated with an example of a complete set of FV

[1] *The specific concentration of a constituent* is the ratio between the mass of the constituent and the mass of the moist air it is mixed into.

prognostic equations that conserve mass, entropy, total energy, and angular momentum in an adiabatic and friction-free atmosphere. Furthermore, Section 3 provides two examples of pioneering mass conserving hydrostatic dynamical cores in spherical geometry, which are based on FV techniques. By a *dynamical core*, we mean a computer code for the numerical integration of the system of meteorological equations governing the dynamics of the atmosphere. Roughly speaking, the dynamical core approximates the solution to the meteorological equations on resolved scales, while parameterizations represent subgrid-scale processes and other processes not included in the dynamical core (THUBURN [2006]). However, in tests of dynamical cores, one includes those dissipation terms, which are needed for smooth and stable integrations. Furthermore, Section 3 includes a discussion of a few remaining issues, such as the so-called mass-wind inconsistency in in-line and off-line FV tracer transport applications, and possibilities of extensions to non-hydrostatic models are briefly discussed. Finally, Section 4 includes a brief summary of the main issues presented in this chapter.

1.1. The exact cell-integrated continuity equation

In this section, an "exact" discretized cell-integrated continuity equation is derived. This is introduced as a pre-requisite and reference for the approximate 2D and three-dimensional (3D) FV schemes to be presented in Sections 2 and 3, respectively. It is exact in the sense explained above. It is derived from assumed exact integrals along assumed exact trajectories, which are determined from given exact 3D fields of density and velocity during a time interval Δt from t to $t + \Delta t$. No further assumptions are made, apart from a simplifying one of no vertical shear of the horizontal velocity in each discrete model layer.

Define Eulerian grid cells as the arrival cell indicated to the right in Fig. 1.1 in a cartesian coordinate system (x, y, h) so that the grid length along the x-axis is Δx, the grid spacing along the y-axis is Δy, and h is a terrain following height-based vertical coordinate defined as $h = z - z_s$, where z is the height above mean sea level and z_s is the height of the surface of the Earth. Surfaces with h equal to a constant $h_{k+1/2}$ separate the grid-cell layers in the vertical. The "½" in the index refers to the Lorenz vertical staggering of the variables (LORENZ [1960]). The "half-levels" are located in between "full-levels" $h_k = 1/2\,(h_{k+½} + h_{k-½})$ with integer index k, where point values of mass and velocity variables traditionally have been located. Thus, the height difference between the bottom and the top of the Eulerian grid cell centered at level k, which is considered in Fig. 1.1, is $\Delta_k h = h_{k+½} - h_{k-½}$.

To derive the FV version of the continuity equation, we need to integrate along exact trajectories ending at the boundaries of the arrival cell at time $t + \Delta t$ and originating from the boundaries of the corresponding departure cell at time t. In Fig. 1.1, the departure cell is shown as the irregular cell to the left. Only four of the trajectories are shown in the figure. The exact velocity fields, supposed to be given during the whole time interval Δt from t to $t + \Delta t$, determine a trajectory ending at any of the points inside or at the boundaries of the arrival cell. We now define an additional auxiliary *vertical coordinate* ξ for a particle: a Lagrangian vertical coordinate (STARR [1945]), which per definition is constant along its 3D trajectory. We choose the Lagrangian coordinate ξ of a particle,

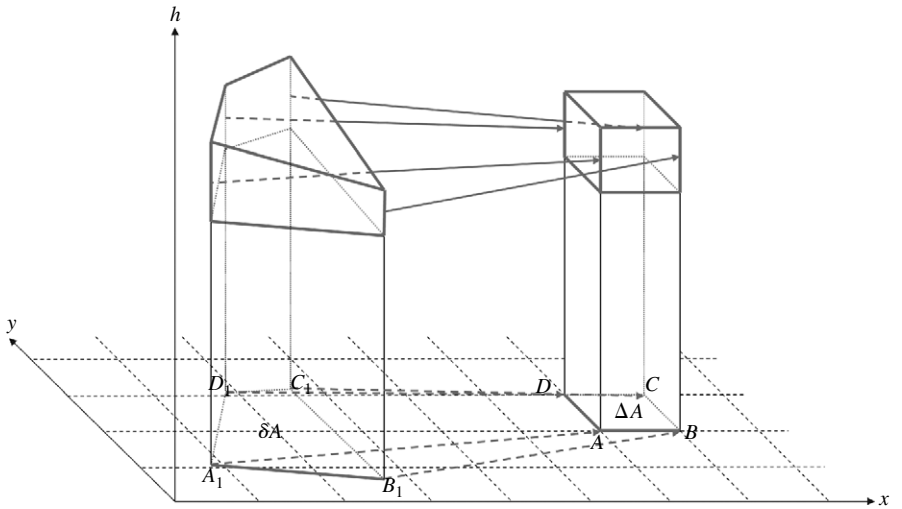


FIG. 1.1 Conceptual sketch showing a cell that is moving with the flow in a Lagrangian model layer during a time-step Δt. To the left is shown the cell at time t (the so-called departure cell). The horizontal velocity $\vec{V}$ within the model layer is assumed independent of height so that the cell walls, which initially at time t are vertical, remain vertical. The cell ends up at time $t + \Delta t$ as the horizontally regular Eulerian grid cell (the so-called arrival cell) shown in the vertical column to the right. Just four trajectories are shown. The projections on a horizontal plane are shown in more detail in Fig. 1.2. (See also color insert).

that is moving with the 3D flow during the time-step, to be equal to its h value in or at the boundary of the arrival cell. Thus, the trajectories constitute a vertical coordinate system, which is defined only in the time interval from t to $t + \Delta t$. Obviously, in this coordinate system, the vertical velocity of a particle is zero:

$$\dot{\xi} = \frac{\mathrm{d}\xi}{\mathrm{d}t} = 0. \tag{1.1}$$

Here, a simplifying assumption is made, namely that the horizontal wind $\vec{V}$ is independent of height within *the Lagrangian model layer*, i.e., the layer enclosing all the trajectories which are ending inside or at the boundary of the arrival cell. Thus, as indicated in Fig. 1.1, vertical columns that move with the horizontal wind in the layer will remain vertical. Mathematically, it implies a simplifying separation of the vertical and horizontal integrations to be performed in the layer. A column may, of course, still change its thickness $\delta_k h$ due to horizontal convergence or divergence. The trajectories in Fig. 1.2, which are ending at the corners of the arrival cell, originate from the corner of the departure cell. For simplicity of the sketch, it is assumed that the horizontal velocity field is such that the trajectories and lines between neighboring corners in the departure cell are straight, i.e., the vertical faces of the departure cell in Fig. 1.1 are plane. Note that since trajectories ending at the boundaries of the arrival cells are shared by neighboring cells, it follows that the departure cells, as does the arrival cells, fill out the entire integration domain without any cracks in between.

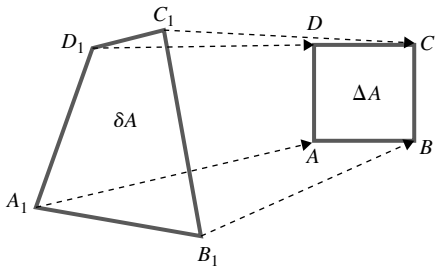


FIG. 1.2 Horizontal projections of the arrival cell (A, B, C, D) at time $t + \Delta t$ with area ΔA and the corresponding upstream departure cell (A_1, B_1, C_1, D_1) at time t with area δA. This figure corresponds to a view from above at the departure and arrival cells in Fig. 1.1.

The differential flux form of the continuity equation in the ξ-coordinate system becomes

$$\frac{\partial \rho}{\partial t} = -\nabla_\xi \cdot \rho \vec{V} - \frac{\partial \rho \dot{\xi}}{\partial \xi}, \tag{1.2}$$

where ρ is the density of moist air and $\vec{V}$ is the horizontal velocity. To obtain the continuity equation for a regular vertical column, integrate Eq. (1.2) vertically over the Lagrangian model layer. The result is

$$\frac{\partial \tilde{\rho}_k \, \delta_k h}{\partial t} = -\nabla_\xi \cdot \tilde{\rho}_k \, \delta_k h \, \vec{V}_k, \tag{1.3}$$

where Eq. (1.1) has been used and $\tilde{\rho}$ is the vertical mean density:

$$\tilde{\rho}_k = \frac{1}{\delta_k h} \int_{\delta_k h} \rho \, \mathrm{d}z.$$

To obtain the cell-integrated continuity equation, integrate Eq. (1.3) horizontally over the area of the arrival grid cell. After application of the Gauss's divergence theorem, we get

$$\Delta A \frac{\partial \overline{\left(\tilde{\rho}_k \, \delta_k h\right)}}{\partial t} = -\sum_{i=1}^{4} \left(< \left(\tilde{\rho}_k \, \delta_k h\right) \vec{V}_k > \cdot \vec{n} \; \Delta l \right)_i, \tag{1.4}$$

where $\Delta A = \Delta x \Delta y$ is the horizontal area of the grid cell and

$$\overline{\left(\tilde{\rho}_k \, \delta_k h\right)} = \frac{1}{\Delta A} \iint_{\Delta x \Delta y} \left(\tilde{\rho}_k \, \delta_k h\right) \, \mathrm{d}x \mathrm{d}y \tag{1.5}$$

is the horizontal mean value of $\tilde{\rho}_k \, \delta_k h$ in the Eulerian grid cell. In Eq. (1.4), $\vec{n}_i$ is a unit vector normal to the ith face of the cell pointing outward, and Δl_i is the length of the

face equal to either Δx or Δy. $\left(\tilde{\rho}_k\right)_i$, $(\delta_k h)_i$, and $(\vec{V}_k)_i$ are instantaneous values at the cell face i, and the angle brackets represent averages in the x- or y-direction over the cell faces. The next step is to integrate over the time-step Δt, between t and $t + \Delta t$, which results in

$$\Delta A\left(\overline{\left(\tilde{\rho}_k\ \delta_k h\right)}^{+} - \overline{\left(\tilde{\rho}_k\ \delta_k h\right)}\right) = -\Delta t \sum_{i=1}^{4} \left(\overline{\overline{< \left(\tilde{\rho}_k\ \delta_k h\right)\vec{V}_k >}} \cdot\ \vec{n}\ \Delta l\right)_i \tag{1.6}$$

Here, the plus-sign in superscript indicates the updated value and the double bar refers to the time average over Δt. Each term on the right-hand side of Eq. (1.6) represents the mass transported through one of the four Eulerian cell faces into the cell during the time-step. Each term involves integrals over the cell face in question and over the time-step. The integral in time over the time-step may be performed in space along the trajectories terminating on the Eulerian cell face in question, cell face AB for instance (see Fig. 1.2). Thus, this term in Eq. (1.6) is computed as a surface integral of $\left(\tilde{\rho}_k\ \delta_k h\right)$ over the area between the Eulerian cell face AB, the two backward trajectories, AA_1 and BB_1 originating from the two end points of the Eulerian cell face and the respective face of the departure cell A_1B_1. That is, the mass inflow through the southern (or lower) face in Fig. 1.2 is equal to the integral of $\left(\tilde{\rho}\,\delta_k h\right)$ over the area marked A_1ABB_1 in the figure. Writing this integral as $\iint\limits_{A_1B_1BA} \tilde{\rho}_k\,\delta_k h\ \mathrm{d}x\,\mathrm{d}y$, Eq. (1.6) may be rewritten as

$$\begin{aligned}
\overline{\left(\tilde{\rho}_k\ \delta_k h\right)}^{+}\Delta A &= \iint\limits_{ABCD} \left(\tilde{\rho}_k\ \delta_k h\right)\,\mathrm{d}x\,\mathrm{d}y + \iint\limits_{A_1B_1BA} \left(\tilde{\rho}_k\ \delta_k h\right)\,\mathrm{d}x\,\mathrm{d}y \\
&\quad + \iint\limits_{A_1ADD_1} \left(\tilde{\rho}_k\ \delta_k h\right)\,\mathrm{d}x\,\mathrm{d}y + \iint\limits_{D_1DCC_1} \left(\tilde{\rho}_k\ \delta_k h\right)\,\mathrm{d}x\,\mathrm{d}y - \iint\limits_{B_1BCC_1} \left(\tilde{\rho}_k\ \delta_k h\right)\,\mathrm{d}x\,\mathrm{d}y \\
&= \iint\limits_{A_1B_1C_1D_1} \left(\tilde{\rho}_k\ \delta_k h\right)\,\mathrm{d}x\,\mathrm{d}y.
\end{aligned} \tag{1.7}$$

Here, the mass inflows through the remaining three cell faces are included in the second line by similar integrals. The first term on the right-hand side is

$$\Delta A\overline{\left(\tilde{\rho}_k\ \delta_k h\right)} = \iint\limits_{ABCD} \left(\tilde{\rho}_k\ \delta_k h\right)\,\mathrm{d}x\,\mathrm{d}y,$$

i.e., the original mass in the Eulerian grid cell at time t. Thus, as illustrated in Fig. 1.2, the sum of the first four terms on the right-hand side of Eq. (1.7), representing the original mass in the Eulerian grid cell, the inflow through the southern, the western, and the northern cell face is compensated partly by the outflow through the eastern cell face, represented by the fifth negative term in Eq. (1.7). The result is the integral on the second right-hand side of Eq. (1.7) that represents the mass in the Lagrangian departure

cell $A_1B_1C_1D_1$. Denoting the departure cell area as δA (Fig. 1.2), the result may be written as

$$\iint\limits_{A_1B_1C_1D_1} \left(\tilde{\rho}_k\ \delta_k h\right)\,\mathrm{d}x\,\mathrm{d}y = \overline{\tilde{\rho}_k\ \delta_k h}\,\delta A,$$

and we obtain finally

$$\overline{\left(\tilde{\rho}_k\ \Delta_k h\right)}^{+}\Delta A = \overline{\left(\tilde{\rho}_k\ \delta_k h\right)}\,\delta A. \tag{1.8}$$

This is a prognostic equation predicting the mass in the arrival area at $t+\Delta t$, $\overline{\left(\tilde{\rho}_k\ \delta_k h\right)}^{+}\Delta A$, from the mass in the departure area at time t, $\overline{\left(\tilde{\rho}_k\ \delta_k h\right)}\,\delta A$. Note that no information is needed between t and $t+\Delta t$ and recall that in the arrival area (exactly in the center of the area) the Lagrangian model layer coincide with the Eulerian cell so that $\delta_k h = \Delta_k h$. Thus, the right-hand side of Eq. (1.8) can be determined by an integration of $\left(\tilde{\rho}_k\ \delta_k h\right)$ over the departure area and Eq. (1.8) becomes

$$\overline{\left(\tilde{\rho}_k\right)}^{+} = \frac{1}{\Delta_k h\ \Delta A}\iint\limits_{A_1B_1C_1D_1}\left(\tilde{\rho}_k\ \delta_k h\right)\mathrm{d}x\,\mathrm{d}y = \frac{1}{\Delta_k V}\iiint\limits_{\delta_k V}\rho_k\ \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z, \tag{1.9}$$

or

$$\overline{\left(\tilde{\rho}_k\right)}^{+}\Delta_k V = \iiint\limits_{\delta_k V}\rho_k\ \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z. \tag{1.10}$$

$\overline{\left(\tilde{\rho}_k\right)}^{+}\Delta_k V$ is the updated mass in the Eulerian arrival grid cell at time $t+\Delta t$. According to Eq. (1.10) it is equal to the mass in the upstream departure cell at time t. Thus, the exact discrete cell-integrated continuity equation (Eq. (1.8)) is simply a cell-integrated analog to the well-known grid-point semi-Lagrangian continuity equation (Robert [1969, 1981, 1982]) that presently is used in most operational meteorological models. Contrary to the grid-point version, the cell-integrated equation is inherently mass conservative. It fulfills exactly our definition of a locally mass conserving scheme as the updated mass in an Eulerian arrival grid cell is exactly the mass in the upstream departure cell. It is easily shown by a summation of Eq. (1.8) over the entire integration domain, with assumed periodic lateral boundary conditions, that it also implies global mass conservation. The analogy to the grid-point semi-Lagrangian continuity equation shows that an alternative way to derive Eq. (1.8) would be to set up the mass conservation law directly for FV on a Lagrangian form and then integrate that form over Δt. The mass in a FV $\delta_k V$ considered at time t is

$$M_{\delta_k V} = \iiint\limits_{\delta_k V}\rho_k\ \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z. \tag{1.11}$$

The mass conservation law for this FV, which is supposed to move with the flow without any mass flux through its boundaries, is

$$\frac{\mathrm{d}\, M_{\delta_k V}}{\mathrm{d}t} = 0. \tag{1.12}$$

When integrated in time from t to $t + \Delta t$, Eq. (1.12) gives Eq. (1.8), which as shown above leads to Eq. (1.10). The reason for presenting the more complicated derivation starting from the Eulerian flux form of the mass conservation law (1.2) is that some numerical FV schemes, the so-called *flux-form schemes*, are based on the flux form (1.2), whereas others, the so-called *Lagrangian schemes*, are based on the Lagrangian form (1.12). The purpose of the present derivation was to show that in the case of exact trajectories and exact mass integrals over the relevant volumes, the flux-form Eq. (1.2) is equivalent to the Lagrangian form (1.12). When, as it is usually the case, a flux-form scheme becomes different from a Lagrangian scheme, it is due to different approximations to the trajectories defining the departure volume and different approximations to the upstream mass integrals. A measure of accuracy for both types of schemes should therefore be how close they are to the ideal "exact" scheme. That is, how close the approximate departure volume is to the real, exact one and how close the exact mass integral over the exact departure volume is to the approximate mass integral over the approximate departure volume. In other words, how accurate the local mass conservation is.

1.2. Longtime step schemes and combinations with semi-implicit time-stepping

The reason for the recent renewed interest in FV methods in meteorological modeling was the observation of a significant lack of global mass conservation in numerical models using the grid-point version of the semi-Lagrangian scheme unless an unphysical so-called mass-fixer, which restores the total mass globally after each time step, is used. There is an arbitrariness in the way these mass-fixing algorithms repeatedly restore global mass conservation without ensuring any local mass conservation, i.e., without fulfilling a continuity equation for the mass that is transported locally between the Eulerian grid cells of the model each time-step (Machenhauer and Olk [1997]). Without such a mass-fix, a significant drift in the global mass was observed (Bates, Higgins and Moorthi [1995]), and even with a mass-fixer, it seems likely that significant local errors are developed (Machenhauer and Olk [1997]). Nevertheless, the reason for the popularity of the grid-point semi-Lagrangian schemes has been its almost unconditional absolute stability, which in practice eliminates the advective Courant-Fredrichs-Levy (CFL) time-step restriction. This property is utilized in most operational meteorological models in combination with a semi-implicit treatment of the gravity wave terms in the primitive equations, which eliminates the fast wave CFL time-step restriction. Then, in principle, the length of the time-steps in a combined semi-implicit semi-Lagrangian model can be chosen solely based on accuracy considerations. This is extremely important in meteorological models where any gain by an increased time-step can be utilized to increase the realism of parameterized physical processes and/or the spatial resolution of the model

grid. According to general operational experience, such improvements have practically always led to an increase in accuracy. As should be expected from the experience with the grid-point semi-Lagrangian schemes, the recently developed cell-integrated semi-Lagrangian schemes are also (almost) unconditionally stable (LAURITZEN [2007]), eliminating in practice the advective CFL time-step restriction. It has furthermore recently been shown that fast waves in cell-integrated semi-Lagrangian models can be stabilized by a combination with a semi-implicit time extrapolation scheme. This has been demonstrated by MACHENHAUER and OLK [1997] for a simple one-dimensional (1D) mass and momentum or mass and total energy conserving model and LAURITZEN, KAAS and MACHENHAUER [2006] for shallow water models and by LAURITZEN, KAAS, MACHENHAUER and LINDBERG [2008] for a complete 3D mass conserving model. An alternative method, which has been used in finite difference grid-point models to stabilize the fast waves, is the so-called split-explicit time-stepping. However, this possibility was abandoned by MACHENHAUER and OLK [1997] for FV models because when splitting the system of continuous equations into an advective part, which should use large time-steps, and an adjustment gravity wave part, which should use short time-steps, it was found that neither of the sub systems were conserving momentum or total energy. Consequently, these invariants for the full system could not be conserved exactly in any FV version.

As mentioned above, Section 3 describes two mass conserving quasi-hydrostatic dynamical cores, both combined with comprehensive physical parameterization packages. One of these dynamical cores described in LAURITZEN, KAAS, MACHENHAUER and LINDBERG [2008] is a semi-implicit version using large time steps for all variables, while the other one described in LIN [2004] and COLLINS, RASCH, BOVILLE, HACK, MCCAA, WILLIAMSON, KIEHL, BRIEGLEB, BITZ, LIN, ZHANG and DAI [2004] uses an explicit time-stepping scheme. The latter model uses explicit, relatively small time-steps for the dynamical core but large time-steps for the transport of all tracer species (including water vapor) and for physical parameterizations.

2. Transport schemes in one and two dimensions

In meteorological models, a FV method for the continuity is based on the exact cell-integrated continuity equation and obviously it should be approximated as accurately as possible. As discussed in Section 1, the vertical and horizontal problems can be separated in a consistent way considering Lagrangian cells moving with vertical walls along three dimensional trajectories. Consequently, only horizontal integrals of vertically integrated mass distributions are needed in the solution of the continuity equation. So in case of a flux-form Eulerian scheme, the fluxes through the four cell faces can be determined by horizontal integrals (as described in connection with Eq. (1.6)), and for the departure cell-integrated semi-lagrangian (DCISL) scheme, direct integrations over the horizontal departure area approximating the true departure area can be performed (as indicated in Eq. (1.8)). Hence, by using this approach, one can directly apply 2D FV schemes for the 3D problem. Alternatively, flux-form schemes may be extended to 3Ds by including vertical advection through the top and bottom surfaces of Eulerian grid cells. Similarly,

the 3D DCISL scheme would perform a 3D integral over the Lagrangian departure cell. However, following these fully 3D approaches would become very complicated if one aims at a numerically efficient and mass conserving integration.

Because of the general applicability of 2D solutions to the continuity equation, this section, beside basic 1D formulations, is devoted to the fully 2D schemes. Compared with the large number of mass conserving transport schemes published in the general fluid dynamical literature, there are many fewer schemes that have been used or are applicable in real meteorological applications on the sphere. Here, we mostly concentrate on the subset that is potentially applicable in a wide range of atmospheric models. Therefore, descriptions of the vast majority of the hundreds of transport schemes developed in computational fluid dynamics in general are excluded. For a more general review of FV methods, see e.g., LEVEQUE [2002] and EUMARD, GALLOUËT and HERBIN [2000].

Before discussing the different FV schemes used in the atmospheric sciences, it is important to realize which properties a transport scheme ideally should possess. An overview of these properties is provided in Section 2.1. The FV schemes presented in this overview use sub grid representations at time t in order to make the forecast at time $t + \Delta t$. The most frequently used sub grid representations and associated filters ensuring some of the properties listed in Section 2.1 are introduced in Section 2.2. This is followed, in Section 2.3, by an overview of the different types of FV methods applied in 2D problems. Section 2.4 briefly describes some – mostly recent – local mass conservation fixers for semi-Lagrangian models which can be considered closely related to FV semi-Lagrangian schemes. Aiming at enhanced accuracy, Section 2.5 discusses the possibilities to include extra prognostic variables in addition to the cell-mean values. The so-called flux-limiter methods have been popular approaches to maintain attractive shape-preserving properties. A brief introduction to these methodologies, which are complementary to the filtering methods mentioned in Section 2.2, is given in Section 2.6. Finally, Section 2.7 provides some concluding remarks on the basic FV transport schemes in 1 and 2Ds.

2.1. Desirable properties

The equation subject to the toughest requirements is probably the continuity equation for tracers such as moisture, the spatial distribution of which includes sharp gradients. RASCH and WILLIAMSON [1990] have defined seven desirable properties for transport schemes: accuracy, stability, computational efficiency, transportivity, locality, conservation, and shape-preservation. In addition to the seven desirable properties defined by RASCH and WILLIAMSON [1990], even more desirable properties have emerged in the literature, e.g., consistency, compatibility, and preservation of constancy. The perfect scheme would have all the desirable properties listed above under all conditions but, in practice, no single method is advantageous under all conditions.

2.1.1. Accuracy

The high-accuracy property is, of course, the primary aim for any numerical method, and all the desirable properties listed above, apart from the efficiency requirement, are part of the overall accuracy. Note that for a flow with shocks or sharp gradients, the formal order

of accuracy in terms of Taylor series expansions does not necessarily guarantee a high level of accuracy. Part of the accuracy is also the rate of convergence of the numerical algorithm.

Widely used measures of accuracy in the meteorological community for idealized test cases are the standard error measures l_1, l_2, and l_∞ (e.g., WILLIAMSON, DRAKE, HACK, JAKOB, SWARZTRAUBER, [1992]):

$$l_1 = \frac{I\left(|\psi - \psi_E|\right)}{I\left(|\psi_E|\right)}, \tag{2.1}$$

$$l_2 = \frac{\left\{I\left[(\psi - \psi_E)^2\right]\right\}^{1/2}}{\left\{I\left[(\psi_E)^2\right]\right\}^{1/2}}, \text{ and} \tag{2.2}$$

$$l_\infty = \frac{\max\left[|\psi - \psi_E|\right]}{\max\left[|\psi_E|\right]}, \tag{2.3}$$

where $I(\cdot)$ denotes the integral over the entire domain, ψ is the numerical solution, and ψ_E is the exact solution if it exists. In case an exact solution does not exist, ψ_E is a high-resolution reference solution. l_1 and l_2 are the measures for the global "distance" between ψ and ψ_E, and l_∞ is the normalized maximum deviation of ψ from ψ_E over the entire domain. In addition to these error measures, the normalized maximum and minimum values of ψ are also used to indicate errors related to overshooting and undershooting.

To evaluate the accuracy of new schemes, several idealized advection test cases have been formulated. The interscheme comparison, however, is often made difficult by the fact that different authors use different test cases and/or different error measures. The test problems can be divided into two categories. Firstly, translational passive advection tests where distributions are transported by prescribed non-divergent winds that, ideally, translate the initial distribution without distorting it; these test cases involve the entire domain. Secondly, deformational test cases which focus on part of the domain such as an initial distribution being deformed by a vortex. Recently, NAIR and JABLONOWSKI [2007] combined these two types of test cases into one.

Probably, the most commonly used idealized test case in the meteorological literature is the solid body rotation of a cosine cone and/or a slotted cylinder. In cartesian geometry, the test case is described in, e.g., ZALESAK [1979] and BERMEJO and STANIFORTH [1992], and the spherical version is test case 1 of the suite of test cases by WILLIAMSON, DRAKE, HACK, JAKOB and SWARZTRAUBER [1992]. The analytic solution to this problem is simply the translation of the initial distribution along a circle in cartesian geometry and a great circle in the spherical case. It is an important part of accuracy that the advection schemes can transport distributions across the singularities of the numerical grids without distortion and imposing severe time-step limitations. DRAKE, HACK, JAKOB, SWARZTRAUBER and WILLIAMSON [1992] suggested that the cosine bell is transported along the equator and across the poles with a slight offset to avoid any symmetry. Note, however, that away from the poles, advection along these great circles is almost along coordinate axis for conventional latitude-longitude grids that, in general, favor the advection scheme.

Passive advection of scalars using the solid body rotation test case only addresses the ability of the scheme to translate a distribution without distorting it. Other commonly used test cases are based on a deformational flow, for example the swirling shear flow test in cartesian geometry considered by DURRAN [1999, Section 5.7.4], which is specified in terms of a periodically reversing time-dependent velocity field. Hence, after one period, the exact solution is the initial distribution. It could, however, be speculated that some errors introduced during the first half period are cancelled when the wind field reverses. Other deformational flow test cases, to which the exact solution is known throughout the time of integration, are defined in SMOLARKIEWICZ [1982] (analytical solution is given in CÔTÉ, STANIFORTH and PUDYKJEWICZ [1987]) and ARMENGAUD and HOURDIN [1999]. The idealized cyclogenesis problem described by DOSWELL [1984], to which the analytic solution is known, has been used for scalar-advection tests by several authors. For example, the non-smooth deformational flow vortex defined on a tangent plane (e.g., RANČIĆ [1992], HÓLM [1995], CÔTÉ, NAIR and STANIFORTH [1999a]). A version was formulated for the sphere by NAIR, SCROGGS and SEMAZZI [2002] and NAIR and MACHENHAUER [2002]. It is a smooth deformational flow test case that consists of two symmetric vortices, one over each pole. This test case has been combined with a translational wind field in NAIR and JABLONOWSKI [2007] to form a test case (where the analytical solution is known) that simultaneously challenges schemes with respect to deformation and translation.

2.1.2. Stability

The stability property ensures that the solution does not "blow up" during the time of integration. Usually, the stability of Eulerian methods is governed by the CFL condition, which in 1D is given by

$$\max \left| \frac{u \Delta t}{\Delta x} \right| \leq 1, \tag{2.4}$$

where u is the velocity and Δx the grid interval. Hence, a fluid parcel may not travel more than one grid interval during one time-step. This overly restrictive time-step limitation is usually alleviated in semi-Lagrangian methods and can be replaced by the less severe Lipschitz convergence criterion

$$\left| \frac{\partial u}{\partial x} \right| \Delta t < 1, \tag{2.5}$$

(BENOIT, PUDYKIEWICZ and STANIFORTH [1985]; KUO and WILLIAMS [1990]), which guarantees that parcel trajectories do not cross during one time-step and ensures the convergence of the trajectory algorithm (a multi dimensional extension of Eq. (2.5) is given in BENOIT, PUDYKIEWICZ and STANIFORTH [1985]). Hence, in semi-Lagrangian models, the time-step can be chosen for accuracy and not for stability because of the lenient stability condition.

For global models based on a conventional latitude-longitude grid, the efficiency and stability of the advection schemes are often challenged by the convergence of the meridians near the poles, and special care must be taken in the vicinity of the poles. Alternatively,

the problem can be tackled by using other types of grids that do not have these singularities or at least reduce the effect of them, for example, the icosahedral-hexagonal grid used operationally by the German Weather Service (e.g., ARAKAWA, MINTZ and SADOURNY [1968], WILLIAMSON [1968], THUBURN [1997], MAJEWSKI, LIERMANN, PROHL, RITTER, BUCHHOLD, HANISCH, PAUL, WERGEN and BAUMGARDNER [2002]), and the cubed sphere approach originally introduced by SADOURNY [1972] which, after having remained dormant for many years, has become a very active research topic (e.g., IACONO, PAOLUCCI and RONCHI [1996], MESINGER and RANČIĆ, PURSER [1996], MCGREGOR [1996], ISKANDRANI, TAYLOR and TRIBBIA [1997], LOFT, NAIR and THOMAS [2005]). These grids are more isotropic than conventional latitude-longitude grids, i.e., all cells have nearly the same size, contrarily to latitude-longitude grids, where the areas decrease as aspect ratios increase toward the poles (this effect can, however, be alleviated by using a Gaussian-reduced grid in which the number of longitudes decrease toward the poles).

2.1.3. Computational efficiency

Computing resources are limited and, given the complexity of geophysical fluid dynamics, the algorithms should be computationally efficient in order to allow for high-resolution runs and/or a large number of prognostic variables. Efficiency is, however, hard to measure objectively. One measure for the efficiency of an algorithm is the number of elementary mathematical operations or the total number of floating-point operations per second (FLOPS) used by the algorithm. The advantage of counting FLOPS is that it can be done without using a computer and is, therefore, a machine-independent measure. But the number of FLOPS only captures one of several dimensions of the efficiency issue. The actual program execution involves subscripting, memory traffic and countless other overheads. In addition, different computer architectures favor different kinds of algorithms and compilers optimize code differently. Measuring efficiency in terms of the execution time on a specific platform can be misleading for a user on another computer platform. Weather prediction and climate models are often executed on massively parallel distributed memory computers where the efficiency is partly determined by the amount of communication between the nodes. This becomes increasingly important if the resolution is held fixed while the number of distributed memory processors is increased. Hence, the parallel programmer is concerned about algorithms being local, thus minimizing the need for communication between the nodes. Nevertheless, a very important measure of efficiency is probably the level of simplicity of the algorithm.

Since models include an increasing number of tracers, an important aspect of the efficiency is how much of the transport algorithm can be reused for additional tracers. Obviously, if the entire transport algorithm must be repeated for each additional tracer, such an algorithm would not be attractive in modern transport models that include hundreds of tracers. In semi-Lagrangian models, for example, the computation of trajectories need only be computed once and can be reused for all tracers (e.g., DUKOWICZ and BAUMGARDNER [2000]).

Thus, the computational cost of a given model depends not only on the number of FLOPS involved in the production of say one model day; it depends also to a high degree on the computer architecture on which the model is run. The optimization of a given model intended for operational application on a given platform is often an extensive

and complicated work for an experienced programmer, and the result will vary with the ingenuity of the programmer. The new algorithms presented here are often developed on an experimental basis by scientists who are not specialized programmers, and therefore, they are usually far from an optimized code suited for operational applications. It is, therefore, not fair to uncritically compare the computational cost of such new FV algorithms with traditional well-optimized algorithms. For this reason, information on the computational costs of the new algorithms are most often not available in the literature. When they are available, they should be considered with the reservations stated here and should only be considered as possible maximum computational costs, which most likely could be reduced for operational applications.

It should be noted that it could be misleading to compare computational efficiency and accuracy of two algorithms at the same spatial and temporal resolution. A scheme might be computationally inefficient at a given resolution compared to other schemes but have an accuracy that other schemes would need a much finer resolution in order to achieve (the opposite situation is, of course, also possible). In other words, ideally one should consider the ratio between computational cost and accuracy when comparing numerical schemes. That would enable one to select the scheme where one pays as little as possible computationally for the highest level of accuracy.

2.1.4. Transportivity and locality

The transportive and local property guarantee that information is transported with the characteristics and that only adjacent grid values affect the forecast at a given point. For FV schemes, one aspect of the local property is the degree of local mass-conservation that we define as follows. Since the mass enclosed in an area moving with the flow is conserved in the absence of sources and sinks, the degree to which the effective departure area of the numerical scheme coincides with the exact departure area is a measure for the local mass conservation of a given scheme. Another aspect of local mass conservation is the degree to which the reconstruction of the subgrid-scale distribution is local. For example, near sharp gradients, it is important that the gradient is not weakened during the process of reconstructing the subgrid-scale distribution, i.e., the reconstruction should be local.

2.1.5. Shape-preservation

The shape of a distribution undergoing pure advection should ideally be preserved in the numerical solution. For general velocity fields, the shape of the distribution may be altered in the form of new extrema. In such situations, the numerical scheme should reproduce only the physical extrema without creating spurious numerical extrema. These spurious numerical extrema especially cause problems in situations where the advection scheme produces negative mixing ratios (or concentrations) or when the values are above the maximum possible. Negative mixing ratios or mixing ratios above a physical threshold value are unphysical and would most likely cause a breakdown in physical parameterizations. If a numerical scheme inherently prevents negative undershoots in mixing ratios (or concentrations), it is termed positive-definite (or positivity preserving), if it preserves gradients, then the scheme is monotone, and if artificial oscillations are prevented, it is termed nonoscillatory. All these properties are, of course, interrelated and,

constitutes together the shape-preservation property. The very popular spectral methods are well known for producing "wiggles" (also known as Gibb's phenomena) near sharp gradients and are, therefore, a typical example of a monotonicity-violating and oscillatory numerical method.

2.1.6. Conservation

Ideally, all global integral invariants of the corresponding continuous problem should be conserved for any kind of flow. For long simulations, the conservation properties become increasingly important as numerical sources, and sinks can degrade the accuracy and alter global balance budgets significantly over time. Hence, for climate models, the FV methods are very attractive given their inherent conservation properties. However, a numerical model can only maintain a small number of analogous invariant properties constant, and some choice must be made as to which conserved quantities are to be conserved in the numerical model. For a comprehensive discussion of this issue, see THUBURN [2006]. Probably, the most important property to conserve for a continuity equation is the first moment, i.e., mass.

2.1.7. Consistency

The consistency property is less frequently discussed in the literature. Notable exceptions are JÖCKEL, VON KUHLMANN, LAWRENCE, STEIL, BRENNINKMEIJER, CRUTZEN, RASCH and EATON [2001] and BYUN [1999]. This property concerns the coupling between the continuity equation for air as a whole and for individual tracer constituents. In the continuous case, the flux-form continuity equation for a constituent with specific concentration q,

$$\frac{\partial}{\partial t}(q\rho) + \nabla \cdot \left(\vec{v} q \rho\right) = 0 \tag{2.6}$$

degenerates to

$$\frac{\partial}{\partial t}(\rho) + \nabla \cdot \left(\vec{v} \rho\right) = 0 \tag{2.7}$$

for $q = 1$. This should ideally be the case numerically as well. If the two equations are solved using the same numerical method, on the same grid and using the same time-step, the consistency is guaranteed. However, in reality, in practical applications of FV transport schemes, the settings are often inconsistent in this sense. This is definitely the case in offline tracer transport models. The consistency property, or rather the lack of it (referred to as the mass-wind inconsistency), will be discussed in detail in Section 3.

2.1.8. Compatibility

The compatibility property was defined by SCHÄR and SMOLARKIEWICZ [1996] for Eulerian schemes, and the definition is here extended also to include semi-Lagrangian schemes. As the consistency property, it concerns the relationship between continuity equations. Equations (2.6) and (2.7) imply

$$\frac{\mathrm{d}q}{\mathrm{d}t} = 0, \tag{2.8}$$

which states that the constituent mixing ratio is conserved along the characteristics of the flow. Compatible transport is when the discretization of Eq. (2.6) is consistent with the advective form (Eq. (2.8)) so that the predicted mixing ratio q^{n+1}, which in a flux-form setting is recovered from $(q\rho)^{n+1}$, is limited by the mixing ratios in the Eulerian cells from which the mass departs. The compatibility property is graphically illustrated in Fig. 2.1.

2.1.9. Preservation of constancy

Another desirable property is the ability of the scheme to preserve a constant tracer field for a non-divergent flow. For traditional semi-Lagrangian methods based on Eq. (2.8), a constant distribution is trivially conserved since the divergence of the velocity field does not appear in the prognostic equation (for a review of traditional semi-Lagrangian methods, see, e.g., STANIFORTH and COTÉ [1991]). In fact, for any velocity field, the traditional semi-Lagrangian method preserves a constant mixing ratio. For FV methods where the divergence appears explicitly since tracer mass, and not mixing ratio, is the prognostic variable, it is not automatic that a constant field is preserved for a nondivergent velocity field. Non-conservation of constant fields may cause error problems and even instability, see Section 3.3.2.

2.1.10. Preservation of linear correlations between constituents

Another desirable property identified by LIN and ROOD [1996] is that a numerical scheme should ideally preserve tracer correlations since correlations carry fundamental information on atmospheric transport. This is particularly important in chemical atmospheric models where the relative concentrations of constituents are crucial for the speed and

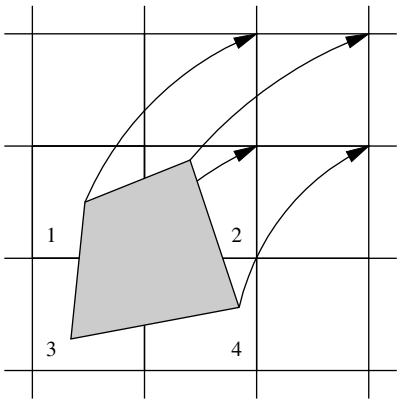


FIG. 2.1 A graphical illustration of the compatibility property. The arrows show the trajectories for the cell vertices. The shaded area is the departure cell that, after one time-step, ends up at the regular grid as depicted by the arrows. A finite-volume scheme predicts the change in total mass in the Eulerian cell $\overline{(q\rho)}^{n+1}$, which is the mass enclosed in the departure cell (shaded area). Since the mixing ratios are preserved along parcel trajectories, the mixing ratio in the arrival cell $\overline{q}^{n+1}$ should be within the range of the mixing ratios at the departure points. For the situation depicted on the figure, the compatibility condition is $\min\left(\overline{q}_1^n, \overline{q}_2^n, \overline{q}_3^n, \overline{q}_4^n\right) \leq \overline{q}^{n+1} \leq \max\left(\overline{q}_1^n, \overline{q}_2^n, \overline{q}_3^n, \overline{q}_4^n\right)$, where $\overline{q}_i$ denotes the average mixing ratio in the cell numbered i, $i = 1.4$, on the figure.

balances of chemical reactions. It is possible to construct transport schemes that maintain spatially constant linear correlations between tracers exactly, see, e.g., Lin and Rood [1996].

2.2. *Subgrid-cell distributions of the prognostic variables*

In all FV schemes presented here – flux based as well as Lagrangian types – it is necessary to determine the subgrid-cell distribution from the surrounding cell averages in order to make accurate estimates of the fluxes through the Eulerian cell walls or mass enclosed in the upstream departure cell. Therefore, 1 and 2D reconstructions are discussed before the actual schemes are introduced.

2.2.1. *1D subgrid-cell reconstructions*

Several 1D methods for reconstructing the subgrid distribution have been published in the literature. The simplest subgrid representation is a piecewise constant function followed, in complexity, by a piecewise linear representation (Van Leer [1974]). Both methods are computationally cheap, optionally monotonic, and positive-definite (the piecewise constant method is shape-preserving by default) but, on the other hand, excessively damping and therefore not suited for long runs at coarser resolutions. To reduce the dissipation to a tolerable level, the subgrid-cell representation must be polynomials of at least second degree. Requirements of computational efficiency put an upper limit to the order of the polynomials used, which explains why the predominant choice is second order.

Let the walls of the ith cell be located at x_i and x_{i+1} and denote the cell width $\Delta x_i = x_{i+1} - x_i$. The coefficients of the subgrid-cell reconstruction polynomials are determined by imposing constraints. Apart from the basic requirement of mass conservation within each grid cell, the choice of constraints is not trivial. Probably the simplest parabolic fit is obtained by requiring that the polynomial

$$p_i(x) = (a_0)_i + (a_1)_i\, x + (a_2)_i\, x^2, \quad x \in [x_i, x_{i+1}] \tag{2.9}$$

not only conserves mass in the ith grid cell

$$\int_{x_i}^{x_{i+1}} p_i(x)\mathrm{d}x = \Delta x_i \overline{\psi}_i, \quad \psi = \rho, \rho q \tag{2.10}$$

but also in the two adjacent cells:

$$\int_{x_{i+1}}^{x_{i+2}} p_i(x)\mathrm{d}x = \Delta x_{i+1} \overline{\psi}_{i+1}, \tag{2.11}$$

$$\int_{x_{i-1}}^{x_i} p_i(x)\mathrm{d}x = \Delta x_{i-1} \overline{\psi}_{i-1}, \tag{2.12}$$

(Laprise and Plante [1995]). By substituting Eq. (2.9) into Eqs. (2.10), (2.11), and (2.12) and by evaluating the analytic integrals, a linear system results that can easily be solved for the three unknown coefficients $(a_0)_i$, $(a_1)_i$, and $(a_2)_i$ (Laprise and Plante [1995]). When performing this operation for all cells, a global piecewise-parabolic representation is obtained. The method is only locally of second order since it is not necessarily continuous across cell borders. This method is referred to as the piecewise parabolic method 1 (PPM1).

An alternative way of constructing the parabolas, which ensures a globally continuous distribution if no filters are applied, is the piecewise-parabolic method of Woodward and Colella [1984] (hereafter referred to as PPM2). PPM2 has been reviewed in the context of meteorological modeling in Carpenter, Droegemeier, Hane and Woodward [1990]. Instead of requiring that $p_i(x)$ conserves mass in adjacent cells, the constraint is that the polynomial equals prescribed west and east cell-edge values, $p_i^W = p_i(x_i)$ and $p_i^E = p_i(x_{i+1})$, respectively, at the cell edges. p_i^E is computed with a cubic polynomial fit (see Woodward and Colella [1984] for details). For an equidistant grid, the result is

$$p_i^W = \frac{7}{12}\left(\overline{\psi}_{i-1} + \overline{\psi}_i\right) - \frac{1}{12}\left(\overline{\psi}_{i+1} + \overline{\psi}_{i-2}\right), \tag{2.13}$$

(for a nonequidistant grid, see Colella and Woodward [1984]). The east cell-border value, p_i^E, is simply an index shift of the formula for p_i^W

$$p_i^E = p_{i+1}^W. \tag{2.14}$$

It is convenient to use the cell average, $\overline{\psi}_i$, and p_i^W and p_i^E to define the ith parabola, instead of using $(a_0)_i$, $(a_1)_i$, and $(a_2)_i$. The equivalent formula for $p_i(x)$ is given by

$$p_i\left(\xi^x\right) = \overline{\psi}_i + \left(\delta p^x\right)_i \xi^x + \left(p_6^x\right)_i \left[\frac{1}{12} - \left(\xi^x\right)^2\right], \tag{2.15}$$

where $(\delta p^x)_i$ is the mean slope

$$\left(\delta p^x\right)_i = p_i^E - p_i^W, \tag{2.16}$$

$\left(p_6^x\right)_i$ is the "curvature"

$$\left(p_6^x\right)_i = 6\overline{\psi}_i - 3\left(p_i^W + p_i^E\right), \tag{2.17}$$

and ξ^x is the nondimensional position defined by

$$\xi^x = \frac{x - x_i}{\Delta x_i} - \frac{1}{2}. \tag{2.18}$$

PPM2 uniquely defines the parabolas and Eq. (2.15) guarantees that the global subgrid distribution is continuous across cell borders. ZERROUKAT, WOOD and STANIFORTH [2002] found in passive advection tests that using the PPM2 for the subgrid-cell reconstructions (where the parabolas were continuous across cell borders) results in more accurate solutions compared with PPM1 (in which the distribution is not necessarily continuous across cell borders).

Instead of using the PPM1 or PPM2, ZERROUKAT, WOOD and STANIFORTH [2002] derived a piecewise cubic method for the reconstruction of the subgrid-cell distributions. PPM2 is a special case of the piecewise cubic method. Of course, any kind of reconstruction that is mass conserving can be used, e.g., rational functions as used in the transport scheme of XIAO, YABE, PENG and KOBAYASHI [2002] and the parabolic spline method (PSM) recently developed by ZERROUKAT, WOOD and STANIFORTH [2006]. In idealized advection tests, ZERROUKAT, WOOD and STANIFORTH [2007] found that using the PSM for the subgrid-scale reconstructions in their scheme generally leads to more accurate results than when using PPM2. This is despite the fact that, in terms of operation count, PSM is 60% more efficient than PPM2. However, at present, the most widespread subgrid-cell reconstruction method is PPM2.

Without further constraining the coefficients of the parabolas, it is not guaranteed that the subgrid-scale reconstruction preserves monotonicity or positive definiteness (GODUNOV [1959]). A simple monotonic filter was proposed by COLELLA and WOODWARD [1984], and is explained in Fig. 2.2. For local extrema, the filter is similar to the quasi-monotonic filter by BERMEJO and STANIFORTH [1992] for traditional semi-Lagrangian advection schemes, i.e., the subgrid-scale distribution is reduced to a constant when there is a local extrema in the cell averages (Fig. 2.2a). This severe clipping can significantly reduce the accuracy as idealized advection tests have shown (compare CISL-N with CISL-M and CCS-N with CCS-M in Table 2.1). Clearly, one would like to retain the higher order polynomial in the situation depicted in Fig. 2.2a while not altering the treatment of the situation in Fig. 2.2b.

LIN and ROOD [1996] modified the COLELLA and WOODWARD [1984] monotonic filter so that the monotonic filter only applies to undershooting and does not interfere with any of the overshooting (referred to as semi-monotonic filter). The semi-monotonic filter can further be modified so that it only prevents negative undershooting, whereby it becomes a positive-definite filter. Since these filters avoid the severe clipping of overshoots, the application of these filters shows a dramatic increase in accuracy in idealized advection tests compared with the monotonic filter described in the previous paragraph (CISL-P and CCS-P in Table 2.1). Other filters with more relaxed constraints, but which are computationally more efficient, can be found in LIN [2004]. However, all these filters are still not fully satisfactory since they do not interfere with all types of spurious undershooting and overshooting.

As mentioned, the filter should not interfere with local extrema as the one in Fig. 2.2a but still apply the monotonic filter in the situation depicted on Fig. 2.2b (similarly for undershooting). That is what the filter of SUN, YEH, SUN and SUN [1996] for traditional semi-Lagrangian schemes is designed to do. Through a series of logical statements, the filter detects local extrema and does not alter the high-order subgrid-scale reconstruction

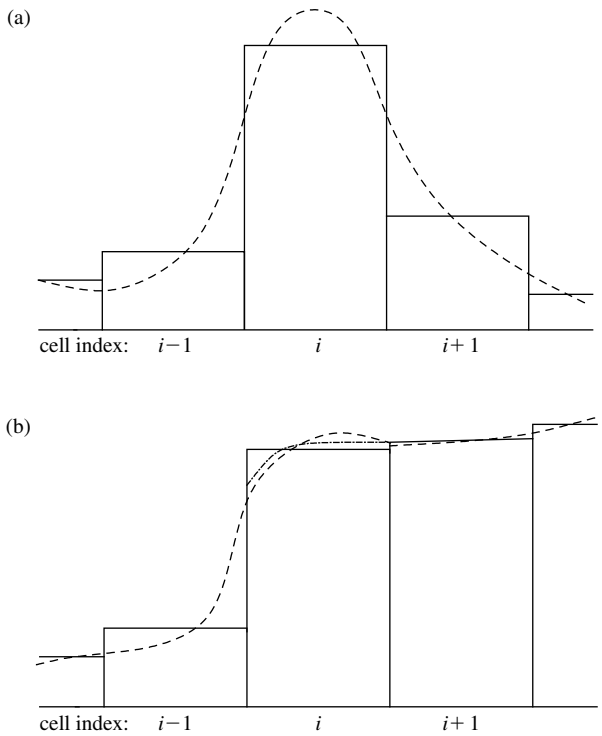


FIG. 2.2 A graphical illustration of the basic monotonic filter of COLELLA and WOODWARD [1984]. Solid lines show the cell averages, and the dashed line is the unmodified piecewise parabolic fit. (a) The situation in which the parabola in cell i is a local extrema. The monotonic filter sets the parabola equal to a constant in cell i. (b) The situation when $\overline{\psi}_i$ is in between $(p_E)_i$ and $(p_W)_i$, but is sufficiently close to one of the edge values that the parabola takes values outside the range of the surrounding cell averages, i.e., when $|(p_W - p_E)_i| \geq \left|(p_6^x)_i\right|$. In this situation, $(p_E)_i$ is reset and the gradient at the east cell wall is set to zero thereby guaranteeing monotonicity of the polynomial in cell i (dash-dotted line) or vice versa.

where these nonspurious extrema are located. As pointed out by NAIR, CÔTÉ and STANIFORTH [1999a], this filter, however, is still unsatisfactory near strong gradients, where the unmodified subgrid-scale distribution exhibits $2\Delta x$ noise. In such a situation, for example, the semi-monotonic filter of LIN and ROOD [1996] or the filter of SUN, YEH and SUN [1996] does not filter the noise satisfactorily (Fig. 2.3). To deal with such situations (and others) while still maintaining non-spurious extrema, ZERROUKAT, WOOD and STANIFORTH [2005] proposed a more advanced filter that, in the situations shown on Figs. 2.2 and 2.3, consecutively reduces the order of the fitting polynomials until none of the spurious overshooting and undershooting depicted on the Figs. 2.2 and 2.3 appear. Then, the severe clipping of physical "peaks" is eliminated and grid-scale noise is removed without introducing excessive numerical damping. Contrarily to the monotonic filter of COLELLA and WOODWARD [1984], this filter can improve the accuracy compared

TABLE 2.1

Error norms for the schemes of ZERROUKAT, WOOD and STANIFORTH [2005] (SLICE), NAIR and MACHENHAUER [2002] (CISL), and NAIR, SCROGGS and SEMAZZI [2002] (CCS) for test case 1 in WILLIAMSON, DRAKE, HACK, JAKOB and SWARZTRAUBER [1992]. $\hat{\alpha}$ is the angle between the axis of solid body rotation and the polar axis of the spherical coordinate system. Hence $\hat{\alpha} = 0$ is solid body rotation along the equator and $\hat{\alpha} = \pi/2$ advection across the poles. The error measures for $\hat{\alpha} = \pi/3$ are from LAURITZEN, KAAS and MACHENHAUER [2006]. "N" denotes no filter, "M" the monotonic filter, and "P" the positive-definite filter used in the respective schemes. Note that the monotonic filter and subgrid-scale reconstructions in SLICE are different from the other schemes (see text for details)

	$\hat{\alpha} = 0$			$\hat{\alpha} = \pi/2$			$\hat{\alpha} = \pi/3$		
Schemes	l_1	l_2	l_∞	l_1	l_2	l_∞	l_1	l_2	l_∞
SLICE-N	0.046	0.029	0.022	0.079	0.049	0.042	—	—	—
SLICE-M	0.038	0.024	0.017	0.058	0.040	0.037	—	—	—
CISL-N	0.052	0.035	0.032	0.063	0.046	0.048	0.075	0.051	0.083
CISL-P	0.025	0.025	0.031	0.059	0.045	0.048	0.043	0.082	0.076
CISL-M	0.094	0.091	0.108	0.084	0.084	0.109	0.077	0.089	0.18
CCS-N	—	—	—	0.054	0.042	0.065	0.051	0.039	0.076
CCS-P	0.036	0.034	0.042	0.051	0.041	0.065	0.033	0.034	0.077
CCS-M	—	—	—	0.076	0.082	0.129	0.070	0.086	0.186

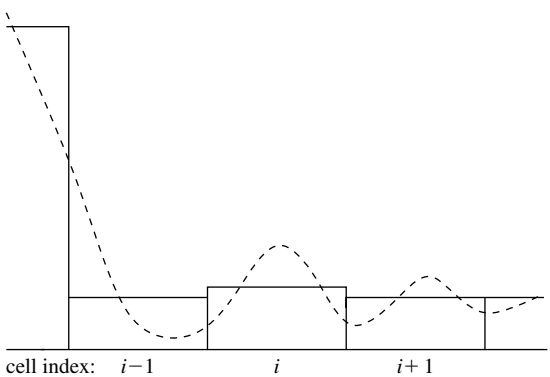


FIG. 2.3 A situation in which the unmodified subgrid-cell reconstruction exhibits strong Gibbs phenomena. The semimonotonic filter of LIN and ROOD [1996] would set the polynomials in cell $i-1$ and $i+1$ equal to the cell average, but would not modify the polynomial in cell i that, in this situation, is a spurious overshoot.

to the unfiltered high-order solution (see Semi-Lagrangian inherently conserving and efficient (SLICE)-N and SLICE-M in Table 2.1). A similar filter has also been developed for the PSM (ZERROUKAT, WOOD and STANIFORTH [2006]).

2.2.2. 2D subgrid-cell reconstructions

As for the 1D case, 2D linear reconstructions exist (e.g., DUKOWICZ and BAUMGARDNER [2000] and SCROGGS and SEMAZZI [1995]), but, in general, they introduce too much numerical damping for meteorological applications. The PPM in 1D can be directly

extended to 2Ds as done by RANČIĆ [1992], i.e., in terms of a fully 2D subgrid-cell reconstruction

$$p_{i,j}(x, y) = (a_1)_{i,j} + (a_2)_{i,j}\, x + (a_3)_{i,j}\, x^2 + (a_4)_{i,j}\, y + (a_5)_{i,j}\, y^2 + (a_6)_{i,j}\, xy + (a_7)_{i,j}\, xy^2 + (a_8)_{i,j}\, x^2 y + (a_9)_{i,j}\, x^2 y^2, (x, y) \in [x_i, x_{i+1}] \times \left[y_j, y_{j+1}\right]. \tag{2.19}$$

This fully biparabolic fit involves the computation of nine coefficients, so nine constraints are needed to determine the coefficient values. Apart from the conservation of mass within each cell

$$\iint\limits_{\Delta A_{i,j}} p_{i,j}(x, y)\mathrm{d}x\mathrm{d}y = \overline{\psi}_{i,j}\Delta A_{i,j}, \tag{2.20}$$

the other eight constraints chosen by RANČIĆ were formulated in terms of the four corner values of $p_{i,j}(x, y)$ and the average of $p_{i,j}(x, y)$ along the four cell walls. The corner point scalar values were computed by fitting 2D third-order polynomials using the 16 cell averages surrounding the corner point in question. The average along the cell walls was computed using $\overline{\psi}$ along a line perpendicular to the cell wall in question. For additional details, see RANČIĆ [1992].

The computational cost of the approach taken by Rančić can be reduced significantly by using a quasi-biparabolic subgrid-cell representation. Contrarily to fully biparabolic fits, the quasi-biparabolic representation does not include the "diagonal" terms and simply consists of the sum of two 1D parabolas, one in each coordinate direction. Using the form (Eq. (2.15)) for the parabolas, the quasi-biparabolic subgrid-cell representation is given by

$$p_i\left(\xi^x, \xi^y\right) = \overline{\psi}_{i,j} + \left(\delta p^x\right)_{i,j}\xi^x + \left(p_6^x\right)_{i,j}\left[\frac{1}{12} - \left(\xi^x\right)^2\right] + \left(\delta p^y\right)_{i,j}\xi^y + \left(p_6^y\right)_{i,j}\left[\frac{1}{12} - \left(\xi^y\right)^2\right], \tag{2.21}$$

where $(\delta p^x)_{i,j}$, $\left(p_6^x\right)_{i,j}$, $(\delta p^y)_{i,j}$, and $\left(p_6^y\right)_{i,j}$ are the coefficients of the parabolic functions in each coordinate direction (MACHENHAUER and OLK [1998]). This representation reduces the computational cost of the subgrid-cell reconstruction significantly but, of course, does not include variation along the diagonals of the cells.

By using 1D filters that prevent undershoot and overshoot to the parabolas in each coordinate direction, monotonicity-violating behavior can be reduced but not strictly eliminated in 2Ds. In case of negative values at the cell boundaries of both unfiltered 1D parabolic representations, even larger negative values may be present in one or more of the cell corners when the 1D representations are added. The monotone and positive-definite filters eliminate only the negative values at the boundaries and not the possible negative corner values. As a result, small negative values can appear even after the application of a monotonic filter (e.g., LIN and ROOD [1996] and NAIR and MACHENHAUER [2002]). To eliminate these negative values an additional filter must be applied.

2.3. *Different schemes in 2Ds*

As mentioned above, different approaches can be used to estimate the integral over the departure cell. These can be divided into two main categories:

- Semi-Lagrangian schemes in which the integral over the departure cell is approximated explicitly. These schemes to be described in Section 2.3.1 are referred to as DCISL schemes. DCISL schemes come in two types: fully 2D schemes and cascade schemes in which the approximation of the upstream integral is divided into two steps where each substep applies 1D methods.
- Flux-based schemes in which the fluxes through the Eulerian arrival cell walls are approximated. These schemes are described in Section 2.3.2. As for the DCISL schemes, there are two types of conceptually different schemes of this category: schemes based on a sequential operator splitting (often referred to as time-splitting) and schemes based on direct estimation of the 2D fluxes.

It is important to note, as was also pointed out in the introduction, that DCISL and flux-based FV schemes are conceptually equivalent since they both estimate the mass in the departure cell. However, as will be illustrated, this is, in practice, done in quite different ways.

The following overview of these two categories will mainly focus on recent developments in DCISL schemes since these have not yet been introduced in textbooks or general overview articles. For these schemes, a stability analysis is performed. Furthermore, the level of local mass conservation, i.e., the accuracy of the approximation to the exact departure area in different DCISL schemes and one flux-based method, is investigated.

2.3.1. *DCISL schemes*

The semi-Lagrangian scheme can either be based on backward or forward trajectories (or equivalently downstream and upstream trajectories), i.e., by considering parcels arriving or departing from a regular grid, respectively. The majority of semi-Lagrangian schemes are based on backward trajectories because it is usually simpler to interpolate/remap from a regular to a distorted mesh. However, forward trajectory cascade schemes and the downstream version of the schemes in LAPRISE and PLANTE [1995] are exceptions to this. The deformed grid resulting from tracking the parcels moving with the flow is referred to as the *Lagrangian grid,* while the stationary and regular grid is referred to as the *Eulerian grid*. The curve resulting from tracking a latitude moving with the flow is referred to as a *Lagrangian latitude*. Similarly for a *Lagrangian longitude*.

The choice of trajectory algorithm is crucial for the accuracy of DCISL schemes. Traditional semi-Lagrangian schemes employ backward trajectories that are computed with an implicit iterative algorithm also known as the second-order implicit midpoint method (see, e.g., CÔTÉ and STANIFORTH [1991]). This trajectory algorithm does not include the acceleration. Several schemes that include estimates of the acceleration in the trajectory computations have been proposed (e.g., HORTAL [2002], MCGREGOR [1993], LAURITZEN, KAAS and MACHENHAUER [2006] – see Section 3).

Using backward trajectories, the 2D discretization of Eq. (1.8) leads to the CISL scheme

$$\overline{\psi}^{+}\Delta A = \overline{\psi}\delta A, \quad \psi = \rho, \rho q, \tag{2.22}$$

where

$$\overline{\psi} = \frac{1}{\delta A}\iint_{\delta A} \psi\,(x, y)\,\mathrm{d}A \tag{2.23}$$

is the integral mean value of $\psi\,(x, y)$ over the irregular departure cell area δA, and $\overline{\psi}^{+}$ is the mean value of $\psi^{+}\,(x, y)$ over the regular arrival cell area ΔA (see Fig. 2.4). The approximation of the integral on the right-hand side of Eq. (2.22) employs two steps: firstly, defining the geometry of the departure cell that involves the computation of parcel trajectories; secondly, performing the remapping, i.e., computing the integral over the departure cell using some reconstruction of the subgrid distribution at the previous time-step. The geometrical definition of the departure cell and the complexity of the subgrid-scale distribution are crucial for the efficiency and accuracy of the scheme.

For realistic flows and for time-steps obeying the Lipschitz criterion (see Section 2.1), the upstream cells deform into simply connected but non-rectangular and possibly locally concave geometric patterns. The question is how to integrate $\psi\,(x, y)$ efficiently over such a complex area.

2.3.1.1. Fully 2D DCISL schemes In 1D, there is very little ambiguity on how to approximate the upstream cell, but in 2Ds, it is much more complicated and several approaches have been suggested in the literature. In Fig. 2.5, the arrival and departure cells in cartesian geometry for three different DCISL schemes are shown.

Rančić [1992] defines the departure cell as a quadrilateral by tracking backward the cell vertices A, B, C, and D and connecting them with straight lines (Fig. 2.5(a)). The

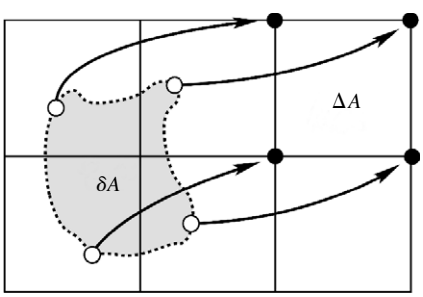


Fig. 2.4 The regular arrival cell with area ΔA and the irregular departure cell (shaded area) with area δA in the continuous case for a generic upstream DCISL scheme. Using the figure of speech in Laprise and Plante [1995], the departure–arrival cell relationship is conceptually equivalent to throwing a fishing net upstream to fetch the mass enclosed into a area that will, after one time-step, end up at the regular mesh. The arrows are the parcel trajectories from the departure points (open circles), which arrive at the regular cell vertices (filled circles).

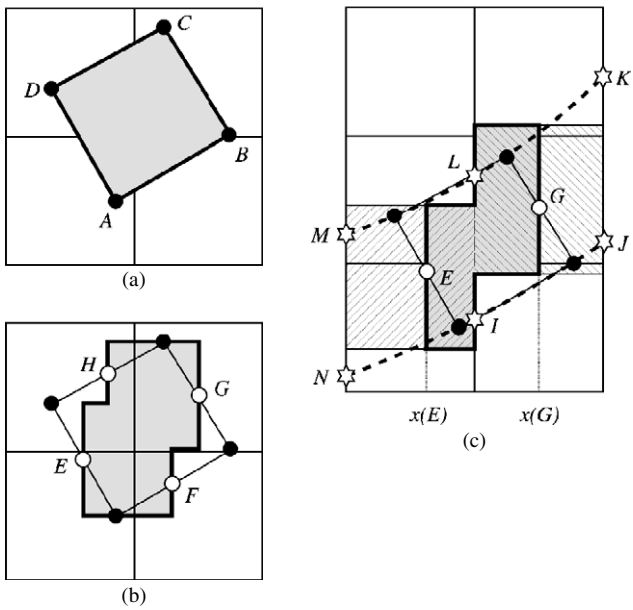


FIG. 2.5 The departure cell (shaded area) when using the scheme of (a) RANČIĆ [1992], (b) MACHENHAUER and OLK [1998] scheme, and (c) the cascade scheme of NAIR, SCROGGS and SEMAZZI [2002]. The filled circles are the departure points, and open circles the midpoints between the departure points, and asterisks are the intermediate grid points which are used to define the intermediate cells in the cascade scheme (crosshatched area).

vertices are not necessarily aligned with the coordinate axis, which leads to some algorithmic complexity for the evaluation of the upstream integral. The integral over the departure area is, in the situation depicted in Fig. 2.5a, decomposed into four subintegrals, i.e., the integral over the areas defined by the overlap between the departure cell and the Eulerian cells. Thus, one has to perform analytic integrals over many possible cases of shapes of subdomains, which makes the computer code rather cumbersome. In addition, the subgrid-scale distribution used by RANČIĆ was a piecewise-biparabolic representation which, being fully 2D, is quite expensive to compute in itself. The combination of the complex geometry of the departure cell and the fully 2D subgrid-cell representation makes the scheme approximately 2.5 times less efficient than the traditional semi-Lagrangian advection scheme using bicubic Lagrange interpolation (RANČIĆ [1992]). This, and the fact that the scheme has not been extended to spherical geometry, has hindered the scheme for widespread use in the meteorological community.

In order to speed up the remapping process, MACHENHAUER and OLK [1998] simplified both the geometry of the departure cell and the subgrid-scale distribution. The departure cell is defined as a polygon with sides parallel to the coordinate axis (Fig. 2.5(b)). The sides parallel to the x-axis are at the y-values of the departure points, and the sides parallel to the y-axis pass through E, F, G, and H, located halfway between the departure points. Hence, the area of the departure cell is identical to the area of the RANČIĆ [1992] scheme. Since the sides of the departure cell are parallel to the coordinate axis, the evaluation of the upstream integral is greatly simplified. By using the

pseudobiparabolic subgrid-scale distribution (see Eq. (2.21)) and accumulated parabola-coefficients along latitudes (see NAIR and MACHENHAUER [2002] for details), the integral over the departure cell can be computed much more efficiently compared to the approach taken by RANČIĆ [1992]. For advection in cartesian geometry, NAIR and MACHENHAUER [2002] reported a 10% overhead with this scheme compared with the traditional semi-Lagrangian scheme.

Note that the departure areas in Fig. 2.5 completely cover the entire integration area without overlaps or cracks, which is crucial to an upstream DCISL scheme, otherwise the total mass is not conserved. For a downwind cell-integrated scheme using forward trajectories, it is, however, not necessary that the arrival cells span the entire domain of integration in order to have global mass conservation. Using the figure of speech of LAPRISE and PLANTE [1995], a downstream cell-integrated scheme is equivalent to throwing dust contained in little buckets (regular departure cells) into the wind and watching it later fall into bins (regular Eulerian cells). Contrary to upstream DCISL schemes, where one integrates over a particular departure cell, a downstream cell-integrated scheme keeps track of the contribution to each regular Eulerian cell from the irregular arrival cells. As long as all the mass in each arrival cell is redistributed with a mass conservative method, mass is conserved even if the neighboring arrival cells overlap. This is taken advantage of in the scheme of LAPRISE and PLANTE [1995], which probably uses the simplest cell geometry of all the schemes presented here. The arrival cell is defined as a rectangle where the edges have the same orientation as the regular cells (see Fig. 2.6). This is achieved by tracing the traverse motion of cell edges and not the cell vertices. Hereby, the arrival cell retains the orthogonality and orientation of the regular departure cell. Note, however, that only two cells share edges, while, if cell vertices are tracked,

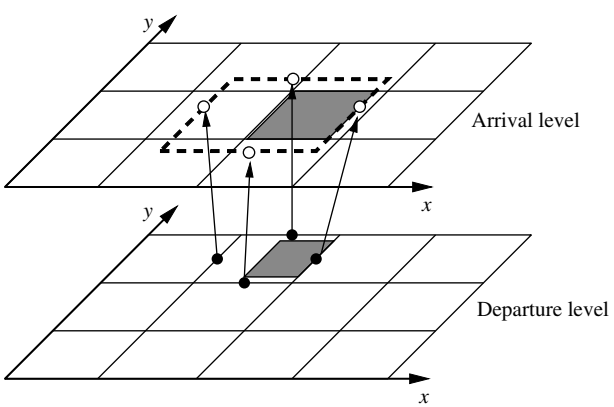


FIG. 2.6 A graphical illustration of the downstream version of the cell-integrated schemes of LAPRISE and PLANTE [1995]. The filled circles are the departure points that are at the edge centers of the regular departure cell. The arrows connect the departure points with the respective arrival points (unfilled circles). The dashed rectangle is the arrival cell which edges have the same orientation as the departure cell. In a downstream cell-integrated scheme, the amount of mass that arrives at a regular Eulerian cell is computed, i.e., the integral over the area (shaded area at the departure level) that arrives at the intersection between the regular Eulerian cell and the arrival cell (shaded area at the arrival level). Similarly for the remaining intersections with Eulerian cells.

cell vertices are shared by four cells. Consequently, one must compute twice as many trajectories compared to a downstream scheme tracking cell vertices.

The actual integral in the downstream scheme of LAPRISE and PLANTE [1995] is not performed at the arrival level since that would require the reconstruction of the subgrid-scale distributions from irregular and overlapping arrival cell averages. The integral is performed at the departure level over the part of the departure cell that, after one time-step, "falls" into a particular regular Eulerian cell (see Fig. 2.6). The intersection between the arrival cell and a particular Eulerian cell is always a rectangular region with sides parallel to the coordinate axis, which simplifies the integration process significantly. Consequently, the downstream scheme of LAPRISE and PLANTE [1995] is approximately twice as fast as the RANČIĆ [1992] scheme, even though both schemes use fully 2D subgrid-scale reconstructions. The schemes of LAPRISE and PLANTE [1995] and RANČIĆ [1992] have not been extended to spherical geometry.

2.3.1.2. Cascade DCISL schemes Using the so-called cascade method, originally developed for non-conservative interpolation by PURSER and LESLIE [1991], the 2D upstream integral can also be approximated by splitting it into two 1D steps. The basic idea is to track backward (forward) the Eulerian grid and then apply 1D integrals, firstly along the Eulerian (Lagrangian) longitudes or latitudes and secondly along the Lagrangian (Eulerian) latitudes or longitudes (see Fig. 2.7). To obtain inherent mass conservation, the interpolation in the original cascade interpolation method must be replaced with the PPM (COLELLA and WOODWARD [1984]) or some other mass-conservative remapping method. Contrary to fully 2D DCISL schemes, the cascade approach is equally suited for downstream and upstream trajectories, or equivalently, the 1D remapping methods are equally suited for remapping from a distorted as from a regular 1D grid. However, to facilitate the comparison with fully 2D DCISL schemes, we assume upstream trajectories in the discussion of the cascade schemes, although some of the schemes initially were formulated for downstream trajectories.

The cascade method can be divided into three steps. Firstly, given the departure points, the application of 1D remappings is prepared by computing an intermediate grid. It is crucial to the cascade technique that the intermediate grid is well defined, i.e., that there should not be multiple intersections between Lagrangian latitudes (longitudes) and Eulerian longitudes (latitudes) (see, e.g., Fig. 1 in NAIR, CÔTÉ and STANIFORTH [1999a]). Therefore, in spherical coordinates, it cannot be applied very near the poles. Secondly, a 1D remapping of mass from the regular Eulerian cells to the intermediate grid cells is performed. Thirdly, the mass on the intermediate grid is remapped to the departure cells.

We start out by considering the conservative cascade DCISL scheme of NAIR, SCROGGS and SEMAZZI [2002]. In this scheme, the departure cells are defined as polygons with sides parallel to the coordinate axis as in the MACHENHAUER and OLK [1998] scheme. In each 1D cascade step, the PPM2 is used. Compared to the MACHENHAUER and OLK [1998] scheme, the departure cell geometry is defined somewhat differently (see Fig. 2.5(c)). Two of the sides parallel to the y-axis, $x = x(E)$ and $x = x(G)$, are defined as in the MACHENHAUER and OLK [1998] scheme, and the remaining two sides parallel to the y-axis are at the Eulerian longitude $x = x_i$. The sides parallel to the x-axis are determined from the intermediate Lagrangian grid points I, J, K, L, M, and

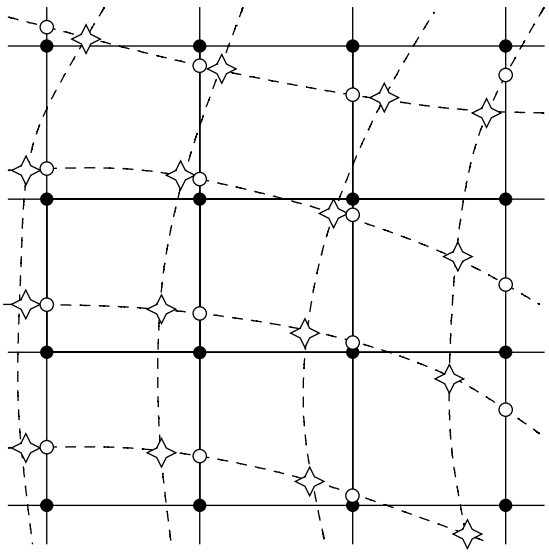


FIG. 2.7 Graphical illustration of the cascade interpolation method introduced by PURSER and LESLIE [1991]. Solid lines are the regular Eulerian grid and the dashed lines are the Lagrangian grid. Here we consider an upstream scheme, hence, the Lagrangian grid end up at the Eulerian grid when moving with the flow over one time-step. The intermediate grid is defined by the crossings between the Eulerian longitudes and the Lagrangian latitudes (unfilled circles). The nonconservative cascade interpolation method proceeds as follows. Perform a 1D interpolation from the Eulerian grid to the intermediate grid, i.e., interpolate along Eulerian longitudes from the filled circles to the unfilled circles. Hereafter interpolate from the intermediate grid to the Lagrangian grid, i.e., from the unfilled circles to the asterisks. Ideally one should interpolate along the curved Lagrangian latitude, but in the original cascade interpolation scheme of PURSER and LESLIE [1991], the x-coordinate is used as the position variable for interpolation along the Lagrangian latitude, which corresponds to approximating the Lagrangian latitude with line segments parallel to the x-axis.

N defined as $y = \frac{1}{2}[y(I) + y(J)]$, $y = \frac{1}{2}[y(K) + y(L)]$, $y = \frac{1}{2}[y(L) + y(M)]$, and $y = \frac{1}{2}[y(I) + y(N)]$, respectively. The y-values of the intermediate points are determined by cubic Lagrange interpolation between the y-values of four adjacent departure points along the *Lagrangian latitude* (dashed line in Fig. 2.5(c)). The Lagrange weights for computing the intersections can be efficiently evaluated using the algorithm outlined in the Appendix in PURSER and LESLIE [1991]. The upstream integral is computed by a remapping in the north-south direction from the Eulerian cells to the intermediate cells (crosshatched rectangular regions on Fig. 2.5(c)), followed by a remapping along the *Lagrangian latitudes* from the intermediate cells to the departure cells. Hence, the first remapping is along the Eulerian longitude passing through the Eulerian cell centers. Since the second remapping uses the x-coordinate as the independent variable, it is along line segments parallel to the x-axis. Without any a priori knowledge of the flow, there is no argument for not reversing the order of the directional sweeps, i.e., first to remap along the Eulerian latitude and then along the Lagrangian longitude. As discussed in some detail in LAURITZEN [2007], the order of the directional sweeps is not symmetric, and hence there is a directional bias built into the cascade approach. However, in neither

of the cascade schemes presented here has this been reported to be a problem. A symmetric version of the cascade scheme can easily be constructed, for example, by alternating between the sweep directions, i.e., by using Lagrangian longitudes and Eulerian latitudes at even time-steps and Lagrangian latitudes and Eulerian longitudes at odd time-steps.

Since the two remappings are 1D and that the intermediate grid can be efficiently computed, the NAIR, SCROGGS and SEMAZZI [2002] scheme is more than twice as efficient as the fully 2D scheme of MACHENHAUER and OLK [1998]. Cascade methods are equally suited for upstream and downstream trajectories. For example, the NAIR, SCROGGS and SEMAZZI [2002] scheme formulated for backward trajectories has also been extended to forward trajectories in NAIR, SCROGGS and SEMAZZI [2003].

On equidistant cartesian grids, the conservative cascade scheme developed by NAIR, SCROGGS and SEMAZZI [2002] is very similar to the one of RANČIĆ [1995], although the way they are presented in the respective articles is very different. RANČIĆ [1995] formulated his scheme without explicit reference to areas by assigning mass to nodes or *mass-points*. The scheme is identical to the PURSER and LESLIE [1991] cascade interpolation but with the two 1D Lagrange interpolation sweeps replaced with PPM2. Although Rančić did not make explicit reference to areas in his formulation, the scheme can, however, be interpreted in terms of areas: in each 1D sweep, the mass nodes represent the mass enclosed in cells with walls located midway between the mass nodes, and the remapping is along line segments which are parallel to the coordinate axis. Hence, by formulating Rančić scheme for upstream trajectories, the only differences between the NAIR, SCROGGS and SEMAZZI [2002] and Rančić scheme are the choice of points for which the trajectories are computed and the order of the 1D sweeps. Where NAIR, SCROGGS and SEMAZZI [2002] track cell vertices as they are transported by the flow, Rančić used cell centers; and where NAIR, SCROGGS and SEMAZZI [2002] remaps first along Eulerian longitudes, the upstream version of Rančić's scheme remaps along the Eulerian latitudes first. Hence, in principle, these schemes are identical and only differing in implementation details when considering a cartesian equidistant grid. However, it is not clear if Rančić's scheme can be extended to non-equidistant grids, and hence it has not been extended to spherical geometry, whereas the NAIR, SCROGGS and SEMAZZI [2002] scheme has been extended to spherical geometry using two different approaches (NAIR, SCROGGS and SEMAZZI [2002] and NAIR [2004]).

In the cascade schemes discussed so far, the second sweep is along Lagrangian latitudes (longitudes) that are defined by line segments parallel to the x-axis (y-axis). In the continuous case, the Lagrangian latitude (longitude) is a curve, and one should ideally remap mass along such a curve. ZERROUKAT, WOOD and STANIFORTH [2002] refined the approaches described so far by performing the second sweep along a continuous piecewise linear line that more accurately represents the curved Lagrangian latitude (longitude). The scheme is called the SLICE scheme and is described next.

The remapping procedure used in SLICE is graphically illustrated on Fig. 2.8. As in the scheme of MACHENHAUER and OLK [1998] and the cascade scheme of NAIR, SCROGGS and SEMAZZI [2002], the cell vertices are tracked backward. The corresponding departure points are connected with straight lines to define *Lagrangian longitudes* and *latitudes*. Regular intermediate cells are defined by the intersections between the Lagrangian longitudes and the Eulerian latitudes that pass through the center of the

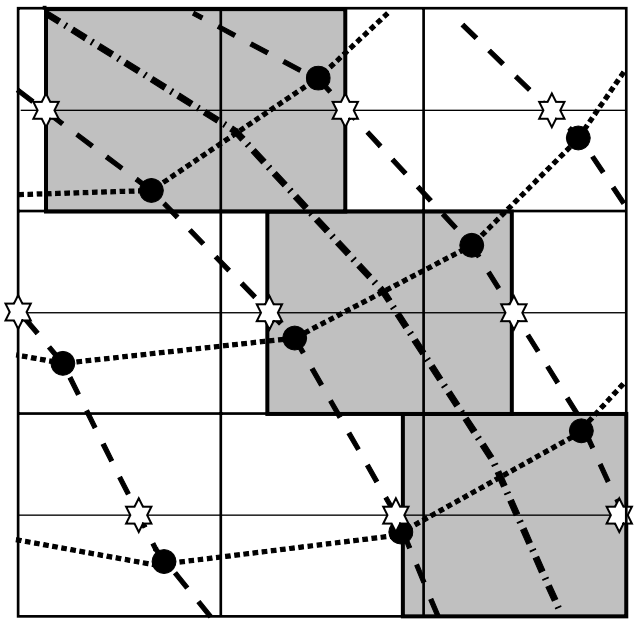


FIG. 2.8 A graphical illustration of the remappings in the Semi-Lagrangian inherently conserving and efficient scheme. The filled circles are the departure points corresponding to the cell vertices. The dotted (dashed) lines are the Lagrangian latitudes defined by connecting the departure points which arrive along the same latitude with straight line segments. The shaded areas are the intermediate areas that are defined by the crossings between the Lagrangian longitudes and the Eulerian latitudes passing through the center of the Eulerian cells (thin lines). The crossings are marked with asterisk. First the mass is remapped from the Eulerian cells to the intermediate cells. The dash-dotted line is the line along which the cumulative distance function is defined, and is used for the second remapping.

cells. Similarly to the NAIR, SCROGGS and SEMAZZI [2002] scheme, the cell averages are mapped from the Eulerian cells to the regular intermediate cells defined by the intersections (see shaded area in Fig. 2.8). As mentioned in the previous paragraph, the remapping from the intermediate cells to the departure cells is quite different from the NAIR, SCROGGS and SEMAZZI [2002] scheme. The second remapping is performed along the Lagrangian longitude that is defined by a continuous piecewise linear line (dash-dot line in Fig. 2.8). A cumulative distance function along the Lagrangian longitude is used to define the Eulerian and Lagrangian north-south cell walls for the second remapping (see ZERROUKAT, WOOD and STANIFORTH [2002] for details). Hereby the independent coordinate for the second sweep is defined along continuous piecewise linear lines that, in principle, are more accurate than line segments parallel to the coordinate axis. However, the intermediate cells have walls parallel to the coordinate axis (east-west walls of shaded on Fig. 2.8). Hence the mass used in the second sweep is only approximately along the piecewise linear Lagrangian longitude (dashed lines on Fig. 2.8). For the 1D remappings, SLICE applies a piecewise cubic method (ZERROUKAT, WOOD and STANIFORTH [2002]) or the PSM (ZERROUKAT, WOOD and STANIFORTH [2007]).

A great potential of cascade schemes is that they may be extended to 3Ds without excessive computational cost and algorithmic complexity. For example, 3D Lagrangian interpolation requires $O(o^3)$ operations, where o is the formal order of accuracy of the interpolator, while cascade schemes require $O(o)$ operations (e.g., PURSER and LESLIE [1991]). Cascade schemes retain their simplicity in higher dimensions, whereas fully higher dimensional DCISL schemes increase rapidly in complexity as the number of dimensions is increased. Cascade interpolation has not only been applied in semi-Lagrangian advection schemes but have also been used for remapping state variables between the regular latitude-longitude grid and the cubed-sphere grid in a conservative and monotone manner (LAURITZEN and NAIR [2007]).

In some situations, the cascade DCISL schemes get a more accurate subgrid-scale representation compared to the fully 2D DCISL scheme of MACHENHAUER and OLK [1998]. The latter scheme uses a 2D reconstruction that does not include variation along the diagonals of the cells. In the cascade schemes, the second sweep is along Lagrangian latitudes (longitudes). Hence, in situations in which the Lagrangian latitudes (longitudes) are sloping toward north-east or north-west and significant variation is along these Lagrangian latitudes (longitudes), the cascade schemes get some of the diagonal variation that is eliminated by the subgrid-scale reconstructions used in MACHENHAUER and OLK [1998]. This is clearly demonstrated when comparing the error measures for the solid body advection for the flow orientation parameter $\hat{\alpha} = \pi/3$ in Table 2.1 (see CCS and CISL). In this situation, the distribution is far away from the poles and the transport is nearly along celldiagonals.

2.3.1.3. Degree of local mass conservation To understand to which extent the different DCISL schemes are local, a test case using an analytic flow field involving translation, rotation, and divergence has been constructed. We define the degree of mass "locality" of the schemes as their ability to approximate the domain of the exact upstream departure cell over which the mass is integrated. Note that the domain of dependence is larger than the upstream departure cell area. The domain of dependence is the area from which information is needed to construct the subgrid-cell representations in the Eulerian cells overlapped by the departure cell.

The analytic wind field is given by

$$\begin{aligned} u(x, y) &= u_0 + D_0 x - R_0 y, \\ v(x, y) &= v_0 + D_0 y + R_0 x, \end{aligned} \tag{2.24}$$

where $(u_0, v_0) = 54\frac{\mathrm{m}}{\mathrm{s}} \times (\cos(10^\circ), \sin(10^\circ))$, $D_0 = -0.0023\,/\mathrm{s}$, and $R_0 = 0.0029\,/\mathrm{s}$. The time-step used for the test is $\Delta t = 120\,\mathrm{s}$ and the grid-point spacing is $\Delta x = \Delta y = 5000\,\mathrm{m}$. These values have been estimated from typical forecast values near strong baroclinic developments obtained with the operational 5×5 km high-resolution limited-area model (HIRLAM) forecasting system run at the Danish Meteorological Institute. However, the HIRLAM D_0 and R_0 values have been multiplied by a factor 10 in order to visualize the effect of divergence and rotation. The "exact" trajectories and departure cell are shown on Fig. 2.9 ("exact" refers to a 18-digit precise computation of the departure

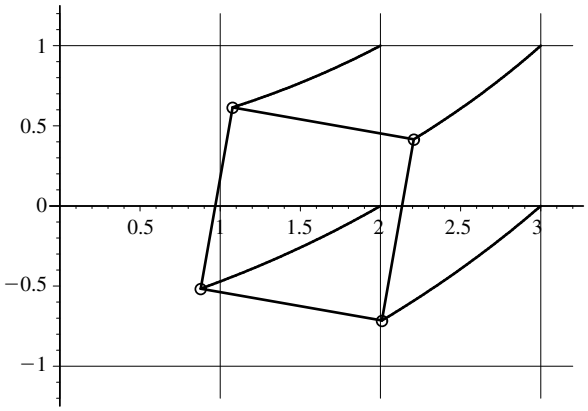


FIG. 2.9 'Exact' departure cell and backward trajectories (curved lines) for the analytic velocity field consisting of a translational, divergent, and rotational part. The unfilled circles are the departure points computed with the trajectory algorithm of LAURITZEN, KAAS and MACHENHAUER [2006]. The values on the x- and y-axis are in units of 5000 m.

points using a Fehlberg fourth–fifth order Runge-Kutta method). Three error measures are used to measure the degree of local mass conservation:

- The area of the departure cell normalized by the exact departure cell area.
- The area located outside the exact departure cell normalized by the exact departure area.
- The area located inside the exact departure cell normalized by the exact departure area.

Figure 2.10 shows the departure cells of the different DCISL schemes and the exact departure cell for the parameters listed above. For this flow field, the departure cell is a polygon with straight-line walls, and hence the departure area of scheme of RANČIĆ [1992] is exact if exact trajectories are used.

In Table 2.2, the error measures for the degree of local mass conservation are shown for three different DCISL schemes. The trajectory algorithm of LAURITZEN, KAAS and MACHENHAUER [2006] has been used for the computation of the departure points. The deviation from unity of the first error measure (column 1 in Table 2.2) for the MACHENHAUER and OLK [1998] scheme is due to the fact that the departure points are not exact. Had the trajectories been exact, the MACHENHAUER and OLK [1998] scheme would have had the first error measure equal to one. With respect to the first error measure, all DCISL schemes are equally accurate for this particular test case. The cascade schemes are more local than the fully 2D scheme of MACHENHAUER and OLK [1998] in terms of the two remaining error measures. Hence, there is less of the departure cell located outside the exact departure cell and less of the exact departure area that is not included in the schemes departure area. Of the two cascade schemes, the SLICE scheme is most accurate for this particular flow case. Note that the order of the 1D sweeps is reversed in the SLICE scheme compared to the NAIR, SCROGGS and SEMAZZI [2002] scheme.

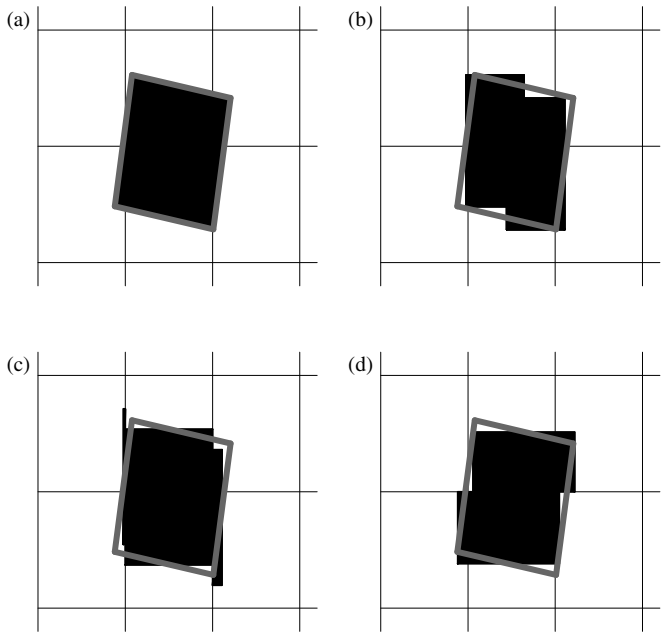


FIG. 2.10 The departure cells (dark area) when using the scheme of (a) RANČIĆ [1992], (b) MACHENHAUER and OLK [1998], (c) the cascade scheme of NAIR, SCROGGS and SEMAZZI [2002] and (d) the cascade scheme of ZERROUKAT, WOOD and STANIFORTH [2002] for the idealized test case for assessing the degree of local mass-conservation. The departure areas are based on the departure points computed with the trajectory scheme of LAURITZEN, KAAS and MACHENHAUER [2006]. The grey lines are the "exact" departure cell walls.

TABLE 2.2
Error measures for the degree of local mass conservation for the schemes of MACHENHAUER and OLK [1998] (CISL), NAIR, SCROGGS and SEMAZZI [2002] (CCS), and ZERROUKAT, WOOD and STANIFORTH [2005] (SLICE) for the analytic flow field described in Section 2

Scheme	Departure area/(exact departure area)	(Area outside exact departure area)/(exact departure area)	Area mising inside exact departure area)/(exact departure area)
CISL	1.0009	0.1124	0.1124
CCS	1.0009	0.0813	0.0805
SLICE	1.0009	0.0778	0.0769

It is important to note that the above example does by no means substitute a general analysis including a statistically large number of departure cells in realistic flows. Therefore, one should not use the analysis to draw general conclusions on the relative accuracy of the three DCISL schemes. For instance, part of the advantage of the NAIR, SCROGGS and SEMAZZI [2002] scheme over the MACHENHAUER and OLK [1998] scheme is, in the case shown, due to the fact that the flow is convergent. Consequently, the departure cell

consists of three rectangles, so there are two "jumps" in the north and south walls, respectively. The MACHENHAUER and OLK [1998] scheme only has one "jump" in the north and south wall. For a divergent flow field, this advantage would no longer appear. Similarly, the direction of the cascade sweeps influences the degree of local mass conservation.

In Section 2.3.2.1, the "locality" of the flux-based scheme of LIN and ROOD [1996] is assessed on the present test case. The results in Fig. 2.11 show the actual areas of information needed to obtain a forecast using that transport scheme and the wind field in Eq. (2.24). It can be seen in this case that the effective departure area is substantially more spread out than those of the DCISL schemes in Fig. 2.10.

2.3.1.4. Stability analysis Although the PPM is a widely used numerical method, there has not been performed a Von Neumann stability analysis of that method, as far the

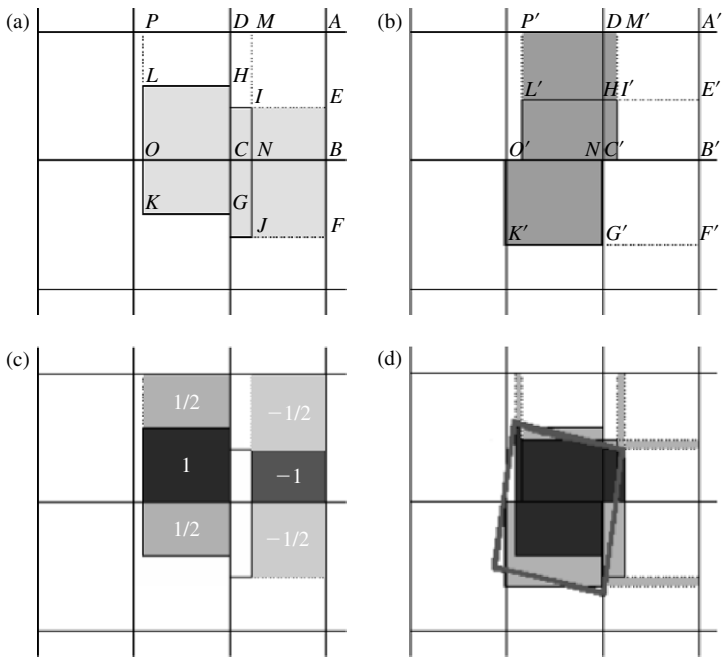


FIG. 2.11 A graphical illustration of the LIN and ROOD [1996] scheme for the idealized test case for assessing the degree of local mass conservation. The arrival cell is the north-eastern most regular grid cell in all plots. The capital letters on (a) and (b) refer to the vertices located south-west of the letter in question except for J' and N' that refer to the vertice to the south-east of the letter in question. The notation $\overline{ABCD}$ will refer to the average value in the cell with vertices at A, B, C, and D. (a) and (c) illustrate $X_C\left[\tfrac{1}{2}\left(\overline{\psi}^n + \overline{\psi}_{AY}\right)\right]$, where $\overline{\psi}_{AY}$ (yellow area) is computed using an advective operator. (a) Following the conceptual illustration of LEONARD, LOCK and MACVEAN [1996], $X_C\left[\tfrac{1}{2}\left(\overline{\psi}^n + \overline{\psi}_{AY}\right)\right]$ is given by $\tfrac{1}{2}\left(\overline{DCOP} + \overline{HGKL}\right) - \tfrac{1}{2}\left(\overline{ABNM} + \overline{EFJI}\right)$. (c) shows the cell averages with weight one (dark blue), half (light blue), minus one (red), and minus half (light red), for the contribution from X_C. (b) Similarly for Y_C, we get that $Y_C\left[\tfrac{1}{2}\left(\overline{\psi}^n + \overline{\psi}_{AX}\right)\right] = \tfrac{1}{2}\left(\overline{BF'G'C} + \overline{N'J'K'O'}\right) - \tfrac{1}{2}\left(\overline{AE'H'D} + \overline{M'I'L'P'}\right)$ and the green area is $\overline{\psi}_{AX}$. (d) shows the final forecast with the same coloring as in (c). The red rectangle is the exact departure area. (See also color insert).

authors are aware. Rather the stability of the schemes has been demonstrated numerically. Here, a stability analysis of the PPM2 in 1 and 2Ds using DCISL schemes is given and is further detailed in LAURITZEN [2007].

First, consider the 1D situation in which all the DCISL schemes are identical. The stability analysis and the notation used here are similar to that used in the stability analysis of traditional grid-point semi-Lagrangian schemes presented in BATES and MCDONALD [1982]. If we assume a constant flow u (without loss of generality assume u positive), then the west cell wall of cell i, located at $x_i = i\Delta x$, departs from

$$(x_i)_* = x_i - u\Delta t. \tag{2.25}$$

Similarly for the right cell wall. Let integer p be such that $(x_i)_*$ is located in between $(i - p - 1)\Delta x$ and $(i - p)\Delta x$, and define

$$\alpha = \frac{u\Delta t}{\Delta x} - p. \tag{2.26}$$

For a constant flow and if a piecewise constant subgrid-cell distribution is used, then the forecast is given by

$$\overline{\psi}_i^+ = (1 - \alpha)\overline{\psi}_{i-p} + \alpha\overline{\psi}_{i-p-1}, \tag{2.27}$$

which is identical to the traditional semi-Lagrangian grid-point scheme using linear Lagrange interpolation under the assumption that the grid-point values represent cell averages. Assume a solution in the form

$$\psi_i = \psi^0\Gamma^n \exp\left(\hat{i}kx\right), \tag{2.28}$$

where Γ is the complex amplification factor, ψ^0 is the initial amplitude, k is the wave number, and $\hat{i}$ is the imaginary unit. Since a cell-integrated scheme is based on cell averages,

$$\overline{\psi}_i = \int_{i\Delta x}^{(i+1)\Delta x} \psi^0\Gamma^n \exp\left(\hat{i}kx\right)\mathrm{d}x \tag{2.29}$$

must be evaluated and the resulting expression substituted into Eq. (2.28). It may easily be shown that the squared modulus of the amplification factor can be written as

$$|\Gamma|^2 = 1 - 2\alpha(1 - \alpha)(1 - \cos k\Delta x). \tag{2.30}$$

This is the same result as would have been obtained for a traditional semi-Lagrangian grid-point scheme using linear Lagrange interpolation (see, e.g., BATES and MCDONALD [1982]). For all resolvable wavelengths, the scheme is stable $|\Gamma|^2 \leq 1$ as long as $0 \leq \alpha \leq 1$. By definition, α is within that range, and hence the 1D DCISL scheme using a piecewise constant subgrid-scale reconstruction is unconditionally stable.

Using the PPM2 for the subgrid-scale distribution with no filters and assuming a constant wind field, the forecast can be written as

$$\begin{aligned}\overline{\psi}_i^+ = &\frac{1}{12}\alpha^2(1-\alpha)\overline{\psi}_{i-p-3} - \frac{1}{12}\alpha(1+7\alpha)(1-\alpha)\overline{\psi}_{i-p-2} \\ &- \frac{1}{3}\alpha\Big(4\alpha^2 - 5\alpha - 2\Big)\overline{\psi}_{i-p-1} - \frac{1}{3}(1-\alpha)\Big(4\alpha^2 - 3\alpha - 3\Big)\overline{\psi}_{i-p} \\ &- \frac{1}{12}\alpha(1-\alpha)(8-7\alpha)\overline{\psi}_{i-p+1} + \frac{1}{12}\alpha(1-\alpha)^2\,\overline{\psi}_{i-p+2}. \end{aligned} \tag{2.31}$$

By performing a Von Neuman stability analysis, it may be shown, after some algebra, that the squared modulus of the amplification factor can be written as

$$\begin{aligned}|\Gamma|^2 = 1 &+ \frac{2}{9}\alpha^2\Big(4\alpha^2 - 4\alpha - 7\Big)(1-\alpha)^2 - \frac{8}{9}\alpha^2\Big(4\alpha^2 - 4\alpha - 5\Big)(1-\alpha)^2\cos k\Delta x \\ &+ \frac{1}{9}\alpha^2\Big(50\alpha^2 - 50\alpha - 39\Big)(1-\alpha)^2\cos^2 k\Delta x \\ &- \frac{2}{9}\alpha^2\Big(19\alpha^2 - 19\alpha - 7\Big)(1-\alpha)^2\cos^3 k\Delta x \\ &+ \frac{1}{9}\alpha^2\Big(14\alpha^2 - 14\alpha - 1\Big)(1-\alpha)^2\cos^4 k\Delta x + \frac{2}{9}\alpha^3(1-\alpha)^3\cos^5 k\Delta x. \end{aligned} \tag{2.32}$$

Since DCISL schemes approximate the integral over the departure area explicitly, the amplification factors in Eq. (2.31) and Eq. (2.32) are not a function of p. Figure 2.12 shows $|\Gamma|^2$ for the four shortest wavelengths for the DCISL scheme using piecewise constant subgrid-cell reconstructions and PPM2, and for comparison, the squared modulus of the amplification factor for the traditional semi-Lagrangian scheme using cubic interpolation. Apart from the $2\Delta x$-wave, the higher order subgrid-cell reconstruction leads to a much less damping scheme compared with the lowest order scheme (as expected). The shortest resolvable wave ($2\Delta x$-wave) is, however, severely damped with the DCISL scheme based on PPM2, which might explain why schemes based on PPM2 do not exhibit excessive noise problems near sharp gradients even without applying filters. It can be demonstrated numerically that the scheme is unconditionally stable for all wavelengths when $0 \leq \alpha \leq 1$ (which is satisfied by definition).

The above analysis is directly extended to 2Ds. Assume a constant flow field (u, v) where the velocity components are positive, let p and q be integers such that the southwest vertice of cell (i, j) is located in the Eulerian cell with indices $(i - p - 1, j - r - 1)$ (see Fig. 7 in Bates and Mcdonald [1982]), α is defined in Eq. (2.26) and

$$\beta = \frac{v\Delta t}{\Delta y} - r. \tag{2.33}$$

Here, only the fully 2D schemes of Machenhauer and Olk [1998] and the cascade scheme of Nair, Scroggs and Semazzi [2002] are considered. Note, however, that for a constant flow field the cascade scheme of Nair, Scroggs and Semazzi [2002] and

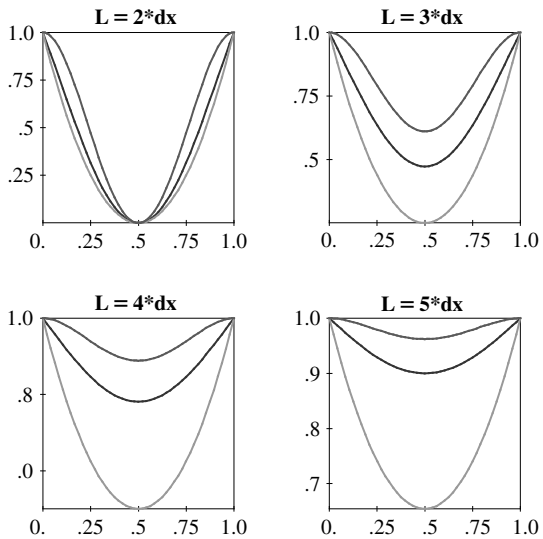


FIG. 2.12 Squared modulus of the amplification factor as a function of α for the (a) $2\Delta x$, (b) $3\Delta x$, (c) $4\Delta x$, and (d) $5\Delta x$ waves. Red and green lines are for the DCISL scheme using PPM2 and piecewise constant subgrid-cell representation, respectively. For comparison, the squared modulus of the amplification factor for the traditional semi-Lagrangian scheme based on cubic Lagrange interpolation (blue line) is shown as well. (See also color insert).

ZERROUKAT, WOOD and STANIFORTH [2002] are identical, apart from the order of the polynomial used for the subgrid-scale reconstructions. When using piecewise constant subgrid-scale reconstructions, the explicit forecast formula for all DCISL schemes is given by

$$\begin{aligned}\overline{\psi}^{+}_{i,j} = {} & (1-\alpha)(1-\beta)\overline{\psi}_{i-p,j-r} + \alpha(1-\beta)\overline{\psi}_{i-p-1,j-r} \\ & + \beta(1-\alpha)\overline{\psi}_{i-p,j-r-1} + \alpha\beta\overline{\psi}_{i-p-1,j-r-1}. \end{aligned} \tag{2.34}$$

Again, the formula is equivalent to the forecast for the traditional semi-Lagrangian scheme using bilinear Lagrange interpolation under the assumption that grid-point values represent cell averages (see, e.g., BATES and MCDONALD [1982]). Assuming a solution in the form

$$\psi_{i,j} = \psi^0 \Gamma^n \exp\left[\hat{i}\,(kx + ly)\right], \tag{2.35}$$

where k and l are the components of the wave number vector, then the mean value of the solution over cell (i, j) is given by

$$\overline{\psi}_{i,j} = \int_{j\Delta y}^{(j+1)\Delta y} \left\{ \int_{i\Delta x}^{(i+1)\Delta x} \psi^0 \Gamma^n \exp\left[\hat{i}\,(kx + ly)\right] \mathrm{d}x \right\} \mathrm{d}y. \tag{2.36}$$

Substituting Eq. (2.36) into Eq. (2.34), the complex amplification factor can be written as

$$\Gamma = \left\{1 - \alpha\left[1 - \exp\left(-\hat{i}k\Delta x\right)\right]\right\}\left\{1 - \beta\left[1 - \exp\left(-\hat{i}l\Delta y\right)\right]\right\}\exp\left[-\hat{i}\,(p\Delta x + r\Delta y)\right]. \tag{2.37}$$

It may be easily verified that $|\Gamma|^2 \leq 1$ for $0 \leq \alpha \leq 1$ and $0 \leq \beta \leq 1$, hence the 2D scheme using piecewise constant subgrid-cell reconstructions is unconditionally stable.

For the higher order schemes, Maple software has been used to compute the explicit forecast formulas and for performing the stability analysis. The explicit formula for the forecast when using the scheme of MACHENHAUER and OLK [1998] and NAIR, SCROGGS and SEMAZZI [2002] can be written as a weighted sum

$$\overline{\psi}^{+}_{i,j} = \sum_{h=-2}^{3}\sum_{g=-2}^{3} C_{h,g}(\alpha, \beta)\overline{\psi}^{n}_{i-p-1+h,\,j-r-1+g} \tag{2.38}$$

The coefficients are listed in Tables 2.3 and 2.4 for the MACHENHAUER and OLK [1998] and NAIR, SCROGGS and SEMAZZI [2002] schemes, respectively. The formula for the squared modulus of the amplification factor resulting from the Von Neumann stability analysis is too lengthy to display here. Instead, plots of $|\Gamma|^2$ for selected wave numbers are shown (see Fig. 2.13). See LAURITZEN [2007] for further details. The cascade scheme of NAIR, SCROGGS and SEMAZZI [2002] is slightly more damping than the fully 2D scheme of MACHENHAUER and OLK [1998] for the shortest resolvable traverse wave (Fig. 2.12(a)), and the situation is vice versa for the next shortest resolvable traverse wave (Fig. 2.12(b)) as well as longer wavelengths. It has been verified numerically that the squared modulus of the amplification factor is less than or equal to unity for $0 \leq \alpha \leq 1$ and $0 \leq \beta \leq 1$, i.e., the 2D DCISL schemes are unconditionally stable and less diffusive compared to a traditional semi-Lagrangian scheme based on bi cubic interpolation.

2.3.1.5. Extension of DCISL schemes to spherical geometry The singularities on the sphere are one of the main challenges for transport schemes formulated on conventional latitude-longitude grids, and most algorithms require a certain amount of "engineering" to tackle the pole problem, which often reduces the efficiency and simplicity of the algorithms. As already mentioned, the number of schemes developed in cartesian geometry is significantly larger than the number of schemes formulated for spherical geometry. For the DCISL schemes discussed here, only the scheme of MACHENHAUER and OLK [1998], NAIR, SCROGGS and SEMAZZI [2002], and ZERROUKAT, WOOD and STANIFORTH [2002] have been extended to the sphere.

In cartesian geometry, the most accurate approximation to a departure cell, given the departure points, is the polygon resulting from connecting the departure points with straight lines. Similarly, in spherical geometry, the cells defined by connecting the departure points with great circle arcs seem as the optimal choice. But as in cartesian geometry, integrating along the optimal curves leads to complicated and computationally expensive algorithms. Therefore, as in the cartesian case, the area approximation must be simplified.

TABLE 2.3

The coefficients $C_{h,g}$ written in "matrix" format for the explicit forecast formula in case of a constant wind field when using the scheme of MACHENHAUER and OLK [1998] (see Eq. (2.38)). The index h is in the first column and the second index g is in the first row

h, g	-2	-1	0	1	2	3
-2	0	0	$\frac{1}{12}\alpha^2\beta(1-\alpha)$	$\frac{1}{12}\alpha^2(1-\alpha)(1-\beta)$	0	0
-1	0	0	$-\frac{1}{12}\alpha\beta(1-\alpha)(1+7\alpha)$	$-\frac{1}{12}\alpha(1-\alpha)\times$ $(1-\beta)(1+7\alpha)$	0	0
0	$\frac{1}{12}\alpha\beta^2(1-\beta)$	$-\frac{1}{12}\alpha\beta(1-\beta)\times$ $(1+7\beta)$	$-\frac{1}{3}\alpha\beta\times$ $\left(4\alpha^2-5\alpha+4\beta^2-5\beta-1\right)$	$-\frac{1}{3}\alpha\,(1-\beta)\times$ $\left(4\alpha^2-5\alpha+4\beta^2-3\beta-2\right)$	$-\frac{1}{12}\alpha\beta\times$ $(1-\beta)\,(8-7\beta)$	$\frac{1}{12}\alpha\beta\,(1-\beta)^2$
1	$\frac{1}{12}\beta^2\,(1-\alpha)\,(1-\beta)$	$-\frac{1}{12}\beta\,(1-\beta)\times$ $(1-\alpha)\,(1+7\beta)$	$-\frac{1}{3}\beta\,(1-\alpha)\times$ $\left(4\alpha^2-3\alpha+4\beta^2-5\beta-2\right)$	$-\frac{1}{3}\,(1-\alpha)\,(1-\beta)\times$ $\left(4\alpha^2-3\alpha+4\beta^2-3\beta-3\right)$	$-\frac{1}{12}\beta\,(1-\alpha)\times$ $(1-\beta)\,(8-7\beta)$	$\frac{1}{12}\beta\,(1-\alpha)\,(1-\beta)^2$
2	0	0	$-\frac{1}{12}\alpha\beta\,(1-\alpha)\,(8-7\alpha)$	$-\frac{1}{12}\alpha\,(1-\alpha)\times$ $(1-\beta)\,(8-7\alpha)$	0	0
3	0	0	$\frac{1}{12}\alpha\beta\,(1-\alpha)^2$	$\frac{1}{12}\alpha\,(1-\alpha)^2\,(1-\beta)$	0	0

TABLE 2.4

The coefficients $C_{h,g}$ written in "matrix" format for the explicit forecast formula in case of a constant wind field when using the scheme of NAIR, SCROGGS and SEMAZZI [2002] (see Eq. (2.38)). The index h is in the first column and the second index g is in the first row.

	-2	-1	0	1	2	3
-2	$\frac{1}{144}\alpha^2\beta^2(1-\alpha)\times$ $(1-\beta)$	$-\frac{1}{144}\alpha^2\beta(1-\alpha)\times$ $(1-\beta)\,(1+7\beta)$	$-\frac{1}{36}\alpha^2\beta(1-\alpha)\times$ $\left(4\beta^2-5\beta-2\right)$	$-\frac{1}{36}\alpha^2(1-\alpha)(1-\beta)\times$ $\left(4\beta^2-3\beta-3\right)$	$-\frac{1}{144}\alpha^2\beta(1-\alpha)\times$ $(1-\beta)\,(8-7\beta)$	$\frac{1}{144}\alpha^2\beta(1-\alpha)\times$ $(1-\beta)^2$
-1	$-\frac{1}{144}\alpha\beta^2(1-\alpha)\times$ $(1-\beta)(1+7\alpha)$	$\frac{1}{144}\alpha\beta(1-\alpha)\times$ $(1-\beta)(1+7\alpha)\times$ $(1+7\beta)$	$\frac{1}{36}\alpha\beta(1-\alpha)\times$ $(1+7\alpha)\times$ $\left(4\beta^2-5\beta-2\right)$	$\frac{1}{36}\alpha(1-\alpha)\times$ $(1-\beta)(1+7\alpha)\times$ $\left(4\beta^2-3\beta-3\right)$	$\frac{1}{144}\alpha\beta(1-\alpha)\times$ $(1-\beta)\,(8-7\beta)\times$ $(1+7\alpha)$	$-\frac{1}{144}\alpha\beta(1-\alpha)\times$ $(1-\beta)^2(1+7\alpha)$
0	$-\frac{1}{36}\alpha\beta^2(1-\beta)\times$ $\left(4\alpha^2-5\alpha-2\right)$	$\frac{1}{36}\alpha\beta(1-\beta)\times$ $\left(4\alpha^2-5\alpha-2\right)\times$ $(1+7\beta)$	$\frac{1}{9}\alpha\beta\times$ $\left(4\alpha^2-5\alpha-2\right)\times$ $\left(4\beta^2-5\beta-2\right)$	$\frac{1}{9}\alpha(1-\beta)\times$ $\left(4\alpha^2-5\alpha-2\right)\times$ $\left(4\beta^2-3\beta-3\right)$	$\frac{1}{36}\alpha\beta(1-\beta)\times$ $\left(4\alpha^2-5\alpha-2\right)\times$ $(8-7\beta)$	$-\frac{1}{36}\alpha\beta(1-\beta)^2\times$ $\left(4\alpha^2-5\alpha-2\right)$
1	$-\frac{1}{36}\beta^2(1-\alpha)\times$ $(1-\beta)\times$ $\left(4\alpha^2-3\alpha-3\right)$	$\frac{1}{36}\beta(1-\alpha)(1-\beta)\times$ $(1+7\beta)\times$ $\left(4\alpha^2-3\alpha-3\right)$	$\frac{1}{9}\beta(1-\alpha)\times$ $\left(4\alpha^2-3\alpha-3\right)\times$ $\left(4\beta^2-5\beta-2\right)$	$\frac{1}{9}(1-\alpha)(1-\beta)\times$ $\left(4\alpha^2-3\alpha-3\right)\times$ $\left(4\beta^2-3\beta-3\right)$	$\frac{1}{36}\beta(1-\alpha)(1-\beta)\times$ $\left(4\alpha^2-3\alpha-3\right)\times$ $(8-7\beta)$	$-\frac{1}{36}\beta(1-\alpha)\times$ $(1-\beta)^2\times$ $\left(4\alpha^2-3\alpha-3\right)$
	-2	-1	0	1	2	3
-2	$\frac{1}{144}\alpha^2\beta^2(1-\alpha)\times$ $(1-\beta)$	$-\frac{1}{144}\alpha^2\beta(1-\alpha)\times$ $(1-\beta)\,(1+7\beta)$	$-\frac{1}{36}\alpha^2\beta(1-\alpha)\times$ $\left(4\beta^2-5\beta-2\right)$	$-\frac{1}{36}\alpha^2(1-\alpha)(1-\beta)\times$ $\left(4\beta^2-3\beta-3\right)$	$-\frac{1}{144}\alpha^2\beta(1-\alpha)\times$ $(1-\beta)\,(8-7\beta)$	$\frac{1}{144}\alpha^2\beta(1-\alpha)\times$ $(1-\beta)^2$

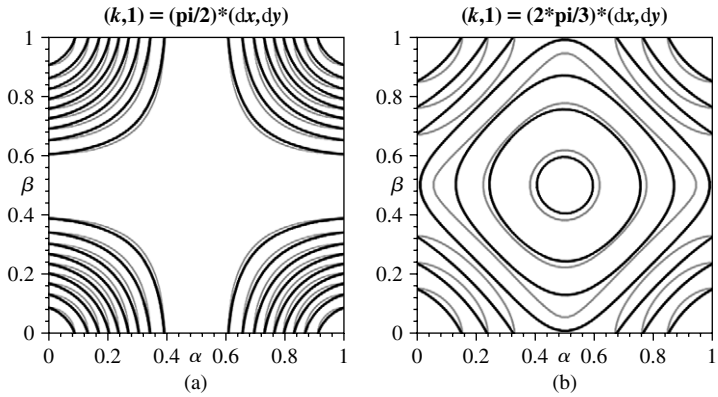


FIG. 2.13 Squared modulus of the amplification factor as a function of (α, β) for (a) $(L_x, L_y) = 2\,(\Delta x, \Delta y)$ and (b) $(L_x, L_y) = 3\,(\Delta x, \Delta y)$, where L_x is the wavelength in the x-direction and similarly for L_y. Black contours are for the scheme of MACHENHAUER and OLK [1998] and grey contours the scheme of NAIR, SCROGGS and SEMAZZI [2002]. The contour-interval is 0.1 and contours start at 0.9 at the corners and decrease toward the center of the plot. The two schemes show similar damping properties.

The MACHENHAUER and OLK [1998] scheme is extended to spherical geometry by using the μ-grid originally introduced by MACHENHAUER and OLK [1996] (NAIR and MACHENHAUER [2002]). The μ-grid is a latitude-longitude grid in which the latitude θ is replaced by $\mu = \sin(\theta)$ (see Fig. 2.14). This transformation is area preserving, and the μ-grid is essentially a cartesian grid where the latitude grid lines are no longer equidistant. The departure cells are defined as quadrilaterals on the (λ, μ)-plane exactly as in cartesian geometry, i.e., the cell walls that in cartesian geometry were parallel to x and y isolines are parallel to the longitudes and latitudes on the μ-grid, respectively. Hence, away from the poles, this transformation is invariant in the sense that the corresponding upstream integrals and departure cells take exactly the same form as in cartesian geometry. Since the algorithm is formally equivalent on the μ-grid, only minor modifications of the algorithm in cartesian geometry are needed away from the poles. In the vicinity of the poles, however, approximating the cells with straight-line walls on the μ-grid is a poor approximation of the cells on the spherical latitude-longitude grid (see, e.g., Fig. 4c in NAIR and MACHENHAUER [2002]). Especially the exact north and south walls are deviating significantly from linearity on the (λ, μ)-grid, and hence some "engineering" is needed. Local tangent planes at the poles are introduced for more accurate cell-approximations. The areas in which the tangent planes are used are referred to as the *polar caps*. Ideally, the integration should be performed along straight lines on the tangent planes. Instead, more latitudes are introduced in the polar caps, and the coordinates of the cell vertices on the tangent plane are transformed into (λ, μ)-coordinates, and thereafter, the integrals are performed along straight lines in the (λ, μ) plane. In the Lagrangian belt containing the pole point (referred to as the *singular belt*), the algorithm breaks down since the Lagrangian cell containing the Eulerian pole is not well defined (see Fig. 2.14). The total mass inside the singular belt can, however, be computed and is distributed among the cells in a mass conservative way using a regular semi-Lagrangian

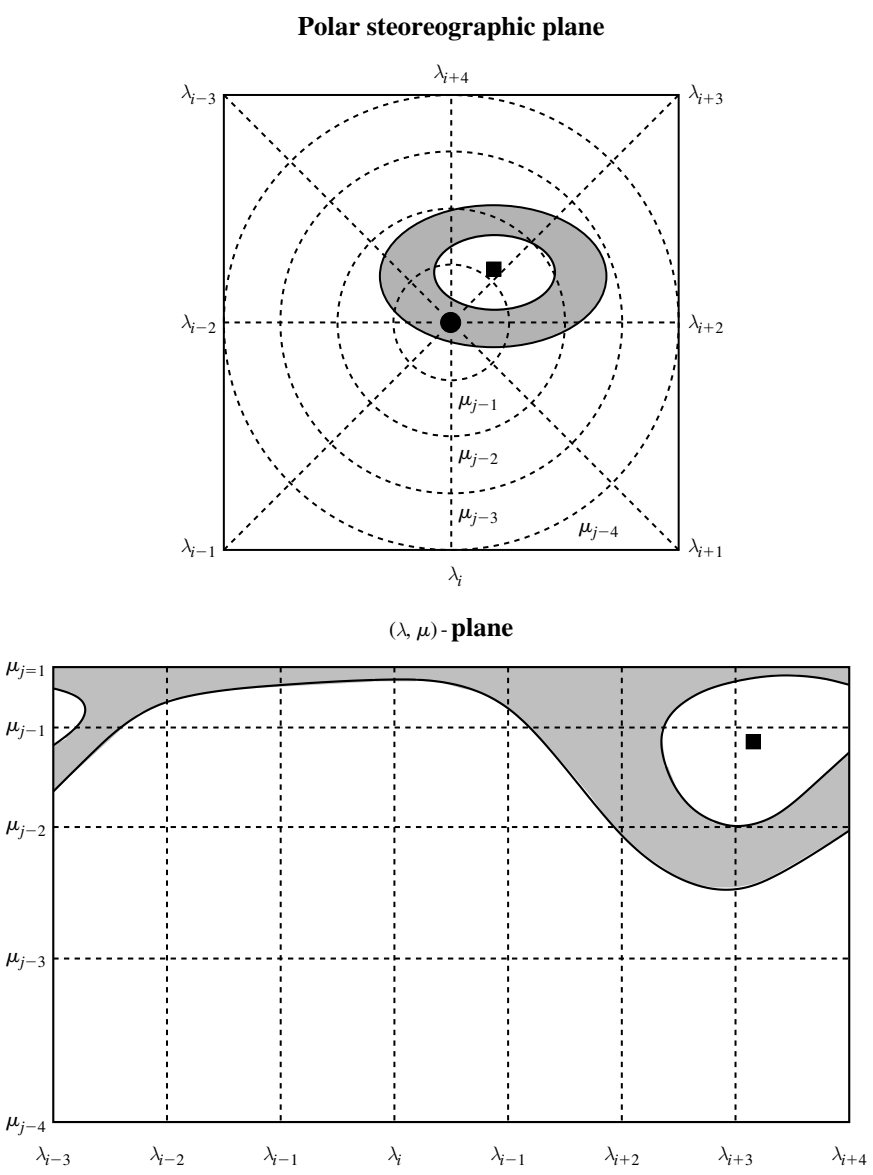


FIG. 2.14 A graphical illustration of the polar cap treatment in the scheme of NAIR and MACHENHAUER [2002]. The upper plot shows the polar stereographic projection of the Eulerian cells (bounded by the dashed lines which are the λ and μ isolines), and the singular belt (shaded region). The singular belt is the set of departure cells bounded by two consecutive Lagrangian latitudes that contain the Eulerian pole point (filled circle). On the lower plot, the Eulerian cells and the singular belt are plotted on the (λ, μ)-plane. Note that the pole point (filled circle on upper plot) is the line $\mu_j = 1$ on the lower plot. The filled square is the Lagrangian pole point.

method. The method computes the densities at the approximate departure cell centers using a quasi-bicubic interpolation. These values are used as weights for distributing the total mass in the singular belt among the cells, i.e., the point value of the density at a given departure cell center, normalized by the sum of all the departure cell point values, determines the fraction of the total mass which the cell in question is attributed (see NAIR and MACHENHAUER [2002] for additional details).

The cascade scheme of NAIR, SCROGGS and SEMAZZI [2002] has been extended to the sphere using the μ grid as well but with two different treatments of the polar cap. The first method used cascade interpolation throughout the spherical domain except for the Lagrangian belts over the Eulerian poles, but the scheme was limited by the meridional Courant number, which must be less than unity in the version presented in NAIR, SCROGGS and SEMAZZI [2002]. Later, the scheme was adapted to large meridional Courant numbers by using the cascade approach away from the polar caps and by using the fully 2D scheme of NAIR and MACHENHAUER [2002] over the polar caps (NAIR [2004]).

Away from the *poles*, the extension of the cascade remapping method from cartesian geometry to the μ-grid is straightforward. The computation of the intermediate grid, or equivalently, the crossings of the Lagrangian latitudes and Eulerian longitudes are computed using cubic Lagrange interpolation on the μ-grid. The mass is transferred to the intermediate grid and from there to the Lagrangian grid, exactly as in the cartesian case, but simply on the μ-grid. Since the meridional Courant number is less than unity, the only problematic zone with ill-defined cells (singular belt) is made up by the cells north of the first Lagrangian latitude that ends up at the first Eulerian latitude circle after one time-step (similarly for the Southern hemisphere). As in the case of the NAIR and MACHENHAUER [2002] scheme, the total mass in the singular belt can be computed and the total mass can be redistributed to the individual cells as explained above. Contrary to the NAIR and MACHENHAUER [2002] scheme, the cascade scheme does not use high-resolution polar belts. Only the singular belts are treated differently from the rest of the domain.

For general applications, this restriction on the meridional Courant number is a severe limitation. NAIR [2004] suggested the use of the efficient cascade method of NAIR, SCROGGS and SEMAZZI [2002] away from the poles and the polar cap treatment of NAIR and MACHENHAUER [2002] in the zones where the cascade method would break down (north of the Lagrangian latitude closest to the Lagrangian pole point and similarly for the Southern hemisphere). Hereby the severe meridional Courant number restriction is alleviated.

The SLICE scheme is extended to spherical geometry by using a regular latitude-longitude grid (ZERROUKAT, WOOD and STANIFORTH [2004]). The intermediate grid is computed in spherical coordinates by using the great circle approach of NAIR, CÔTÉ and STANIFORTH [1999b] (for details see Section 2b in this reference), which is more efficient though less accurate than the cubic Lagrange interpolation on the μ-grid used in NAIR, CÔTE and STANIFORTH [2002]. The cascade method breaks down when not all Lagrangian longitudes intersect an Eulerian latitude. Consequently, there are some intermediate cells that are ill defined, i.e., the intermediate cell walls are not both well

defined. Consider the situation in which the western intermediate cell wall exists but not the eastern one, while the east cell wall of intermediate cell $(i + 1, j)$ is well defined. After the first cascade sweep, the total mass between the west wall of intermediate cell (i, j) and the east wall of intermediate cell $(i + 1, j)$ is known. The total mass is split in two and allocated to the nearest Lagrangian mass centers used in the second cascade sweep. This is not accurate but the redistribution is mass conservative. Some of the lost accuracy is recovered with a "post fix" procedure similar to the singular belt treatment in NAIR and MACHENHAUER [2002]. Here, however, the mass in the entire polar cap is redistributed mass-conservatively using Lagrange weights.

For solid body rotation over the poles, the error measures shown in Table 2.1 do not indicate a superior method for treating the pole problem. Even though SLICE uses a subgrid-reconstruction that, under the assumption that no filters are invoked, is one order of magnitude higher than PPM2 used in the scheme of NAIR and MACHENHAUER [2002] and NAIR, SCROGGS and SEMAZZI [2002], SLICE is not superior with respect to error measures l_1 and l_2. This also suggests that the polar treatment reduces the accuracy of the scheme.

2.3.2. Flux-based FV schemes in 2Ds

Up to the beginning of the 1990s, when RANČIĆ [1992] presented the first semi-Lagrangian DCISL scheme, all FV schemes used in meteorology were flux-based in nature.[2] In flux-based schemes, the prognostic equation for the volume-specific scalar ψ is obtained as a sum of estimates of inward and outward fluxes in the Eulerian grid cell. Generally assuming an Eulerian grid cell to be a polygon with L faces, the differential FV prognostic equation (1.6) for the total "mass" for this particular cell can be written as follows:

$$\overline{\psi}^{n+1} \Delta A = \overline{\psi}^{n} \Delta A + \sum\nolimits_{l=1}^{L} m_l^L, \tag{2.39}$$

where $\overline{\psi}$ is the cell average "density," n is the time step index, ΔA is the area, and m_l^L is the total inward mass flux integrated over a time step for face l. m_l^L is defined negative if the net flow through face l is outward and positive for inward fluxes. The conservation is ensured if m_l^L is unique for face l, i.e., the mass that leaves a cell through face l is exactly gained in the neighboring cells sharing face l.

As pointed out by HIRSCH [1990], FV schemes of the type in Eq. (2.39) were introduced by GODUNOV [1959], and they were first used in meteorological applications by CROWLEY [1968]. Since then, the schemes have gradually evolved with increasing sophistication, and they have been used extensively in recent decades in both meteorology and oceanography. We will not go into great detail regarding the entire historical development of flux-based FV methods and their application on the sphere. Instead the focus is on some aspects that are important for understanding the methodology and how it has evolved into the most modern schemes.

To introduce the basic ideas behind flux-based FV schemes, consider at first the continuity equation in 1D, x, without any source terms. Define the flux convergence

[2]The process of integrating over the departure cell, or equivalently, the remapping or rezoning of mass between two grids, was, however, studied already in the 1970's (e.g., HIRT, AMSDEN and COOK [1974]).

operator

$$X_C\left(u, \Delta t, \Delta x; \overline{\psi}^n\right) = \left[F_w^n + F_e^n\right]/\Delta x \tag{2.40}$$

for a given Eulerian cell with extension Δx. Indices w and e indicate the left ("western") and right ("eastern") cell boundary, and F is the time-integrated mass flux related to the flow speed u, i.e., F corresponds to the flux "m" in the multidimensional case (Eq. (2.39)). Since the total fluxes depend on the flow speed, the time-step Δt, and Δx, this is also the case for X_C. Implicitly X_C also depends on the Eulerian cell averages since they are used for the reconstruction of the subgrid-cell distributions. For schemes where also higher order moments or cell face values are prognostic variables (e.g., PRATHER [1986], XIAO and YABE [2002]), the subgrid-cell representation also depends on these moments or values. In the limit as Δt and Δx approach zero, the operator X_C divided by Δt is the FV approximation to the term $-\partial(u\psi)/\partial x$. Expressed in terms of the flux convergence operator, the 1D version of Eq. (2.39) becomes

$$\overline{\psi}_i^{n+1} = \overline{\psi}_i^n + X_{i_C}(\overline{\psi}^n) \tag{2.41}$$

for the Eulerian grid cell i, omitting the obvious dependence on u, Δt, and Δx for brevity. The operator X_C redistributes mass between the Eulerian grid cells. By definition, application of the X_C operator does not change the total mass in the integration domain since the left-flux of cell i always cancels the right-flux of cell $i-1$ ($F_{i-1_e} = -F_{i_w}$). In all FV schemes, the fluxes are obtained as integrals – or as approximations to integrals – of the total mass in the length interval being "swept though" the face within the time-step Δt. This leads to the following "general" equation for X_C:

$$X_C\left(\overline{\psi}^n\right) = \frac{1}{\Delta x}\left[\int_{x_w - u_w^* \Delta t}^{x_w} \psi^n \mathrm{d}x - \int_{x_e - u_e^* \Delta t}^{x_e} \psi^n \mathrm{d}x\right]$$
$$\left(= \frac{1}{\Delta x}\left[F_w + F_e\right]\right), \tag{2.42}$$

where u_w^* is the effective advection speed for the area (interval) 'swept though' the left face within the time-step from $n\Delta t$ to $(n+1)\Delta t$. Similarly u_e^* is for the right face. Note, that the integrant ψ^n in (2.42) in each grid cell (i) is an analytic function of x determined as a constant, the cell mean value $\overline{\psi}_i^n$, plus a subgrid-scale deviation from this mean value. The deviation inside grid cell (i) is determined by $\overline{\psi}_i^n$ and cell mean values of surrounding cells, in the case of a parabolic representation, of two cells on each side. As mentioned in Section 2.2, the accuracy of FV schemes will depend on how accurate the fluxes are estimated. Assuming exact effective advective speeds, the accuracy of the scheme will therefore depend only on the order of the ψ subcell representation. To maintain mass conservation, obviously u_{i-1e}^* must equal u_{iw}^* for any pair $(i-1, i)$ of grid cells. For flux-based transport schemes, the advective speed used to estimate the mass flux F_l^n is the flow speed at the spatial location of face l. Ideally this flow speed should be a time mean value over Δt. In practice it is evaluated as a local forecast or a simple

extrapolation valid at time $(n + ½)\Delta t$. For a 1D flow with no spatial variations and where Eq. (2.42) is used to determine X_C, Eq. (2.41) becomes identical to a 1D DCISL scheme. The reader is referred to Section 1 for a general description of the analogy between DCISL and flux-based schemes. One of the main differences between the flux-based FV schemes and the DCISL schemes is that the departure points is identified from trajectories (i.e., characteristics) estimated iteratively in both time and space to better include the effect of spatial as well as temporal variations in the flow. A similar accurate estimation of the true departure area is generally not part of the flux-based schemes.

The generalization of the 1D FV schemes to 2Ds can be done in two fundamentally different ways: via a more or less direct estimation of the 2D-fluxes – to be described below in Section 2.3.2.3 – or via *operator splitting*. In the operator splitting method Section 2.3.2.1 – the transport problem is split into a combination of operators in each of the two coordinate directions.

2.3.2.1. Operator splitting Consider for simplicity only cartesian x-y coordinates. In this case, the individual conservative flux convergence operators are in the x-direction (Eq. (2.40)) and in the y-direction,

$$Y_C\left(v, \Delta t, \Delta y; \overline{\psi}^n\right) = \left[G_s^n + G_n^n\right]/\Delta y, \tag{2.43}$$

where v is the spatially varying speed in the y-direction, Δy is the grid extension in the y-direction while s and n denote the lower ("southern") and upper ("northern") face of the grid cell. G denotes the fluxes related solely to the translations in the y-direction.

A simple-minded operator splitting is where the fluxes in each direction are treated independently as simultaneous 1D fluxes:

$$\overline{\psi}_{i,j}^{n+1} = \overline{\psi}_{i,j}^{n} + X_{i,jC}\left(\overline{\psi}^n\right) + Y_{i,jC}\left(\overline{\psi}^n\right), \tag{2.44}$$

where i and j is the spatial index in the x- and y-direction, respectively. This scheme is inherently mass-conserving, and as noted by Leonard, Lock and Macvean [1996], it is also stable when the flux is calculated using a first-order – or so-called donor cell – method. However, as shown by Leith [1965], a scheme of a type in Eq. (2.44) is unstable when second-order polynomials are used for the subgrid-cell representation in Eq. (2.42) and in the corresponding expression for flux convergence in the y-direction,

$$Y_C\left(\overline{\psi}^n\right) = \frac{1}{\Delta y}\left[\int_{y_s - v_s^* \Delta t}^{y_s} \psi^n \mathrm{d}y - \int_{y_n - v_n^* \Delta t}^{y_n} \psi^n \mathrm{d}y\right]$$
$$\left(= \frac{1}{\Delta y}\left[G_s + G_n\right]\right), \tag{2.45}$$

where v^* is the effective advection speed in the y-direction.

It is not surprising that the simple-minded update becomes unstable: the effective departure area being split into two separate areas in the upstream x and y directions. What is needed to achieve stability is a sequential flux splitting instead of the simultaneous flux

splitting in (2.44). This means that the transport problem is first solved in one coordinate direction and the resulting field is subsequently transported in the transverse coordinate direction. Thereby the transport through a certain cell wall is determined from both the velocity parallel to the cell wall, by the first transport step, and from the velocity normal to the cell wall, by the second transport step. As a result an unbroken and more realistic departure area is obtained. Inspired by the notation in LEONARD, LOCK and MACVEAN [1996], consider first the intermediate transport problem in the x-direction (omitting for simplicity the grid cell indexing):

$$\overline{\psi}_{CX} = \overline{\psi}^{n} + X_C\left(\overline{\psi}^{n}\right). \tag{2.46}$$

By definition, the total mass is conserved after this intermediate forecast. The subsequent second update step in the y-direction becomes

$$\overline{\psi}^{n+1}_{CXY} = \overline{\psi}_{CX} + Y_C\left(\overline{\psi}_{CX}\right), \tag{2.47}$$

or equivalently (by inserting Eq. (2.46) into Eq. (2.47)):

$$\overline{\psi}^{n+1}_{CXY} = \overline{\psi}^{n} + X_C\left(\overline{\psi}^{n}\right) + Y_C\left(\overline{\psi}_{CX}\right). \tag{2.48}$$

Here the argument $\overline{\psi}_{CX}$ to the Y_C operator is obtained as integrals of the subgrid-cell representation in the y-direction of the $\overline{\psi}_{CX}$ field. Since only conservative operators have been applied, the total mass is unchanged.

The algorithm in Eq. (2.48) introduces a directional bias. Therefore, in practical applications, it has been common procedure to alternate between the directional splitting in (2.48) and the opposite sequential splitting:

$$\overline{\psi}^{n+1}_{CYX} = \overline{\psi}^{n} + Y_C\left(\overline{\psi}^{n}\right) + X_C\left(\overline{\psi}_{CY}\right) \tag{2.49}$$

with

$$\overline{\psi}_{CY} = \overline{\psi}^{n} + Y_C\left(\overline{\psi}^{n}\right). \tag{2.50}$$

The operator split schemes in Eq. (2.48)/Eq. (2.49) are also referred to as *time-split schemes*. Alternatively, instead of alternating Eq. (2.48) and Eq. (2.49), one can, of course, combine the operators and define a spatially symmetric conservative scheme as

$$\begin{aligned}\overline{\psi}^{n+1} &= \tfrac{1}{2}\left(\overline{\psi}^{n+1}_{CXY} + \overline{\psi}^{n+1}_{CYX}\right)\\ &= \overline{\psi}^{n} + X_C\left[\tfrac{1}{2}\left(\overline{\psi}^{n} + \overline{\psi}_{CY}\right)\right] + Y_C\left[\tfrac{1}{2}\left(\overline{\psi}^{n} + \overline{\psi}_{CX}\right)\right].\end{aligned} \tag{2.51}$$

Schemes of the general type Eq. (2.48)/Eq. (2.49)/Eq. (2.51) have been presented in several papers, e.g., TREMBACK, POWELL, COTTON and PIELKE [1987], BOTT [1989, 1992].

There is a fundamental time split error – often referred to as lumpiness – associated with schemes of this type. This splitting error is a consequence of the fact that a constant ψ-field is generally not conserved with these schemes as it should in a spatially but divergence free flow (i.e. a flow with $\partial u/\partial x = -\partial v/\partial y \neq 0$). This lake of constancy conservation of the scheme (2.46)–(2.47) arises because $\overline{\psi}_{CX}$ in (2.47) is a result of not only an advection but also a deformation in the x-coordinate direction. This is illustrated in Fig. 2.15 for the operator splitting scheme in Eqs. (2.46) and (2.47). The original constant density ψ_0 is indicated by the level of shading in the left two columns in the left panel (grid cells with x-index, $i-1$ and $i-2$). The effect of the operation (Eq. (2.46)), i.e., $\overline{\psi}_{CX}$, is shown for the cells with x-index i (only) as the level of shading. The effective deformation in the x-direction for the arrival cell marked i, j is a convergence $-D_0$ leading to increased values with darker shading. Around the arrival cell $i, j-2$, no deformation takes place in the x-direction and the shading is therefore unchanged. Finally, in the arrival cell $i, j-3$ deformation (i.e. divergence) is leading to a decreased value and consequently a lighter shading. The horizontal dotted lines in the right panel indicate the limits of the subsequent transport in the y-direction through the "southern" and "northern" walls, respectively, of cell i, j. Since the y-extension of this area is less than one (i.e., divergence D_0). However, since the air that is transported into the arrival cell i, j comes from a place with density less than ψ_0 the final forecasted value $\overline{\psi}^{n+1}_{CYX}$ (not shown) in cell i, j will end up being less than the original value ψ_0 which it should have retained.

The operator splitting error problem leads to serious error growth for transport by deformational flows. Petschek and Libersky [1975] showed that a kind of numerical instability is associated with the time-splitting. In simulations of highly compressible fluids, the splitting error seems to be of less importance (e.g., Woodward and Colella [1984], Carpenter, Droegemeier, Hane and Woodward [1990], Colella [1990]). However, in most geophysical applications, the splitting error must be explicitly dealt with to obtain sufficiently accurate simulations.

Leonard, Lock and Macvean [1996] and Lin and Rood [1996], independently, introduced essentially the same technique to eliminate the splitting error. The following derivation leads to the same expressions as those originally presented by these authors. Here, however, the focus is on the motivation behind the basic idea: ensuring the contribution from flow deformations to the final forecasted value are excluded in the initial transports parallel to the cell faces. The first ingredient is to define those contributions to $X_c \approx -\Delta t\, \partial\big(u\overline{\psi}\big)/\partial x$ and $Y_c \approx -\Delta t\, \partial\big(v\overline{\psi}\big)/\partial y$ that are related to flow deformations in each direction ($TDx \approx -\Delta t\, \overline{\psi}\partial u/\partial x$ and $TDy \approx -\Delta t\, \overline{\psi}\partial v/\partial y$) and those related to advection ($TAx \approx -\Delta t\, u\partial\overline{\psi}/\partial x$ and $TAy \approx -\Delta t\, v\partial\overline{\psi}/\partial y$). For given conservative flux convergences, one can at first define either the deformation or the advection contributions and then determine the other pair using the following relationships:

$$\begin{aligned} X_C\Big(\overline{\psi}^n\Big) &= TAx + TDx, \\ Y_C\Big(\overline{\psi}^n\Big) &= TAy + TDy. \end{aligned} \tag{2.52}$$

The advective updates are defined in terms of TAx and TAy:

$$\begin{aligned} \overline{\psi}_{AX} &= \overline{\psi}^n + TAx, \\ \overline{\psi}_{AY} &= \overline{\psi}^n + TAy. \end{aligned} \tag{2.53}$$

Lin [2004] defined initially deformations by the following centered approximations (here expressed in cartesian coordinates):

$$\begin{aligned} TDx &= -\Delta t \overline{\psi}^n \frac{u_e^* - u_w^*}{\Delta x}, \\ TDy &= -\Delta t \overline{\psi}^n \frac{u_n^* - u_s^*}{\Delta y}. \end{aligned} \tag{2.54}$$

The associated advective contributions to the forecasts are then defined from the conservative flux-form operators as follows:

$$TAx = X_C - TDx, \tag{2.55}$$

$$TAy = Y_C - TDy, \tag{2.56}$$

respectively. The terms TAx and TAy are referred to as "advective" since they are approximations to $-\Delta t u \partial \overline{\psi}^n / \partial x$ and $-\Delta t v \partial \overline{\psi}^n / \partial y$, respectively.

When initially advective contributions are defined, e.g., as in Lin and Rood [1996], one will typically define them as:

$$\begin{aligned} TAx &= \overline{\psi}_{AX} - \overline{\psi}^n = \frac{1}{\Delta x} \int_{x_w - \overline{u}^* \Delta t}^{x_e - \overline{u}^* \Delta t} \psi^n \mathrm{d}x - \overline{\psi}^n \approx -\Delta t \overline{u}^* \frac{\partial \psi^n}{\partial x}, \\ TAy &= \overline{\psi}_{AY} - \overline{\psi}^n = \frac{1}{\Delta y} \int_{y_s - \overline{v}^* \Delta t}^{y_n - \overline{v}^* \Delta t} \psi^n \mathrm{d}y - \overline{\psi}^n \approx -\Delta t \overline{v}^* \frac{\partial \psi^n}{\partial y}, \end{aligned} \tag{2.57}$$

where $\overline{u}^* = 1/2(u_w^* + u_e^*)$ and $\overline{v}^* = 1/2(v_n^* + v_s^*)$. In this case,

$$\begin{aligned} TDx &= X_C - TAx \\ &= \frac{1}{\Delta x} \left[\int_{x_w - u_w^* \Delta t}^{x_w} \psi^n \mathrm{d}x - \int_{x_e - u_e^* \Delta t}^{x_e} \psi^n \mathrm{d}x \right] + \overline{\psi}^n - \frac{1}{\Delta x} \int_{x_w - \overline{u}^* \Delta t}^{x_e - \overline{u}^* \Delta t} \psi^n \mathrm{d}x \\ &= \frac{1}{\Delta x} \left[\int_{x_w - u_w^* \Delta t}^{x_w} \psi^n \mathrm{d}x + \int_{x_w}^{x_e} \psi^n \mathrm{d}x + \int_{x_e}^{x_e - u_e^* \Delta t} \psi^n \mathrm{d}x \right] - \frac{1}{\Delta x} \int_{x_w - \overline{u}^* \Delta t}^{x_e - \overline{u}^* \Delta t} \psi^n \mathrm{d}x \\ &= \frac{1}{\Delta x} \left[\int_{x_w - u_w^* \Delta t}^{x_e - u_e^* \Delta t} \psi^n \mathrm{d}x - \int_{x_w - \overline{u}^* \Delta t}^{x_e - \overline{u}^* \Delta t} \psi^n \mathrm{d}x \right] \\ &\approx -\Delta t \psi^n \frac{u_e^* - u_w^*}{\Delta x} \end{aligned} \tag{2.58}$$

with a similar expression for TDy.

Once the advective update in the x-direction $\overline{\psi}_{AX}$ has been calculated from (2.53) and (2.57), one can consider a provisional conservative transport of this update in the y-direction:

$$\overline{\psi}_{CYAX} = \overline{\psi}_{AX} + Y_C\left(\overline{\psi}_{AX}\right). \tag{2.59}$$

Note that generally

$$\iint\limits_{totaldomain} \overline{\psi}_{CYAX}\mathrm{d}x\mathrm{d}y = \iint\limits_{totaldomain} \overline{\psi}_{AX}\mathrm{d}x\mathrm{d}y \neq \iint\limits_{totaldomain} \overline{\psi}^{n}\mathrm{d}x\mathrm{d}y, \tag{2.60}$$

where the last integral represents the total mass. To achieve mass conservation, we need to re-add the TDx field:

$$\begin{aligned}\overline{\psi}_{YX}^{n+1} &= \overline{\psi}_{CYAX} + TDx \\ &= \overline{\psi}_{AX} + Y_C\left(\overline{\psi}_{AX}\right) + TDx \\ &= \overline{\psi}_{CX} + Y_C\left(\overline{\psi}_{AX}\right) \\ &= \overline{\psi}^{n} + X_C\left(\overline{\psi}^{n}\right) + Y_C\left(\overline{\psi}_{AX}\right).\end{aligned} \tag{2.61}$$

Like the (2.48) forecast the (2.61) forecast is conservative since it only includes conservative operators. However, the difference is that the field $\overline{\psi}_{AX}$, which is finally transported in the north-south direction into the i, j-cell, has been advected but not deformed. With $\psi^n \equiv \psi_0$ in (2.61) it is easily verified that

$$\overline{\psi}_{YX}^{n+1} = \psi_0\left(1 - \Delta t\left[\frac{\Delta u}{\Delta x} + \frac{\Delta v}{\Delta y}\right]\right) = \psi_0 \quad \text{if} \left[\frac{\Delta u}{\Delta x} + \frac{\Delta v}{\Delta y}\right] \equiv 0.$$

Here $\Delta u = u_e - u_n$ and $\Delta v = v_n - v_s$. Thus, a constant is exactly conserved if the discretized divergence $\left[\frac{\Delta u}{\Delta x} + \frac{\Delta v}{\Delta y}\right] \equiv 0$.

As for Eq. (2.48), a directional bias is introduced by Eq. (2.61). This can be compensated by averaging the update $\overline{\psi}_{YX}$ in Eq. (2.61) with the equivalent update $\overline{\psi}_{XY}$ in the opposite direction leading to the symmetric expression:

$$\begin{aligned}\overline{\psi}^{n+1} &= \tfrac{1}{2}\left(\overline{\psi}_{YX} + \overline{\psi}_{XY}\right) \\ &= \overline{\psi}^{n} + X_C\left[\tfrac{1}{2}\left(\overline{\psi}^{n} + \overline{\psi}_{AY}\right)\right] + Y_C\left[\tfrac{1}{2}\left(\overline{\psi}^{n} + \overline{\psi}_{AX}\right)\right].\end{aligned} \tag{2.62}$$

An alternative and slightly cheaper approach than Eq. (2.62) is to alternate between $\overline{\psi}_{YX}$ and $\overline{\psi}_{XY}$ in each second time-step.

Considering the forecast in Eq. (2.61) as an example, it is noteworthy to observe that the contributions to the divergence term $-\Delta t\overline{\psi}(\Delta u/\Delta x + \Delta v/\Delta y)$ are determined from field values at different locations for each of the two directions. It is obvious that a contribution comes from $\overline{Y}_C(\overline{\psi}_{AX})$ at a location upstream in the y-direction, and due

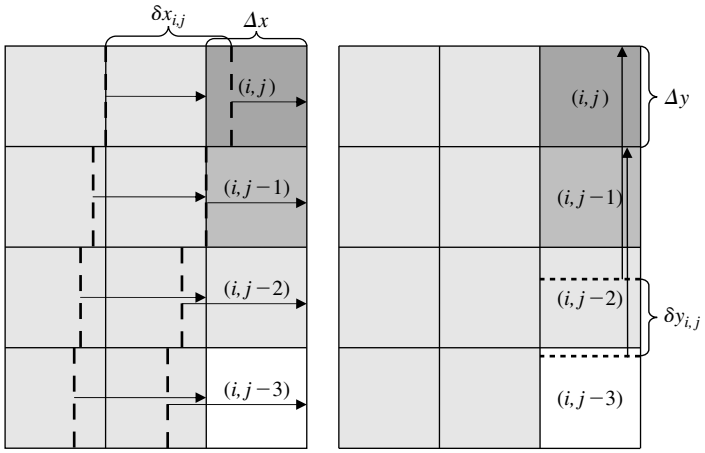


FIG. 2.15 Schematic illustration of the mass conservative but not constancy conserving time splitting in Eqs. (2.46)/(2.47) for a nondivergent flow. The left panel illustrates the intermediate forecast of $\overline{\psi}_{CX}$ for the ith column of grid cells. The upstream departure areas $\delta x = \Delta x + \Delta t\left(u_e^* - u_w^*\right)$ arriving in the ith column are indicated with dashed lines. The shading in column i indicates the level of flow deformation $(-(\delta x - \Delta x) / (\Delta t \Delta x))$ related to the flow in the x-direction *only*, with dark shading indicating strong "convergence" and light shading "divergence." The right panel illustrates the final forecast (Eq. (2.47)) of ψ_{CXY}^{n+1} for grid cell (i, j). Here, the upstream departure area $\delta y_{i,j} = \Delta y + \Delta t\left(v_e^{*i,j} - v_w^{*i,j}\right)$ is illustrated with dotted lines. The shading in the right panel is identical to that in the left, i.e., the final forecast in (i, j) is not indicated with shading. It is obvious, however, that it would be lighter than the original shading in the two grid columns $j-1$ and $j-2$. Thus, the scheme is not constancy conserving. Note that since flow is nondivergent, we have

$$D \approx -\left(\delta x_{i,j} - \Delta x\right) / (\Delta t \Delta x) - \left(\delta y_{i,j} - \Delta y\right) / (\Delta t \Delta y) = 0.$$

to the definition of $\overline{Y}_C$ ($\overline{\psi}_{AX}$) the ultimate origin of this contribution is upstream in both the y- and x-directions as it should be according to the "exact solution". Contrary to this, the term *TDx* gives a contribution to the divergence term that is based on field values at a different location: if *TDx* is chosen as the primary definition and the definition in Eq. (2.54) is used, it is based on the value $\overline{\psi}^n$ at the location of the arrival Eulerian cell, and if *TAx* is chosen as the primary definition and (2.58) apply, it is based on the value $\overline{\psi}^n$ at a location $\Delta t \overline{u}^*$ upstream in the x-direction. Similar arguments apply to $\overline{\psi}_{XY}$, and therefore the forecast in Eq. (2.62) includes some (small) contributions from the $\overline{\psi}^n$ field either at the location of the Eulerian arrival cell (when *TDx* and *TDy* are chosen as primary definition) or upstream in each of the two directions (when *TAx* and *TAy* are primaries). Fig. 2.11 shows the actual departure area for a forecast using the transport schemes by LIN and ROOD [1996] and the wind field in Eq. (2.24). It can be seen that for this particular case, there is a considerable spread out of the departure area relative to the DCISL schemes in Fig. 2.10. A careful inspection of Fig. 2.11 shows that the net departure area is displaced systematically towards "south-east" as compared to the "exact departure area". This is related to curvature of the trajectories in Fig. 2.9. As for the DCISL schemes the shift of mass inside grid cells, due to calculated sub-grid-scale deviations from the cell average value $\overline{\psi}^n$ in cells crossed by the boundary of the departure area, may result in a larger or a smaller effective departure area from which mass is contributed to $\overline{\psi}^{n+1}$,

compared to the departure area shown in Fig. 2.11. However, in any case the global mass is conserved and the sub-grid-scale distributions serve to reduce any artificial damping.

The approaches by LEONARD, LOCK and MACVEAN [1996] and LIN and ROOD [1996] to eliminate the operator – or time – splitting error may have been inspired by the earlier techniques introduced by BOTT [1993] and EASTER [1993]. BOTT [1993] applied flux limiters (see Section 2.6) to obtain monotonicity-preserving transport, but otherwise, the basic idea is the same as explained above. One fundamental difference is, however, that the BOTT [1993] scheme does not permit time-steps exceeding the CFL criterion such as the schemes by LEONARD, LOCK and MACVEAN [1996] and LIN and ROOD [1996].

EASTER [1993] introduced an alternative way of eliminating the splitting error in the original positive-definite scheme by BOTT [1989]. His approach is equivalent to that by LEONARD, LOCK and MACVEAN [1996] and LIN and ROOD [1996]: the 1D conservative transport in the first direction of the operator splitting will generally change the fluid density due to fluid deformations and due to pure advection. For tracer mixing ratio, only pure advection along tracer gradients can change the value. By eliminating the deformational part, one can isolate mixing ratio transport and subsequently during the transport in the second direction include the deformational contribution in a consistent way: the estimated total deformation in each of the two directions for a given cell is based on the local flow around this cell.

Schemes of the type presented above have been quite popular in recent years (e.g., RASCH and LAWRENCE [1998], LIN [2004]) and have and will be been implemented into atmospheric models ranging from meso-scale models (e.g., SKAMAROCK, KLEMP, DUDHIA, GILL, BARKER, WANG and POWERS [2007]) to general circulation models GCMs (e.g., ADCROFT, CAMPIN, HILL and MARSHALL [2004]. Note, however, that by far most applications have been offline, i.e., passive advection of tracers in models using a different scheme for the solution of the continuity equation in the dynamical core. For such models, the need for special attention in relation to the mass-wind inconsistency problem – see Section 3 – is often even more important than in the online models mentioned here.

As shown by LIN and ROOD [1996], the flux-based schemes derived above will normally lead to conservation of linear correlations between the mixing ratios q_a and q_b of two tracers a and b, i.e., if $\overline{q}_b^n = \alpha \overline{q}_a^n + \beta$ then $\overline{q}_b^{n+1} = \alpha \overline{q}_a^{n+1} + \beta$ where α and β are constants. This is because the flux convergences and transports of the types of Eqs. (2.42), (2.45), and (2.54) satisfy the general linear relationships:

$$\begin{aligned} Z(\psi + \beta) &= Z(\psi) + Z(\beta), \\ Z(\alpha\psi) &= \alpha Z(\psi). \end{aligned} \tag{2.63}$$

The above relationships will in general also apply to DCISL schemes based on, e.g., the PPM method.

The conservation of linear relationships between different tracers is an attractive feature in, e.g., chemical modeling because it prevents artificial chemical reactions in idealized situations where the mixing ratio within a domain of one tracer can be expressed as a linear function of another. Note, however, that for schemes where the upstream subgrid-cell representation is forced positive-definite or monotonic, Eq. (2.63) is generally not fulfilled. It should also be mentioned that no such thing as linear relationships exist

between tracers in nature. Generally, the flux-based schemes (and the DCISL schemes) do, therefore, not conserve the local relative concentrations or mixing ratios between two tracers. In particular, this is the case in regions where one tracer has a reasonably smooth behavior while the other is dominated by a sharp variation of the spatial gradient.

In practice, the mixing ratio for a tracer is used as prognostic variable in many flux-based schemes. HOURDIN and ARMENGAUD [1999] used a scheme quite similar to that by LIN and ROOD [1996], where a few requirements on the spatial behavior of the mixing ratio were sufficient to ensure both monotonicity and positive definiteness. Generally, schemes based on mixing ratio should conserve a constant mixing ratio, although enforcement of positive definiteness and monotonicity in such schemes may deteriorate correlations between mixing ratios. Note that schemes based solely on mixing ratio will generally not conserve the total mass of the tracer unless care is taken to conserve total mass of the air in a consistent way. If flux-limiters (see Section 2.6) or a priori constraints on the subcell representation are applied solely to mixing ratios, one will lose mass conservation unless special additional constraints are imposed. In meteorological models the atmospheric air contain more or less water vapour so *specific concentration* (see definition page 3) is used instead of *mixing ratio*, defined relative to dry air.

2.3.2.2. Stability of operator-split, flux-based schemes The operator-split, flux-based schemes have often been subject to a CFL criterion (i.e., $\Delta t < \Delta x/u^*$) in 1D. One can identify two main reasons for this:

- The upstream subgrid-cell representation needed to estimate the face fluxes was defined locally from the grid cells neighboring the target Eulerian cell. This means that longer time steps led to extrapolation (and not aggregation) of information and hence the CFL criterion. It was not considered to apply "semi-Lagrangian" thinking.
- Some schemes are quite heavily hooked up on localized flux limiters. This makes it difficult to generalize into far upstream constraints. Examples are the otherwise popular and accurate schemes by BOTT [1989, 1992, 1993].

As pointed out by LEONARD [1994], there are, however, no immediate scientific reasons to limit the integration domain to the neighboring grid cells for schemes where the face fluxes are based on pure upstream integrals of "mass".[3] If the subgrid-cell information is defined everywhere without extrapolation, the face fluxes consist of interpolation/aggregation of information and they will be stable as shown in the 1D PPM case for the DCISL schemes (Section 2.3.1.4).

The stability of low-order versions of 2D flux-based semi-Lagrangian schemes was cursorily investigated by LIN and ROOD [1996].[4] Later, LAURITZEN [2007] made a detailed stability analysis of both higher and lower order versions of the LIN and ROOD

[3] However, depending on the machine architecture and the actual scheme it may in some case be more efficient to require that characteristics depart from the immediate neighbouring cells because then there is no need for searching.

[4] Note that equations (A.5) and (B.1) in LIN and ROOD [1996] are missing some terms, but that the correct formulas have the same generic form. See the Appendix in LAURITZEN [2007].

[1996] class of schemes as well as a conceptual analysis to explain the results. The LIN and ROOD [1996] class of schemes is given by Eq. (2.62), and the different schemes in the class are formed by varying the order of the 1D operators that are applied. The flux-form operators applied to the terms in the square brackets in Eq. (2.62) are referred to as *outer* operators, whereas the operators used for $\overline{\psi}_{AX}$ and $\overline{\psi}_{AY}$ are referred to as the *inner* operators. When the *inner* and *outer* operators differ, LAURITZEN [2007] showed in a linear Von Neumann stability analysis that increased damping (or weak instability) may result, but this spurious damping disappears when the operators are identical (similarly for phase errors). This is due to the fact that for Courant numbers larger than unity there can be contributions to the forecast not originating from the Lagrangian departure area as they physically should. If the operators are identical and under the assumptions applied in a Von Neumann stability analysis, the LIN and ROOD [1996] scheme becomes formally identical to the NAIR, SCROGGS and SEMAZZI [2002] and ZERROUKAT, WOOD and STANIFORTH [2002] schemes, and therefore only includes information from the departure cell. For more details, see LAURITZEN [2007].

2.3.2.3. Explicit estimation of the 2D fluxes Most flux-based schemes have used the technique of time- or operator splitting described in Section 2.3.2.1, and the approach has proven to be very efficient and economic. There will, however, always be a slight inconsistency since the splitting prevents integration of the exact departure area. To reduce these problems, several papers (e.g., DUKOWICZ and RAMSHAW [1979], SMOLARKIEWICZ [1984], BELL, DAWSON and SHUBIN [1988], COLELLA [1990], DUKOWICZ and KODIS [1987], SMOLARKIEWICZ and GRABOWSKI [1990], RASCH [1994], LEONARD [1994], HÓLM [1995], DUKOWICZ and BAUMGARDNER [2000]) have investigated the possibility of constructing fully 2D flux-based schemes. In these schemes, one aims directly at an estimate of the transport in the "cross-directions", which was taken care of by the sequential approach in the operator split methods.

The fully 2D flux-based schemes are similar to fully 2D DCISL schemes since they – more or less directly – are based on estimates of integrals over upstream areas to obtain the mass interactions with all the neighboring grid cells. One example is the scheme by HÓLM [1995] in cartesian geometry, which for a given grid cell is based on four unique fluxes in the x-direction, the y-direction, and the two cross-directions (see Fig. 2.16).

The scheme proposed by RASCH [1994] appears to be somewhat different. This scheme is based on an upwind biased stencil of points that are used to define an upstream spatial interpolation of the same type as that in semi-Lagrangian models.[5] However, making use of certain symmetry rules in the upstream polynomials, this interpolation can be formulated in the traditional flux form (Eq. (2.39)) for a given Eulerian grid cell. In other words, constraints on the polynomial coefficients ensure that the implied fluxes are unique for each face.

Contrary to the operator-split flux-form FV schemes, it is complicated to circumvent the CFL criterion for fully 2D flux-based schemes. Referring to the discussion in Section 1 on the analogies between flux-form and DCISL schemes, a fully 2D semi-Lagrangian flux-form scheme would in fact be the same as a DCISL scheme.

[5] Note, however, that the scheme by RASCH [1994] is Eulerian and therefore subject to a CFL criterion.

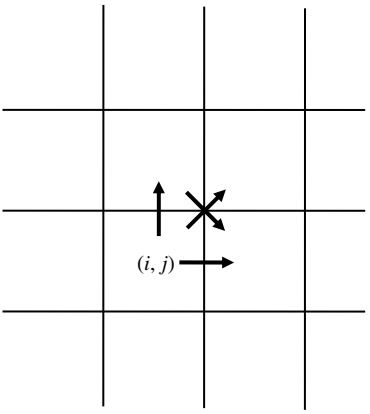


FIG. 2.16 Schematic illustration of the four unique fluxes needed in fully 2D flux-based finite-volume schemes.

2.4. Locally mass conserving semi-Lagrangian grid point methods

Recently, a few locally mass conserving upstream or downstream interpolating semi-Lagrangian grid point schemes have been proposed in the literature. Mass conservation has been achieved via local modifications of the polynomial interpolations in such a way that the total mass is always conserved. Effectively the prognostic grid point variable in such schemes is the average density in Eulerian grid cells, and therefore these schemes can be considered special types of semi-Lagrangian FV schemes.

We have already – in Section 2.3.1.1 – mentioned the scheme by LAPRISE and PLANTE [1995] where a downstream semi-Lagrangian scheme was modified along these lines. The scheme by RASCH [1994] – although Eulerian – is an example of an upstream mass conserving scheme based on modifications of the polynomial coefficients. More recently, KAAS [2008], COTTER, FRANK, REICH [2007] and REICH [2007] have proposed upstream and downstream grid point semi-Lagrangian schemes which are locally mass conserving. The basic idea is to modify the upstream or downstream polynomial interpolation coefficients.

For an upstream traditional semi-Lagrangian scheme – following KAAS [2008] – these coefficients can be considered area (or volume in the 3D case) weights transferring information from Eulerian grid points to the different irregularly spaced neighboring semi-Lagrangian departure points. The original weights in this remapping are modified by that fraction, which ensures that the sum of the weights given off by a given Eulerian grid point to all the surrounding departure points is equal to the unique area (volume) represented by this grid point. Hereby a local mass conservation is achieved when the prognostic variable is density. The forecasted densities (in the arrival Eulerian grid points), including the effects of divergence, are equal to the modified upstream interpolated values divided by the unique area (volume) represented by the arrival Eulerian grid point.

For downstream schemes as in COTTER, FRANK, REICH [2007] and REICH [2007], the procedure is opposite and somewhat more tricky since it is the distribution weights that are modified. Also REICH [2007] discusses the possibility of introducing modified upstream semi-Lagrangian weights.

The domain of dependence of locally mass conserving semi-Lagrangian schemes is comparable to that in pure DCISL schemes and depends on the actual polynomial accuracy used for the upstream interpolations/subgrid-cell representations. However, the degree of local mass conservation is higher in the DCISL schemes since the departure cell area is close to the true departure area, while in the locally mass conserving semi-Lagrangian schemes, mass is extracted from grid cells in a larger domain.

2.5. Additional prognostic variables

To improve the subgrid-cell representation needed to estimate the mass fluxes, one may introduce additional prognostic variables. VAN LEER [1977] (scheme IV) and PRATHER [1986] used traditional second-order polynomials to represent the spatial distribution and used both gradients and curvatures as additional prognostic variables to define these polynomials. This allowed for the formulation of a formally very accurate scheme that conserved second-order moments. Note, however, the arguments by THUBURN [2006] that it is not desirable to conserve second moments since one can only conserve the resolved part. Furthermore, the PRATHER [1986] scheme is very computationally demanding both in terms of CPU and memory requirements. Therefore, it has not been popular in "real" applications.

More recently, schemes have been introduced (e.g., XIAO and YABE [2002]), where not only the cell mean values but also the values and gradients at the cell interfaces are prognostic variables. These additional prognostic variables have the same role as the moments introduced by PRATHER [1986]: the reduction of the loss of information (damping) associated with the spatial remappings that are fundamental to all FV schemes. The proposed new schemes are highly accurate as the scheme by PRATHER [1986], but at significantly reduced computational cost, particularly in terms of CPU usage. The scheme has been further improved and generalized to 2 and 3Ds using directional splitting (XIAO, YABE, PENG and KOBAYASHI [2002], PENG, XIAO, OHFUCHI and FUCHIGAMI [2005]). To describe the basics behind the new so-called conservative semi-Lagrangian schemes based on rational functions (CSLR) schemes, consider transport in 1D and assume that we know the cell mean value $\overline{\psi}$ and the "west" and "east" interface values, ψ_w and ψ_e. From this information, one has three degrees of freedom to construct the subgrid-cell representation at a given time step n. One possible choice of functions could of course be the PPM. However, XIAO, YABE, PENG and KOBAYASHI [2002] found that rational functions with second-order polynomials gave better results at less numerical cost. However, the rational functions used have a built-in singularity that causes problems unless special care is taken. In the CSLR scheme, this singularity can appear when a local maximum or minimum is transported. According to XIAO, YABE, PENG and KOBAYASHI [2002], the problem can be dealt with by introducing a small machine-dependent constant that prevents division by zero at the singularity. In the 2D case, the update of the $\overline{\psi}$ values is performed as standard flux-form integrals, Eqs. (2.42) and (2.45), of the rational

functions. The cell interface values are updated using standard semi-Lagrangian upstream interpolation based on the rational functions followed by the relevant change related to the divergence of the flow. The new schemes are stable and efficient, but, of course, they will be more memory demanding since an additional prognostic variable is introduced. It is anticipated that the basic idea behind the CSLR schemes is so powerful that it will be adopted in many future integration schemes used in atmospheric models.

We see no fundamental problems in applying the powerful technique of including cell interface values as additional prognostic values in fully 2D and cascade DCISL schemes although the authors are not aware of any specific attempts along this direction.

2.6. Flux limiters

FV schemes based on polynomial unfiltered subcell representations do not, in general, fulfill requirements such as positive definiteness and monotonicity. In particular, numerical oscillations often develop near discontinuities or large variability in gradients. In Section 2.2.1, it was described how it is possible to introduce different filters or constraints on the subgrid-cell representations to reduce or eliminate these problems. In most cases – with the filter by ZERROUKAT, WOOD and STANIFORTH [2005] as an exception – the applications of such filters tend to reduce the accuracy of the schemes because of the implied clippings and smoothings of the subcell scale polynomials. We can denote these filters *a priori* filters because they are introduced before the estimation of fluxes or the upstream cell integrations. It is, however, also possible to introduce a posteriori corrections – often referred to as flux limiters – of the fluxes to ensure fulfillment of the desired properties. This type of flux corrected transport (FCT) filters was introduced by BORIS and BOOK [1973] and by ZALEZAK [1979]. The basic idea behind the classical FCT is to perform a local mixing of the fluxes obtained from a high-order scheme (which is accurate but violate the desired properties) with fluxes from a low-order highly diffusive scheme (which fulfill the properties), e.g., a simple so-called upstream scheme. The procedure is – for each cell interface – to modify the local fluxes of the diffusive scheme as much as possible toward the fluxes in the high-order scheme without exceeding the magnitude of this flux and without creating new local maxima or minima in the neighboring cells; i.e., the local fluxes are changed differently at all interfaces under the constraint that the change in neighboring cell values do not lead to changed sign of gradients in the neighboring interfaces.

Several different types of flux limiter approaches have been presented in the literature to obtain positive definiteness, e.g., BOTT [1989], or monotonicity, e.g., SMOLARKIEWICZ and GRABOWSKI [1990], BOTT [1992], RASCH [1994], HÓLM [1995], XUE [2000]. For some schemes, such as the schemes by BOTT [1989], BOTT [1992], the flux limiters are inherent parts of the basic flux calculations. HÓLM [1995] was the first to apply flux limiters directly to the fluxes in fully 2D flux-based schemes.

Although one can argue that the specific flux limiters used will be somewhat arbitrary from a physical point of view, such filters can improve the performance of transport schemes significantly at a reasonable cost although there are some logical statements and "max/min" functions involved in the algorithms. As mentioned above, limiters enforcing positive definiteness will generally not ensure conservation of mixing ratios between

different tracers. Furthermore, when flux limiters are applied in operator-split schemes permitting long time-steps, the flux limiters cannot guarantee strict multidimensional shape conservation in flows with strong deformation (LEONARD, LOCK and MACVEAN [1996]). An exception is the limiter of SKAMAROCK [2006].

2.7. Concluding remarks

Two fundamentally different FV methods in 2Ds are being used in meteorology: DCISL schemes and the flux-based methods. In DCISL schemes, the forecast for a given Eulerian cell is based on an integral over an isolated area approximating the exact upstream departure area. This means that DCISL schemes are quite direct approximations to the exact forecast that is an integral of the exact subgrid representation over the exact departure area. In flux-based methods, the forecast is obtained as the net flux of mass through each of the faces of the Eulerian cell. For each face of the cell, this flux is shared with a neighboring Eulerian cell and it is determined as an integral over the area swept through the actual face during one time-step. Although less direct than DCISL schemes, the flux-based methods also approximate an integral over the exact departure area. Therefore, the two methods are equivalent and the accuracy of both will depend on the order of the subgrid representation being integrated and the effective approximation to the "true" departure cell.

For DCISL as well as flux-based schemes, the operations related to two directions can be separated or split. For DCISL schemes, this is referred to as cascade integration, and for the flux-based schemes, it is termed operator- or time-splitting. The advantage of the splitting is that only 1D subgrid representations and integrations are required which makes these schemes considerably more efficient.

Although the flux-based schemes are generally quite accurate and conserve mass (or any integral invariant) locally, higher order subgrid representations, i.e., high accuracy, will generally violate conservation of shape, i.e., the schemes become nonmonotonic or nonpositive-definite. A number of constraints to reduce or eliminate such problems can be applied to the subgrid representations entering the upstream integrals. It is also possible to apply a posteriori corrections (e.g., so-called flux limiters) to the forecast that reduce or eliminate these problems.

DCISL schemes are by construction semi-Lagrangian and not subject to any advective CFL criterion that limits the maximum possible time-step apart from the requirement of the departure cells being well defined. In contrast to this, many traditional flux-based schemes are formulated to allow only transport over a maximum distance of one grid cell within one time-step, i.e., the Courant number must be less than unity to obtain stability. For operator-split flux-form schemes, it is, however, possible to extend the integration domain thereby avoiding the CFL criterion.

In the original time-split flux-form schemes, the lake of conservation of constant density fields in non-divergent (but deforming) flow, caused a splitting error, the so-called nudging error, The introduction of combined advective-conservastive flux-form schemes circumvented this problem. In realistic case studies with DCISL schemes no error like the flux-form nudging error has been reported, although the DCISL schemes do not conserve exactly a constant field in non-divergent flow. Such constancy conservation will be

obtained in upstream DCISL schemes only if the departure area δA is exactly equal to the arrival area ΔA, or in other words if the discretized Lagrangian divergence $D = \frac{\partial A - \Delta A}{\Delta t \Delta A}$ is exactly zero. This should be the case in any reasonably smooth divergence free flow if the determination of the departure area was exact. In praxis the departure area is determined only approximately. In present schemes it is dependent on the accuracy of the backward trajectories from the corner points of the arrival area and the strait line approximation to the exact connections between the departure corner points. LAPRISE and PLANTE [1995] found in an idealized solid body rotation experiment an accumulated relative error $(\partial A - \Delta A)/\Delta A$ over 10 time steps of up to 1%. It seems reasonably to assume that this number will be reduced substantially with a trajectory computation like that used by LAURITZEN, LINDBERG, KAAS and MACHENHAUER [2008], that takes accelerations into account. This should be investigated further. Furthermore in realistic flows it must be expected that accumulation of errors will occur less frequent and cancellation of errors will reduce the problem, if any.

For the schemes considered here, the exact departure area in deforming flows is better represented and integrated over in the DCISL schemes than in flux-based schemes. Therefore, with respect to the schemes for effective approximation to the departure area, one may conclude that DCISL schemes are generally more accurate than flux-based schemes.

One may anticipate that further developments of FV methods will include introduction of additional prognostic values and gradients at the cell interfaces as was recently proposed.

3. FV models

As stated previously, FV methods are well suited for the numerical simulation of conservation laws. This is demonstrated in Section 3.1 where a complete set of FV prognostic equations, that conserve exactly mass, entropy, total energy, and angular momentum in an adiabatic, friction-free and quasi-hydrostatic atmosphere, is derived. A numerical model based on this set of FV conservation laws, a so-called complete set of conservation laws (CSCL) model, remains to be realized. However, assuming forcing terms and Eulerian vertical discretization as in an existing operational primitive equation model, it is shown how such a prognostic system may be set up. The advection of all invariants is supposed to be calculated by an explicit, absolutely stable DCISL time-stepping scheme borrowed from LAURITZEN, KAAS, MACHENHAUER and LINDBERG [2008]. Here, in each time-step, mass and other invariants are transported conservatively along Lagrangian surfaces determined as in LAURITZEN, KAAS, MACHENHAUER and LINDBERG [2008] by 3D, so-called hybrid trajectories that are horizontally upstream (determined from the horizontal wind field) and vertically downstream (determined indirectly by the condition of hydrostatic balance).

In Sections 3.2 and 3.3, respectively, two recently developed quasi-hydrostatic dynamical cores in spherical coordinates are described, namely, the global NCAR-FFSL (National Center for Atmospheric Research – Flux Form Semi-Lagrangian) dynamical core (COLLINS, RASCH, BOVILLE, HACK, MCCAA, WILLIAMSON, KIEHL, BRIEGLEB, BITZ, LIN, ZHANG and DAI [2004]) and the limited area HIRLAM-DCISL dynamical core (LAURITZEN, KAAS, MACHENHAUER and LINDBERG [2008]). They are pioneering

examples of the two different types of FV dynamical cores developed in the meteorological modeling community. The former is based on the flux-form (Eq. (1.2)) and the latter on the Lagrangian form (Eq. (1.8)) of the continuity equation. These dynamical core examples of FV dynamical cores developed recently in the meteorological modeling community. In both dynamical cores, the continuity equation is solved by absolutely stable FV advection methods, which ensure exact mass conservation. In the HIRLAM DCISL, the remaining primitive equations are solved with finite-difference methods that do not ensure exact conservation of additional integral invariants. In the NCAR-FFSL, additionally potential temperature and absolute potential vorticity are conserved for adiabatic friction-free flow. In the HIRLAM-DCISL dynamical core, the DCISL advection scheme is combined with a semi-implicit time-stepping, thereby allowing large time-steps for all variables at the expense of solutions to elliptic Helmholtz equations each time-step (ROBERT [1969, 1981, 1982]). In the NCAR-flux-form semi-Lagrangian (FFSL) dynamical core, an explicit flux-based advection scheme is used, which means that shorter time-steps must be used for the dynamical variables, while advection of tracers (including water vapor) and physical parameterization can be predicted with long time-steps. Both dynamical cores have been coupled with comprehensive physical parameterization packages. In Section 3.4, the properties of the dynamical cores are discussed.

The main part of this section is dealing almost exclusively with complete quasi-hydrostatic atmospheric models. However, relevant aspects of online and off-line applications are taken up in Section 3.5, and finally in Section 3.6, possibilities of extensions to nonhydrostatic models are briefly discussed.

3.1. A complete set of FV conservation laws for a quasi-hydrostatic atmosphere

As shown by MACHENHAUER [1994], an explicit FV general circulation model, which conserves exactly a maximum number of fundamental integral invariants, may be formulated. Let it as above be called a CSCL model. In this section, the prognostic equations are derived and a possible explicit time-stepping procedure is presented. Finally, the feasibility of such a model is discussed.

3.1.1. The continuous primitive equations

Consider the continuous equations, the so-called primitive equations, for a general pressure-based terrain-following vertical coordinate $\eta\,(p,\, p_s)$ as formulated, for example, for the European Center for Medium Weather Research (ECMWF) integrated forecast system (IFS) model (SIMMONS and BURRIDGE [1981]) and the HIRLAM (KÄLLÉN [1996]) operational atmospheric models. The prognostic equations are:
the quasi-horizontal momentum equation

$$\frac{\mathrm{d}\vec{V}}{\mathrm{d}t} = -\nabla\phi - \alpha\nabla p - f\vec{k}\times\vec{V} + \vec{P}_{\vec{V}} + \vec{K}_{\vec{V}}, \tag{3.1}$$

the thermodynamic equation

$$c_p\frac{\mathrm{d}T}{\mathrm{d}t} = \alpha\omega + (P_T + K_T)c_p, \tag{3.2}$$

the continuity equation for moist air

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial p}{\partial \eta}\right)+\frac{\partial p}{\partial \eta}\nabla\cdot\vec{V}+\frac{\partial p}{\partial \eta}\frac{\partial \dot{\eta}}{\partial \eta}=0, \tag{3.3}$$

and *the moisture equation*

$$\frac{\mathrm{d}q_v}{\mathrm{d}t}=P_{q_v}+K_{q_v}, \tag{3.4}$$

which in combination with Eq. (3.3) may be written as

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(q_v\frac{\partial p}{\partial \eta}\right)+\left(q_v\frac{\partial p}{\partial \eta}\right)\nabla\cdot\vec{V}+\left(q_v\frac{\partial p}{\partial \eta}\right)\frac{\partial \dot{\eta}}{\partial \eta}=\frac{\partial p}{\partial \eta}\left(P_{q_v}+K_{q_v}\right). \tag{3.5}$$

The hybrid vertical coordinate $\eta\,(p,\,p_s)$, introduced by SIMMONS and BURRIDGE [1981], is a monotonic function of pressure p and surface pressure p_s such that

$$\eta(0,\,p_s)=0 \quad \text{and} \quad \eta(p_s,\,p_s)=1.$$

Here t is time, $\vec{V}$ is the horizontal wind vector, q_v is the specific humidity, ∇ is the horizontal gradient operator along η-surfaces, ϕ is the geopotential, $\vec{k}$ is the vertical upward unit vector, f is the Coriolis parameter ($f=2\Omega\sin\varphi$, where Ω is the angular velocity of the Earth and φ is the latitude), ω is the p-coordinate vertical velocity ($\omega=\mathrm{d}p/\mathrm{d}t$), α is the specific volume, and ρ is the density of moist air determined by *the ideal gas equation*

$$\alpha=\frac{1}{\rho}=\frac{R_dT_v}{p}. \tag{3.6}$$

R_d is the gas constant for dry air and T_v is the virtual temperature defined by $T_v=T\left[1+\left(\frac{1-\varepsilon}{\varepsilon}\right)q_v\right]$, where T is the absolute temperature, $\varepsilon=R_d/R_v$, and R_v is the gas constant for water vapor. c_p is the specific heat of moist air defined by $c_p=c_{pd}(1+(\delta-1)\,q_v)$, where $\delta=c_{pv}/c_{pd}$, c_{pv} and c_{pd} are the specific heat of water vapor and dry air, respectively. The geopotential ϕ, which appears in Eq. (3.1), is defined by the diagnostic *hydrostatic equation*:

$$\frac{\partial \phi}{\partial \eta}=R_dT_v\frac{\partial \ln\,p}{\partial \eta}. \tag{3.7}$$

The P-forcing terms in Eqs. (3.1), (3.2), and (3.4) represent the contributions of the parameterized physical processes while the k-forcing terms represent the parameterized horizontal diffusion. The P-terms may be specified as for the ECMWF model:

$$\vec{P}_{\vec{V}}=-g\left(\frac{\partial p}{\partial \eta}\right)^{-1}\frac{\partial \vec{J}_{\vec{V}}}{\partial \eta}, \tag{3.8}$$

$$c_PP_T=Q_R+Q_L+Q_D-g\left(\frac{\partial p}{\partial \eta}\right)^{-1}\left[\frac{\partial J_S}{\partial \eta}-c_{pd}T\,(\delta-1)\,\frac{\partial J_{q_v}}{\partial \eta}\right], \tag{3.9}$$

$$P_q=S_{q_v}-g\left(\frac{\partial p}{\partial \eta}\right)^{-1}\frac{\partial J_{q_v}}{\partial \eta}, \tag{3.10}$$

where $\vec{J}_{\vec{V}}$, J_S, and J_{q_v} represent net parameterized vertical fluxes of momentum, dry static energy $c_{pd}T + \phi$, and moisture. Q_R, Q_L, and Q_D represent heating due, respectively, to radiation, internal phase changes, and internal dissipation of kinetic energy associated with the $\vec{P}_{\vec{v}}$-term. S_{q_v} denotes the rate of change of q_v due to rain and snowfall. Comprehensive physical forcing packages have been developed for the calculation of the P- and k-terms in operational primitive equation models as the IFS and HIRLAM. Assuming that such a package is available, it is convenient to express the forcing of the CSCL model in terms of P- and k-terms.

3.1.2. Vertical discretization

In the formulation of the FV CSCL model, it is convenient to make use of the traditional Eulerian hybrid sigma-pressure vertical discretization which is used widely, e.g., in the ECMWF and the HIRLAM models. This means that the model atmosphere is divided into NLEV layers which are defined by the pressures at the interfaces between them (the "half levels"):

$$p_{k+1/2} = A_{k+1/2} + B_{k+1/2}\, p_s \tag{3.11}$$

for $0 \leq k \leq \text{NLEV}$. The pressure thickness of the model layers is denoted as $\Delta_k p = p_{k+1/2} - p_{k-1/2}$. The coefficients $A_{k+1/2}$ and $B_{k+1/2}$ are constants whose values completely define the vertical η coordinate.

The finite difference analog, to the hydrostatic Eq. (3.7) for the geopotential thickness of a single layer and for the air mass from surface and up to a half level, is given by

$$\phi_{k+1/2} - \phi_{k-1/2} = -R_d (T_v)_k \, (\Delta \ln p)_k \,, \tag{3.12}$$

$$\phi_{k+1/2} = \phi_s + R_d \sum_{l=k+1}^{\text{NLEV}} (T_v)_l \, (\Delta \ln p)_l \,, \tag{3.13}$$

respectively, where

$$(\Delta \ln p)_k = \ln\left(\frac{p_{k+1/2}}{p_{k-1/2}}\right). \tag{3.14}$$

To obtain the geopotential at a full level, the almost universal approach of Simmons and Burridge [1981] is used:

$$\phi_k = \phi_{k+1/2} + \alpha_k R (T_v)_k \,, \tag{3.15}$$

where

$$\alpha_k = \left\{ \begin{array}{ll} \ln 2 & k = 1 \\ 1 - \dfrac{p_{k-1/2}}{\Delta_k p} \, (\Delta \ln p)_k & k = 2, .., \text{NLEV} \end{array} \right\}. \tag{3.16}$$

Following Lauritzen, Kaas, Machenhauer and Lindberg [2008], the "full level" pressures p_k are computed with

$$p_k = p_{k+1/2} \exp(-\alpha_k) \,. \tag{3.17}$$

Away from the upper boundary, $p_k \cong \frac{1}{2}\left(p_{k+1/2} + p_{k-1/2}\right)$. The various FV conservation laws may be derived directly from the primitive equations listed above (as was done in Machenhauer [1994]). However, the more straightforward procedure used in Section 1 to derive the mass conservation law (Eq. (1.12)) will be used here.

3.1.3. Conservation of mass of moist air

The mass in a FV δV with vertical walls, horizontal cross-section δA, and thickness $\delta z = z_1 - z_2$ (see Fig. 3.1) is

$$M_{\delta V} = \iint\limits_{\delta A} \left(\int_{z_2}^{z_1} \rho \, \mathrm{d}z \right) \mathrm{d}x \mathrm{d}y. \tag{3.18}$$

Utilizing the hydrostatic balance $\mathrm{d}p = -g\rho \, \mathrm{d}z$, the inner integral may be written as

$$\int_{z_2}^{z_1} \rho \, \mathrm{d}z = \frac{1}{g} \int_{p_1}^{p_2} \mathrm{d}p = \frac{1}{g} \left(p_2 - p_1\right) = \frac{1}{g} \delta p$$

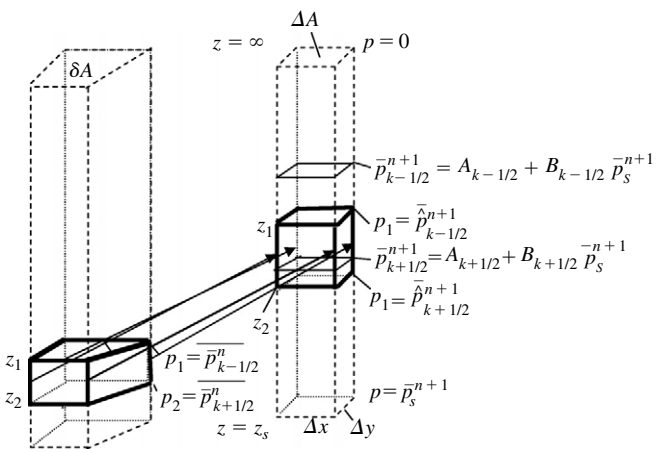


Fig. 3.1 Cell of moist air with vertical walls extending from the height z_1 at pressure p_1 to the height z_2 at pressure p_2, situated in the left column at time $t = n\Delta t$ and in the right column at time $t + \Delta t = (n+1)\,\Delta t$. During a time step, we suppose the cell is moving in an air flow without vertical shear so that its vertical walls remain vertical. Generally, its horizontal cross-section area δA, its thickness $\delta z = z_1 - z_2$, and the corresponding pressure difference $\delta p = p_2 - p_1$ are changing with time during the time-step. At time t the cell is enclosed in model layer k, i.e., $p_1 = \overline{p}^n_{k-1/2} = A_{k-1/2} + B_{k-1/2}\,\overline{p}^n_s$ and $p_2 = \overline{p}^n_{k+1/2} = A_{k+1/2} + B_{k+1/2}\,\overline{p}^n_s$, and at $t + \Delta t$, the cell arrives in the regular grid column (the column to the right) with cross-section area $\Delta A = \Delta x \Delta y$ in a layer, which generally do not coincide with a model layer $((p_1, p_2) \neq \left(\overline{p}^n_{k-1/2}, \overline{p}^n_{k+1/2}\right))$.

so that Eq. (3.18) becomes

$$M_{\delta V} = \frac{1}{g} \iint_{\delta A} \delta p \mathrm{d}x \mathrm{d}y = \frac{1}{g} \overline{\delta p}\, \delta A \tag{3.19}$$

This FV is supposed to move with the flow with vertical walls and without any flux of mass through its boundaries. Thus, the condition for mass conservation is

$$\frac{\mathrm{d}M_{\delta V}}{\mathrm{d}t} = \frac{1}{g} \frac{\mathrm{d}}{\mathrm{d}t} \left(\overline{\delta p}\, \delta A \right) = 0 \tag{3.20}$$

This is similar to Eq. (1.12) except that here the hydrostatic approximation has been applied.

3.1.3.1. 3D trajectories In Section 1.1, a traditional FV Lagrangian approach was applied with an integration of Eq. (1.12) in time along (exact) 3D upstream trajectories, starting at time t from the irregular departure cell and ending at $t + \Delta t$ at the regular arrival cell (an Eulerian grid cell with area ΔA and vertical height difference $\Delta_k h = h_{k+1/2} - h_{k-1/2}$). This resulted in the prognostic Eq. (1.8) that was rewritten finally as Eq. (1.10). A prediction based on Eq. (1.10) would require a 3D integration over the irregular departure FV that would be very complicated and thus inefficient to do in practice. An even more serious objection against using Eq. (1.10) in practice is that it would require the construction of 3D trajectories that would require a priori known vertical velocities. In a quasi-hydrostatic atmosphere, however, the vertical velocity is a diagnostic quantity, which is determined by the diabatic heating and the instantaneous horizontal flow of mass and heat (Richardson [1922]). In pressure coordinates, which is more relevant here, it is even simpler; the pressure vertical velocity $\omega = \mathrm{d}p/\mathrm{d}t$ is diagnostically determined by the continuity equation from just the instantaneous horizontal flow of mass. Thus, the vertical displacements of mass during a time step Δt must be determined as those displacements that ensure re-establishment of hydrostatic equilibrium after the given horizontal displacements of mass. Such considerations led Machenhauer and Olk [1997] to suggest a change in the traditional Lagrangian approach used in Section 1.1. They suggested construction and use of combined backward horizontal and forward vertical trajectories as indicated in Fig. 3.1. This idea of introducing quasi-horizontal Lagrangian trajectories and associated hydrostatically determined vertical velocities was concretized by Lauritzen, Kaas, Machenhauer and Lindberg [2008] in the HIRLAM-DCISL (Section 3.2) as described in the following. They called the combined backward horizontal and forward vertical trajectories hybrid trajectories. Also Lin and Rood [1998, 2004] introduced a so-called "floating Lagrangian control volume vertical coordinate" determined from hydrostatic balance, which in its essence is similar to the Lagrangian trajectories introduced by Machenhauer and Olk [1997], although the vertical displacements and thereby the vertical velocity in their scheme are defined from upstream trajectories determined by horizontal winds at the faces of the arrival Eulerian cell only and not from upstream winds as in DCISL.

The DCISL hybrid trajectories depart at $t = n\Delta t$ from the corner points of the irregular area $\delta_k A^n$ (the departure cell in the left column in Fig. 3.1) with a vertical extent

equal to that of a model layer, i.e., with an averaged pressure difference between its top and bottom equal to $\overline{\Delta_k \overline{p}^{\,n}}^{\delta} = \overline{\overline{p}^{\,n}_{k+1/2} - \overline{p}^{\,n}_{k-1/2}}^{\delta} = \frac{1}{\delta A} \iint\limits_{\delta A} \Delta_k \overline{p}^{\,n} \, \mathrm{d}x \mathrm{d}y$, the horizontal mean of $\Delta_k \overline{p}^{\,n}$ over the irregular departure area $\delta_k A^n$. Note that $\overline{(x)}^{\delta}$ denotes a horizontal mean over an irregular departure cell area $\delta_k A^n$, whereas $\overline{(x)}$ denotes a horizontal mean value over a regular arrival cell area ΔA. The area-averaged full level pressure $\overline{\overline{p}^{\,n}_k}^{\delta}$ in the departure area is determined from $\overline{\overline{p}^{\,n}_{k+1/2}}^{\delta}$ and $\overline{\overline{p}^{\,n}_{k-1/2}}^{\delta}$ analogously to Eq. (3.17). The pressure at the trajectory starting point is interpolated from the $\overline{p}^{\,n}_k$ values in the surrounding grid cells. The trajectories are ending at time $t + \Delta t = (n+1)\,\Delta t$ at the corner points of an arrival cell with horizontal area ΔA located in a regular grid column (the right column in the figure) and in a layer with pressure thickness $\delta_k \overline{\hat{p}}^{\,n+1} = \left(\overline{\hat{p}}_{k+1/2}\right)^{n+1} - \left(\overline{\hat{p}}_{k-1/2}\right)^{n+1}$. The full level pressure in the arrival cell is $\overline{\hat{p}}^{\,n+1}_k = \left(\overline{\hat{p}}_{k+1/2}\right)^{n+1} \exp\left(-\overline{\alpha}^{\,n+1}_k\right)$; and the pressure at the trajectory end point is interpolated from the $\overline{\hat{p}}^{\,n+1}_k$ values in the surrounding grid cells. In general, $\delta_k \overline{\hat{p}}^{\,n+1}$ does not coincide with an Eulerian model layer (as also indicated in the figure).

The Lagrangian FV is supposed to move along the hybrid trajectories so that an integration of Eq. (3.20) from $t = n\Delta t$ to $t + \Delta t = (n+1)\Delta t$ results in the prognostic equation

$$\delta_k \overline{\hat{p}}^{\,n+1} \Delta A = \overline{\Delta_k \overline{p}^{\,n}}^{\delta} \delta_k A^n, \tag{3.21}$$

where the n and $n+1$ superscripts refer to the time levels.

As indicated in Fig. 3.2, NLEV-1 FVs arrive at time $t + \Delta t$ in the same grid column in addition to cell k considered above. These FVs originate from all the other model layers, one FV on top of the other. Here, it is assumed that a FV originating from model level k ends up in the arrival column also as number k from the top (without mixing with the one above and the one below). It is described below how the right-hand side of Eq. (3.21) can be estimated for each of the layers. Once the right-hand sides are known for each of the NLEV layers, $\delta_k \overline{\hat{p}}^{\,n+1}$ can be computed from Eq. (3.21) for each level k, and finally $\overline{\hat{p}}^{\,n+1}_{k-1/2}$ can be determined by summing up the hydrostatic weight of all the cells above:

$$\overline{\hat{p}}^{\,n+1}_{k-1/2} = \sum_{l=1}^{k-1} \delta_k \overline{\hat{p}}^{\,n+1}. \tag{3.22}$$

Hereby the vertical displacement from $\overline{p}^{\,n}_{k-1/2}$ to $\overline{\hat{p}}^{\,n+1}_{k-1/2}$ is determined in a hydrostatically fully consistent way (see Eq. (3.34) below).

Summing up the hydrostatic weight of all the NLEV cells yields the surface pressure

$$\overline{p}^{\,n+1}_s = \sum_{l=1}^{\mathrm{NLEV}} \delta_k \overline{\hat{p}}^{\,n+1}, \tag{3.23}$$

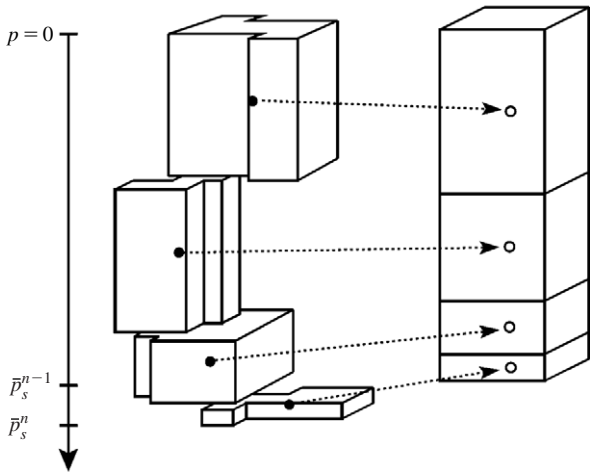


FIG. 3.2 Schematic illustration of the departure and arrival cells which make up the deformed column on the left and the regular column on the right, respectively. The cells move with vertical walls, and the horizontal extension is a polygon. In this figure, the polygon is as in the 2D DCISL scheme of NAIR and MACHENHAUER [2002] but the general idea applies to all DCISL schemes. The filled and unfilled circles indicate the center of mass of the departure and arrival cells, respectively. Note that the vertical levels in the arrival column $\hat{p}_{k+1/2}^{n+1}$ are the ones implied by the advection scheme and not the model levels, $p_{k+1/2}^{n+1}$, based on the hydrostatically determined surface pressure, p_s^{n+1}, and the predefined coefficients (Eq. (3.11)).

from which the pressure at the interfaces between the Eulerian model layers can be determined:

$$\overline{p}_{k-1/2}^{n+1} = A_{k-1/2} + B_{k-1/2}\,\overline{p}_s^{n+1}. \tag{3.24}$$

Now Eq. (3.23) we return to the determination of the right-hand side of Eq. (3.21), $\overline{\Delta_k \overline{p}^{\,n}}^{\delta} \delta A_k^n$. This is an iterative process where each iteration involves two steps: (I) at first the area δA_k^n is determined by constructing hybrid trajectories from the corner points in the irregular departure cell to the corner points of the regular arrival cell. The sides in δA_k^n are defined as the straight lines connecting the corner points. (II) Then, $\overline{\Delta_k \overline{p}^{\,n}}^{\delta} = \overline{\overline{p}_{k+1/2}^n - \overline{p}_{k-1/2}^n}^{\delta} = \frac{1}{\delta A} \iint\limits_{\delta A} \Delta_k \overline{p}^{\,n}\, \mathrm{d}x\mathrm{d}y$, the horizontal mean of $\Delta_k \overline{p}^{\,n}$ over the irregular departure area δA_k^n is computed. Steps I and II are iterated.

3.1.3.2. Trajectory algorithm (I) Several trajectory algorithms have been developed (see Section 2.3); here we choose the hybrid trajectory scheme developed by LAURITZEN, KAAS, MACHENHAUER and LINDBERG [2008], which is used in the HIRLAM-DCISL dynamical core to be described in Section 3.2.

Since the FV is assumed to move with horizontal winds and vertical walls, the problem is 2D. Thus, we need to consider only the projection of the trajectories on a horizontal plane. The horizontal position vectors for the departure point, the arrival point, and the

trajectory midpoint are denoted $\vec{r}_*^{\,n}$, $\vec{r}^{\,n+1}$, and $\vec{r}_{*/2}^{\,n+1/2}$, respectively. The arrival point is defined as

$$\vec{r}^{\,n+1} = \vec{r}_*^{\,n} + \left(\vec{C}_1\left(\vec{V}_*^{\,n}\right) + \vec{C}_2\left(\tilde{\vec{V}}^{n+1}\right)\right). \tag{3.25}$$

For notational clarity, the level number (k) has been suppressed. The trajectory consists of two parts:

(i) $\vec{C}_1\left(\vec{V}_*^{\,n}\right) = \vec{r}_{*/2}^{\,n+1/2} - \vec{r}_*^{\,n}$ is the vector from the departure point to the trajectory midpoint. It depends on $\vec{V}_*^{\,n}$, the horizontal velocity at the departure point at time $t = n\Delta t$. $\vec{C}_1\left(\vec{V}_*^{\,n}\right)$ is determined by one or more terms in a Taylor series expansion about the departure point:

$$\vec{C}_1 = \frac{\Delta t}{2}\vec{V}_*^{\,n} + \sum_{\upsilon=1}^{N-1} \frac{1}{(\upsilon+1)!}\left(\frac{\Delta t}{2}\right)^{\upsilon+1}\left(\frac{\mathrm{d}^\upsilon \vec{V}}{\mathrm{d}t^\upsilon}\right)_*^{n}, \tag{3.26}$$

where N is the order of the expansion.

(ii) $\vec{C}_2\left(\tilde{\vec{V}}^{n+1}\right) = \vec{r}^{\,n+1} - \vec{r}_{*/2}^{\,n+1/2}$ is the vector from the trajectory midpoint to the arrival point. It depends on $\tilde{\vec{V}}^{n+1}$, a horizontal velocity at the arrival point extrapolated in time to $t + \Delta t = (n+1)\,\Delta t$. The time extrapolation, defined by $\tilde{\vec{V}}^{n+1} = 2\vec{V}^{\,n} - \vec{V}^{\,n-1}$. $\vec{C}_2\left(\tilde{\vec{V}}^{n+1}\right)$, is determined by one or more terms in a Taylor series expansion about the arrival point:

$$\vec{C}_2 = \frac{\Delta t}{2}\tilde{\vec{V}}^{n+1} - \sum_{\upsilon=1}^{N-1} \frac{1}{(\upsilon+1)!}\left(-\frac{\Delta t}{2}\right)^{\upsilon+1}\left(\frac{\mathrm{d}^\upsilon \tilde{\vec{V}}}{\mathrm{d}t^\upsilon}\right)^{n+1}. \tag{3.27}$$

In the HIRLAM-CISL dynamical core, the first two terms in the Taylor series are included $(N = 2)$, and thus, estimates of the acceleration $\dfrac{\mathrm{d}\vec{V}}{\mathrm{d}t}$ are taken into account. The acceleration is approximated with $\dfrac{\mathrm{d}\vec{V}}{\mathrm{d}t} \approx \vec{V}\cdot\nabla\vec{V}$ (McGregor [1993]). It follows from Eq. (3.25) that the departure point is given by

$$\vec{r}_*^{\,n} = \vec{r}^{\,n+1} - \left(\vec{C}_1\left(\vec{V}_*^{\,n}\right) + \vec{C}_2\left(\tilde{\vec{V}}^{n+1}\right)\right). \tag{3.28}$$

3.1.3.3. Upstream integral (II) An "upstream integration"

$$\overline{\Delta_k \overline{p}^{\,n}}^{\delta} = \overline{\overline{p}^{\,n}_{k+1/2} - \overline{p}^{\,n}_{k-1/2}}^{\delta} = \frac{1}{\delta A}\iint\limits_{\delta A} \Delta_k \overline{p}^{\,n}\,\mathrm{d}x\mathrm{d}y \tag{3.29}$$

determines the horizontal mean of $\Delta_k \overline{p}^{\,n}$ over the irregular departure area $\delta_k A^n$. It may be estimated by one of the DCISL methods described in Section 2. In HIRLAM-DCISL

(Lauritzen, Kaas, Machenhauer and Lindberg [2008]), two alternative methods are available, the method of Nair and Machenhauer [2002] and that of Nair, Scroggs and Semazzi [2002]. For each of the model layers, all the departure areas $\delta_k A^n$ cover the entire integration domain without overlaps or cracks. Consequently, it follows from Eq. (3.21) that mass is conserved both locally and globally.

3.1.3.4. Iteration In order to determine the departure point from Eq. (3.28), we need to iterate steps I and II (Sections 3.1.3.2 and 3.1.3.3) since, to start with, the pressure at the end point of the trajectory (at a corner of the regular arrival cell) is not known, and also the horizontal positions of the start point of the trajectory (the corner point of the departure cell) are unknown. Generally, for each grid cell, only the trajectory ending at the south-western corner point needs to be determined since adjacent cells share vertices.

The first guess (iteration number $v = 1$)

1. The winds in model layer k are extrapolated in time to time level $n + 1$ and interpolated to the arrival point, the south-western corner of a grid cell. The result is $\left(\tilde{\vec{V}}_k^{n+1}\right)_{SW}$. It is used to determine a first guess $\vec{C}_2$-value: $\left(\vec{C}_2\right)_k^1$.
2. A first guess $\vec{C}_1$-value $(\vec{C}_1)_k^1$ is determined using the time level n winds $\overline{\vec{V}}_k^n$, interpolated also to the arrival point.
3. Thus, the first-guess departure point $\left(\vec{r}_*^n\right)^1$ is determined from Eq. (3.28) using the first guess C's: $\left(\vec{r}_*^n\right)^1 = \vec{r}^{n+1} - \left(\left(\vec{C}_1\right)_k^1 + \left(\vec{C}_2\right)_k^1\right)$.

Iterations

1. The upstream integral is made using the departure points $\left(\vec{r}_*^n\right)^\upsilon$ and an updated $(\nu+1)^{th}$ guess pressure $\left(\overline{\hat{p}}_{k-1/2}^{n+1}\right)^{\upsilon+1}$ is determined using Eqs. (3.21) and (3.22). Corresponding full level pressures $\left(\hat{p}_k\right)^{\upsilon+1} = \left(\overline{\hat{p}}_{k+1/2}\exp(-\alpha_k)\right)^{\upsilon+1}$ are computed and an interpolation of these to the south-western corner point gives $\left(\hat{p}_k\right)_{SW}^{\upsilon+1}$.
2. Vertical interpolation of $\left(\vec{C}_2\right)_k^\upsilon$ is made to this pressure $\left(\left(\hat{p}_k\right)_{SW}^{\upsilon+1}\right)$ giving $\left(\vec{C}_2\right)_k^{\upsilon+1}$.
3. Interpolate $\left(\vec{C}_1\right)_k^\upsilon$ to a preliminary point $\left(\vec{r}_*^n\right)_{prel.}^{\upsilon+1} = \vec{r}^{\,n+1} - \left(\left(\vec{C}_1\right)_k^\upsilon + \left(\vec{C}_2\right)_k^{\upsilon+1}\right)$ giving $\left(\vec{C}_1\right)_k^{\upsilon+1}$.
4. Determine the $(\upsilon+2)^{\text{th}}$ iteration location of the departure point $\left(\vec{r}_*^n\right)_k^{\upsilon+1} = \vec{r}^{n+1} - \left(\left(\vec{C}_1\right)_k^{\upsilon+1} + \left(\vec{C}_2\right)_k^{\upsilon+1}\right)$.
5. If repeat steps 1 to 4.

Note that this departure point algorithm does not require 3D interpolation. Only 1 and 2D interpolations are used.

3.1.3.5. Values needed for the prediction of other invariants To be used for the prediction of other invariants, the following values are stored each time-step for every grid cell:

Horizontal position of the final departure corner points, determined from Section 3.1.3.4:

$$\vec{r}_*^{\,n} = \vec{r}^{\,n+1} - \left(\vec{C}_1\left(\vec{V}_*^{\,n}\right) + \vec{C}_2\left(\tilde{\vec{V}}^{n+1}\right)\right). \tag{3.30}$$

These determine the areas $\delta_k A^n$ of the departure cells.

Full level pressure $\overline{\left(\overline{p}_k^{\,n}\right)}_*^{\delta}$ averaged over the departure area $\delta_k A^n$, determined from Eqs. (3.16) and (3.17)*:*

$$\overline{\left(\overline{p}_k^{\,n}\right)}_*^{\delta} = \overline{\overline{p}_{k+1/2}^{\,n} \exp\left(-\overline{\alpha}_k^{\,n}\right)}^{\delta}, \tag{3.31}$$

where

$$\overline{\alpha}_k^{\,n} = \left\{ \begin{array}{ll} \ln 2 & k = 1 \\ 1 - \frac{\overline{p}_{k-1/2}^{\,n}}{\overline{\Delta_k p}^{\,n}} \ln\left(\frac{\overline{p}_{k+1/2}^{\,n}}{\overline{p}_{k-1/2}^{\,n}}\right) & k = 2, .., \mathrm{NLEV} \end{array} \right\}. \tag{3.32}$$

Mean pressure at the top of arrival cells, determined from Eq. (3.22):

$$\overline{\hat{p}}^{\,n+1}_{k-1/2} = \sum_{l=1}^{k-1} \delta_k \overline{\hat{p}}^{\,n+1}. \tag{3.33}$$

They determine together with Eq. (3.31) a mean value of *the vertical pressure velocity* $\omega = \mathrm{d}p/\mathrm{d}t$ of the cell, moving along the trajectories:

$$\overline{(\omega_k)}^{n+1/2}_{*/2} = \frac{1}{\Delta t}\left(\overline{\hat{p}}^{\,n+1}_k - \overline{\left(\overline{p}_k^{\,n}\right)}_*^{\delta}\right). \tag{3.34}$$

This may be used for parameterizations.

The mean surface pressure is determined from Eq. (3.23):

$$\overline{p}_s^{\,n+1} = \sum_{l=1}^{\mathrm{NLEV}} \delta_k \overline{\hat{p}}^{\,n+1}, \tag{3.35}$$

which determine the pressure at the top of the Eulerian cells (3.24):

$$\overline{p}_{k-1/2}^{\,n+1} = A_{k-1/2} + B_{k-1/2}\, \overline{p}_s^{\,n+1}. \tag{3.36}$$

3.1.4. Conservation of mass of passive tracers

The mass of a passive tracer with specific concentration q_i in a FV δV with vertical walls, horizontal cross-section δA, and thickness $\delta z = z_1 - z_2$ (see Fig. 3.1) is

$$M_{q_i \delta V} = \iint\limits_{\delta A} \left(\int_{z_2}^{z_1} q_i \rho \, \mathrm{d}z \right) \mathrm{d}x \mathrm{d}y. \tag{3.37}$$

Utilizing again the hydrostatic balance $\mathrm{d}p = -g\rho \, \mathrm{d}z$, the inner integral may be written as

$$\int_{z_2}^{z_1} q_i \rho \, \mathrm{d}z = \frac{1}{g} \int_{p_1}^{p_2} q_i \mathrm{d}p = \frac{1}{g} \widehat{q_i} \, (p_2 - p_1) = \frac{1}{g} \widehat{q_i} \delta p,$$

where $\widehat{q_i}$ is the pressure averaged specific concentration. Hereby, Eq. (3.37) becomes

$$M_{q_i \delta V} = \frac{1}{g} \iint\limits_{\delta A} \widehat{q_i} \delta p \mathrm{d}x \mathrm{d}y = \frac{1}{g} \overline{\widehat{q_i} \delta p} \, \delta A. \tag{3.38}$$

The FV mass conservation law for this passive tracer, which is supposed to move with the flow, with vertical walls, and without any flux of mass through its boundaries, is then

$$\frac{\mathrm{d}M_{q_i \delta V}}{\mathrm{d}t} = \frac{1}{g} \frac{\mathrm{d}}{\mathrm{d}t} \left(\overline{\widehat{q_i} \delta p} \, \delta A \right) = 0. \tag{3.39}$$

The FV is supposed to move along the hybrid trajectories determined for the continuity equation so that an integration of Eq. (3.39) from $t = n\Delta t$ to $t + \Delta t = (n + 1) \, \Delta t$ results in the prognostic equation

$$\left(\overline{\widehat{q_i}}^{\delta} \right)_k^{n+1} \delta_k \overline{\hat{p}}^{n+1} \, \Delta A = \overline{\left(\overline{\widehat{q_i}}^{\Delta} \right)_k^n \Delta_k \overline{p}^n}^{\delta} \, \delta A_k^n, \tag{3.40}$$

where $\widehat{x}^{\delta}$ and $\widehat{x}^{\Delta}$ denote vertical mean values over $\delta_k \hat{p}$ and $\Delta_k p$, respectively. Here and in the following, we make the *discretization assumption*: the horizontal mean over the arrival area of a product is equal to the product of the horizontal mean values of the factors.

3.1.5. Conservation of mass of water vapor

Apart from forcing terms, the derivation of the discretized prognostic equation for water vapor is identical to the above for passive tracers. The mass of water vapor with specific humidity q_v in a FV δV with vertical walls, horizontal cross-section δA, and pressure thickness δp is

$$M_{q_v \delta V} = \frac{1}{g} \iint\limits_{\delta A} \widehat{q_v} \delta p \mathrm{d}x \mathrm{d}y = \frac{1}{g} \overline{\widehat{q_v} \delta p} \, \delta A. \tag{3.41}$$

The FV mass conservation law for water vapor is then

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\widehat{\overline{q_v}^{\delta p}}\,\delta A\right) = \overline{\left(\widehat{P}_{q_v} + \widehat{K}_{q_v}\right)\delta p\,\delta A}^{\delta}, \tag{3.42}$$

and the discretized prediction equation becomes

$$\left(\overline{\widehat{q}_v}^{\delta}\right)_k^{n+1} \delta_k \overline{\hat{p}}^{n+1}\,\Delta A = \overline{\left(\overline{\widehat{q}_v}^{\Delta}\right)_k^{n} \Delta_k \overline{p}^{n}}^{\delta}\,\delta A_k^n + \Delta t\,\overline{\left(\overline{\widehat{P}_{q_v}}^{\Delta} + \overline{\widehat{K}_{q_v}}^{\Delta}\right)_k^{n+1/2} \delta_k \overline{p}^{n+1/2}}^{\delta_{n+1/2}}\,\delta A_k^{n+1/2} \tag{3.43}$$

For simplicity, here and in the following, an instantaneous forcing is assumed to work on the FV at time $t + \Delta t/2$ when it is at the midpoint of the trajectory. Of course, it should ideally be averaged along the trajectory. In reality, it might be convenient to treat the forcing as in an existing semi-Lagrangian model. Thus, in the current HIRLAM-DCISL, for instance, the physics are added at time level $n+1$ at the arrival cell as in HIRLAM.

3.1.6. Conservation of total energy

The total energy E is the sum of the internal energy E_i, the potential energy E_p, and the kinetic energy E_k.

The internal energy

$$E_i = \int_{z_2}^{z_1} c_v T\,\rho\,\mathrm{d}z = 1/\mathrm{g}\int_{p_1}^{p_2} c_v T \mathrm{d}p. \tag{3.44}$$

The potential energy

$$\begin{aligned} E_p &= \int_{z_2}^{z_1} gz\,\rho\,\mathrm{d}z = \int_{p_1}^{p_2} z\,\mathrm{d}p \\ &= z_2 p_2 - z_1 p_1 - \int_{z_1}^{z_2} p\,\mathrm{d}z \\ &= z_2 p_2 - z_1 p_1 + 1/\mathrm{g}\int_{p_1}^{p_2} RT\,\mathrm{d}p, \end{aligned} \tag{3.45}$$

where integration by parts as well as the equation of state $p = \rho RT$ has been used.

The "total potential" energy is then

$$\begin{aligned} E_i + E_p &= z_2 p_2 - z_1 p_1 + 1/\mathrm{g}\int_{p_1}^{p_2} (c_v + R)T\,\mathrm{d}p \\ &= 1/\mathrm{g}(\delta(\phi\,p) + \int_{p_1}^{p_2} c_p T\,\mathrm{d}p), \end{aligned} \tag{3.46}$$

where $\delta(\phi\, p) = \phi_2 p_2 - \phi_1\, p_1$ with $\phi = gz$ and, in addition, the relation $c_p = R + c_v$ has been used. Here, respectively, c_p, c_v, and R are the specific heat capacity at constant pressure, the specific heat capacity at constant volume, and the individual gas constants, all for moist air.

The kinetic energy is defined as

$$E_k = 1/(2\mathrm{g}) \int_{p_1}^{p_2} \left(u^2 + v^2\right) \mathrm{d}p. \tag{3.47}$$

Hence, *the total energy* becomes

$$E = E_i + E_p + E_k = 1/\mathrm{g}(\delta(\phi\, p) + \int_{p_1}^{p_2} c_p T \mathrm{d}p + 1/2 \int_{p_1}^{p_2} \left(u^2 + v^2\right) \mathrm{d}p). \tag{3.48}$$

Introducing vertical mean values and including the level index k, Eq. (3.48) may be written as

$$E = 1/\mathrm{g}(\delta_k(\phi\, p) + \left(\widehat{c}_p \widehat{T}\right)_k \delta_k p + (\tfrac{1}{2}) \left(\widehat{u}_k^2 + \widehat{v}_k^2\right) \delta_k p), \tag{3.49}$$

where we have assumed that $\widehat{c}_p T = \widehat{c}_p \widehat{T}$ and $\widehat{u}^2 = \widehat{u}\,\widehat{u}$ and similarly for the northward component v. This is consistent with assuming that the variables are independent of pressure in the layer.

The total energy of a FV δV with vertical walls and horizontal cross-section δA and thickness $\delta z = z_2 - z_1$ (see Fig. 3.1) is then $\overline{E}\delta A$, where $\overline{(\,)} = \frac{1}{\delta A} \iint_{\delta A} (\,)\, \mathrm{d}x\mathrm{d}y$. Finally, we can construct *the FV total energy conservation law:*

$$\frac{\mathrm{d}\left(\overline{E}\delta A\right)}{\mathrm{d}t} = \iiint_{\delta V} \left(\vec{V} \cdot \left(\vec{P}_{\vec{V}} + \vec{K}_{\vec{V}}\right) + c_p \left(P_T + K_T\right)\right) \mathrm{d}p\, \mathrm{d}x\, \mathrm{d}y, \tag{3.50}$$

where the right-hand side follows from a derivation directly from the primitive equations (see MACHENHAUER [1994])

The FV is supposed to move along the hybrid trajectories determined from the continuity equation so that an integration of Eq. (3.50) from $t = n\Delta t$ to $t + \Delta t = (n+1)\Delta t$ results in the prognostic equation

$$\begin{aligned}
&\delta_k \left(\overline{\hat{\phi}}^{\delta} \overline{\hat{p}}^{\delta}\right)^{n+1} + \left(\left(\overline{\widehat{c}_p}^{\delta} \overline{\widehat{T}_k}^{\delta}\right) + 1/2 \left(\left(\overline{\widehat{u}_k}^{\delta}\right)^2 + \left(\overline{\widehat{v}_k}^{\delta}\right)^2\right) \delta_k \overline{\hat{p}}^{\delta}\right)^{n+1} \Delta A \\
&= \overline{\Delta_k \left(\overline{\phi}^{\delta} \overline{p}^{\delta}\right)^n + \left(\left(\overline{\widehat{c}_p}^{\Delta} \overline{\widehat{T}_k}^{\Delta}\right) + 1/2 \left(\left(\overline{\widehat{u}_k}^{\Delta}\right)^2 + \left(\overline{\widehat{v}_k}^{\Delta}\right)^2\right) \Delta_k \overline{p}^{\Delta}\right)^{n\delta}}^{\delta} \delta A_k^n \\
&\quad + \overline{\Delta t \left(\overline{\vec{V}} \cdot \left(\overline{\vec{P}}_{\vec{V}}^{\Delta} + \overline{\vec{K}}_{\vec{V}}^{\Delta}\right) + \overline{\widehat{c}_p}^{\Delta} \left(\overline{P_T}^{\Delta} + \overline{K_T}^{\Delta}\right)\right)^{n+1/2} \delta_k \overline{p}^{n+1/2}}^{\delta_{n+1/2}} \delta A_k^{n+1/2}.
\end{aligned} \tag{3.51}$$

3.1.7. Conservation of entropy

The specific entropy is $S = c_p \ln\theta$, where θ is the potential temperature defined by $\theta = T\,(p/p_0)^{R/c_p}$ and $p_0 = 1000\,hPa$. Note that here R and c_p are for moist air and not, as usual, for dry air. The total entropy of a FV δV with vertical walls and horizontal cross-section δA and thickness $\delta p = p_2 - p_1$ (see Fig. 3.1) is then

$$S_{\delta V} = \frac{1}{g}\iiint\limits_{\delta V} c_p \ln\theta \,\mathrm{d}p\,\mathrm{d}x\,\mathrm{d}y. \tag{3.52}$$

Introducing vertical and horizontal mean values and at the same time including the level number k, Eq. (3.52) may be written as

$$S_{\delta V} = \frac{1}{g}\overline{\left(\widehat{c}_p \ln\widehat{\theta}\right)_k}\,\delta_k p\,\delta A_k, \tag{3.53}$$

from which we get *the* FV *entropy conservation law:*

$$\frac{\mathrm{d}S_{\delta V}}{\mathrm{d}t} = \iiint\limits_{\delta V}\left(\frac{c_p\,(P_T + K_T)}{T} + \left\{\left(c_{pv} - c_{pd}\right)\ln T + (R_v - R_d)\ln(p/p_0)\right\}\left(P_{q_v} + K_{q_v}\right)\right)\mathrm{d}p\,\mathrm{d}x\,\mathrm{d}y, \tag{3.54}$$

where again the right-hand side follows from a derivation directly from the primitive equations (see MACHENHAUER [1994]).

The FV is supposed to move along the hybrid trajectories determined for the continuity equation so that an integration of Eq. (3.54) from $t = n\Delta t$ to $t + \Delta t = (n+1)\,\Delta t$ results in the prediction equation

$$\left(\overline{\widehat{c}_p}^{\,\delta} \ln\overline{\widehat{\theta}}^{\,\delta}\right)_k^{n+1} \delta_k \overline{\widehat{p}}^{\,n+1}\,\Delta A = \overline{\left(\overline{\widehat{c}_p}^{\,\Delta} \ln\overline{\widehat{\theta}}^{\,\Delta}\right)_k^{n}\,\Delta_k \overline{p}^{\,n}\;\delta A_k^n}^{\,\delta} + \Delta t\,\overline{\left(\frac{\widehat{c}_p^{\Delta}\left(\widehat{P}_T^{\Delta} + \widehat{K}_T^{\Delta}\right)}{\widehat{T}} + \left\{\left(\widehat{c}_{pv}^{\Delta} - \widehat{c}_{pd}^{\Delta}\right)\ln\widehat{T}^{\Delta} + \left(\widehat{R}_v^{\Delta} - R_d\right)\ln\left(p_k/p_0\right)\right\}\right.}$$
$$\left.\times\left(\widehat{P}_{q_v}^{\Delta} + \widehat{K}_{q_v}^{\Delta}\right)\right)^{n+1/2}\delta_k p^{n+1/2}}^{\,\delta_{n+1/2}}\;\delta A_k^{n+1/2}. \tag{3.55}$$

Here, a further "*discretization assumption*" is made: the mean value of $\ln x$ over the arrival area is set equal to the logarithm of the mean value.

3.1.8. Conservation of angular momentum

The absolute angular momentum per unit mass of air is

$$m = (\Omega a\cos\varphi + u)a\cos\varphi \tag{3.56}$$

or

$$m = \Omega a^2 \cos^2\varphi + ua\cos\varphi, \tag{3.57}$$

where u is the eastward component of velocity, a is the radius of the Earth (for simplicity and as usual assumed constant), and φ is the latitude.

The absolute angular momentum of the mass in a FV δV with vertical walls and horizontal cross-section δA and thickness $\delta p = p_1 - p_2$ (see Fig. 3.1) is then

$$m_{\delta V} = \frac{1}{g}\iiint_{\delta A}\int_{p_1}^{p_2} (\Omega a\cos\varphi + u)a\cos\varphi \,\mathrm{d}p\,\mathrm{d}x\,\mathrm{d}y. \tag{3.58}$$

Introducing vertical and horizontal mean values and at the same time adding the level number k, Eq. (3.58) may be written as

$$m_{\delta V} = \frac{1}{g}\left(\overline{\widehat{m}_k\,\delta_k p}\,\delta A_k\right) = \frac{1}{g}\left(\overline{\left(\Omega a^2\cos^2\varphi_k + \widehat{u}_k a\cos\varphi_k\right)\delta_k p}\,\delta A_k\right) \tag{3.59}$$

So, *the absolute angular momentum conservation law* becomes

$$\frac{\mathrm{d}(m_{\delta V})}{\mathrm{d}t} = \frac{1}{g}\iiint_{\delta A}^{p_2}_{p_1}\left(-\left(\frac{\partial\phi}{\partial\lambda} + R_d T_v \frac{\partial}{\partial\lambda}(\ln p)\right) + (P_u + K_u)a\cos\varphi\right)\mathrm{d}p\,\mathrm{d}x\,\mathrm{d}y. \tag{3.60}$$

The FV is supposed to move along the hybrid trajectories determined for the continuity equation so that an integration of Eq. (3.60) from $t = n\Delta t$ to $t + \Delta t = (n+1)\Delta t$ results in the prognostic equation

$$\begin{aligned}
&\left(\Omega a^2\cos^2\varphi + \left(\overline{\widehat{u}_k}^{\delta}\right)^{n+1} a\cos\varphi\right)\delta_k\overline{\widehat{p}}^{n+1}\Delta_k A \\
&\quad = \overline{\left(\Omega a^2\cos^2\varphi + \left(\overline{\widehat{u}_k}^{\Delta}\right)^{n} a\cos\varphi\right)\Delta_k\overline{p}^{n}}^{\delta}\ \delta A_k^n \\
&\quad + \Delta t\left\{-\left(\frac{\partial\widehat{\phi}^{\Delta}}{\partial\lambda} + R_d\widehat{T}_v^{\Delta}\frac{\partial}{\partial\lambda}\left(\ln\widehat{p}^{\Delta}\right)\right) + \left(\widehat{P}_u^{\Delta} + \widehat{K}_u^{\Delta}\right)\right\}a\cos\varphi\,\delta_k p^{\delta_{n+1/2}}\ \delta A_k^{n+1/2}.
\end{aligned} \tag{3.61}$$

3.1.9. Choice of invariants

As mentioned by Thuburn [2006], the continuous adiabatic frictionless equations have an infinite number of invariants. In CSCL, we have chosen to fulfill those conservation laws which are fundamental for the dynamics and thermodynamics of the atmosphere, namely the basic conservation laws from which the primitive equations are derived. It should be mentioned that other invariants might substitute for some of those selected above. One obvious example is to replace the conservation of angular momentum by conservation of Ertel potential vorticity. In this case, we would still have a complete set of prognostic equations. An advantage of using angular momentum is that it leads to a direct separation of the u and v contributions to kinetic energy.

3.1.10. Explicit integration procedure

Together with the diagnostic hydrostatic equation (Eq. (3.13)), the discretized FV versions of the conservation laws for (1) mass of moist air (Eq. (3.21)), (2) mass of passive tracers (Eq. (3.40)), (3) mass of water vapor (Eq. (3.43)), (4) total energy (Eq. (3.51)), (5) entropy (Eq. (3.55)), and (6) angular momentum (Eq. (3.61)) constitute a complete prognostic system equivalent to the primitive equations. Initially, the following quantities should be given: Eulerian grid cell area averaged surface pressure $\overline{p}_s$ for each vertical grid column, grid cell averaged values of temperature $\overline{\widetilde{T}}_k^{\Delta}$, specific humidity $\left(\overline{\widetilde{q}}_v^{\Delta}\right)_k$, specific concentration of passive tracers $\left(\overline{\widetilde{q}}_i^{\Delta}\right)_k$, and eastward and northward horizontal velocity components, $\overline{\widetilde{u}}_k^{\Delta}$ and $\overline{\widetilde{v}}_k^{\Delta}$. These would be the history carrying variables. Explicit time-stepping with such a system would be relatively easy. At first the continuity equation (Eq. (3.21)) is solved as described in Section 3.1.3. The outcome, summarized in Section 3.2.1, is the grid cell averaged surface pressure $\overline{p}_s^{n+1}$. In addition, the hybrid trajectories needed for the transport of all the other invariants and diagnostic values of ω (which might be needed in the physical parameterization package) are determined. Next step is to solve the continuity equations for water vapor (Eq. (3.43)) and passive tracers (Eq. (3.40)) giving the updated cell averaged prognostic variables $\left(\overline{\widetilde{q}}_v^{\delta}\right)_k^{n+1}$ and $\left(\overline{\widetilde{q}}_i^{\delta}\right)_k^{n+1}$. The cell averaged values, $\delta_k \overline{\hat{p}}$, $\left(\overline{\widetilde{q}}_v^{\delta}\right)_k^{n+1}$, and $\left(\overline{\widetilde{q}}_i^{\delta}\right)_k^{n+1}$, over the Lagrangian cells, which originally are transported into a vertical Eulerian column, must be remapped into the Eulerian cells, giving $\Delta_k \overline{p}$, $\left(\overline{\widetilde{q}}_v^{\Delta}\right)_k^{n+1}$, and $\left(\overline{\widetilde{q}}_i^{\Delta}\right)_k^{n+1}$. Next, the conservation law for entropy (Eq. (3.55)) is solved giving (after some algebra) the updated cell averaged prognostic variables $\overline{\widetilde{T}}_k^{\delta}$. Again, the cell-averaged Lagrangian values $\overline{\widetilde{T}}_k^{\delta}$ of the cells must be remapped into the Eulerian cells giving $\overline{\widetilde{T}}_k^{\Delta}$. Next, the conservation law of angular momentum (Eq. (3.61)) is solved giving (after some algebra) the updated cell averaged prognostic variables $\overline{\widetilde{u}}_k^{\delta}$. Again, they must be remapped into the Eulerian cells giving $\overline{\widetilde{u}}_k^{\Delta}$. Finally, the conservation law of total energy (Eq. (3.51)) is solved giving (after some algebra and vertical remapping) the updated cell averaged prognostic variables $\overline{\widetilde{v}}_k^{\Delta}$.

3.1.11. Feasibility of a CSCL model

To the author's knowledge, a dynamical core that includes FV versions of all the conservation laws considered here has not yet been realized in spite of the "fact" that (as mentioned in the introduction) it may be expected that a simultaneous exact conservation of all the fundamental physical invariants valid for the atmosphere will result in a particular fast convergence to any "true" solution. The reason for not realizing such a system seems to be difficulties with the application of any of the popular fast-wave-stabilizing techniques, i.e., the semi-implicit or the split-explicit technique, which would eliminate fast wave CFL restrictions on the time-step. Machenhauer and Olk [1997] succeeded in the construction of two 1D shallow water semi-implicit semi-Lagrangian dynamical cores, one that conserves mass and total energy and another that conserves

mass and angular moment (see Section 3.2.2). A 1D shallow water system has just two prognostic variables. Therefore, just two invariants can be conserved exactly. In both cases, the fast wave CFL restriction on the time-step was eliminated and large time steps could be used. However, it seems difficult to extend this to 2 and 3Ds (see Section 3.2.3). Another possibility would be the application of the split-explicit technique. However, as already noted in Section 1, this possibility was abandoned by MACHENHAUER and OLK [1997] for FV models because when splitting the system of continuous equations into an advective part (which should use large time steps) and an adjustment gravity wave part (which should use short time steps), it was found that neither of the sub systems was conserving momentum or total energy. Consequently, it seems unlikely that these invariants could be conserved exactly for the full system in any FV version. In the two examples of dynamical cores with FV techniques, which are described in the following two sections, the continuity equations are solved with the FV technique so that mass is conserved exactly. In the NCAR-FFSL also potential temperature is conserved for adiabatic and friction-free flow. In the HIRLAM-DCISL, a semi-implicit time-stepping is implemented, thereby allowing large time-steps for all variables at the expense of solutions to elliptic Helmholtz equations. This has been feasible because just the continuity equation is solved with the FV technique, while the other primitive equations are kept in their original form, i.e., Eq. (3.1) with u and v and Eq. (3.2) with T as prognostic variables. Furthermore, a special "predictor-corrector" approach (see Section 3.2.3.2) has been used successfully in the semi-implicit continuity equation. In the other system, the NCAR-FFSL, an explicit time-stepping scheme, is used. Consequently, shorter time-steps have to be used for the dynamical core. Tracers (including water vapor) and physical parameterization can, however, be updated with long time-steps. Such a time-stepping procedure would, of course, be possible also in a CSCL model, which then, most likely, would be comparable to the NCAR-FFSL in efficiency.

3.2. The HIRLAM-DCISL with a departure cell-integrated semi-implicit semi-Lagrangian dynamical core

The HIRLAM-DCISL, described in details by LAURITZEN, KAAS, MACHENHAUER and LINDBERG [2008], is a pioneering example of a FV model based on the Lagrangian form of the continuity equation (Eq. (1.8)). The continuity equations for moist air, water vapor, cloud water, and miscellaneous passive tracers are updated each time-step using a DCISL FV scheme while the remaining prognostic equations are in finite difference form and solved using a traditional upstream semi-Lagrangian scheme. It has been developed from the HIRLAM system, (KÄLLÉN [1996] and UNDÉN et al. [2002]). The new HIRLAM-DCISL uses the same horizontal C-grid (ARAKAWA and LAMB [1977]) and vertical LORENZ [1960] staggering of variables as the HIRLAM model (see Section 3.1.2). Also the lateral boundary relaxation scheme is the same. HIRLAM-DCISL is the first model that combines a FV semi-Lagrangian integration scheme with a semi-implicit treatment of gravity wave terms. Thus, this semi-implicit version is absolutely stable as long as the trajectories do not cross (Lipschitz criterion), which in practice means that it runs stably with relatively long time-steps, similar to those used by HIRLAM, and still sufficiently small compared to the time scale of weather system developments. In

Section 3.2.1, we shall introduce at first an explicit version of HIRLAM-DCISL and then, in Sections 3.2.2 and 3.2.3, the changes needed to make it semi-implicit are discussed.

3.2.1. Explicit HIRLAM-DCISL

The explicit continuity equation for moist air is solved for each model layer as described in Section 3.1.3. (see Eq. (3.21)). Hybrid trajectories determine the irregular upstream departure area $\delta_k A^n$, and an "upstream integration" determines the horizontal mean of $\Delta_k \overline{p}^n$ over the departure area $\delta_k A^n$ (3.29). Here $\Delta_k \overline{p}^n$ is defined as

$$\Delta_k \overline{p}^n = \overline{p}^n_{k+1/2} - \overline{p}^n_{k-1/2}. \tag{3.62}$$

The departure cells are the same for all tracers, including water vapor, and Lagrange interpolations between the hybrid trajectory departure points determine the departure points for temperature T and the velocity components u and v. In HIRLAM-DCISL, two alternative upstream integration methods are available, the method of Nair and Machenhauer [2002] and that of Nair, Scroggs and Semazzi [2002]. The mean top pressures of the arrival cells $\overline{\hat{p}}^{n+1}_{k-1/2}$ are determined hydrostatically from Eq. (3.22), i.e., from the Lagrangian pressure thicknesses $\delta_k \overline{\hat{p}}^{n+1}$ in Eq. (3.33). Together with Eq. (3.62), these values determine a mean value of the vertical pressure velocity $\omega = \mathrm{d}p/\mathrm{d}t$ along the trajectory (Eq. (3.34)). This ω is consistent with the hydrostatic assumption and the horizontal flow, contrary to the inconsistent vertical velocities, based on partly Eulerian solutions to the continuity equation, which are applied in traditional semi-Lagrangian models such as HIRLAM. ω is used in the thermodynamic equation (Eq. (3.2)) in the energy conversion term $\frac{\alpha\omega}{c_p} = \frac{R_d T_v}{c_p}\frac{\omega}{p}$, which is approximated with

$$\Delta t \left[\left(\frac{R_d T_v}{c_p}\frac{\omega}{p}\right)_k\right]^{n+1} = \frac{R_d}{c_p}\left[T_v^n + \tilde{T}_v^{n+1}\right]_k \left[\frac{\overline{\hat{p}}^{n+1}_k - \overline{\left(\overline{p}^n_k\right)_*}^{\delta}}{\overline{\hat{p}}^{n+1}_k + \overline{\left(\overline{p}^n_k\right)_*}^{\delta}}\right]. \tag{3.63}$$

The hydrostatic mean surface pressure (Eq. (3.23)) is the weight of all NLEV model layers above the surface:

$$\overline{p}^{n+1}_s = \sum_{l=1}^{\mathrm{NLEV}} \delta_k \overline{\hat{p}}^{n+1}, \tag{3.64}$$

determining the top pressure of Eulerian cells (Eq. (3.24))

$$\overline{p}^{n+1}_{k-1/2} = A_{k-1/2} + B_{k-1/2}\,\overline{p}^{n+1}_s. \tag{3.65}$$

The explicit continuity equations for passive tracers (Eq. (3.40)) and water vapor (Eq. (3.43)) are

$$\left(\overline{\widetilde{q}_i}^{\delta}\right)^{n+1}_k \delta_k \overline{\hat{p}}^{n+1} \Delta A = \overline{\left(\overline{\widetilde{q}_i}^{\Delta}\right)^n_k \Delta_k \overline{p}^n}^{\delta} \delta A^n_k \tag{3.66}$$

and

$$\left(\overline{\overline{q}_v}^{\delta}\right)_k^{n+1} \delta_k \hat{\overline{p}}^{n+1} \Delta A = \overline{\left(\overline{\overline{q}_v}^{\Delta}\right)_k^{n} \Delta_k \overline{p}^{n}}^{\delta} \delta A_k^n + \Delta t \overline{\left(\overline{\overline{P}_{q_v}}^{\Delta} + \overline{\overline{K}_{q_v}}^{\Delta}\right)_k^{n+1/2} \delta_k \overline{p}^{n+1/2}}^{\delta_{n+1/2}} \delta A_k^{n+1/2}, \quad (3.67)$$

respectively, determine updated specific concentrations, $\left(\overline{\overline{q}_i}^{\delta}\right)_k^{n+1}$ and $\left(\overline{\overline{q}_v}^{\delta}\right)_k^{n+1}$, in Lagrangian arrival cells $\left(\delta V = \delta \hat{p} \Delta V\right)$ from $\left(\overline{\overline{q}_i}^{\Delta}\right)_k^{n}$ and $\left(\overline{\overline{q}_v}^{\Delta}\right)_k^{n}$ plus Eq. (3.62). Finally, the updated specific concentrations, $\left(\overline{\overline{q}_i}^{\Delta}\right)_k^{n+1}$ and $\left(\overline{\overline{q}_v}^{\Delta}\right)_k^{n+1}$, in the Eulerian cells $(\Delta V = \Delta p \Delta A)$ are determined from $\left(\overline{\overline{q}_i}^{\delta}\right)_k^{n+1}$ and $\left(\overline{\overline{q}_v}^{\delta}\right)_k^{n+1}$ by 1D vertical remappings.

The discretized explicit momentum and thermodynamic equations are straightforward grid-point semi-Lagrangian and finite difference approximations to Eqs. (3.1) and (3.2), respectively (see KÄLLÉN [1996] and UNDÉN et al. [2002]), except that in the thermodynamic equation the consistent energy conversion term is approximated consistently with (3.63). Regarding the addition of the physics in Eq. (3.67): since DMI-HIRLAM adds the physics at the arrival level (no averaging along the trajectory), that procedure was also adopted in HIRLAM-DCISL. Of course, it should ideally be done as indicated in Eq. (3.67).

3.2.2. 1D semi-implicit CSCL shallow water models

MACHENHAUER and OLK [1997] made a preliminary study, in which a successful implementation of a semi-implicit scheme was made in two different cell-integrated versions of the simple 1D shallow water model. One version conserves mass and momentum and another version conserves mass and total energy. The momentum and continuity equations for the 1D shallow water model are, respectively,

$$\frac{\mathrm{d}u}{\mathrm{d}t} + g\frac{\partial h}{\partial x} = 0 \quad \text{and} \quad (3.68)$$

$$\frac{\mathrm{d}h}{\mathrm{d}t} + h\frac{\partial u}{\partial x} = 0, \quad (3.69)$$

where u is velocity (constant with height), h height (of the fluid surface), and x distance. A periodic domain is assumed $0 \leq x \leq L$. The implementation of a semi-implicit scheme in the cell-integrated model versions will be compared with the traditional approach in a traditional finite difference grid-point model based on Eqs. (3.68) and (3.69). The traditional explicit semi-Lagrangian prediction equations are

$$u_{\exp}^{n+1} = u^n - \frac{\Delta t g}{\Delta x}(\delta(h))^{n+1/2}, \quad (3.70)$$

$$h_{\text{exp}}^{n+1} = h^n - \frac{\Delta t}{\Delta x}(h\delta u)^{n+1/2}, \tag{3.71}$$

where δ indicates a finite difference operator. Like in HIRLAM, we choose a second-order centered finite difference. A semi-implicit system corresponding to this system is obtained simply by averaging $(n+1)\Delta t$ and $n\Delta t$ values of the pressure gradient term and the linear part of the divergence term ($H\delta u$) along the trajectories instead of taking them at $(n+1/2)\Delta t$ at the midpoint of the trajectory as in Eqs. (3.70) and (3.71). The resulting equations may be written as

$$u^{n+1} = u_{\text{exp}}^{n+1} - \frac{\Delta t g}{2\Delta x}\left[\delta h^{n+1} + \delta h^n - 2(\delta(h))^{n+1/2}\right], \tag{3.72}$$

$$h^{n+1} = h_{\text{exp}}^{n+1} - \frac{\Delta t H}{2\Delta x}\left[\delta u^{n+1} + \delta u^n - 2(\delta(u))^{n+1/2}\right]. \tag{3.73}$$

We note that a linear version of Eqs. (3.70) and (3.71), linearized around a state at rest with a mean fluid height H, has gravity wave solutions. These solutions are characterized by purely divergent velocity fields. The height field and the divergence field in these solutions oscillate with a frequency $v \approx 2\pi/\sqrt{gH}$ driven by an oscillating pressure gradient force $-g\partial h/\partial x$ and divergence $-H\partial u/\partial x$, respectively. This explains intuitively why the implicit system, obtained by averaging these terms, can be expected to have stable gravity wave solutions. The system (Eqs. (3.72) and (3.73)) is absolutely stable (as long as the trajectories do not cross).

We may write the system as

$$u^{n+1} = q_1 - \frac{\Delta t g}{2\Delta x}\delta h^{n+1}, \tag{3.74}$$

$$h^{n+1} = q_2 - \frac{\Delta t H}{2\Delta x}\delta u^{n+1}. \tag{3.75}$$

In q_1 and q_2, the terms which do not depend on values at $(n+1)\Delta t$ have been collected. Applying the δ operator on Eq. (3.75) and substituting in Eq. (3.74) gives

$$u^{n+1} - \frac{gH\Delta t^2}{4\Delta x^2}\delta^2 u^{n+1} = q_1 - \frac{g\Delta t}{2\Delta x}\delta q_2. \tag{3.76}$$

This is an elliptic equation which can be solved to give u^{n+1} and then (3.75) can be used to determine h^{n+1}.

The fact that the elliptic equation is with constant coefficients, a so-called Helmholtz equation, means that it is relatively easy and fast to solve. In operational semi-implicit multi-level models as HIRLAM, a series of elliptic equations must be solved. This reduces the advantage of large time-steps. It is therefore important that the elliptic equations in any new implementation of the semi-implicit scheme are kept as simple and fast to solve as possible. The strategy of Machenhauer and Olk [1997] for the present

models and later that of LAURITZEN, KAAS, MACHENHAUER, LINDBERG, [2006, 2008] for the HIRLAM-DCISL has been to do the semi-implicit implementation in the FV model in such a way that the resulting elliptic equation becomes similar to that of the traditional model it replaces.

Now let us derive the FV models corresponding to Eqs. (3.70) and (3.71). The mass, momentum, and total energy in a volume of length δx are, respectively,

$$M_{\delta x} = \rho \int_x^{x+\delta x} \left(\int_0^h z \mathrm{d}z \right) \mathrm{d}x = \rho \overline{h}\ \delta x, \tag{3.77}$$

$$m_{\delta x} = \rho \int_x^{x+\delta x} \left(u \int_0^h \mathrm{d}z \right) \mathrm{d}x = \rho \overline{uh}\, \delta x, \text{ and} \tag{3.78}$$

$$E = \rho \int_x^{x+\delta x} \int_0^h (gz + (1/2)(u^2)) \mathrm{d}z \mathrm{d}x = (1/2)\rho \left(g\overline{h^2} + \overline{hu^2} \right) \delta x, \tag{3.79}$$

where ρ is a constant density (mass per unit length). Then the conservation laws for mass, momentum, and total energy are

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\overline{h}\delta x\right) = 0, \tag{3.80}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\overline{uh}\delta x\right) = -g\delta x \overline{\frac{\partial}{\partial x}\left(\frac{1}{2}h^2\right)} = -\frac{g}{2}\delta\left(h^2\right), \text{ and} \tag{3.81}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\left(g\overline{h^2} + \overline{hu^2}\right)\delta x\right) = -g\delta x \overline{\frac{\partial}{\partial x}\left(\frac{1}{2}h^2 u\right)} = -\frac{g}{2}\delta\left(h^2 u\right), \tag{3.82}$$

respectively, and after integration over a time-step from $t = n\Delta t$ to $t + \Delta t = (n+1)\Delta t$, we get the corresponding discretized conservation laws

$$\overline{h}^{n+1}\Delta x = \overline{h}^{n}\delta x, \tag{3.83}$$

$$\overline{uh}^{n+1}\Delta x = \overline{uh}^{n}\delta x - g\frac{\Delta t}{2}\delta\left(h^2\right)^{n+1/2}, \quad \text{and} \tag{3.84}$$

$$\left(g\overline{h^2} + \overline{hu^2}\right)^{n+1}\Delta x = \left(g\overline{h^2} + \overline{hu^2}\right)^{n}\delta x - g\frac{\Delta t}{2}\delta\left(h^2 u\right)^{n+1/2} \tag{3.85}$$

A summation over all grid cells of each of these equations and application of the periodic boundary condition show that mass, momentum, and total energy are globally conserved.

We have used a slightly simplified notation compared to that introduced in Section 3.1. On the right-hand side, we have, for example, simplified $\overline{\overline{u}^{n}\overline{h}^{n}}^{\delta}$ to $\overline{uh}^{n}$. Also, the left-hand sides are assumed to be discretized like in Section 3.1 so that here $\overline{uh}^{n+1}\Delta x$ stands for $\overline{u}^{n+1}\overline{h}^{n+1}\Delta x$ and $\left(g\overline{h^2}+\overline{hu^2}\right)^{n+1}\Delta x$ stands for $\left(g\left(\overline{h}^{n+1}\right)^2+\overline{h}^{n+1}\left(\overline{u}^{n+1}\right)^2\right)\Delta x$. Eqs. (3.83) and (3.84) constitute a complete set of prognostic equations, which we may call "the momentum set" and Eqs. (3.83) and (3.85) constitute another complete set of prognostic equations, which we may call "the energy set." Explicit time-stepping can be performed with both set of equations as was the case for the complete system of 3D FV conservation laws considered in the Section 3.1. The explicit time-stepping scheme is absolutely stable with regard to advection but only conditional stable with regard to gravity waves. The strategy of MACHENHAUER and OLK [1997] was to duplicate as far as possible the implementation of the semi-implicit scheme done above in the corresponding traditional model. Several problems are encountered when trying to do this.

The first problem is that the divergence, which we want to average over the two time levels $n\Delta t$ and $(n+1)\Delta t$, is not explicit in the cell-integrated continuity equation, Eq. (3.83), as it is in the traditional one, Eq. (3.70). This can be dealt with by using the Lagrangian expression for divergence.

$$D=\frac{1}{\delta_i x}\frac{\mathrm{d}\delta_i x}{\mathrm{d}t} \tag{3.86}$$

where here $\delta_i x$ is an infinitesimal small length. A finite difference approximation to this expression is

$$D=\frac{1}{\Delta x}\frac{\Delta x-\delta x}{\Delta t}. \tag{3.87}$$

Isolating δx in Eq. (3.87) and inserting it in Eq. (3.83) gives

$$\overline{h}^{n+1}=\overline{h}^{n}-\Delta t\overline{h}^{n}D=\overline{h}^{n}-\overline{h}^{n}\frac{\Delta x-\delta x}{\Delta x}. \tag{3.88}$$

Noting that approximately $\Delta x-\delta x=\Delta t\delta u^{n+1/2}$ where $\delta u^{n+1/2}$ is the velocity increment over the cell at time $(n+1/2)\Delta t$, we get finally

$$\overline{h}^{n+1}_{\mathrm{exp}}=\overline{h}^{n}-\Delta t\overline{h}^{n}\frac{\delta u^{n+1/2}}{\Delta x}. \tag{3.89}$$

This expression is similar to Eq. (3.71) and can be used in the same way as Eq. (3.71) in the implementation of the semi-implicit scheme. Doing so, the semi-implicit equation becomes

$$\overline{h}^{n+1}=\overline{h}^{n}-\frac{\Delta t}{\Delta x}\left(\overline{h}^{n}\right)'\delta u^{n+1/2}-\frac{\Delta t H}{2\Delta x}\left(\delta u^{n+1}+\delta u^{n}\right). \tag{3.90}$$

The second problem is that both the momentum and total energy equation are nonlinear quantities in the basic variables $\overline{u}^{n+1}$ and $\overline{h}^{n+1}$. We shall see how that becomes a problem

and how MACHENHAUER and OLK [1997] dealt with it. The explicit momentum equation (Eq. (3.84)) may be written as

$$\overline{uh}^{n+1}_{\exp}\,\Delta x = \overline{uh}^{n}\,\Delta x - \Delta t\Big(UH + \overline{(uh)'}^{n}\Big)\delta u^{n+1/2} - g\frac{\Delta t}{2}\delta\Big(h'^2 + 2h'H\Big)^{n+1/2}, \tag{3.91}$$

where we have used $\overline{(uh)}^{n} = \overline{(uh)'}^{n} + UH$ and $\delta x = \Delta x - \Delta t\delta u^{n+1/2}$. The corresponding semi-implicit equation becomes

$$\begin{aligned}\overline{uh}^{n+1} = \overline{uh}^{n}\,\Delta x - \frac{\Delta t}{\Delta x}\overline{(uh)'}^{n}\delta u^{n+1/2} - g\frac{\Delta t}{2\Delta x}\delta\Big(h'^2\Big)^{n+1/2} \\ - \frac{\Delta t H}{2\Delta x}\Big(\Big(U\Big(\delta u^{n+1} + \delta u^{n}\Big) + g\Big(\delta h^{n+1} + \delta h^{n}\Big)\Big)\Big).\end{aligned} \tag{3.92}$$

The two semi-implicit equations (Eqs. (3.90) and (3.92)) may now be written in the form

$$\overline{h}^{n+1} = \overline{h}^{n+1}_{\exp} - \frac{\Delta t H}{\Delta x}\Big(\delta u^{n+1} + \delta u^{n} - 2\delta u^{n+1/2}\Big) \tag{3.93}$$

and

$$\overline{(uh)}^{n+1} = \overline{(uh)}^{n+1}_{\exp} - \frac{\Delta t H}{\Delta x}\Big(U\Big(\delta u^{n+1} + \delta u^{n} - 2\delta u^{n+1/2}\Big) + g\Big(\delta h^{n+1} + \delta h^{n} - 2\delta h^{n+1/2}\Big)\Big). \tag{3.94}$$

As for the explicit momentum system (Eqs. (3.83) and (3.84)), a summation over all grid cells of each of these equations and application of the periodic boundary condition show that also the semi-implicit system conserves mass and momentum globally. This follows from the fact that the sum of the semi-implicit corrections to the explicit updated values is zero.

To proceed as for the traditional system in Eqs. (3.72) and (3.73), $(uh)^{n+1}$ in Eq. (3.94) must be linearized. At first it is expanded as

$$(uh)^{n+1} = \Big(\big(U + u'\big)\big(H + h'\big)\Big)^{n+1} = UH + \big(u'\big)^{n+1}H + \big(h'\big)^{n+1}U + \big(u'h'\big)^{n+1}. \tag{3.95}$$

In order to make it linear in $\big(u'\big)^{n+1}$ and $\big(h'\big)^{n+1}$, we approximate $\big(u'h'\big)^{n+1}$ with $\big(u'h'\big)^{n}$ getting

$$(uh)^{n+1} \cong UH + u'^{n+1}H + h'^{n+1}U + \big(u'h'\big)^{n}. \tag{3.96}$$

When Eq. (3.96) is inserted in Eq. (3.94), it becomes (after using Eq. (3.93) and some algebra)

$$u'^{n+1} = \frac{1}{H}\left((uh)^{n+1}_{\exp} - Uh^{n+1}_{\exp} - \big(u'h'\big)^{n} - \frac{\Delta t g}{\Delta x}\Big(\delta h^{n+1} + \delta h^{n} - 2\delta h^{n+1/2}\Big)\right) - U. \tag{3.97}$$

This equation is in the same form as the traditional velocity equation (Eq. (3.72)). So, using it together with Eq. (3.93), we can establish a system, similar to the traditional system (Eqs. (3.74) and (3.75)):

$$u'^{n+1} = \tilde{q}_1 - \frac{\Delta t g}{2\Delta x}\delta h^{n+1}, \tag{3.98}$$

$$h^{n+1} = \tilde{q}_2 - \frac{\Delta t H}{2\Delta x}\delta u^{n+1}, \tag{3.99}$$

where terms not depending on values at time level $(n+1)\Delta t$ are collected in $\tilde{q}_1$ and $\tilde{q}_2$, respectively. Applying the operator $\delta(.)$ on Eq. (3.99) and substituting in Eq. (3.98), a Helmholtz equation like Eq. (3.76) is obtained:

$$u'^{n+1} - \frac{gH\Delta t^2}{4\Delta x^2}\delta^2 u^{n+1} = \tilde{q}_1 - \frac{g\Delta t}{2\Delta x}\delta\tilde{q}_2. \tag{3.100}$$

The solution to this equation determines u'^{n+1} and Eq. (3.101) determines h^{n+1}, which may then be used to determine the semi-implicit correction terms in Eqs. (3.93) and (3.94).

Machenhauer and Olk [1997] derived a similar semi-implicit model for the cell-integrated energy system. As for the momentum system, a summation over all grid cells of each of these equations and application of the periodic boundary condition showed that the semi-implicit corrections to the explicit updated values become zero so that also the semi-implicit energy system conserves mass and total energy globally.

In this case, the question is if the prognostic variable in Eq. (3.76), the total energy $g\overline{h^2} + \overline{hu^2}$, can be linearized as we did with the momentum. To show that, we first expand it and then approximate nonlinear terms in perturbations with their values at time $n\Delta t$.

$$\begin{aligned} g\overline{h^2}^{n+1} + \overline{hu^2}^{n+1} &= \overline{gh^2 + hu^2}^{n+1} \\ &= \overline{g(H+h')^2 + (H+h')\,(U+u')^2}^{n+1} \\ &= 2HU\overline{u'}^{n+1} + (UU + 2gH)\,\overline{h'}^{n+1} + \overline{\left(gh'^2 + Hu'^2 + 2Uh'u' + h'u'^2\right)}^{n} \\ &\quad + gH^2 + HU^2. \end{aligned} \tag{3.101}$$

Apparently, one gets something that might work as it did for the momentum equation. The main question is if the coefficient $2HU$ in front of $\overline{u'}^{n+1}$ is sufficiently large. Of course, it will not work if $U = 0$, i.e., if the mean zonal flow is zero. In a realistic flow, however, both H and U are relatively large compared to the perturbations $\overline{u'}^{n+1}$. In test runs with such a flow, the resulting semi-implicit energy system did work satisfactorily, even with a time-step 50 times larger than the CFL maximum.

However, for the full multilevel CSCL system considered in Section 3.1, a corresponding linearization of the total energy $E = 1/g(\delta_k(\phi\, p) + (\widehat{c_p}\widehat{T})_k\delta_k p + 1/2$

$\left(\widehat{u}_k^2 + \widehat{v}_k^2\right)\delta_k p$), defined in Eq. (3.48), cannot be expected to work. Therefore, the present approach cannot be expected to lead to a semi-implicit system that works satisfactory. The reason is that, in the full CSCL system, the explicit angular momentum equation determines the zonal velocity u and the explicit energy equation determines the meridional velocity v. Thus, the question is how big the coefficient in front of v' will be in a linearized expression for E. It is easily seen that the term in question is $\left(\Delta_k p^{ref} V\right) v_k^{'}$, where we have used $\delta_k p = \left(\Delta_k p^{ref} + \delta_k p\right)$ and $v_k = V + v_k^{'}$. Thus, the question is whether the coefficient $\left(\Delta_k p^{ref} V\right)$ is large enough. Generally, the answer is no as a time independent V in most places will be close to zero. Thus, most likely the approach of MACHENHAUER and OLK [1997] cannot be extended to the full CSCL system – at least not in a system including the total energy conservation law. An efficient full semi-implicit CSCL system may be possible and may be developed, eventually. However, it will require the invention of a new way to transform the explicit system to a semi-implicit system. The classical approach, described above, goes back to ROBERT [1969]. It was developed for a system consisting of the momentum equations (or the vorticity and the divergence equations), the continuity equation, and the thermodynamic equation. A new approach, if possible, should be based directly on modifications of the explicit CSCL system.

3.2.3. The semi-implicit version of HIRLAM-DCISL

As a consequence of the conclusions in the last paragraph of the preceding section, LAURITZEN, KAAS, MACHENHAUER and LINDBERG [2006, 2008] decided to develop a semi-implicit version of the HIRLAM model in which u and v are kept as prognostic variables, and just the continuity equation is implemented in FV form. In this case, no linearization of the prognostic variables is needed. In the present section, we present the derivation of a semi-implicit system of prognostic equations based on the system of explicit equations in Section 3.2.1. The derivation is described in more details in LAURITZEN, KAAS, MACHENHAUER and LINDBERG [2006, 2008]. Here we concentrate on deviations from the traditional derivation procedure used, e.g., for HIRLAM. The traditional procedure for deriving the elliptic equations associated with the baroclinic HIRLAM model is for the central parts the same as for the 1D shallow water models considered in Section 3.2.2.

First, the explicit system is made semi-implicit by time averaging certain right-hand side terms in the discretized primitive equations between time levels $n\Delta t$ and $(n+1)\,\Delta t$.

These linearized terms are:

1. the linearized pressure gradient force in the momentum equation (Eq. (3.1)), i.e., $-\nabla G_k$, which depends on temperature and surface pressure
2. two linearized divergence terms: (1) one in the FV continuity equation (Eq. (3.21)) and (2) one in the energy conversion term $\alpha\omega$ in the thermodynamic equation (Eq. (3.2)), i.e., $-\frac{R_d}{c_{pd}}\left(\frac{T}{p_{k+1/2}}\right)^{ref} \sum_{l=1}^{k} (\Delta_l p)^{ref}\, D_l^{n+1/2}$ in Eq. (3.120).

Secondly, the formula for the surface pressure and temperatures at time level $(n+1)$ from the semi-implicit continuity and thermodynamic equations, respectively, are inserted in

the formula for the linearized pressure gradient force, $-\nabla G_k$, in the momentum equations. Finally, the divergence operator, $\nabla \cdot (\ )$, is applied to the momentum equation resulting in a set of coupled elliptic equations with updated divergence as an independent variable. The vertically coupled equations are separated into a set of vertically decoupled shallow water Helmholtz equations via diagonalization. The final solution to the elliptic system determines the semi-implicit corrections to the explicit solutions for all the prognostic variables.

3.2.3.1. The linearized pressure gradient force The explicit semi-Lagrangian momentum equation at model level k is

$$\frac{\left(\vec{V}_k^{n+1}\right)_{\exp} - \vec{V}_k^n}{\Delta t} = -\nabla\phi_k - R_d(T_v)_k \nabla \ln p_k - f\vec{k} \times \vec{V}_k + \left(\vec{P}_{\vec{V}} + \vec{K}_{\vec{V}}\right)_k . \tag{3.102}$$

The pressure gradient force

$$\vec{F}_k = -\nabla\phi_k - \frac{R_d(T_v)_k}{p_k}\nabla p_k = -\nabla\left(\phi_s + R_d \sum_{l=k+1}^{\mathrm{NLEV}} (T_v)_l (\Delta \ln p)_l + R_d \alpha_k (T_v)_k\right) - R_d(T_v)_k \nabla \ln p_k \tag{3.103}$$

is linearized as in HIRLAM:

$$-\nabla G_k = -\nabla\left(\phi_s + R_d \sum_{l=k+1}^{\mathrm{NLEV}} (T_v)_l \left(\Delta \ln p^{ref}\right)_l + R_d \alpha_k^{ref} (T_v)_k\right) - \frac{R_d T^{ref}}{p_s^{ref}} \nabla p_s, \tag{3.104}$$

where T^{ref} and p_s^{ref} are constant reference temperature and constant surface pressure, respectively. $\left(\Delta \ln p^{ref}\right)_k$ and α_k^{ref} are defined by Eqs. (3.14) and (3.16) with the "half-level" pressures obtained from Eq. (3.11) by choosing $p_s = p_s^{ref}$.

After temporal averaging of $-\nabla G_k$, the semi-implicit momentum equation may be written as in HIRLAM:

$$\left[\vec{V} + \frac{\Delta t}{2}\nabla G - f_0 \vec{k} \times \vec{V}\right]_k^{n+1} = \left(\vec{R}_{\vec{V}}\right)_k , \tag{3.105}$$

where $\left(\vec{R}_{\vec{V}}\right)_k$ represents explicit terms. In the traditional HIRLAM derivation of the elliptic system, a substitution in the linearized pressure gradient force, $-\nabla G_k^{n+1}$, of the updated surface pressure from the continuity equation and the updated temperatures from the thermodynamic equations is performed. Thereby the linearized geopotential, G_k^{n+1}, is expressed in terms of the divergence, D_k^{n+1}. The HIRLAM-DCISL derivation proceeds similarly. Here, just those parts involving the FV continuity equation will be dealt with. These are the parts which deviate from the traditional derivation.

3.2.3.2. The semi-implicit DCISL continuity equations The discretization of the explicit continuity equation for moist air was discussed in Section 3.2.1. The derivation of the semi-implicit continuity equation considered here is a direct extension of the derivation for the 1D shallow water models in Section 3.2.2. Defining the discretized Lagrangian divergence,

$$D_k^{n+1/2} = \frac{1}{\Delta A}\frac{\Delta A - \delta A_k^n}{\Delta t} = \frac{1}{\Delta t}\left(1 - \frac{\delta A_k^n}{\Delta A}\right) \tag{3.106}$$

and substituting $\frac{\delta_k A^n}{\Delta A}$ from the explicit continuity equation (Eq. (3.21)) in the form

$$\overline{\left(\delta_k \hat{p}\right)}_{\exp}^{n+1} = \overline{\Delta_k p}^n \frac{\delta A_k^n}{\Delta A}, \tag{3.107}$$

it may be written as

$$\begin{aligned}\overline{\left(\delta_k \hat{p}\right)}_{\exp}^{n+1} &= \overline{\Delta_k p}^n - \Delta t \overline{\Delta_k p}^n D_k^{n+1/2} \\ &= \overline{\Delta_k p}^n - \Delta t \left(\overline{\Delta_k p}^n\right)' D_k^{n+1/2} - \Delta t (\Delta_k p)^{ref} D_k^{n+1/2}.\end{aligned} \tag{3.108}$$

Treating the linear term as a temporal average, the ("ideal") semi-implicit continuity equation results

$$\overline{\delta_k \hat{p}}^{n+1} = \overline{\left(\delta_k \hat{p}\right)}_{\exp}^{n+1} - \frac{\Delta t}{2}(\Delta_k p)^{ref}\left(D_k^{n+1}\left(\vec{V}_k^{n+1}\right) + D_k^n\left(\vec{V}_k^n\right) - 2D_k^{n+1/2}\right) \tag{3.109}$$

or

$$\overline{\delta_k \hat{p}}^{n+1} = \overline{\left(\delta_k \hat{p}\right)}_{\exp}^{n+1} - \frac{\Delta t}{2}(\Delta_k p)^{ref}\left(D_k^{n+1}\left(\vec{V}_k^{n+1}\right) - D_k^{n+1}\left(\tilde{\vec{V}}_k^{n+1}\right)\right), \tag{3.110}$$

where $D_k^{n+1}\left(\tilde{\vec{V}}_k^{n+1}\right)$ is defined as the Lagrangian divergence for the last part of the hybrid trajectory

$$D_k^{n+1}\left(\tilde{\vec{V}}_k^{n+1}\right) = \frac{1}{\Delta A}\frac{\Delta A - \delta A_k^{n+1/2}}{\Delta t/2} \tag{3.111}$$

and $D_k^n\left(\vec{V}_k^n\right)$ is defined as the Lagrangian divergence for the first part of the hybrid trajectory

$$D_k^n\left(\vec{V}_k^n\right) = \frac{1}{\Delta A}\frac{\delta A_k^{n+1/2} - \delta A_k^n}{\Delta t/2}, \tag{3.112}$$

(see Fig. 3.3). Thus,

$$D_k^n\left(\vec{V}_k^n\right) + D_k^{n+1}\left(\tilde{\vec{V}}_k^{n+1}\right) = 2D_k^{n+1/2}. \tag{3.113}$$

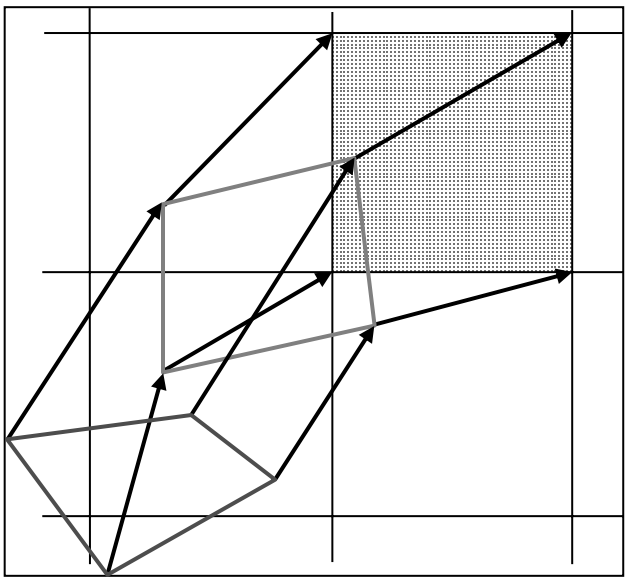


FIG. 3.3 Illustrating the different areas in Eqs. (3.113) and (3.114): δA_k^n (*red*), the departure area at time $n\Delta t$, $\delta A_k^{n+1/2}$ (*green*), the "mid-way" area at time $(n+1/2)\,\Delta t$, and δA_k^{n+1} (*blue*), the arrival area at time $(n+1)\,\Delta t$. (See also color insert).

This was used to derive Eq. (3.110) from Eq. (3.109). Note that corresponding to Eq. (3.112), $D_k^{n+1}\left(\vec{V}_k^{n+1}\right)$ is defined as

$$D_k^{n+1}\left(\vec{V}_k^{n+1}\right) = \frac{1}{\Delta A}\frac{\delta A_k^{n+1/2} - \delta A_k^{n+1}}{\Delta t/2} = \frac{1}{\Delta A}\frac{\delta A_k^{n+3/2} - \delta A_k^{n+1/2}}{\Delta t}, \tag{3.114}$$

where the last expression is centered about time level $n+1$.

In order to proceed with the derivation of the semi-implicit system, the Lagrangian divergence (Eq. (3.114)) should be expressed as a function of the velocity components. In 1D, it was straightforward (see Eq. (3.89)), but in 2Ds, it is complicated although not impossible. However, LAURITZEN, KAAS and MACHENHAUER [2006], LAURITZEN, KAAS, MACHENHAUER and LINDBERG [2008], found that the resulting elliptic equations would be much more complicated than the elliptic equations associated with the traditional HIRLAM system, and therefore, it would be more time consuming to solve. This would significantly reduce the efficiency of the semi-implicit model version. Therefore, they decided to use instead a predictor-corrector approach, which results in elliptic equations in the same form as in HIRLAM. The predictor-corrector approach applied to Eq. (3.110) gives finally the semi-implicit continuity equation for moist air:

$$\begin{aligned}\overline{\delta_k \hat{p}}^{n+1} = {} & \overline{\left(\delta_k \hat{p}\right)}_{\exp}^{n+1} - \frac{\Delta t}{2}(\Delta_k p)^{ref}\left(\mathcal{D}_k^{n+1}\left(\vec{V}_k^{n+1}\right) - D\left(\tilde{\vec{V}}_k^{n+1}\right)\right) \\ & + \frac{\Delta t}{2}(\Delta_k p)^{ref}\overline{\left[\mathcal{D}_k^n\left(\vec{V}_k^n\right) - D_k^n\left(\vec{V}_k^n\right)\right]}_*^{\delta}\frac{\delta A_k^n}{\Delta A},\end{aligned} \tag{3.115}$$

where the discretized Eulerian divergence $\mathcal{D}_k^{n+1}\left(\vec{V}_k^{n+1}\right)$ is defined in the C-grid in spherical coordinates as

$$\mathcal{D}_k^n\left(\vec{V}_k^n\right) = \frac{1}{a\cos\varphi}\left[\frac{\delta_\lambda u_k^n}{\Delta\lambda} + \frac{\delta_\varphi\left(v_k^n\cos\varphi\right)}{\Delta\varphi}\right] = \frac{\delta_\lambda u_k^n}{\Delta x} + \frac{1}{\cos\varphi}\frac{\delta_\varphi\left(v_k^n\cos\varphi\right)}{\Delta y}, \tag{3.116}$$

λ is the longitude, φ is the latitude, $\Delta x = a\cos\varphi\Delta\lambda$, and $\Delta y = a\Delta\varphi$.

By replacing $D_k^{n+1}(\vec{V}_k^{n+1})$ in Eq. (3.110) with the discretized Eulerian divergence $\mathcal{D}_k^{n+1}(\vec{V}_k^{n+1})$, as done in Eq. (3.115), an elliptic equation in the same form as in the traditional HIRLAM system results. However, if just this replacement was done, the scheme would be inconsistent since $\mathcal{D}_k^{n+1}\left(\vec{V}_k^{n+1}\right)$ is different from $D_k^{n+1}\left(\vec{V}_k^{n+1}\right)$ (in fact small-scale noise would develop and it would result in an instability). Therefore, a correction term is added, the last term in Eq. (3.115), which corrects for the error introduced in the first term of Eq. (3.115). The correction term is equal to the error $D_k^{n+1}\left(\vec{V}_k^{n+1}\right) - \mathcal{D}_k^{n+1}\left(\vec{V}_k^{n+1}\right)$ introduced at time level $n+1$, but it is computed in the subsequent time-step when $D_k^n\left(\vec{V}_k^n\right)$ is known. (The current $D_k^n\left(\vec{V}_k^n\right)$ is then equal to $D_k^{n+1}\left(\vec{V}_k^{n+1}\right)$ from the previous time step.) Note that the bar over the last term in Eq. (3.115) indicates a spatial average over the departure area δA_k^n, i.e., the same departure area as the one used to calculate $\overline{\left(\delta_k\hat{p}\right)}_{\exp}^{n+1}$. In practice, $\frac{\Delta t}{2}\left(\Delta_k p\right)^{ref}\left[\mathcal{D}_k^n\left(\vec{V}_k^n\right) - D_k^n\left(\vec{V}_k^n\right)\right]$ can be added to the term $\Delta_k p^n$ as the first operation in each time-step. This means that only one upstream integration is needed to evaluate the first and the last term on the right-hand side of Eq. (3.115). The correction term is necessary because the discretized Eulerian divergence $\mathcal{D}_k^{n+1}\left(\vec{V}_k^{n+1}\right)$, defined in Eq. (3.116), corresponds to a discretized Lagrangian divergence (see Fig. 3.4a), which is different from $D_k^{n+1}\left(\vec{V}_k^{n+1}\right)$, defined in Eq. (3.114). This difference is illustrated in Fig. 3.4.

As demonstrated in Section 3.1.3.3, the explicit continuity equation (Eq. (3.21)) conserves mass both locally and globally. Since the correction terms in Eq. (3.115), which correct the explicit prediction, consist of linear divergence terms, integration over the entire integration area become zero if the Lagrangian and the Eulerian divergence both are zero at the boundaries or if the integration area is global. Consequently, with these assumptions fulfilled, the semi-implicit continuity equation also conserves global mass. It is the impression from preliminary tests that the semi-implicit correction terms generally are small compared to the explicit local mass changes, so it is our impression that the local mass conservation is only slightly modified by the semi-implicit corrections.

The explicit continuity equation for a passive tracer was derived in Section 3.1.4. The result was Eq. (3.40) which may be written as

$$\left(\overline{\left(\overline{q_i^\delta}\right)_k^{n+1}\delta_k\hat{p}^{n+1}}\right)_{\exp} = \overline{\left(\left(\overline{\overline{q_i}^\Delta}\right)_k^n\Delta_k\overline{p}^n\right)_*}^\delta\frac{\left(\delta A_k^n\right)}{\Delta A}.$$

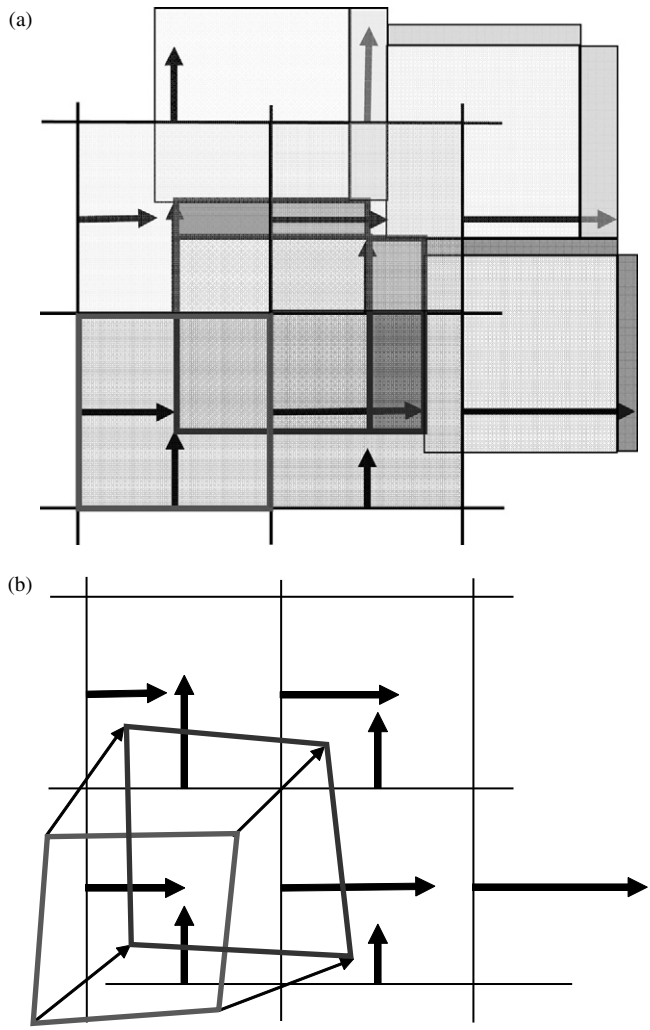


FIG. 3.4 Panel A: Illustrating the Lagrangian divergence $\mathcal{D}_k^n\left(\vec{V}_k^n\right) = \frac{1}{\Delta A}\frac{\delta A_k^{n+1}-\Delta A}{\Delta t}$, which corresponds to the Eulerian divergence (Eq. (3.116)). The periphery of a regular departure area, ΔA, is marked red and the periphery around its arrival area, δA_k^{n+1}, is marked blue. Additional departure and arrival areas for three neighbor cells are shown. Note that the areas δA_k^{n+1} do not cover the whole domain; there are cracks between them.

Panel B: Illustrating the Lagrangian divergence $D_k^n\left(\vec{V}_k^n\right) = \frac{1}{\Delta A}\frac{\delta A_k^{n+1/2}-\delta A_k^{n-1/2}}{\Delta t}$, where the departure area $\delta A_k^{n-1/2}$ is marked red and the arrival area $\delta A_k^{n+1/2}$ is marked blue. Obviously, $D_k^n\left(\vec{V}_k^n\right)$ is generally different from $\mathcal{D}_k^n\left(\vec{V}_k^n\right)$ in A. In both panels: Black arrows are velocity components in the C-grid. (See also color insert).

The variable on the left-hand side is the weight of the cell-integrated tracer mass per unit horizontal area (see Section 3.1.4). For a nonpassive tracer, source and sink terms should be added on the right-hand side of the equation. In a semi-implicit model, semi-implicit correction terms must be added to the explicit predicted tracer weights in order to make them consistent with the predicted moist air weights. The corrected semi-implicit tracer prediction equation must be identical to that for moist air (Eq. (3.115)) for $\left(\overline{\widehat{q}_i}\right)_k^n \equiv 1$. This means that the semi-implicit tracer continuity equation must be

$$\begin{aligned}\overline{\left(\widehat{q}_i^{\delta}\right)_k^{n+1}\,\delta\hat{p}_k^{n+1}} &= \left(\overline{\left(\overline{\widehat{q}_i}^{\delta}\right)_k^{n+1}\,\delta_k\overline{\hat{p}}^{n+1}}\right)_{\exp} \\ &\quad - \frac{\Delta t}{2}\left(\left(\overline{\widehat{q}_i}^{\Delta}\right)_k\Delta_k\overline{p}\right)^{ref}\left(\mathcal{D}_k^{n+1}\left(\vec{V}_k^{n+1}\right) - D\left(\tilde{\vec{V}}_k^{n+1}\right)\right) \\ &\quad + \frac{\Delta t}{2}\left(\left(\overline{\widehat{q}_i}^{\Delta}\right)_k\Delta_k\overline{p}\right)^{ref}\overline{\left[\mathcal{D}_k^n\left(\vec{V}_k^n\right) - D_k^n\left(\vec{V}_k^n\right)\right]_*}^{\delta}\frac{\delta A_k^n}{\Delta A}.\end{aligned}$$

3.2.3.3. The semi-implicit energy conversion term Also a dependence on divergence in the thermodynamic equation (Eq. (3.2)) needs to be temporally averaged in the semi-implicit model. Specifically, it is the energy converting term $\frac{\alpha\omega}{c_p} = \frac{R_d T_v}{c_p}\frac{\omega}{p}$, approximated in the explicit model with Eq. (3.63), which is divergence dependent. To isolate this dependence $\overline{\omega_k}^{n+1/2}$, given by Eq. (3.34), is expanded as follows:

$$\overline{[\omega_k]}_{\exp}^{n+1/2} = \frac{1}{\Delta t}\left(\overline{\hat{p}}_k^{n+1} - \overline{\left(\overline{p}_k^n\right)_*}^{\delta}\right) = \frac{1}{\Delta t}\left(\exp\left(-\alpha_k^{n+1}\right)\sum_{l=1}^{k}\overline{\delta_l\hat{p}}^{n+1} - \overline{\left(\overline{p}_k^n\right)_*}^{\delta}\right), \tag{3.117}$$

where Eqs. (3.17) and (3.22) have been used. When then $\overline{\delta_l\hat{p}}^{n+1}$ is substituted from Eq. (3.108), the result is

$$\begin{aligned}\overline{[\omega_k]}_{\exp}^{n+1/2} &= \frac{1}{\Delta t}\left(\exp\left(-\alpha_k^{n+1}\right)\sum_{l=1}^{k}\left(\overline{\Delta_l p}^n - \Delta t\,\overline{\Delta_l p}^n\,D_l^{n+1/2}\right) - \overline{\left(\overline{p}_k^n\right)_*}^{\delta}\right) \\ &= \frac{1}{\Delta t}\left(\exp\left(-\alpha_k^{n+1}\right)\sum_{l=1}^{k}\overline{\Delta_l p}^n - \overline{\left(\overline{p}_k^n\right)_*}^{\delta}\right) - \exp\left(-\alpha_k^{n+1}\right)\sum_{l=1}^{k}\overline{\Delta_l p}^n\,D_l^{n+1/2}.\end{aligned} \tag{3.118}$$

Thus, the explicit energy converting term may be written as

$$\left[\left(\frac{R_d T_v\omega}{c_p p}\right)_k\right]_{\exp}^{n+1}$$

$$= \frac{R_d(T_v)_k^{n+1/2}}{\left(c_p\right)_k^{n+1/2}\overline{p}_k^{n+1/2}}\overline{[\omega_k]}_{\exp}^{n+1/2} = \frac{R_d\,(T_v)_k^{n+1/2}}{\left(c_p\right)_k^{n+1/2}\exp\left(-\alpha_k^{n+1/2}\right)\overline{p}_{k+1/2}^{n+1/2}}\overline{[\omega_k]}_{\exp}^{n+1/2}$$

$$= \frac{R_d(T_v)_k^{n+1/2}}{\left(c_p\right)_k^{n+1/2}\exp\left(-\alpha_k^{n+1/2}\right)\overline{p}_{k+1/2}^{n+1/2}}\frac{1}{\Delta t}\left(\exp\left(-\alpha_k^{n+1}\right)\sum_{l=1}^{k}\overline{\Delta_l p}^{n} - \overline{\left(\overline{p}_k^n\right)}_*^{\delta}\right)$$

$$- \frac{R_d(T_v)_k^{n+1/2}\exp\left(-\alpha_k^{n+1}\right)}{\left(c_p\right)_k^{n+1/2}\exp\left(-\alpha_k^{n+1/2}\right)\overline{p}_{k+1/2}^{n+1/2}}\sum_{l=1}^{k}\overline{\Delta_l p}^{n} D_l^{n+1/2}, \tag{3.119}$$

where again Eqs. (3.17) and (3.22) have been used. When the last term is linearized about a reference temperature T^{ref} and a reference surface pressure p_s^{ref}, the result is

$$\left[\left(\frac{R_d T_v \omega}{c_p p}\right)_k\right]_{\exp}^{n+1}$$

$$= \frac{R_d\,(T_v)_k^{n+1/2}}{\left(c_p\right)_k^{n+1/2}\exp\left(-\alpha_k^{n+1/2}\right)\overline{p}_{k+1/2}^{n+1/2}}\frac{1}{\Delta t}\left(\exp\left(-\alpha_k^{n+1}\right)\sum_{l=1}^{k}\overline{\Delta_l p}^{n} - \overline{\left(\overline{p}_k^n\right)}_*^{\delta}\right)$$

$$- \sum_{l=1}^{k}\left(\overline{\frac{R_d\,(T_v)_k^{n+1/2}\exp\left(-\alpha_k^{n+1}\right)}{\left(c_p\right)_k^{n+1/2}\exp\left(-\alpha_k^{n+1/2}\right)\overline{p}_{k+1/2}^{n+1/2}}\Delta_l p}^{\,n}\right)' D_l^{n+1/2}$$

$$- \frac{R_d}{c_{pd}}\left(\frac{T}{p_{k+1/2}}\right)^{ref}\sum_{l=1}^{k}(\Delta_l p)^{ref} D_l^{n+1/2}. \tag{3.120}$$

Treating the linear term as a temporal average, we finally get the ("ideal") semi-implicit energy conversion term

$$\left(\frac{R_d T_v \omega}{c_p p}\right)_k^{n+1} = \left[\left(\frac{R_d T_v \omega}{c_p p}\right)_k\right]_{\exp}^{n+1}$$

$$- \frac{R_d}{c_{pd}}\left(\frac{T}{p_{k+1/2}}\right)^{ref}\sum_{l=1}^{k}(\Delta_l p)^{ref}\left(D_l^{n+1}\left(\vec{V}_l^{n+1}\right) + D_l^n\left(\vec{V}_k^n\right) - 2D_l^{n+1/2}\right) \tag{3.121}$$

or

$$\left(\frac{R_d T_v \omega}{c_p p}\right)_k^{n+1} = \left[\left(\frac{R_d T_v \omega}{c_p p}\right)_k\right]_{\exp}^{n+1}$$

$$- \frac{R_d}{c_{pd}}\left(\frac{T}{p_{k+1/2}}\right)^{ref}\sum_{l=1}^{k}(\Delta_l p)^{ref}\left(D_l^{n+1}\left(\vec{V}_k^{n+1}\right) - D\left(\tilde{\vec{V}}_l^{n+1}\right)\right), \tag{3.122}$$

where again $D\left(\tilde{\vec{V}}_l^{n+1}\right)$ is defined by Eq. (3.111).

In order to obtain the same uncomplicated elliptic equations as in HIRLAM, the predictor-corrector approach is again utilized. This changes Eq. (3.122) to

$$\left(\frac{R_d T_v \omega}{c_p p}\right)_k^{n+1} = \left[\left(\frac{R_d T_v \omega}{c_p p}\right)_k\right]_{\exp}^{n+1}$$

$$- \frac{R_d}{c_{pd}}\left(\frac{T}{p_{k+1/2}}\right)^{ref}\left[\sum_{l=1}^{k}(\Delta_l p)^{ref}\left(\mathcal{D}_l^{n+1}\left(\vec{V}_l^{n+1}\right) - D\left(\tilde{\vec{V}}_l^{n+1}\right)\right)\right.$$

$$\left. - \sum_{l=1}^{k}(\Delta_l p)^{ref}\overline{\left[\mathcal{D}_l^n\left(\vec{V}_k^n\right) - D_l^n\left(\vec{V}_k^n\right)\right]}_*^{\delta}\frac{\delta A_l^n}{\Delta A}\right]. \tag{3.123}$$

3.3. The NCAR-FFSL dynamical core

This section describes main features of the FV dynamical core, included in the NCAR Community Atmospheric Model (CAM 3.0) description, Collins, Rasch, Boville, Hack, Mccaa, Williamson, Kiehl, Briegleb, Bitz, Lin, Zhang and Dai [2004], Chapter 3.3. It is the pioneering example of a meteorological FV model based on the flux form (Eq. (1.2)) of the continuity equation. It was initially developed and used at the NASA Data Assimilation Office for data assimilation, numerical weather prediction, and climate simulations. The dynamical core is quasi-hydrostatic, global, and formulated for traditional latitude-longitude coordinates.

3.3.1. The 3D transport scheme

The quasi-horizontal transport of air mass, tracer mass, and potential temperature is based on a 2D FV FFSL scheme developed by Lin and Rood [1996] and Lin and Rood [1997]. This Eulerian scheme of the operator splitting or time splitting type, described in Section 2.3.2, is among the most modern flux-based FV schemes. For the sake of the following description of the dynamical core, it is convenient to rewrite the prediction equation, Eq. (2.52), with the notations of Lin and Rood [1997].

At first, the following standard finite difference δ and average $\overline{(\,)}$ operators are defined

$$\begin{aligned}\delta_\sigma q &= q\left(\sigma + \frac{\Delta\sigma}{2}\right) - q\left(\sigma - \frac{\Delta\sigma}{2}\right),\\ \overline{q}^{\sigma} &= \frac{1}{2}\left[q\left(\sigma + \frac{\Delta\sigma}{2}\right) + q\left(\sigma - \frac{\Delta\sigma}{2}\right)\right].\end{aligned} \tag{3.124}$$

The conservation law for a density-like variable Q is

$$\frac{\partial Q}{\partial t} + \nabla\cdot\left(Q\vec{V}\right) = 0. \tag{3.125}$$

As explained in Section 2.3.2, the FFSL scheme involves the application of 1D flux convergence operators and advective operators, successively applied along the two horizontal coordinate directions, in such a way that the scheme becomes both *conservative* and *constancy preserving*.

The 1D flux convergence operators F and G are defined as

$$\begin{aligned} F\big(u^*, \Delta\tau; Q^n\big) &= -\frac{\Delta\tau}{a\Delta\lambda\cos\varphi}\delta_\lambda\big[X\big(u^*, \Delta\tau; Q^n\big)\big], \\ G\big(v^*, \Delta\tau; Q^n\big) &= -\frac{\Delta\tau}{a\Delta\varphi\cos\varphi}\delta_\varphi\big[\cos\varphi Y\big(v^*, \Delta\tau; Q^n\big)\big]. \end{aligned} \tag{3.126}$$

F and G updates Q for one time-step in the zonal (λ) and meridional (φ) directions, respectively. Here, X and Y, the time-averaged fluxes of Q in the zonal (λ) and meridional (φ) directions, respectively, are defined as

$$\begin{aligned} X\big(u^*, \Delta\tau; Q^n\big) &\cong \frac{1}{\Delta\tau}\int_t^{t+\Delta\tau} uQ\,dt - hot \cong u^*\big(Q^n\big)^*_\lambda, \\ Y\big(v^*, \Delta\tau; Q^n\big) &\cong \frac{1}{\Delta\tau}\int_t^{t+\Delta\tau} vQ\,dt - hot \cong v^*\big(Q^n\big)^*_\varphi, \end{aligned} \tag{3.127}$$

where u^* and v^* are predicted time-centered velocity components at $t + \Delta\tau/2$ in C-grid positions at the east and south face of the cell, respectively. $(Q^n)^*_\lambda$ is determined by an upstream integral

$$\big(Q^n\big)^*_\lambda = \frac{1}{a\cos\varphi\Delta\lambda}\int_0^{u^*\Delta\tau} Q^n a\cos\varphi d\lambda, \tag{3.128}$$

with a corresponding expression for $(Q^n)^*_\varphi$.

Thus, to approximate the time-averaged fluxes across the cell faces the time-centered winds, u^* and v^*, and the cell averaged field, Q^n, at time level n are required. Furthermore, for modeling cross-stream advection in the zonal and meridional directions, respectively, the *advective* flux operators $\tilde{f}$ and $\tilde{g}$ are introduced. They are defined in Collins, Rasch, Boville, Hack, Mccaa, Williamson, Kiehl, Briegleb, Bitz, Lin, Zhang and Dai [2004] in terms of the corresponding F and G operators. Here, $\tilde{f}$ is defined as

$$\tilde{f}\big(u^*, \Delta\tau; Q^n\big) = F\big(u^*, \Delta\tau; Q^n\big) + \frac{\Delta\tau}{a\Delta\lambda\cos\varphi}\delta_\lambda u^*, \tag{3.129}$$

with a corresponding definition for $\tilde{g}$. With these definitions, the following prognostic equation (corresponding to Eq. (2.52)) results

$$\begin{aligned} Q^{n+1} = Q^n &+ F\left[u^*, \Delta\tau; Q^n + \frac{1}{2}\tilde{g}\big(v^*, \Delta\tau; Q^n\big)\right] \\ &+ G\left[v^*, \Delta\tau; Q^n + \frac{1}{2}\tilde{f}\big(u^*, \Delta\tau; Q^n\big)\right] \end{aligned} \tag{3.130}$$

or

$$Q^{n+1} = Q^n + F\big[u^*, \Delta\tau; Q^{\varphi}\big] + G\big[v^*, \Delta\tau; Q^{\lambda}\big], \tag{3.131}$$

where

$$Q^{\varphi} = Q^n + \frac{1}{2}\tilde{g}\big(v^*, \Delta\tau; Q^n\big) \quad \textit{and} \quad Q^{\lambda} = Q^n + \frac{1}{2}\tilde{f}\big(u^*, \Delta\tau; Q^n\big). \tag{3.132}$$

As in the HIRLAM-DCISL, the fluxes are assumed to be along 3D trajectories; however, here the trajectories are line-segment parallel to the coordinate axes. The final *vertical displacements* after each time-step are determined so that hydrostatic balance is maintained. A *Lagrangian vertical coordinate* ξ, is introduced (see Section 3.1.3.1), which per definition is constant along the 3D trajectories. The quasi-horizontal flow along such coordinate surfaces is 2D and relative to the coordinate surfaces, where the vertical velocity is zero as expressed in Eq. (1.1). The governing Eulerian equations, presented below in Section 3.3.2, are therefore without vertical advection terms.

In the present setup, the Eulerian vertical discretization defining the vertical extend of the Eulerian grid cells is similar to the hybrid sigma-pressure discretization (Simmons and Burridge [1981]) described in Section 3.1.2, except that here the top of the model atmosphere is at a constant pressure p_∞. Thus, the pressure at a η-model-surface is $p^n_{k+1/2} = p_\infty + A_{k+1/2} + B_{k+1/2}p^n_s$ and the vertical "pressure thickness" of an Eulerian grid cell is $\delta_k p^n = \Delta_k A + \Delta_k B\, p^n_s$. As illustrated in Fig. 3.1, the transport of air during a time-step, ending up in an Eulerian grid column, is effectuated by a Lagrangian cell, with initial pressure thickness $\delta_k p^n = p^n_{k+1/2} - p^n_{k-1/2}$. The Lagrangian cell is moving with the 3D flow and is ending up with a pressure thickness $\delta_k \hat{p}^n = \hat{p}^n_{k+1/2} - \hat{p}^n_{k-1/2}$. Its pressure ($\hat{p}^{n+1}_{k+1/2}$) is determined hydrostatically from the weight of the cells arriving above in the same Eulerian grid column. The scheme is globally conservative; however, it is less locally conservative than the DCISL schemes. Thus, as illustrated in Fig. 2.15(d), the FFSL scheme uses information from an area that is somewhat dispersed compared to the exact departure area.

3.3.2. The governing equations

Neglecting physical forcing terms, the governing continuous quasi-hydrostatic equations in spherical latitude-longitude coordinates are:

The hydrostatic balance equation on the form

$$\delta_k p = -g\rho\delta_k z, \tag{3.133}$$

which shows that the variable in the continuity equation, $\delta_k p$, is the weight per unit horizontal area in the layer of air with thickness $\delta_k z$.

The continuity equation or in other words the conservation law for mass is written as

$$\frac{\partial}{\partial t}\delta_k p + \frac{1}{a\cos\varphi}\left[\frac{\partial}{\partial\lambda}(u_k\delta_k p) + \frac{\partial}{\partial\varphi}(v_k\delta_k p\cos\varphi)\right] = 0. \tag{3.134}$$

Similarly, the mass conservation law for tracer species (including water vapor) is

$$\frac{\partial}{\partial t}\left((q_i)_k\,\delta_k p\right) + \frac{1}{a\cos\varphi}\left[\frac{\partial}{\partial\lambda}\left(u_k(q_i)_k\,\delta_k p\right) + \frac{\partial}{\partial\varphi}\left(v_k\,(q_i)_k\,\delta_k p\cos\varphi\right)\right] = 0. \tag{3.135}$$

The thermodynamic equation or the conservation law for potential temperature is

$$\frac{\partial}{\partial t}(\theta_k\delta_k p) + \frac{1}{a\cos\varphi}\left[\frac{\partial}{\partial\lambda}(u_k\theta_k\delta_k p) + \frac{\partial}{\partial\varphi}(v_k\theta_k\delta_k p\cos\varphi)\right] = 0, \tag{3.136}$$

where $\theta_k = (T_v)_k\left(\frac{p_k}{1000\,hPa}\right)^{\frac{R_d}{c_{pd}}}$ is the (virtual) potential temperature.
The momentum equations are used on the so-called "vector invariant form:"

$$\frac{\partial}{\partial t}u_k = \eta_k v_k - \frac{1}{a\cos\varphi}\left[\frac{\partial}{\partial\lambda}((e_{kin})_k + \phi_k - \nu D_k) + \frac{1}{\rho}\frac{\partial p_k}{\partial\lambda}\right], \tag{3.137}$$

$$\frac{\partial}{\partial t}v_k = -\eta_k u_k - \frac{1}{a}\left[\frac{\partial}{\partial\varphi}((e_{kin})_k + \phi_k - \nu D_k) + \frac{1}{\rho}\frac{\partial p_k}{\partial\varphi}\right], \tag{3.138}$$

where D_k is the divergence, defined as in Eq. (3.146), ν is the coefficient for an optional divergence damping, and the absolute vorticity η_k is

$$\eta_k = 2\Omega\sin\varphi + \frac{1}{a\cos\varphi}\left[\frac{\partial v_k}{\partial\lambda} - \frac{\partial(u_k\cos\varphi)}{\partial\varphi}\right]. \tag{3.139}$$

Here, Ω is the angular velocity of the earth. Finally, the kinetic energy $(e_{kin})_k$ is defined as

$$(e_{kin})_k = \frac{1}{2}\left((u_k)^2 + (v_k)^2\right). \tag{3.140}$$

3.3.3. Time-stepping

In the model, the dynamics and the NCAR CAM physics are time-split as in HIRLAM in the sense that all prognostic variables are updated sequentially, at first by the dynamics and then by the physics. The time-stepping is fully explicit with subcycling over small time-steps, $\Delta\tau = \Delta t/m$, within the 2D dynamics. The number of subcycles needed to stabilize the fast gravity waves is m. To avoid excessive small time-steps due to the convergence of the meridians near the poles, a polar Fourier filter, which filters out the shortest zonal waves, is applied to u^* and v^* and certain tendency terms in the prognostic equations. The transport for tracers, however, can take a much larger time-step Δt equal to the interval between the physics updates. In the present setup, the cells are transported along the Lagrangian surfaces during the long tracer time-steps, starting initially at the beginning of the first small time-step as a model layer with "pressure thickness" $\Delta_k p^n = \Delta_k A + \Delta_k B\,p_s^n$, without any remapping of the variables to the Eulerian model levels. So during these m short time-steps, the transport is fully Lagrangian. Only at the end of each tracer time-step, a remapping takes place. This is done to avoid excessive smoothing caused by too frequent vertical remapping after each small time-step.

3.3.3.1. Conservative predictions The *prognostic variables*, the cell averaged values $\delta_k p$, $(q_i)_k$, θ_k, u_k, and v_k, are updated by the use of the prognostic equations (Eqs. (3.134) – (3.138)), but only the Eqs. (3.134)–(3.136) for the density-related variables are in the proper flux-form (Eq. (3.125)) and are integrated directly by the FFSL prediction equation (Eq. (3.130)). Thus, the dynamical core conserves exactly mass of air, tracer mass, including water vapor (apart from evaporation and condensation), and potential temperature (in adiabatic friction free flow). The integration of the momentum equations, (Eqs. (3.137) and (3.138)) is discussed in Section 3.3.3.2.

At the start of time-step n (time $n\Delta\tau$), the prognostic variables, $\delta_k p$, $(q_i)_k$, θ_k, u_k, and v_k, are given in the D-grid as indicated in Fig. 3.5. In addition, the advective winds u^* and $v^*$are needed for the update of $\delta_k p$, $(q_i)_k$, and θ_k. So at first, these are updated to time $(n+1/2)\Delta\tau$on the C-grid using the momentum prediction equation (Eq. (3.144)). When they are available, Eqs. (3.134)–(3.136) can be advanced one time-step using the FFSL prediction form (Eq. (3.130)). u^* and v^* are not history carrying variables. They are overwritten after being used. When the continuity equation has been solved, the updated pressure thickness of each Lagrangian layer, $\overline{(\delta_k \hat{p})}^{n+1}$, determines the pressure of the Lagrangian surfaces by summing up the hydrostatic weight of all the cells above:

$$\overline{\hat{p}}^{n+1}_{k-1/2} = p_\infty + \sum_{l=1}^{k-1} \overline{\delta_l \hat{p}}^{n+1}. \tag{3.141}$$

Summing up the hydrostatic weight of all the NLEV Lagrangian layers yields the surface pressure:

$$\overline{p}^{n+1}_s = p_\infty + \sum_{l=1}^{\mathrm{NLEV}} \overline{\delta_l \hat{p}}^{n+1}. \tag{3.142}$$

This is needed for determination of the pressure at the interfaces between the Eulerian model layers:

$$\overline{p}^{n+1}_{k-1/2} = p_\infty + A_{k-1/2} + B_{k-1/2}\, \overline{p}^{n+1}_s, \tag{3.143}$$

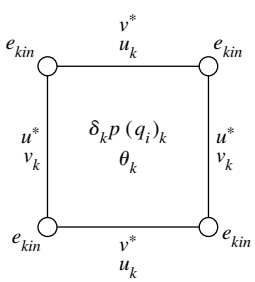


FIG. 3.5 Schematic stencil of the location of variables in finite volume schemes for the Arakawa C- (indicated with superscript *) and D-grids.

which are needed for the physical parameterization and the vertical remapping of the prognostic variables. The density-like prognostic variables are given after each large time-step, Δt, as mean values over the Lagrangian layers, $\overline{\delta_k \hat{p}}^{n+1}$. To be used for physical parameterization, they must be remapped on the Eulerian model layers $\overline{\Delta_k p}^{n+1}$ (determined from Eq. (3.143)).

3.3.3.2. Integration of the momentum equation Inspired by the papers of SADOURNY [1972] and ARAKAWA and LAMB [1981], a discretization of the momentum equations has been achieved that results in conservation of the absolute vorticity. Since also mass and potential energy are conserved and all three invariants are consistently transported, it might be expected that approximately the same will be the case for potential vorticity.

The resulting prognostic equations are

$$u^{n+1} = u^n + \Delta\tau \left\{ Y\left(v^*, \Delta\tau; \eta^\lambda\right) - \frac{1}{a\cos\varphi\,\Delta\lambda}\delta_\lambda[(e_{kin})^* - \nu D^*] + \widehat{P}_\lambda \right\},$$
$$v^{n+1} = v^n - \Delta\tau \left\{ X\left(u^*, \Delta\tau; \eta^\varphi\right) - \frac{1}{a\Delta\varphi}\delta_\varphi\,[(e_{kin})^* - \nu D^*] + \widehat{P}_\varphi \right\}, \tag{3.144}$$

where $(e_{kin})^*$, the upstream-biased kinetic energy (defined in the four corners of the D-grid (Fig. 3.5)), is formulated as

$$(e_{kin})^* = \frac{1}{2}\left\{ X\left(\overline{u^*}^\varphi, \Delta\tau; u^n\right) \text{`} + Y\left(\overline{v^*}^\lambda, \Delta\tau; v^n\right)\right\} \tag{3.145}$$

and

$$D^* = \frac{1}{a\cos\varphi}\left[\frac{\delta_\lambda u^{n+1}}{\Delta\lambda} + \frac{\delta_\varphi\left(\cos\varphi\, v^{n+1}\right)}{\Delta\varphi}\right]. \tag{3.146}$$

The FV mean pressure gradient terms $\widehat{P}_\lambda$ and $\widehat{P}_\varphi$ in Eq. (3.144) are computed by the method presented in LIN [1997], which eliminates a long-standing problem in terrain-following coordinates, i.e., the inaccuracy caused by different truncation errors in the two terms that the pressure gradient force traditionally are split into.

The velocity components u_k and v_k are given after each large time at Lagrangian levels, $\hat{p}_k$. Like the remaining prognostic variables, they need to be remapped to Eulerian levels. An accurate and conservative remapping procedure has been developed. The current remapping version, described in detail in COLLINS, RASCH, BOVILLE, HACK, MCCAA, WILLIAMSON, KIEHL, BRIEGLEB, BITZ, LIN, ZHANG and DAI [2004], conserves exactly mass, momentum, and total energy.

3.4. Properties of the dynamical cores

3.4.1. Conservation properties

With proper boundary conditions, the HIRLAM-DCISL dynamical core conserves exactly global mass of moist air and tracers, including water vapor, liquid water, and

solid water, if included (apart from evaporation and condensation). Also the NCAR-FFSL dynamical core conserves exactly these global masses. However, in addition, it conserves globally potential temperature (in adiabatic friction-free flow) and absolute vorticity (in adiabatic friction-free flow). Thus, the NCAR-FFSL dynamical core comes closer than the HIRLAM-DCISL to the ideal CSCL model considered in Section 3.1. However, as we have seen in the idealized tests presented in Section 2, the local conservation is more accurate in the HIRLAM-DCISL transport schemes than in the NCAR-FFSL scheme (compare Figs. 2.10(b) and 2.10(c) with Fig. 2.11(d)). Although the vertical remapping is designed to conserve the total energy, it is not globally conserved in NCAR-FFSL. The horizontal discretization and the use of a "diffusive" transport scheme with monotonicity constraint tend to decrease the kinetic energy and thereby the total energy. Whether this is realistic is difficult to assess. A total energy "fixer" is applied to effectively add the loss in kinetic energy due to "diffusion" back to the model as total potential energy so that the total energy is globally conserved. However, even without the fixer, the loss is found to be very small, less than 2 W/m^2 with a 2-degree resolution and it is found to decrease with increasing resolution. It is stated in Collins, Rasch, Boville, Hack, Mccaa, Williamson, Kiehl, Briegleb, Bitz, Lin, Zhang and Dai [2004] that in the future it may considered to use the total energy as a transported prognostic so that the total energy could be automatically conserved. In the HIRLAM-DCISL dynamical core, no total energy "fixer" is applied.

3.4.2. FV Lagrangian pressure gradient force

As mentioned above, a particular feature of the NCAR-FFSL model is the Eulerian expression for the FV mean pressure-gradient force (Lin [1997]), which is used in the model. It eliminates a long-standing problem: the inaccuracy caused by different truncation errors in the two terms that the pressure gradient force traditionally is split into in terrain-following coordinates. A similar finite volume Lagrangian expression for the mean pressure gradient force along the trajectories during a time-step is suggested by Lauritzen, Kaas, Machenhauer and Lindberg [2008]. It may be used in any DCISL model. It has not yet been implemented in HIRLAM-DCISL; however, it is expected that it may lead to increased accuracy, even though the linearized pressure gradient force (Eq. (3.104)), used to derive the semi-implicit correction terms, is based on the two-term expression of the pressure gradient force. The proposed Lagrangian mean pressure gradient force $PGF_{\Delta s'}$ along the sloping trajectory $\Delta s'$is easily computed from the pressure of the arrival cell, at the end of the trajectory, $\left(\overline{\hat{p}}_k^{n+1}\right)_{\text{exp}}$ and the pressure of the departure cell, at the start of the trajectory, $\overline{\left(\overline{p}_k^n\right)}_*^{\delta}$. Note that these are the same pressures that are used to define ω in Eq. (3.117). The proposed expression is

$$PGF_{\Delta s'} = \left[\left(-\frac{1}{\rho}\frac{\partial p}{\partial s'}\right)_k\right]_{\text{exp}}^{n+1/2} = -\frac{1}{\overline{\overline{\rho}}}\frac{\left(\overline{\hat{p}}_k^{n+1}\right)_{\text{exp}} - \overline{\left(\overline{p}_k^n\right)}_*^{\delta}}{\Delta s'}$$

$$= -\frac{1}{\overline{\overline{\rho}}}\frac{\left(\overline{\hat{p}}_k^{n+1}\right)_{\text{exp}} - \overline{\left(\overline{p}_k^n\right)}_*^{\delta}}{\Delta s}\cos(\vartheta). \tag{3.147}$$

$\overline{\overline{\rho}}$ is an approximation of the mean density along the trajectory. Δs is the horizontal distance between the midpoints of the departure and the arrival cells, and $\Delta s'$ is the corresponding distance along the sloping trajectory. $\cos(\vartheta) = \frac{\Delta s}{\Delta s'}$ defines the slope of the trajectory. In order to determine the horizontal component of the pressure gradient force PGF_h, the vertical component PGF_v, which is balanced by gravity g, must be subtracted:

$$PGF_h = \sqrt{(PGF_{\Delta s'})^2 - g^2}. \tag{3.148}$$

3.4.3. Tests of performance

THUBURN [2006] has made an attempt to estimate the relative importance of different conservation laws. It is argued that satisfactory model performance requires spurious sources of a conservable quantity to be much weaker than any true physical sources; for several conservable quantities, the magnitudes of the physical sources are estimated in order to provide benchmarks against which any spurious sources may be measured. A model with weak spurious sources of a conservable quantity compared with the physical sources may in practice produce as accurate forecasts, especially long simulations, as a model which conserves the quantity exactly if the spurious sources are not systematic. However, even if the spurious sources are weak but the spurious sources are systematic, long simulations may be very inaccurate. Of course, if possible, the spurious sources should be estimated relative to the physical sources, but that may be very difficult and it is not enough; it is also necessary to know if the spurious sources are systematic. So, in practice, as when other potential model improvements are considered, it is necessary to carry out a series of real cases that is validated against observations in competition with any model it is supposed to substitute. To the authors' knowledge, such a series of real case tests have not yet been carried out for any of the two FV dynamical cores considered here.

3.4.4. Idealized baroclinic wave test

However, both have been preliminary tested in the idealized baroclinic wave test case of JABLONOWSKI and WILLIAMSON [2006a], in the following called JW06a (a more detailed technical note JABLONOWSKI and WILLIAMSON [2006b] is available). These tests have, as we shall see, confirmed that both dynamical cores work properly, producing in general as realistic baroclinic developments as present day's nonconservative state-of-the-art dynamical cores. It would have been of considerable interest to validate the importance of the mass conservation property of the FV dynamical cores considered here in real case tracer transport simulations. There seems to be no doubt that this property is essential for tracer transport however to the authors' knowledge, it has not yet been verified.

The idealized baroclinic wave test case of JW06a consists of an analytic steady-state zonal solution to the global primitive equations. The steady-state is unstable so that an overlaid perturbation in global reference integrations triggers the development of an idealized baroclinic wave in the northern hemisphere. By day 4, a well-defined wave train is established, and by days 7–9, a significant deepening of the highs and lows takes place before a breakdown by days 20–30 leads to a full circulation in both hemispheres.

3.4.5. Idealized test of NCAR-FFLS

JW06a applied the test to four different global dynamical cores at varying horizontal and vertical resolutions. Namely, the NCAR Eulerian three-time-level semi-implicit spectral transform dynamical core (EUL), the NCAR two-time-level semi-Lagrangian semi-implicit spectral transform dynamical core (SLD), the German Weather Service icosahedral finite-difference three-time-level semi-implicit dynamical core (GME), and finally the NCAR-FFSL FV dynamical core.

3.4.6. Diffusion processes

Before summarizing the performance of FV in the test, it is relevant to list the diffusion processes which had to be included in all the dynamical cores considered in order to ensure stable integrations.

- EUL includes a ∇^4horizontal diffusion on temperature, divergence, and vorticity to control the energy on the smallest resolved scales and a ∇^2 horizontal diffusion on the top three levels to control upward propagating waves. The thermodynamic equation includes a frictional heating term corresponding to the momentum diffusion. It includes also *a posterior* mass fixer applied at every time-step. The three-time-level core includes a time filter to control the $2\Delta t$ time computational mode.
- SLD do not include the ∇^4 and ∇^2 horizontal diffusions; the interpolants control the energy at the smallest scales. Every time-step *posterior* mass and energy fixers are applied. A standard decentering parameter $\varepsilon = 0.2$ is used in the semi-implicit scheme.
- GME includes the ∇^4 and ∇^2 horizontal diffusions as EUL. Neither a mass fixer nor an energy fixer is applied.
- FV do not include explicit $\nabla^4$and ∇^2 horizontal diffusion operators; the horizontal remapping, using a monotonic PPM sub-grid representation, is supposed to control the energy at the smallest scales. An explicit divergence damping is, however, applied. The monotonic and conservative vertical remapping is performed every 10 explicit time-steps. FV employs in addition both a 3-point digital filter in mid-latitudes and an FFT filter in polar regions to control unstable waves in the zonal direction. A *posterior* energy fixer is applied at every time step.

3.4.7. Resolution, time step, run times

After the addition of small perturbation to the unstable steady-state zonal flow in the four dynamical cores, they are run for 30 model days with different horizontal resolutions. The five horizontal resolutions and the corresponding time-steps used in the FV integrations are shown in Table 3.1. The other models were run with five approximately equivalent resolutions. The time-steps used in the integrations with the different dynamical cores are also included in the table. All these integrations were run with 26 standard vertical levels (L26).

JW06a also publish the runtimes for the four dynamical cores at their midrange and second highest resolutions. They are added in Table 3.1. They are meant to serve as a general guide for the computational costs of each model acknowledging that the cost are

TABLE 3.1
Horizontal resolutions (FV), time-step Δt (s) and wall clock time WT (s) for one model day

Resolution (FV) $\Delta\varphi \times \Delta\lambda$	EUL Δt/WT	SLD Δt/WT	GME Δt/WT	FV $\Delta\tau(=\Delta t/10)$/WT
$4^\circ \times 5^\circ$	2400	7200	1600	720
$2^\circ \times 2.5^\circ$	1200	3600	800	360
$1^\circ \times 1.25^\circ$	600/44	1800/24	400/48	180/66
$0.5^\circ \times 0.625^\circ$	300/483	900/271	200/325	90/625
$0.25^\circ \times 0.3125^\circ$	150	450	100	45

hardware-dependent and vary with the ingenuity of the programmer. The runtime data represent the wall clock time (WT) needed to complete one model day on a 32-processer node of an IBM (International Business Machines) Power 4 architecture when using a pure message passing interface parallelization approach. Identical compiler optimization flags were used for all models. No efforts was made to optimize the numerical schemes or to configure the models in their optimal setup, such as selecting an optimal time-step or switching from quadratic to a linear truncation technique in case of SLD. The dynamical cores represent the standard versions in CAM3 (The NCAR Community Atmospheric Model system, version3) and GME.

The FV model is seen to be the most expensive in computational costs. Thus, the conservative property is achieved on the expense of efficiency. At the medium resolution, it is 2.75 times more expensive than the semi-Lagrangian semi-implicit SLD dynamical core. However, it is also more expensive than the two Eulerian semi-implicit dynamical cores at medium resolution: 50% more expensive than EUL and 38% more expensive than GME. At the second highest resolution, it is still the most expensive dynamical core. Its WT has increased by a factor 9.5, whereas those of EUL and SLD increased by a factor of 11.0 and 11.3, respectively. However, GME increased much less, by a factor of only 6.8. So at this resolution, FV is 92% more expensive than GME. The advantage of EUL on the other hand has been reduced, but FV is still 29% more expensive than EUL.

It should be noted that here FV is compared with operational dynamical cores that have been carefully optimized. It must be noted furthermore that with massive parallel computers with a high number of nodes, expected to be more common in the future, the FV explicit code is supposed to gain in relative efficiency due to better parallelization than Eulerian as well as semi-Lagrangian semi-implicit codes with elliptic solvers that involve more data exchange between nodes.

3.4.8. Results of the global simulations

As already mentioned, the JW06a tests indicated that the synoptic performance of the FV (or NCAR-FFSL) dynamical core generally is satisfactory. It is producing as realistic idealized baroclinic developments as the state-of-the-art dynamical cores it is compared with. All four dynamical cores compared are found to converge toward a common

solution. Thus, the second-highest and highest resolution FV L26 surface pressure solutions at day 9 are visually almost indistinguishable (see Fig. 6 in JW06a). Up to day 10, differences between the solutions from different dynamical cores can only be seen at the smallest scales that are most influenced by the diffusive characteristics of the numerical schemes, summarized above. An example is the closed cells in the low-pressure center of the surface pressure fields at day 9 (Fig. 7 in JW06a) of the EUL and GME solutions. They are slightly deeper than those from the FV dynamical core. Such small-scale differences are seen more clearly in the 850 hPa relative vorticity fields (shown in Fig. 8 in JW06a). At day 7, the high-resolution FV dynamical core exhibits a slightly weaker vorticity pattern in comparison with EUL, SLD, and GME at high resolutions. According to JW06a, the slightly weaker vorticity fields are caused by the frequent remappings with monotonicity constraint for every short dynamic time-step in the FV dynamical core. This constraint adds nonlinear intrinsic diffusion in the regions where the monotonicity principle is locally violated. Note that JW06a has shown that the slightly more diffusive solution of the FV dynamical core can be matched very closely by EUL and SLD when increasing their diffusive coefficients. So there is no doubt that the excessive smoothing in FV is caused by the intrinsic diffusion caused by the frequent remappings at the end of every short dynamic time-step. Since the vertical remappings are performed only every 10 dynamic time-step, their smoothing effects are less pronounced.

The small-scale differences between the solutions of the different dynamical core were interpreted as an uncertainty of their individual estimates of the true reference solution. JW06a defined the uncertainty as the maximum root mean square deviation l_2 between a highest and a second highest horizontal resolution surface pressure simulation among all model versions (see JW06a for details). The uncertainty is increasing with the number of days simulated, becoming more and more large scale, until saturation between day 25 and 30, when the l_2 difference is as big as the l_2 difference between two randomly picked global surface pressure fields. Using this uncertainty measure, JW06a found that both two highest resolutions of the FV, the SLD and the EUL dynamical cores, converge within the estimated uncertainty to the true solution, whereas only the highest resolution of the GME dynamical core was found to converge.

3.4.9. Idealized test of HIRLAM-DCISL

Also the HIRLAM-DCISL dynamical core has been tested with the Jablonowski-Williamson test case. As it would be difficult to extend the limited area of the HIRLAM-DCISL dynamical core to a global domain, its domain was made as global as possible and an effort was made to minimize the effects of its boundaries. The active domain was extended meridionally to 80°S – 80°N and zonally to 80°W – 280°E, without changing the zero divergence boundary condition, used in the elliptic system solver, to a periodic boundary condition at the zonal boundaries. The zonal extension was chosen so that the initial perturbation, centered at (20°E, 40°N), which triggers a *main wave* is separated (by exactly 100°) from the western domain boundary where the boundary scheme initially trigger a weak *boundary wave*. Both waves develop into wave trains which become less and less separated, although they move with approximately the same speed toward the east. The usual HIRLAM boundary relaxation scheme is applied in a 6°

wide zone along the boundary inside the active domain. Within this zone, the updated prognostic variables are relaxed toward the initial values with a weight that decrease from 1 at the boundary to zero approximately 6° inside it. To accommodate the DCISL upstream integrations, there is also a halo zone around the active domain in which the prognostic variables are held fixed at the initial values. In order to facilitate a comparison with the global FV dynamical core reference solution, the boundary wave was effectively eliminated from both the HIRLAM-DCISL and the HIRLAM solutions. This was done by utilizing that a completely similar boundary wave is created in simulations without an initial perturbation. The simulations with the boundary wave removed were then compared with the global reference solution. This was done only up to day 8 after which the main wave reached the eastern boundary zone.

3.4.10. Resolution, time-step and runtimes

Two horizontal resolutions were used. The lower resolution, corresponding to the middle global resolution (see Table 3.1), is $\Delta\varphi \times \Delta\lambda = 1.15° \times 1.45°$ and the higher resolution, corresponding to the second highest global resolution, is $\Delta\varphi \times \Delta\lambda = 0.59° \times 0.74°$. In the vertical, 27 levels are placed as in JW06b but with one more level at the top of the model atmosphere to accommodate the zero top pressure of HIRLAM. Like the global SLD dynamical core, the time-steps used for the two horizontal resolutions were 30 and 15 minutes, respectively. On a single NEC SX6 processor using ad hoc coding with almost no optimization, the lowest resolution HIRLAM-DCISL dynamical core is approximately twice as expensive as the corresponding highly optimized reference HIRLAM dynamical core. There is no doubt, however, that the efficiency of the HIRLAM-DCISL dynamical core can be increased considerably by a dedicated optimization.

3.4.11. Diffusion processes

- HIRLAM use decentering with a decentering parameter $\varepsilon = 0.1$. The nonlinear terms in continuity equation, the thermodynamic equation, and the momentum equations are needed at time level $n + 1/2$. As they are potential sources of instability, they are extrapolated from filtered values at time level $n - 1$ as follows: $\psi^{n+1/2} = (3\psi^n - \psi_f^{n-1})/2$, where ψ is any of the nonlinear terms and $\psi_f^{n-1} = \psi^{n-1} + \varepsilon_N[\psi^n - 2\psi^{n-1} + \psi_f^{n-2}]$ with $\varepsilon_N = 0.1$. At the end of each time-step, all prognostic variable, except liquid water, are diffused using an approximate implicit ∇^4 horizontal diffusion with the diffusion coefficient $K = 3.5 \times 10^{14}$ for $\Delta x = 0.5°$ and $\Delta t = 300\,\mathrm{s}$ (see p. 12–13 in Undén 2002). The coefficients are scaled for resolution so that the e-folding time of the $2\Delta x$ wave is the same regardless of resolution (McDonald, 1998). There has been no attempt to tune the diffusion coefficient for the present idealized dry adiabatic simulations. The horizontal diffusion was increased at the uppermost 4 model layers. In addition, the horizontal and vertical interpolations, using cubic Lagrange interpolation, are supposed to control the energy at the smallest scales. No mass and energy fixers are applied.
- HIRLAM-DCISL does not use decentering and filtering of the nonlinear terms, but it was necessary to retain a weak implicit ∇^6 horizontal diffusion on T, u, and v. The horizontal diffusion was increased at the uppermost 4 model layers. In addition, the horizontal and vertical remapping and interpolations, using, respectively, a

positive-definite PPM subgrid representation and cubic Lagrange interpolation, are supposed to control the energy at the smallest scales. No energy fixer is applied

3.4.12. Results of simulations

For both resolutions of HIRLAM and HIRLAM-DCISL, the l_2 difference between the simulation and the simulation of the global highest resolution FV dynamical core ($\Delta\varphi \times \Delta\lambda = 0.25° \times 0.3125°$) was computed. The results showed that up to day 8 both the highest resolution ($\Delta\varphi \times \Delta\lambda = 0.59° \times 0.74°$) simulation of HIRLAM and the highest resolution simulation of HIRLAM-DCISL had converted within the uncertainty of the reference solution. For the lower resolution simulations, the simulation of the HIRLAM-DCISL version had not converted, whereas that of the HIRLAM version had; so the FV version needs higher resolution than the grid-point version for the same level of accuracy (Fig. 3(a) in LAURITZEN, KAAS, MACHENHAUER and LINDBERG [2008]). The explanation seems, as for the global FV dynamical core, to be too heavy smoothing due to the repeated remappings and interpolations. Regarding phase error, the HIRLAM-DCISL simulation is slightly better than the HIRLAM simulation. When using the cascade scheme of NAIR, SCROGGS and SEMAZZI [2002] instead of the fully 2D CISL scheme of NAIR and MACHENHAUER [2002], the accuracy in terms of the l_2 difference is not altered (Fig. 3(b) in LAURITZEN, KAAS, MACHENHAUER and LINDBERG [2008]). An important result of the idealized baroclinic wave tests is that the consistent Lagrangian discretization of the energy conversion term, introduced in Section 3.2.1, is seen clearly to be better with both a smaller l_2 difference and a smaller phase error, than when using the traditional Eulerian discretization (Fig. 3(b) in LAURITZEN, KAAS, MACHENHAUER and LINDBERG [2008]).

HIRLAM-DCISL has also been coupled with the HIRLAM physics package and initial test runs from the initial conditions of a strongly developing extratropical storm have been performed. The mass conserving version ran stably and produced simulations that were quite similar to the reference HIRLAM simulations, except again for slightly more smoothing in the DCISL version. Also for the full-physics run, the Lagrangian discretization of the energy conversion term leads to a more accurate simulation than the traditional discretization. The results of these tests are mentioned in LAURITZEN, KAAS, MACHENHAUER and LINDBERG [2008]. In this paper also a possible cure for the slightly excessive smoothing is suggested, although it was not tested in practice. It is suggested to keep the Lagrangian cells in the Lagrangian model layers for a number of consecutive large semi-implicit time-steps before performing the vertical remapping and interpolation to the Eulerian model layers and levels, just as it is done in the NCAR-FFSL over 10 consecutive small time-steps. In HIRLAM-DCISL, an additional vertical remapping and interpolation must be performed after each of the long semi-implicit time-steps as it is needed for the physical parameterization. Thus, it will not affect the computational efficiency.

3.5. Online and offline applications – The problem of mass-wind inconsistency

An obvious application of FV models such as the NCAR-FFSL and the HIRLAM-DCISL is tracer transport since tracer-mass conservation is important. These quasi-hydrostatic FV models use a pressure-based vertical coordinate and the prognostic variable for tracer

mass is the cell averaged value $\overline{\overline{q_i}}^{\Delta} \Delta\overline{p}$, where $\overline{p}$ is the horizontal mean pressure over the area ΔA of an Eulerian grid cell and $\Delta\overline{p}$ is the pressure thickness of the cell. $\overline{\overline{q_i}}^{\Delta}$ is the Eulerian cell average specific concentration[6] of the tracer in question. Thus, $\overline{\overline{q_i}}^{\Delta} \Delta\overline{p}$ is the weight of tracer mass in the cell per unit horizontal area.[7] At each time-step, the FV model solves at first the continuity equation for air mass. The input to the continuity equation for air mass is the horizontal wind field which together with the hydrostatic balance determines the 3D trajectories along which the air is transported. The output is the updated values of $\Delta\overline{p}$. Hereafter, the continuity equation for each specific tracer mass is solved, using the same trajectories, giving the updated values of $\overline{\overline{q_i}}^{\Delta} \Delta\overline{p}$. Thus, both the predicted air mass and the predicted tracer mass fields are consistent with the "driving" horizontal velocity field. If, the specific concentration q_i is needed (e.g., for a compability check as illustrated in Fig. 2.1) $\overline{\overline{q_i}}^{\Delta} \Delta\overline{p}$ must be divided by $\Delta\overline{p}$. In a model setup where the tracer continuity equation is an integrated part of the dynamical core, this can be done without loss of mass because $\overline{\overline{q_i}}^{\Delta} \Delta\overline{p}$ and $\Delta\overline{p}$ are internally consistent, i.e., they are both computed by the same mass conserving transport and remapping operations.

Often a FV transport scheme is imported into a GCM, which dynamical core does not conserve the mass of air locally. For example, in the Eulerian ECHAM5 (ECmwf/HAMburg, version 5) model, where the vertically integrated mass variable is log p_s the continuity equation for air is solved using the spectral transform method. This model is neither globally nor locally mass conserving, while the tracer transport is performed using the inherently mass-conserving advection scheme of Lin and Rood [1996]. This set up is called an *online coupling*. Here, the FV transport scheme is solved on the same grid as used by the GCM and the horizontal GCM winds $\vec{V}_{GCM}$ needed by the transport scheme are provided by the GCM at every GCM time-step. A problem of such an online coupling is that the GCM predicts its own *air mass field* Δp_{GCM}, which is generally different from the $\Delta\overline{p}$ predicted from *the GCM wind field* by the mass conservative tracer transport scheme with $q_i = 1$ (see Fig. 3.6). This is a manifestation of the so-called *mass-wind inconsistency* discussed in detail by Jöckel, Von Kuhlmann, Lawrence, Steil, Brenninkmeijer, Crutzen, Rasch and Eaton [2001].

The consequences of the mass-wind inconsistency in long online coupled simulations can be severe. Jöckel, Von Kuhlmann, Lawrence, Steil, Brenninkmeijer, Crutzen, Rasch and Eaton [2001] ran a low-resolution FV transport scheme online coupled to a nonmass conserving GCM. For passive tracers initialized at different locations in the atmosphere, the variations in the total mass were up to 70% in a one-year simulation. The amount of artificially (spuriously) created and destroyed mass due to the mass-wind inconsistency is strongly dependent on the vertical gradient of the tracer. Since tracer gradients are usually steepest around the tropopause the problem is large in the tropopause region.

[6] *The specific tracer concentration* q_i is the ratio between the mass of the tracer m_t and the mass of the moist air it is mixed into m_v.

[7] $\overline{\overline{q_i}}^{\Delta} \Delta\overline{p} = \frac{m_t}{m_v} \overline{\overline{\rho}}_v g \Delta z = \frac{m_t}{m_v} \frac{m_v}{\Delta A \Delta z} g \Delta z = g \frac{m_t}{\Delta A}$ = the weight of tracer mass per unit horizontal area.

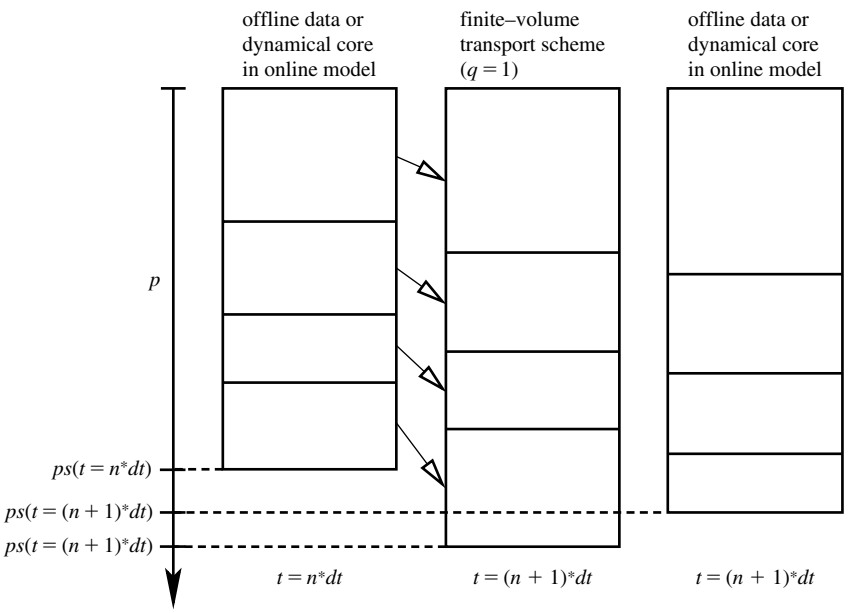


FIG. 3.6 Graphical illustration of the mass-wind inconsistency. The figure shows the location of pressure levels at the beginning of a time-step $t = n\Delta t$ (left), after one time-step $t = (n + 1)\ \Delta t$ using a FV transport scheme (middle) and given by offline data or predicted by the continuity equation of the dynamical core (right), respectively. If the vertical levels implied by the transport scheme and the dynamical core or offline data do not coincide, an inconsistency between the mass and wind fields exists and affects the mass of the tracer advection.

If the wind and pressure data driving a FV transport model are not given at every time-step and may be specified on another grid than used in the FV transport model (typically an archived meteorological data-set such as ECMWF reanalysis (REA), ERA40, or the National Centers for Environmental Prediction (NCEP)/NCAR REA) so that both interpolation in time and space is needed, then the coupling is called *offline.* This is typically the situation in a *chemical transport model*. (It should be noted that this does not apply to the NCAR-FFSL model although this model for the tracer transport uses a large time-step that is equal to an integer number m of the small time-steps, which are used for determining the air mass transport. This is because the fluxes used in the tracer transport during a large time-step is obtained by accumulating fluxes over the m small time-steps.) In an *offline* setting, assimilated analysis or REA data is often provided to the FV transport scheme. Both the driving horizontal wind field V_{REA} and the corresponding mass field Δp_{REA} must be spatially and temporally interpolated to accommodate the grid and time-step used by the transport scheme. In such a situation, the consistency between the mass and wind fields cannot be achieved unless a posteriori consistency correction methods are applied to the *offline* data. This inconsistency with respect to the ECMWF analyses is discussed in TRENBERTH [1991].

One can attempt to restore the mass conservation by altering the specific tracer concentrations a posteriori. JÖCKEL, VON KUHLMANN, LAWRENCE, STEIL, BRENNINKMEIJER,

CRUTZEN, RASCH and EATON [2001] investigated the effects of various a posteriori mass-fixing algorithms. But all these fixers have severe disadvantages such as violation of shape preservation or introduction of non physical transport components.

Instead of altering the specific tracer concentrations a posteriori, one can as well adjust the horizontal velocity field such that the tracer advection equation is consistent with the mass field, i.e., the winds are corrected so that the vertical integrated divergence of mass matches the surface pressure tendency of the meteorological data (for details see PRATHER, MCELROY, WOFSY, RUSSEL and RIND [1987]; ROTMAN, TANNAHILL, KINNISON, CONNELL, BERGMANN, PROCTOR, RODRIGUEZ, LIN, ROOD, PRATHER, RASCH, CONSIDINE, RAMAROSON and KAWA [2001]; CAMERON-SMITH, CONNELL and PRATHER [2002])[8]. This type of restoration algorithm is referred to as a *pressure fixer*. Contrary to the algorithms described in the preceding paragraph, it ensures that constant specific concentrations and mass conservation are retained with this pressure fixer. However, the approach is not completely satisfactory either since "true" wind data are enforced to provide mass-wind consistency. Introduction of a pressure fixer in a semi-Lagrangian FV model would be somewhat different from that used by Eulerian type FV models since divergence is defined directly by the trajectories. Therefore modifications of trajectories are needed in such models to achieve an analogy to the traditional pressure fixer.

The pressure fixer method can be used to indicate the severity of the mass-wind inconsistency problem, i.e., by running a model with and without a pressure fixer and assuming that the pressure fixer does not have a significant effect on the wind field. HOROWITZ, WALTERS, MAUZERALL, EMMONS, RASCH, GRANIER, TIE, LAMARQUE, SCHULTZ, TYNDALL, ORLANDO and BRASSEURET [2003] have run the global Model of Ozone Research version 2 (MOZART-2) with and without a pressure fixer. Near the tropopause (where the vertical gradient of the ozone specific concentration is large) the difference between the two runs was approximately 187 Tg/yr. Assuming that the pressure fixer is perfect, it can be estimated that a spurious source of ozone of 187 Tg/yr is caused by the mass-wind inconsistency problem which is not a negligible amount. For example, the spurious source of ozone is similar in magnitude to the estimated amount of influx of ozone per year to the troposphere from the stratosphere. This is, of course, only an indication of the magnitude of the problem. In order to estimate the systematic spurious sources and sinks a fully consistent model must be run. But the estimates provided by HOROWITZ, WALTERS, MAUZERALL, EMMONS, RASCH, GRANIER, TIE, LAMARQUE, SCHULTZ, TYNDALL, ORLANDO and BRASSEURET [2003] suggest that the mass-wind inconsistency can introduce significant errors.

As discussed above, the problem of performing accurate offline or online tracer transport in a model using a pressure-based vertical coordinate is not limited to the use of an accurate FV tracer transport scheme, but it is also a question of mass-wind consistency. That is consistency between on one hand the mass field $\Delta\bar{p}$ determined by the FV transport scheme and the driving wind field and on the other hand the associated mass field Δp_{REA} or Δp_{GCM}. In offline applications using existing REA data, there is little choice but to use some kind of correction method. However, it is hoped that

[8]Alternatively one may adjust the surface pressure field instead of the horizontal velocity field to achieve consistency (P. JÖCKEL personal communication).

in the future analysis data set with a better inherent mass conservation, produced by mass conserving data-assimilation models, may become available. In an online transport coupling, the consistency can be guaranteed only if the same numerical method is used for the continuity equation of the driving model as for the tracer transport. That is the case if a consistent FV model like the NCAR-FFSL or the HIRHAM-DCISL is used. The problem is that the majority of the GCM models available are based on nonlocally conservative schemes for the air mass continuity equation (e.g., traditional semi-Lagrangian models such as the IFS at ECMWF, the HIRLAM, and the Max-Planck Institute model (ECHAM)). The problem is, as described above, that the changes needed to convert a non-conserving model to a locally mass conserving one, like the change of HIRLAM to the HIRLAM-DCISL, are rather extensive. All discretizations in the model as a whole must be carefully rethought in order to obtain inherent local mass conservation. It involves changes to almost all parts of the model and all the prognostic equations and not only changes to the continuity equations. However, it is necessary in order to guarantee consistency and thereby accurate tracer transport.

3.6. Extensions to nonhydrostatic models

In the quasi-hydrostatic FV dynamical cores considered in Sections 3.1–3.5, the continuity equation was solved in so-called Lagrangian (ξ) vertical coordinates, with, per definition, $\xi =$ constant along 3D trajectories. That is, all transport during a time-step was assumed to be effectuated by Lagrangian finite control volumes moving with the 3D flow, usually in a semi-Lagrangian sense, starting or ending as Eulerian grid cells. The advantage is that in such Lagrangian coordinates, the vertical velocity is zero as expressed in Eq. (1.1), so the transport problem becomes 2D. However, the vertical components of the 3D trajectories still need to be determined. In the quasi-hydrostatic models considered so far, this is done hydrostatically, i.e., the vertical displacement of the Lagrangian control-volume during a time-step was determined by requiring that the arrival cell is in hydrostatic balance. Of course, this approach cannot be used in a nonhydrostatic model, but instead the vertical displacement can be determined directly from the vertical velocity, which in a nonhydrostatic model is an independent prognostic variable. Realizing this one can formulate a DCISL solution to the continuity equation also for the nonhydrostatic case.

We may start from the continuity equation on the form (Eq. (1.8)), which was derived without assuming hydrostatic balance. Using the notations in Section 1.1 Eq. (1.8) is

$$\overline{\left(\tilde{\rho}\ \delta_k h\right)}^{+} \Delta A = \overline{\left(\tilde{\rho}\ \delta_k h\right)}\, \delta_k A. \tag{3.149}$$

This is a prognostic equation when $\left(\tilde{\rho}\ \delta_k h\right)$ is known in the departure area at time t. Unlike what we did in the "exact" case, we now set the height of the Lagrangian surfaces $\xi_{k-1/2}$ and $\xi_{k+1/2}$ equal to the Eulerian surfaces $h_{k-1/2}$ and $h_{k+1/2}$, respectively, in the departure area so that here $\delta_k h = \Delta_k h$. Thus, Eq. (3.149) becomes

$$\overline{\left(\tilde{\rho}\ \delta_k h\right)}^{+} \Delta A = \Delta_k h\, \overline{\left(\tilde{\rho}\right)}\, \delta_k A, \tag{3.150}$$

Now in the nonhydrostatic case, we utilize the prognostic variable w, the vertical velocity, to determine $\delta_k h$. So it is a known constant in Eq. (3.150). Therefore, it may be written as

$$\overline{(\tilde{\rho})}^{+}\,\delta_k h \Delta A = \Delta_k h\,\overline{(\tilde{\rho})}\,\delta_k A, \tag{3.151}$$

or

$$\overline{(\tilde{\rho})}^{+} = \frac{\Delta_k h}{\delta_k h\,\Delta A}\iint\limits_{\delta_k A} \tilde{\rho}\,\mathrm{d}x\,\mathrm{d}y. \tag{3.152}$$

$\overline{(\tilde{\rho})}^{+}\,\delta_k h\,\Delta A$ is the updated mass in the arrival grid cell at time $t + \Delta t$. According to Eq. (3.152), it is equal to the mass in the upstream departure cell at time t, which can be computed simply as a 2D integral over the departure area of the vertical mean density in the Eulerian model layer considered.

This demonstrates that FV methods of the DCISL type, like the one used in HIRLAM-DCISL, can be used also for nonhydrostatic models. It also means that the Lagrangian (ξ) vertical coordinate approach can be used in nonhydrostatic Eulerian flux-type FV models, just as the corresponding quasi-hydrostatic approach was used in the NCAR-FFSL dynamical core.

Of course, alternatively, a nonhydrostatic model may be based on a traditional 3D operator-split flux-form method, with fluxes entering an Eulerian grid cell at both horizontal and vertical faces. However, it becomes rather complicated if one uses the most accurate schemes of the symmetric FFSL type. Thus, Leonard, Lock and Macvean [1996] presents a symmetric 3D scheme. Using a notation similar to that used in Section 2.3.2, it becomes

$$\begin{aligned}\overline{\psi}^{n+1} = \overline{\psi}^{n} &+ X_C\left\{\frac{'1}{6}\left[\left(\overline{\psi}^{n} + \overline{\psi}_{AY} + \overline{\psi}_{AYZ}\right) + \left(\overline{\psi}^{n} + \overline{\psi}_{AZ} + \overline{\psi}_{AZY}\right)\right]\right\}\\ &+ Y_C\left\{\frac{'1}{6}\left[\left(\overline{\psi}^{n} + \overline{\psi}_{AZ} + \overline{\psi}_{AZX}\right) + \left(\overline{\psi}^{n} + \overline{\psi}_{AX} + \overline{\psi}_{AXZ}\right)\right]\right\}\\ &+ Z_C\left\{\frac{'1}{6}\left[\left(\overline{\psi}^{n} + \overline{\psi}_{AX} + \overline{\psi}_{AXY}\right) + \left(\overline{\psi}^{n} + \overline{\psi}_{AY} + \overline{\psi}_{AYX}\right)\right]\right\},\end{aligned} \tag{3.153}$$

where, for example,

$$\overline{\psi}_{AYZ} = \overline{\psi}_{AZ}\left(\overline{\psi}_{AY}\right) = \overline{\psi}_{AY} + Z_A\left(\overline{\psi}_{AY}\right). \tag{3.154}$$

Such schemes have been used extensively; a recent example is the MIT-GCM (Adcroft, Campin, Hill and Marshall [2004]).

4. Summary

Recent developments in FV methods have provided the basis for new meteorological dynamical cores that conserve integral invariants exactly, globally as well as locally. In particular, these new FV methods have been the basis for design of exact mass conserving

tracer transport models. The new technologies are reviewed and the perspectives for the future are discussed.

During about two decades, the traditional semi-implicit and semi-Lagrangian spectral or grid-point dynamical cores have been dominating worldwide in meteorological models applied for weather prediction and climate simulations. They are efficient and otherwise accurate but lack exact mass conservation, which is considered a serious drawback for the hydrometeorological variables as well as an increasing number of chemical variables included in the models. In Section 3, we presented two recently developed pioneering meteorological dynamical cores which potentially solve these problems. These are the semi-implicit cell-integrated semi-Lagrangian limited area dynamical core HIRLAM-DCISL and the global flux-form NCAR-FFSL dynamical core. Each of them extends newly developed 2D FV semi-Lagrangian schemes, described among others in Section 2, to 3D utilizing a common newly developed Lagrangian time-stepping technique building on horizontally upstream and vertically downstream time-steps. Values of certain quantities integrated horizontally over an upstream departure area in an Eulerian model layer are assumed to be transported with vertical walls along 3D trajectories into a Lagrangian layer in a column of Eulerian grid cells. The vertical coordinates of this Lagrangian layer are then determined hydrostatically. As a result, the quantities in question are conserved exactly globally and with high, slightly different accuracy, also locally in both dynamical cores. Idealized tests presented in Section 2 showed that the local conservation is slightly more accurate in the HIRLAM-DCISL transport schemes than in the NCAR-FFSL scheme. With proper boundary conditions, the HIRLAM-DCISL dynamical core conserves the global mass of moist air and tracers exactly, including water vapor, liquid water, and solid water, if included (apart from evaporation and condensation). Also the NCAR-FFSL dynamical core conserves exactly these masses. In addition, it conserves, except for time truncation errors, potential temperature and absolute vorticity (in adiabatic friction-free flow). Thus, the NCAR-FFSL dynamical core comes closer than the HIRLAM-DCISL to the ideal CSCL model considered in Section 3.1. Both FV dynamical cores have been tested and compared with nonconservative dynamical cores in an idealized baroclinic wave test. All dynamical cores considered were found to converge toward a common solution. However, in their present formulation, the FV dynamical cores needed higher resolution than the nonconservative dynamical cores they were compared with for the same level of accuracy. The explanation seems to be a slight smoothing due to the repeated remappings and interpolations. As a possible cure to HIRLAM-DCISL, it is suggested to keep the vertical Lagrangian cells for a number of consecutive large semi-implicit time-steps before performing the vertical remapping to the Eulerian model layers. The idealized as well as other tests with HIRLAM DCISL showed that a consistent Lagrangian discretization of the energy conversion term in the thermodynamic equation leads to a more accurate simulation than the traditional discretization. A further increase in accuracy is expected from a corresponding Lagrangian discretization of the horizontal pressure gradient term in the momentum equation. The idealized intercomparison tests showed that among the dynamical cores the FV ones are the most expensive in computational costs. Thus, the conservative property is achieved at the expense of efficiency. There is no doubt, however, that the efficiency of the present experimental ad hoc coded FV dynamical cores can be increased considerably by a dedicated optimization.

An obvious application of FV models such as the NCAR-FFSL and the HIRLAM-DCISL is tracer transport since tracer-mass conservation is essential. As described in Section 3.5, a FV transport scheme has often been imported into a GCM with a dynamical core that does not conserve mass locally. A problem of such an online coupling is that it leads to the so-called mass-wind inconsistency, which in long simulations can lead to severe errors with large amounts of artificially (spuriously) created or destroyed tracer mass. The only way to avoid completely such errors is to use a complete FV model with exactly the same locally mass conserving algorithms for all tracers and for the moist air.

To facilitate a wider application, the limited area HIRLAM-DCISL may be extended to a global domain. This may be done by using the extension of the horizontal FV schemes used in this dynamical core that have already been developed in spherical latitude-longitude coordinates. Another possibility which may be relevant also for the NCAR-FFSL is to change to a new grid, as the icosahedral-hexagonal grid, which is almost uniform on the sphere. Finally, the possibility of an extension to nonhydrostatic dynamical cores is discussed in Section 3.6. The same extension, as used in the hydrostatic dynamical cores, of the available 2D FV semi-Lagrangian schemes to 3Ds may be used in nonhydrostatic dynamical cores. That is, utilizing horizontally upstream and vertically downstream time-stepping. The only difference is that the vertical displacement of the Lagrangian cells must be determined directly by predicted vertical velocities and not from hydrostatic balance.

5. Acknowledgments

The authors wish to express their gratitude to Dr. Ramachandran D. Nair and Dr. Phil Rasch for helpful suggestions and useful discussions on parts of Section 2. Thanks to Dr. Patrick Jöckel and Rune Graversen for their comments on the ‘mass-wind inconsistency’ section. The third author is grateful to NCAR’s Advanced Study Program and Climate Modeling Section for providing necessary support for this research.

References

Adcroft, A., Campin, J-M., Hill, C., Marshall, J. (2004). Implementation of an atmosphere-ocean general circulation model on the expanded spherical cube. *Mon. Wea. Rev.* **132** (12), 2845–2863.

Arakawa, A. (2000). A personal perspective on the early years of general circulation modeling at UCLA. In: Randall, D.A. (ed.), *General Circulation Model Development. Past, Present, and Future* (Academic Press), pp. 1–65.

Arakawa, A., Lamb, V. (1977). Computational design and the basic dynamical processes of the UCLA general circulation model. *Methods Comput. Phys.* **17**, 173–265.

Arakawa, A., Lamb, V. (1981). A potential enstrophy and energy conserving scheme for the shallow water equations. *Mon. wea. Rev.* **109**, 18–36.

Bates, J.R., Mcdonald, A. (1982). Multiply-upstream, semi-Lagrangian advective schemes: analysis and application to a multi-level primitive equation model. *Mon. Wea. Rev.* **110** (12), 1831–1842.

Bell, J.B., Dawson, C.N., Shubin, G.N. (1988). An unsplit, higher order Godunov method for scalar conservation laws in multiple dimensions. *J. Comput. Phys.* **74**, 1–24.

Bermejo, R. Staniforth, A. (1992). The conversion of semi-Lagrangian advection schemes to quasi-monotone schemes. *Mon. Wea. Rev.* **120** (11), 2622–2632.

Boris, J.P., Book, D.L. (1973). Flux corrected transport. I. SHASTA: a fluid transport algorithm that works. *J. Comput. Phys.* **11**, 38–69.

Bott, A. (1989). A positive definite advection scheme obtained by nonlinear renormalization of the advective fluxes. *Mon. Wea. Rev.* **117** (5), 1006–1016.

Bott, A. (1992). Monotone flux limitation in the area-preserving flux-form advection algorithm. *Mon. Wea. Rev.* **120** (11), 2592–2602.

Bott, A. (1993). The monotone area-preserving flux-form advection algorithm: reducing the time-splitting error in two-dimensional flow fields. *Mon. Wea. Rev.* **121** (9), 2637–2641.

Burridge, D.M., Hasler, J. (1977). A model for medium range forecasting. Adiabatic formulation. ECMWF Tech. Rep. 4, Reading, United Kingdom, pp. 46.

Byun, B.W. (1999). Dynamically consistent formulations in meteorological and air quality models for multiscale atmospheric studies, Part II: mass conservation issues. *J. Atmos. Sci.* **56** (21), 3808–3820.

Cameron-Smith, P., Connell, P., Prather, M. (2002). Pressure fixers: conserving mass in chemical transport models by adjusting the winds. Available at 2002; http://eed.llnl.gov/pfix/.

Carpenter, J.R., Droegemeier, K.K., Woodward, P.R., Hane, C.E. (1990). Application of the piecewise parabolic method (PPM) to meteorological modeling. *Mon. Wea. Rev.* **118** (3), 586–612.

Colella, P. (1990). Multidimensional upwind methods for hyperbolic conservation laws. *J. Comput. Phys.* **87**, 171–200.

Colella, P., Woodward, P.R. (1984). Piecewise parabolic method for gas-dynamical simulations. *J. Comput. Phys.* **54**, 174–201.

Collins, W.D., Rasch, P.J., Boville, B.A., Hack, J.J., Mccaa, J.R., Williamson, D.L., Kiehl, J.T., Briegleb, B., Bitz, C., Lin, S.-J., Zhang, M., Dai, Y. (2004). Description of the NCAR Community Atmosphere Model (CAM3). Technical Report, NCAR/TN- 464+STR, National Center for Atmospheric Research, Boulder, Colorado 80307-3000, pp. 226.

Cotter, C.J., Frank, J., Reich, S. (2007). The remapped particle-mesh semi-Lagrangian advection scheme. *Q. J. R. Meteorol. Soc.* **133**, 251–260.

Crowley, W.P. (1968). Numerical advection experiments. *Mon. Wea. Rev.* **96** (1), 1–11.

Doswell III, C.A. (1984). A kinematic analysis of frontogenesis associated with a nondivergent vortex. *J. Atmos. Sci.* **41** (7), 1242–1248.

Dukowicz, J.K., Baumgardner, J.R. (2000). Incremental remapping as a transport/advection algorithm. *J. Comput. Phys.* **160**, 318–335.

Dukowicz, J.K., Kodis, J.W. (1987). Accurate conservative remapping (rezoning) for arbitrary Lagrangian-Eulerian computations. *SIAM J. Sci. Stat. Comput.* **8**, 305–321.

Dukowicz, J.K., Ramshaw, J.D. (1979). Tensor viscosity method for convection in numerical fluid dynamics. *J. Comput. Phys.* **32**, 71–79.

Durran, D.R. (1999). *Numerical Methods for Wave Equations in Geophysical Fluid Dynamics* (Springer-Verlag), pp. 465.

Easter, R.E. (1993). Two modified versions of Bott's positive-definite numerical advection scheme. *Mon. Wea. Rev.* **121** (1), 297–304.

Eumard, R., Gallouët, T., Herbin, R. (2000). Finite volume methods. In: Ciarlet, P.G., Lions, J.L. (eds.), *Handbook of Numerical Analysis, vol. VII* (Elsevier Science B.V.), pp. 715–1020.

Godunov, S.K. (1959). Finite difference method for numerical computation of discontinous solution of the equations of fluid dynamics. *Math. Sb.* **47**, 271, translated from Russian by I. Bohachevsky.

Hirsch, C. (1990). *Numerical Computation of Internal and External Flows, vol. 2*, Wiley Series in Numerical Methods (John Wiley and Sons), pp. 714.

Hirt, C.W., Amsden, A.A., Cook, J.L. (1974). An arbitrary Lagrangian-Eulerian computing method for all flow speeds. *J. Comput. Phys.* **14**, 227.

Hólm, E.V. (1995). A fully two-dimensional, nonoscillatory advection scheme for momentum and scalar transport equations. *Mon. Wea. Rev.* **123** (2), 536–552.

Horowitz, L.W., Walters, S., Mauzerall, D.L., Emmons, L.K., Rasch, P.J., Granier, C., Tie, X., Lamarque, J., Schultz, M.G., Tyndall, G.S., Orlando, J.J., Brasseuret, G.P. (2003). A global simulation of tropospheric ozone and related tracers: description and evaluation of MOZART, version 2. *J. Geophys. Res.* **108** (D24), 4784.

Hortal, M. (2002). The development and testing of a new two-time-level semi-Lagrangian scheme (SETTLS) in the ECMWF forecast model. *Q. J. R. Meteorol. Soc.* **128**, 1671–1687.

Hourdin, F., Armengaud, A. (1999). The use of finite-volume methods for atmospheric advection of trace species. Part I: test of various formulation in a general circulation model. *Mon. Wea. Rev.* **127** (5), 822–837.

Jablonowski, C. and Williamson, D.L. (2006a). A baroclinic instability test case for atmospheric model dynamical cores. *Q. J. R. Meteorol. Soc.* **132**, 2943–2975.

Jablonowski, C. and Williamson, D.L. (2006b). A baroclinic wave test case for dynamical cores of general circulation models: Model intercomparisons. NCAR Technical Note, NCAR/TN-469+STR, Boulder CO, pp. 89.

Jöckel, P., Von Kuhlmann, R., Lawrence, M.G., Steil, B., Brenninkmeijer, C.A.M., Crutzen, P.J., Rasch, P.J., Eaton, B. (2001). On a fundamental problem in implementing flux-form advection schemes for tracer transport in 3-dimensional general circulation and chemistry transport models. *Q. J. R. Meteorol. Soc.* **127**, 1035–1052.

Kaas, E. (2008). A simple and efficient locally mass conserving semi-Lagrangian transport scheme. *Tellus A*, in press.

Källén, E. (1996). HIRLAM documentation manual, System 2.5. Available from SMHI, Norrköping, Sweden.

Kuo, H., Williams, R.T. (1990). Semi-Lagrangian solutions to the inviscid burgers equation. *Mon. Wea. Rev.* **118** (6), 1278–1288.

Laprise, L.P.R., Plante, A. (1995). A class of semi-Lagrangian integrated-mass (SLIM) numerical transport algorithms. *Mon. Wea. Rev.* **123** (2), 553–565.

Lauritzen, P.H. (2007). A stability analysis of finite-volume advection schemes permitting long time steps. *Mon. Wea. Rev.* **135** (7), 2658–2673.

Lauritzen, P.H., Kaas, E., Machenhauer, B. (2006). A mass-conservative semi-implicit, semi-Lagrangian limited-area shallow water model on the sphere. *Mon. Wea. Rev.* **134** (4), 1196–1212.

Lauritzen, P.H., Kaas, E., Machenhauer, B., Lindberg, K. (2008). A mass-conservative version of the semi-implicit semi-Lagrangian HIRLAM. *Q. J. R. Meteorol. Soc.*, **134** (635), 1583–1595.

Lauritzen, P.H., Nair, R.D. (2008). Monotone and conservative cascade remapping between spherical grids (CaRS): regular latitude-longitude and cubed-sphere grids. *Mon. Wea. Rev.*, **136** (4), 1416–1432.

Leith, C.E. (1965). Numerical simulation of the earth's atmosphere. *Methods Comput. Phys.* **4**, 1–28.

Leonard, B.P. (1994). Note on the von Neumann stability of explicit one-dimensional advection schemes. *Comp. Meth. Appl. Mech. Ens.* **118**, 29–46.

Leonard, B.P., Lock, A.P., Macvean, M.K. (1996). Conservative explicit unrestricted time step multidimensional constancy-preserving advection schemes. *Mon. Wea. Rev.* **124** (11), 2588–2606.

Leslie, M.L., Purser, R.J. (1995). Three-dimensional mass-conserving semi-Lagrangian scheme employing forward trajectories. *Mon. Wea. Rev.* **123** (8), 2551–2566.

Leveque, R.J. (2002). *Finite Volume Methods for Hyperbolic Problems* (Cambridge University Press).

Li, Y., Chang, J.S. (1996). A mass-conservative, positive-definite, and efficient Eulerian advection scheme in spherical geometry and on a non-uniform grid system. *J. Appl. Meteorol.* **35**, 1897–1913.

Lin, S.-J. (1997). A finite-volume integration method for computing pressure gradient forces in general vertical coordinates. *Q. J. R. Meteorol. Soc.* **123**, 1749–1762.

Lin, S.-J. (2004). A "Vertically Lagrangian" finite-volume dynamical core for global models. *Mon. Wea. Rev.* **132** (10), 2293–2307.

Lin, S.-J., Rood, R.B. (1996). Multidimensional flux-form semi-Lagrangian transport schemes. *Mon. Wea. Rev.* **124** (9), 2064–2070.

Lin, S.-J., Rood, R.B. (1997). An explicit flux-form semi-Lagrangian shallow water model on the sphere. *Q. J. R. Meteorol. Soc.* **123**, 2477–2498.

Lin, S.-J., Rood, R.B. (1998). A flux-form semi-Lagrangian general circulation model with a Lagrangian control-volume vertical coordinate. In: *Proc. The Rossby-100 Symp.* (University of Stockholm, Stockholm, Sweden), pp. 220–222.

Lorenz, E.N. (1960). Energy and numerical weather prediction. *Tellus* **12**, 364–373.

Machenhauer, B. (1979). The spectral method. In: Kasahara, A. (ed.), *Numerical Methods Used in Atmospheric Models, vol. 2. GARP Publications Series No.17* (WMO and IGSU, Geneva, Switzerland), pp. 121–275.

Machenhauer, B. (1994). A Note on a Mass-, Energy- and Entropy Conserving Semi-Lagrangian and Explicit Integration Scheme for the Primitive Equations. MPI Workshop on Semi-Lagrangian Methods, Hamburg 8-9 Oct. 1992. **146** (*Max Planck Institute for Meteorology Tech. Rep*, Hamburg, Germany), pp. 73–102.

Machenhauer, B., Olk, M. (1996). On the development of a cell-integrated semi-Lagrangian shallow water model on the sphere. ECMWF Workshop on semi-Lagrangian methods, 6-8 November, Workshop Proceedings, pp. 213–228.

Machenhauer, B., Olk, M. (1997). The implementation of the semi-implicit scheme in cell-integrated semi-Lagrangian models. *Atmos. Ocean.* **35**, 103–126.

Machenhauer, B., Olk, M. (1998). Design of a semi-implicit cell-integrated semi-Lagrangian model. Max Planck Institute for Meteorology Tech. Rep., 265, Hamburg, Germany, pp. 76–85.

Majewski, D., Liermann, D., Prohl, P., Ritter, B., Buchhold, M., Hanisch, T., Paul, G., Wergen, W., Baumgardner, J. (2002). The operational global icosahedral-hexagonal gridpoint model GME: description and high-resolution tests. *Mon. Wea. Rev.* **130** (2), 319–338.

McDonald A. (1998). Default horizontal diffusion coefficients in HIRLAM-4.1. HIRLAM Newsletter No. 31, electronic version available at http://hirlam.knmi.nl/.

Mcgregor, J.L. (1993). Economical determination of departure points for semi-Lagrangian models. *Mon. Wea. Rev.* **121** (1), 221–230.

Mcgregor, J.L. (1996). Semi-Lagrangian advection on a conformal cubic grid. *Mon. Wea. Rev.* **124** (6), 1311–1322.

Moorthi, S., Higgins, R.W., Bates, J.R. (1995). A global multilevel atmospheric model using a vector semi-Lagrangian finite-difference scheme. Part II: version with physics. *Mon. Wea. Rev.* **123** (5), 1523–1541.

Nair, R.D., Côté, J., Staniforth, A. (1999a). Monotonic cascade interpolation for semi-Lagrangian advection. *Q. J. R. Meteorol. Soc.* **125**, 197–212.

Nair, R.D., Côté, J., Staniforth, A. (1999b). Cascade interpolation for semi-Lagrangian advection over the sphere. *Q. J. R. Meteorol. Soc.* **125**, 1445–1468.

Nair, R.D., Thomas, S.J., Loft, R.D. (2005). A discontinuous galerkin transport scheme on the cubed sphere. *Mon. Wea. Rev.* **133** (4), 814–828.

NAIR, R.D. (2004). Extension of a conservative cascade scheme on the sphere to large courant numbers. *Mon. Wea. Rev.* **132** (1), 390–395.

NAIR, R.D., JABLONOWSKI, C. (2007). Moving vortices on the sphere: a test-case for horizontal advection problems. *Mon. Wea. Rev.*, to appear.

NAIR, R.D., SCROGGS, J.S., SEMAZZI, F.H.M. (2003). A forward-trajectory global semi-Lagrangian transport scheme. *J. Comput. Phys.* **190**, 275–294.

NAIR, R.D., SCROGGS, J.S., SEMAZZI, F.H.M. (2002). Efficient conservative global transport schemes for climate and atmospheric chemistry models. *Mon. Wea. Rev.* **130** (8), 2059–2073.

NAIR, R.D., MACHENHAUER, B. (2002). The mass-conservative cell-integrated semi-Lagrangian advection scheme on the sphere. *Mon. Wea. Rev.* **130** (3), 649–667.

PENG, X., XIAO, F., OHFUCHI, W., FUCHIGAMI, H. (2005). Conservative semi-Lagrangian transport on a sphere and the impact on vapor advection in an atmospheric general circulation model. *Mon. Wea. Rev.* **133** (3), 504–520.

PETSCHEK, A.G., LIBERSKY, L.D. (1975). Stability, accuracy, and improvement of Crowley advection schemes. *Mon. Wea. Rev.* **103** (12), 1104–1109.

PRATHER, M., MCELROY, M., WOFSY, S., RUSSEL, G., RIND, D. (1987). Chemistry of the global troposphere: fluorocarbons as tracers of air motion. *J. Geophys. Res.* **92** (D6), 6579–6613.

PRATHER, M.J. (1986). Numerical advection by conservation of second order moments. *J. Geophys. Res.* **91**, 6671–6681.

PUDYKIEWICZ, J., BENOIT, R., STANIFORTH, A. (1985). Preliminary results from a partial LRTAP model based on an existing meteorological forecast model. *Atmos. Ocean.* **22**, 283–308.

PURSER, R.J., LESLIE, L.M. (1991). An efficient interpolation procedure for high-order three-dimensional semi-Lagrangian models. *Mon. Wea. Rev.* **119** (10), 2492–2498.

RANČIĆ, M. (1992). Semi-Lagrangian piecewise biparabolic scheme for two-dimensional horizontal advection of a passive scalar. *Mon. Wea. Rev.* **120** (7), 1394–1406.

RANČIĆ, M. (1995). An efficient, conservative, monotone remapping for semi-Lagrangian transport algorithms. *Mon. Wea. Rev.* **123** (4), 1213–1217.

RANČIĆ, M.R., PURSER, J., MESINGER, F. (1996). A global-shallow water model using an expanded spherical cube. *Q. J. R. Meteorol. Soc.* **122**, 959–982.

RASCH, P.J. (1994). Conservative shape-preserving two-dimensional transport on a spherical reduced grid. *Mon. Wea. Rev.* **122** (6), 1337–1350.

RASCH, P.J., LAWRENCE, M. (1998). Recent development in transport models at NCAR. In: MCHENHAUER, B. (ed.), MPI Workshop on Conservative Transport Schemes Hamburg 2-3 June 1997. **265**, pp. 65–75, (*Max Planck Institute for Meteorology Tech. Rep*, Hamburg, Germany).

RASCH, P.J., WILLIAMSON, D.L. (1990). Computational aspects of moisture transport in global models of the atmosphere. *Q. J. R. Meteorol. Soc.* **116**, 1071–1090.

REICH, S. (2007). An explicit and conservative remapping strategy for semi-Lagrangian advection. *Atmos. Sci. Let.*, in press.

RICHARDSON, L.F. (1922). *Weather Prediction by Numerical Process* (Cambridge University Press).

ROBERT, A.J. (1969). The integration of a spectral model of the atmosphere by the implicit method. In: *Proc. WMO/IUGG Symp. on Numerical Weather Prediction* (Japan Meteorological Agency, Tokyo, Japan), pp. VII-19–VII-24.

ROBERT, A. (1981) A stable numerical integration scheme for the primitive meteorological equations. *Atmos. Ocean.* **19**, 35–46.

ROBERT, A. (1982). A semi-Lagrangian and semi-implicit numerical integration scheme for the primitive meteorological equations. *J. Meteorol. Soc. Jpn.* **60**, 319–325.

RONCHI, C., IACONO, R., PAOLUCCI, P.S. (1996). The "cubed sphere": A new method for the solution of partial differential equations in spherical geometry. *J. Comput. Phys.* **124**, 93–114.

ROTMAN, D.A., TANNAHILL, J.R., KINNISON, D.E., CONNELL, P.S., BERGMANN, D., PROCTOR, D., RODRIGUEZ, J.M., LIN, S.J., ROOD, R.B., PRATHER, M.J., RASCH, P.J., CONSIDINE, D.B., RAMAROSON, R., KAWA, S.R. (2001). Global modeling initiative assessment model: model description, integration, and testing of the transport shell. *J. Geophys. Res.* **106** (D2), 1669–1692.

SADOURNY, R., ARAKAWA, A., MINTZ, Y. (1968). Integration of the non-divergent barotropic vorticity equation with an icosahedral-hexagonal grid for the sphere. *Mon. Wea. Rev.* **96** (6), 351–356.

Sadourny, R. (1972). Conservative finite-difference approximations of the primitive equations on quasi-uniform spherical grids. *Mon. Wea. Rev.* **100** (2), 136–144.

Schär, C., Smolarkiewicz, P.K. (1996). A synchronous and iterative flux-correction formalism for coupled transport. *J. Comput. Phys.* **128**, 101–120.

Scroggs, J.S., Semazzi, F.H.M. (1995). A conservative semi-Lagrangian method for multidimensional fluid dynamics applications. *Numer. Methods Partial Differ. Equ.* **11**, 445–452.

Simmons, A.J., Burridge, D.M. (1981). An energy and angular-momentum conserving vertical finite-difference scheme and hybrid vertical coordinates. *Mon. Wea. Rev.* **109** (4), 758–766.

Skamarock, W.C. (2006). Positive-definite and monotonic limiters for unrestricted-time-step transport schemes. *Mon. Wea. Rev.* **134** (8), 2241–2250.

Skamarock, W.C., Klemp, J.B., Dudhia, J., Gill, D.O., Barker, D.M., Wang, W., Powers, J.G. (2007). A description of the advanced research WRF Version 2. NCAR Tech Note, NCAR/TN–468+STR, Jun. 2005, Upd. Jan. 2007.

Smolarkiewicz, P.K. (1982). The multi-dimensional Crowley advection transport algorithm with small implicit diffusion. *Mon. Wea. Rev.* **110** (12), 1968–1983.

Smolarkiewicz, P.K. (1984). A fully multidimensional positive definite advection transport algorithm with small implicit diffusion. *J. Comput. Phys.* **54**, 325–362.

Smolarkiewicz, P.K., Pudykiewicz, J.A. (1992). A class of semi-Lagrangian approximations for fluids. *J. Atmos. Sci.* **49** (22), 2082–2096.

Smolarkiewicz, P.K., Grabowski, W.W. (1990). The multidimensional positive definite advection transport algorithm: non oscillatory option. *J. Comput. Phys.* **86**, 355–375.

Staniforth, A., Côté, J. (1991). Semi-Lagrangian integration schemes for atmospheric models—a review. *Mon. Wea. Rev.* **119** (9), 2206–2223.

Staniforth, A., Coté, J., Pudykjewicz, J. (1987). Comments on "Swolarkiewicz's Deformational Flow." *Mon. Wea. Rev.* **115** (4), 894–900.

Starr, V.P. (1945). A quasi-Lagrangian system of hydrodynamical equations. *J. Atmos. Sci.* **2** (4), 227–237.

Sun, W.Y., Yeh, K.S., Sun, R.Y. (1996). A simple semi-Lagrangian scheme for advection equations. *Q. J. R. Meteorol. Soc.* **122**, 1211–1226.

Taylor, M., Tribbia, J., Iskandrani, M. (1997). The spectral element method for the shallow water equations on the sphere. *J. Comput. Phys.* **130**, 92–108.

Thuburn, J. (1997). A PV-based shallow-water model on a hexagonal-icosahedral grid. *Mon. Wea. Rev.* **125** (9), 2328–2347.

Thuburn, J. (2006). Some conservation issues for dynamical cores of NWP and climate models. *J. Comput. Phys.*, in press.

Tremback, C.J., Powell, J., Cotton, W.R., Pielke, R.A. (1987). The forward in time upstream advection scheme: extension to higher orders. *Mon. Wea. Rev.* **115** (2), 540–555.

Trenberth, K.E. (1991). Climate diagnostics from global analyses: conservation of mass in ECMWF analyses. *J. Clim.* **4** (7), 707–722.

Per Undén, Laura Rontu, Heikki Järvinen, Peter Lynch, Javier Calvo, Gerard Cats, Joan Cuxart, Kalle Eerola, Carl Fortelius, Jose Antonio Garcia-Moya, Colin Jones, Geert Lenderlink, Aidan McDonald, Ray McGrath, Beatriz Navascues, Niels Woetman Nielsen, Viel Ødegaard, Ernesto Rodriguez, Markku Rummukainen, Rein Rõõm, Kai Sattler, Bent Hansen Sass, Hannu Savijärvi, Ben Wichers Schreur, Robert Sigg, Han The, Aleksander Tijm (2002). Hirlam-5 scientific documentation. SMHI Hirlam-5 Project Tech. Rep., pp. 146 [Available online at http://www-hirlam.knmi.nl.].

Van Leer, B. (1974). Towards the ultimate conservative difference scheme. II. Monotonicity and conservation combined in a second-order scheme. *J. Comput. Phys.* **14**, 361–370.

Van Leer, B. (1977). Towards the ultimate conservative difference scheme. Part IV: a new approach to numerical convection. *J. Comput. Phys.* **23**, 276–299.

Williamson, D.L. (1968). Integration of the barotropic vorticity equation on a spherical geodesic grid. *Tellus* **20**, 642–653.

Williamson, D.L., Drake, J.B., Hack, J.J., Jakob, R., Swarztrauber, R.N. (1992). A standard test set for numerical approximations to the shallow water equations in spherical geometry. *J. Comput. Phys.* **102**, 211–224.

Wooward, P.R., Colella, P. (1984). The numerical simulation of two-dimensional fluid flow with strong shocks. *J. Comput. Phys.* **54**, 115–173.

Xue, M. (2000). High-order monotonic numerical diffusion and smoothing. *Mon. Wea. Rev.* **128**, 2853–2864.

Xiao, F., Yabe, T., Peng, X., Kobayashi, H. (2002). Conservative and oscillation-less atmospheric transport schemes based on rational functions. *J. Geophys. Res.* **107** (D22), 4609, doi: 10.1029/2001JD001532.

Xiao, F., Yabe, T. (2001). Completely conservative and oscillation-less semi-Lagrangian schemes for advection transport. *J. Comput. Phys.* **170**, 498–522.

Yabe, T., Ishikawa, T., Wang, P.Y., Aoki, T., Kodota, Y., Ikeda, F. (1991). A universal solver for hyperbolic-equations by cubic-polynomial interpolation. II. Two- and three-dimensional solvers. *Comput. Phys. Commun.* **66**, 233–242.

Zalesak, S.T. (1979). Fully multidimensional flux-corrected transport algorithm for fluids. *J. Comput. Phys.* **31**, 335–362.

Zerroukat, M., Wood, N., Staniforth, A. (2002). SLICE: a semi-Lagrangian inherently conservative and efficient scheme for transport problems. *Q. J. R. Meteorol. Soc.* **128** (586), 2801–2820.

Zerroukat, M., Wood, N., Staniforth, A. (2004). SLICE-S: A semi-Lagrangian inherently conserving and efficient scheme for transport problems on the sphere. *Q. J. R. Meteorol. Soc.* **130** (602), 2649–2664.

Zerroukat, M., Wood, N., Staniforth, A. (2005). A monotonic and positive-definite filter for a semi-Lagrangian inherently conserving and efficient (SLICE). *Q. J. R. Meteorol. Soc.* **131** (611), 2923–2936.

Zerroukat, M., Wood, N., Staniforth, A. (2006). The parabolic spline method (PSM) for conservative transport. *Int. J. Numer. Methods Fluids* **11**, 1297–1318.

Zerroukat, M., Wood, N., Staniforth, A. (2007). Application of the parabolic spline method (PSM) to a multi-dimensional conservative semi-Lagrangian transport scheme (SLICE). *J. Comput. Phys.* **225** (1), 935–948.

Computational Kernel Algorithms for Fine-Scale, Multiprocess, Longtime Oceanic Simulations

Alexander F. Shchepetkin

Institute of Geophysics and Planetary Physics, University of California, 405 Hilgard Avenue, Los Angeles, CA 90095-1567, USA

James C. McWilliams

Institute of Geophysics and Planetary Physics, University of California, 405 Hilgard Avenue, Los Angeles, CA 90095-1567, USA

Abstract

Progress in computer technology has made it possible to make larger calculations with finer grid-scale resolution, and physical processes that were beyond the reach of coarse-resolution models are now simulated directly. This focuses scientific interest toward more turbulent flow regimes and applications toward more realistic modeling of specific regional configurations.

In this chapter, we examine the numerical design of oceanic modeling codes specifically suited for modern demands. These are compared with traditional "legacy" oceanic general circulation models and with computational fluid dynamics methods for modern engineering applications. Our primary subject is how the numerical algorithms for different aspects of the discretized partial differential equation system – the computational kernel – combine to yield the overall model performance, with particular focus on avoiding destructive interference among algorithmic components.

1. Introduction: integrated kernel design

Oceanic General Circulation Models (OGCMs) (Bleck and Smith [1990], Blumberg and Mellor [1987], Bryan and Cox [1969], Dukowitz and Smith [1994], Griffies, Böning, Bryan, Chassignet, Gerdes, Hasumi, Hirst, Treguier and Webb [2000],

Computational Methods for the Atmosphere and the Oceans
Special Volume (Roger M. Temam and Joseph J. Tribbia, Guest Editors) of
HANDBOOK OF NUMERICAL ANALYSIS, VOL. XIV
P.G. Ciarlet (Editor)
Copyright © 2009 Elsevier B.V.
All rights reserved
ISSN 1570-8659
DOI 10.1016/S1570-8659(08)01202-0

Marshall, Adcroft, Hill, Perelman and Heisey [1997], McWilliams [1996]) have historically been a separate branch of computational fluid mechanics with significantly different choices for numerical methods compared with most engineering computational fluid dynamics (CFD) applications. The main motivation is the need to perform very long – even millennial – simulations over hundreds of thousands of time-steps, which makes it essential to ensure conservation properties for the mean and variance of model fields (Arakawa and Lamb [1977], Lilly [1965]). This typically has led to the choice of discrete algorithms as a combination of basic second-order, centered spatial operators and leapfrog (LF) time-stepping, both because they can easily be made to assure desired conservation properties and because higher-order advection schemes usually do not give better solutions for coarse grids that do not adequately resolve the baroclinic deformation radii (i.e., the "non-eddy-resolving" regime typical for climate studies). Instead, coarse-grid models have to rely on parameterizations of subgrid mesoscale processes to achieve physically correct results (Gent and McWilliams [1990], Griffies, Gnanadesikan, Pacanowski, Larichev, Dukowicz and Smith [1998]). The OGCM codes targeting large-scale circulation were ill-suited for nearshore phenomena due to inaccurate handling of complex geometry, bottom topography, free surface, and bottom boundary layer; coastal model developments took a rather independent route (Casulli and Cheng [1992], Casulli and Cattani [1994], Casulli and Stelling [1998], Casulli [1999]) with more focus on achieving accurate dispersion properties for surface gravity waves, wetting and drying capabilities, etc, with less emphasis on long-term conservation properties and Coriolis-force effects. These coastal codes are characterized by two-time-level time-stepping, upstream-biased, semi-Lagrangian, monotonicity-preserving advection schemes, and sometimes non-hydrostatic effects and unstructured grids (Cheng and Casulli [2001]). The combination of these features makes them more similar to CFD codes than to OGCMs.

During the 1990s, all major OGCMs underwent a substantial redesign in order to take advantage of the rapidly developing computer technology, especially parallel processing. This allowed much larger computational grids, ultimately ones that can close the resolution gap between coastal and regional-global applications. A somewhat paradoxical outcome of this evolution is that the use of parallel codes has become widespread in oceanic modeling, while as yet there has been relatively little overhaul of the numerical methods in their hydrodynamic kernels. Most of the recent model content developments have come in physical parameterizations and peripheral modules for biogeochemical processes. The study of Griffies, Böning, Bryan, Chassignet, Gerdes, Hasumi, Hirst, Treguier and Webb [2000] is an overview of the modern state of OGCMs in climate modeling. Some rare exceptions to the widespread use of classical time-stepping and second-order advection algorithms have been adopted to avoid spurious oscillations and negative concentrations for material tracers (Willebrand, Barnier, Böning, Dieterich, Killworth, LeProvost, Jia, Molines and New [2001]), but only rarely are better advection schemes used for momentum (Dietrich, Lin, Mestas-Nunez and Ko [1997], Dietrich, Marietta and Roache [1987]). Monotonic advection schemes are also used in the context of isopycnic layer models to deal with vanishing layer thickness (Bleck and Smith [1990]).

The code organizational structure in the Modular Ocean Model (MOM) became another de facto standard, adopted by many modelers when their code complexity matured to OGCM status. This approach is encouraged and often justified by the ease of incorporating peripheral modules. However, it led to a widespread "modular vision" of the kernel, and the interaction and, in fact, interference among the algorithmic components was often overlooked. For example, allowing a freesurface in a previously rigid-lid model (Dukowitz and Smith [1994], Killworth, Stainforth, Webb and Paterson [1991]) may result in the loss of conservation and/or constancy-preservation properties of control-volume scheme for tracer advection,[1] which was noticed, mitigated (Griffies, Pacanowski, Schmidt and Balaji [2001]), and eliminated completely (Campin, Adcroft, Hill and Marshall [2004], Marsaleix, Auclair, Herrmann, Estournel, Pairaud and Ulses [2008], Shchepetkin and McWilliams [2005]) only a decade later. If one wants to implement a non-oscillatory advection scheme for tracers, the tracer time-step has to be changed from LF to two-time level algorithm, e.g., predictor-corrector. However, since this change applies to tracers only, it leads to an underutilization of its potential benefit because the time-step size Δt of an OGCM is usually limited by the gravity-wave speed for the first baroclinic mode. The gravity wave behavior arises from an interplay between the momentum and the tracer equations and cannot be improved by refining tracers alone.

There is a common practice of two-stage code development, where a single-processor prototype code is parallelized later, only after being considered sufficiently mature. This is another reason for suboptimal algorithmic choices because considerations of computational efficiency (cost) may be quite different between parallel and non-parallel cases. For example, the treatment of the Coriolis force for the barotropic (i.e., depth-averaged) mode on a C-grid with an alternating-direction method (Bleck and Smith [1990]) or a fixed-point iteration procedure (Higdon [2005]) is straightforward on a single processor, but due to the staggered placement of u- and v-points on a C-grid and the associated interpolation, it results in excessive synchronization and message passing in a parallel implementation. In contrast, for even moderately high spatial and temporal resolution, the associated stability-limiting Courant number is very small ($f\Delta t \ll 1$, where f is the Coriolis frequency). The Coriolis force can be successfully treated in parallel with virtually any explicit, conditionally stable time-stepping algorithm. Another interesting example comes from the experience of parallelization on shared-memory computers: a very efficient code can be obtained by arranging the mathematical operations in such a way that intermediate results are stored in cache-sized private arrays that are reused in as many stages as possible before a global synchronization event takes place. This experience thus stimulates the use of multistage, high-order accurate, wide-stencil algorithms because they naturally allow a higher computational density (i.e., in this context, the ratio of mathematical operations to cache-to-main-memory loads and stores). From this point of view, the recent tendency to develop an abstract Earth-System Modeling Framework (http://www.esmf.ucar.edu), driven primarily by computer scientists, has the danger of decoupling physical-modeling from code-infrastructure decisions, as a further commitment to modular architectures. While this approach may indeed save modelers

[1] The exact cause of this loss and a remedy are considered later in Section 3.1.

labor by providing common code components, it also can have the effect of hiding or even impeding the resolution of the types of algorithmic interferences that are the focus of this chapter.

In our designs for the computational kernel in the Regional Oceanic Modeling System (ROMS) (SHCHEPETKIN and MCWILLIAMS [2005]), we adopt an integrated approach where we try to analyze and take into account all previously known experience, but in such a way that no component from a legacy code is accepted a priori. Rather, we try to identify potential algorithmic interferences and conflicts and their possible reconciliations. This principle encompasses a full range of considerations, from the theoretical analysis of a linearized time-stepping scheme all the way to cross- and within-processor code optimization issues.

The advantage of using higher-order advection schemes for turbulent flows is well understood (LEONARD, LOCK and MCVEAN [1996], ORSZAG [1971], SHCHEPETKIN and MCWILLIAMS [1998]). This approach exposes the primary criterion not as the formal order of accuracy per se (which is merely a Taylor series estimate of the asymptotic convergence rate for smooth functions) but rather as the spectral bandwidth (i.e., the fraction of grid-resolved Fourier components that are correctly represented by the discretized operator). In practice, this translates into downplaying the goal of achieving a uniformly high order of accuracy for all terms in the governing equations – a rather unrealistic hope for a multiscale, multiprocess, nonlinear system anyway – in favor of isolating and removing specific causes of accuracy loss in particular solution regimes. Although ROMS has been successfully used for coarse-resolution climate studies (HAIDVOGEL, ARANGO, HEDSTROM, BECKMANN, RIZZOLI and SHCHEPETKIN [2000]), its main intended applications are medium- or high-resolution simulations with a well-resolved baroclinic deformation radius and strong advective influences. Thus, it is intended to simulate mesoscale, approximately geostrophically balanced currents and eddies, together with nonlinear gravity and inertial waves with similar spatial scales. This downplays the importance of eddy parameterization in comparison with most climate models. However, the need to avoid erroneous vertical mixing, especially across isopycnic surfaces in stably stratified regions, is a high priority for long-term simulations. For this reason, the use of upstream-biased advection schemes in the vertical direction is discouraged. The Δt value is expected to be limited by the phase speeds for barotropic and baroclinic gravity waves (i.e., external and internal modes, respectively), which are different from each other by at least an order of magnitude (barotropic is faster). The first-mode baroclonic speed is usually larger than the advective velocity, although comparable in its order of magnitude. The baroclinic time-step is expected to be much smaller than the inertial period so that the Coriolis force does not impose any additional restriction on Δt. Vertical mixing is always treated implicitly since its transport rate can be much larger than the vertical advective rate.

Taking into account the specifics of this physical regime, we have been developing the kernel code in ROMS to have the features in the following list (cf., Fig. 1.1) that foreshadows the algorithms discussed in more detail below.

- Vertical coordinate: Although ROMS nominally belongs to the vertical-boundary-following model family (i.e., $\sigma(x, y, t)$-coordinate), the code stores the

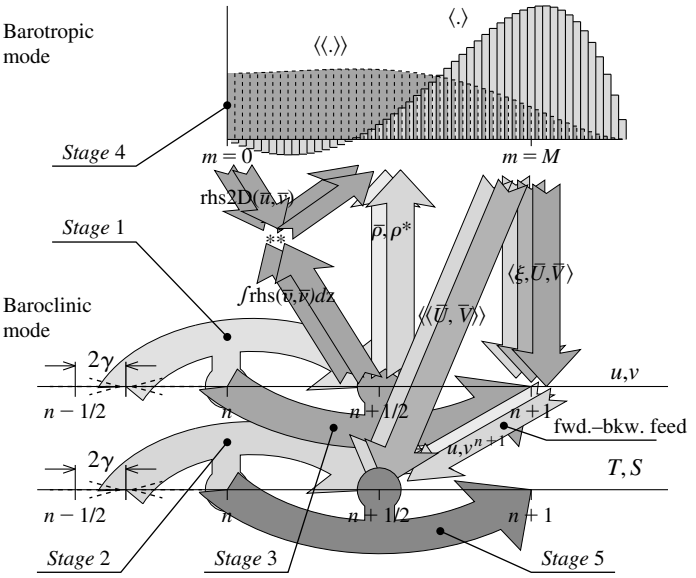


FIG. 1.1 Schematic representation of main time-stepping procedure of ROMS hydrodynamic kernel using LeapFrog – third-order Adams-Moulton (LF–AM3) predictor-corrector step for the baroclinic (3D) mode with mode coupling during the corrector stage. The arcs (curved arrows) represent "steps," i.e., updates of either momenta or tracers that involve computation of right-hand side (r.h.s) terms (shown as circles attached to the arcs). Straight arrows indicate exchange of data between the modes. Each arrow originates at the time when the corresponding variable becomes logically available, regardless of its actual temporal placement. Arcs and arrows are drawn in the sequence that matches the sequence of operations in the actual code: whenever arrows overlap, the operation occurring later corresponds to the arc or arrow on top. Note that labels *Stage 1 . . . Stage 5* correspond to the actual computational stages described in Section 5 of SHCHEPETKIN and MCWILLIAMS [2005]. The four ascending arrows denote the vertically integrated r.h.s. terms for 3D momentum equations; and the two-way, vertically averaged densities, $\overline{\rho}$ and ρ^*, which participate in computation of pressure gradient terms for the barotropic mode (Section 3.2 below). The two descending arrows of smaller size on the left symbolize r.h.s. terms computed from barotropic variables. The asterisks (* *) where the two pairs of ascending and descending arrows meet denote the computation of baroclinic-to-barotropic forcing terms, two smaller arrows ascending diagonally to the right. The five large descending arrows symbolize two-way fast-time-averaged barotropic variables (enclosed in $\langle\,.\,\rangle$ and $\langle\langle\,.\,\rangle\rangle$, Section 3.1 below) for backward coupling; fwd.-bkw. feed stands for FB feedback between momentum and tracer equations – the update of tracers is delayed until the new-time-step velocities u, v^{n+1} become available so that they can participate in computation of r.h.s. terms for tracers; M is mode-splitting ratio – number of barotropic time-steps per one baroclinic. Note that barotropic time-stepping goes slightly beyond ($\sim 25\%$ in the case above) the baroclinic step $[n+1]$; $\gamma = 1/12$ is associated with LF–AM3 algorithm (this is further explained in Fig. 4.1 in Section 4).

height-coordinate transform $z = z(x, y, \sigma)$ as a special array, and, in principle, it can be used as a generalized vertical coordinate.

- Free surface: ROMS is free-surface model with split-explicit time-stepping. The pressure-gradient force (PGF) for the barotropic mode is defined as a variational derivative of vertical integral of the hydrostatic PGF with respect to perturbations in the free-surface elevation $\zeta(x, y, t)$. As a result, the barotropic PGF depends not only on ζ but also on the two differently averaged density fields indicated by two

ascending arrows ρ^* and $\overline{\rho}$ in Fig. 1.1 (Section 3.2) that are computed from three dimensional (3D) fields and held constant during barotropic time-stepping. This insures an accurate and stable split, even with a large ratio between the Δt for the baroclinic and barotropic modes.

- Barotropic averaging: The barotropic variables are averaged in the fast (barotropic) time-step to prevent aliasing of frequencies not resolved by the slow (3D baroclinic) time-stepping. To avoid undesirable damping of resolved frequencies, the fast-time averaging is performed using a specially designed S-shaped filter function (denoted by $\langle\,.\,\rangle$ in Fig. 1.1) that has second-order temporal accuracy for the averaged barotropic prognostic variables, $\langle\zeta, \overline{u}, \overline{v}\rangle$. (A strictly positive-definite averaging yields at most first-order accuracy.)
- Tracer conservation and constancy preservation: To assure these properties when the grid-box control volumes change due to changes in ζ, one must ensure that slow-time volume fluxes are exactly consistent with the changes in ζ as computed with the barotropic mode. Hence, it is not enough to know the final state of $\langle\,.\,\rangle$-averaged barotropic variables at the new time-step; one also needs to have an integral measure of the entire barotropic evolution between two consecutive baroclinic times. This is accomplished by fast-time averaging the barotropic volume flux using a second operator ($\langle\!\langle\,.\,\rangle\!\rangle$ in Fig. 1.1) that is derived from the primary $\langle\,.\,\rangle$ (Section 3.1).
- Barotropic time-stepping: Since the external mode phase speed imposes the dominant CFL restriction on Δt, the generalized forward-backward (FB) step is chosen for barotropic mode. This algorithm consists of a modified Adams-Bashforth update of free surface followed by update of momentum equations where the newly computed ζ participates in the computation of PGF. Unlike the classical FB step, the new algorithm naturally combines with advection and Coriolis terms and has a dissipative leading-order truncation term.
- Baroclinic time-stepping: 3D time-stepping schemes are designed in anticipation of different Courant-number limitations corresponding to different physical processes. The internal gravity-wave speed is expected to be most restrictive although the other limits – advective and Coriolis CFL – are not as distant as in the barotropic mode. A modified predictor-corrector scheme with FB feedback between advection of ρ (via the tracers T and S) and PGF in momentum equations is chosen. It generally maintains temporal third-order accuracy for advection and Coriolis terms to match the accuracy of spatial discretization. The use of FB feedback expands the CFL stability limit for internal waves. A forward Euler step is used for horizontal viscosity and diffusion terms, and an implicit backward step is used for vertical mixing. The overall time-stepping procedure is compatible with both centered and upstream-biased advection, which is important for ROMS where we commonly use a third-order, upstream-biased advection scheme in the horizontal directions for both tracer and momentum, but a centered scheme in the vertical to avoid spurious diffusion due to "rectification" of dissipative truncation terms.
- Temporal stability limits: The time-stepping algorithms are specifically designed for use close to their limiting Courant number for computational stability yet still guarantee a numerically accurate solution. The optimal algorithms are derived by an inverse stability analysis, by writing them with arbitrary coefficients first, then

deriving characteristic equations, and choosing coefficients that yield the desirable characteristic roots. This makes it possible to resolve the phase propagation for both internal and external modes with an accuracy order higher than for each equation taken individually.

- Updating: ROMS's time-stepping utilizes a form where all temporal interpolations are applied to the primitive variables rather than their right-hand-side (r.h.s.) tendencies. This allows us to combine different time-stepping algorithms for different physical terms and reduces memory usage for a more efficient code.
- Baroclinic PGF: This term is discretized with a high-order, density Jacobian scheme based on monotonized cubic polynomial fits for the vertical profiles of ρ and geopotential height z. This scheme preserves most of symmetries of the original Jacobian of BLUMBERG and MELLOR [1987] while dramatically reducing errors in hydrostatic balance.
- Compressible equation of state: Because of seawater's compressibility, most of the vertical change of in situ ρ is due to pressure change. Monotonicity of in situ ρ does not guarantee the absence of spurious oscillations in the interpolated stratification profile; this degrades the accuracy of the PGF scheme and potentially leads to numerical instability. Furthermore, the combination of the Boussinesq approximation and the full equation of state (EOS) is a source of both inaccuracy and mode-splitting error. Therefore, the EOS (JACKETT and MCDOUGALL [1995]) is modified to cancel the bulk compressibilty in in situ ρ to achieve a more consistent Boussinesq approximation (DUKOWICZ [2001]) and reformulated in terms of adiabatic ρ derivatives.
- Advection: ROMS commonly uses a third-order upstream-biased advection in the horizontal direction for both tracer and momentum equations and fourth-order centered advection in the vertical.
- Coriolis and curvilinear metric terms: These are combined with advection of momentum and discretized using an energy-conserving scheme.
- Code architecture: The code architecture is distinct from a modular design (cf., MOM). The architectural design decisions involve optimization in multidimensional space for the model physics, numerical algorithms, and computational performance. As a rule, this results in significantly larger functional units in the code than in more traditional oceanic modeling practice. This is typically beneficial for both exploiting cache locality and minimizing the number of synchronization events in a parallel code.
- Parallelization: ROMS is a parallel code which has both shared- (via OpenMP, www.openmp.org) and distributed-memory (via MPI (GROPP, LUSK, and SKJELLUM [1999])) capabilities, including a possibility of allowing multiple threads within each MPI process. Both OpenMP and MPI options are implemented using two-dimensional subdomain decomposition in horizontal directions.

A detailed description of the components and algorithms of ROMS is outside the scope of this chapter. Instead, we present a comprehensive overview of the kernel algorithms, focusing on algorithm interferences that require special effort to reconcile conflicts so that multiple desired properties can coexist at the same time. Examples of such conflicts

are (i) the barotropic-baroclinic time-splitting scheme (as diagnosed by a linear stability analysis) can interfere with finite-volume mass conservation in slow mode, as well as cause loss of the tracer constancy-preservation property; (ii) linear stability analysis favors FB time-stepping for momentum and tracers over predictor-corrector by the stability vs. computational cost criterion for internal waves alone, but most suitable advection algorithms are two-stage procedures that are more naturally incorporated into a predictor-corrector scheme; (iii) barotropic-baroclinic mode-splitting makes it impossible to satisfy the finite-volume continuity equation on slow baroclinic time during a predictor substep, causing loss of the constancy-preservation property for tracers; (iv) high-order polynomial interpolation requires monotonicity constraints to prevent spurious oscillations if the interpolated field is not smooth on the grid scale, and for ρ, this leads to a monotonicity constraint for stratification that further leads to a redesign of the EOS for seawater; and (v) with modal time splitting, the barotropic time-step requires knowledge of bottom stress related to bottom velocity that is a sum of both types of modes, yet it would be unphysical to remove more than the total momentum within the bottommost grid box per baroclinic time-step while the baroclinic velocity is held constant.

2. Time-stepping: accuracy and linear stability

Oceanic flows in a regime with high Reynolds number can usefully be viewed from the perspective of time-stepping algorithms as satisfying hyperbolic partial differential equations. We consider two simple hyperbolic test systems. One can be called an advection equation,

$$\frac{\partial q}{\partial t} + c\frac{\partial q}{\partial x} = 0, \tag{2.1}$$

and the other a wave system,

$$\frac{\partial \zeta}{\partial t} = -c\frac{\partial u}{\partial x}, \quad \frac{\partial u}{\partial t} = -c\frac{\partial \zeta}{\partial x}. \tag{2.2}$$

Table 4 from Griffies, Böning, Bryan, Chassignet, Gerdes, Hasumi, Hirst, Treguier and Webb [2000] provides a comprehensive summary of time-stepping algorithms used in different oceanic models. These can be subdivided into two major classes. The first class is synchronous schemes where the r.h.s. tendencies for all prognostic variables are computed at the same time and simultaneously used to advance the variables to the next time-step; examples are LF with an Asselin Filter to suppress temporal oscillations, second-order Runge-Kutta (RK2), predictor-corrector (LF with a trapezoidal rule (LF–TR), LF with third-order Adams-Moulton (LF–AM3), second-order Adams-Bashforth with TR predictor-corrector (AB2–TR)), and third-order Adams-Bashforth (AB3) (Durran [1991]). The second class is FB schemes where one variable is advanced then immediately used to advance the other(s),

$$\zeta^{n+1} = \zeta^n - c\Delta t \cdot \frac{\partial u^n}{\partial x}, \quad u^{n+1} = u^n - c\Delta t \cdot \frac{\partial \zeta^{n+1}}{\partial x}, \tag{2.3}$$

where n is a time index. A FB scheme obviously is applicable only to multivariate systems. Almost all OGCMs currently use a synchronous method.

One can easily verify that synchronous time-stepping has identical accuracy and stability limits for the advection equation and wave system (CANUTO, HUSSAINI, QUARTERONI and ZANG [1988], SHCHEPETKIN and MCWILLIAMS [2005]). This typically occurs for $\alpha_{\max} < 0.8$ (where $\alpha \equiv \omega \Delta t$ is the Courant number, and $\omega = ck$ is the frequency for a solution component with wavenumber k) per r.h.s. computation for the most efficient algorithms within this class. This is only half as efficient (as measured by the ratio of stability limit to the number of r.h.s. computations) as a FB scheme with $\alpha_{\max} = 2$. Thus, the commonly used synchronous time-stepping is less than optimal for oceanic modeling because the fastest process, gravity waves, occur as an interplay between momentum and mass as in the wave system. Therefore, we define our primary design goals as (i) to generalize the most used synchronous algorithms (i.e., RK2, LF–TR, LF–AM3, and AB3) by introducing a FB-like feedback and (ii) to generalize FB to higher orders of accuracy. In both cases, the time-stepping algorithm must be accurate and robust even if used close to the α limit for numerical stability.

The methodology employed here is a von Neumann linear stability analysis (DURRAN [1998]) applied in an "inverse" manner to design the algorithm rather than to assess one chosen a priori. We insert adjustable parameters into a time-stepping algorithm, then derive the characteristic equation for the eigenvalues of the step-multiplier matrix, and then solve it as an optimization problem to find parameters that achieve the desired properties. These properties include the order of accuracy and related bandwidth of the resolved frequency spectrum that is accurately represented, the maximum stability limit, the nature of the dominant truncation error term (note that dissipation of fastest, poorly resolved frequencies is preferred over dispersion), and sufficient damping for any computational modes. The method is applied to the spatial Fourier transform of Eq. (2.3),

$$\frac{\partial \zeta}{\partial t} = -i\omega \cdot u, \quad \frac{\partial u}{\partial t} = -i\omega \cdot \zeta, \tag{2.4}$$

with $\omega = ck$. Although it is implicit here that the primitive system is nonlinear, the stability analysis is linear. For example, consider the evolution of a small perturbation to a nonlinear flow described by

$$\frac{\partial \zeta}{\partial t} + V\frac{\partial \zeta}{\partial x} = -c\frac{\partial u}{\partial x}; \quad \frac{\partial u}{\partial t} + V\frac{\partial u}{\partial x} = -c\frac{\partial \zeta}{\partial x}, \tag{2.5}$$

where V is velocity of background flow. An instability of an algorithm applied to Eq. (2.5) would automatically imply an instability of the fully nonlinear system using the same algorithm. Thus, a practical time-stepping algorithm for Eq. (2.5) is always a combination of both a generalized FB step for terms involving the ζ-u interplay and a synchronous algorithm for other terms where a FB step is either not applicable or impractical. Although less critical in its CFL limitation, the synchronous step must be at least conditionally stable. A similar requirement comes from the need for stable treatment of advection and Coriolis force, and the latter is the more restrictive since robustly stable, dissipative, upstream-biased, advection schemes can be used.

2.1. A two-time-level scheme: RK2 with FB feedback

Consider a discrete time-stepping algorithm for Eq. (2.4) with a predictor substep,

$$\begin{aligned} \zeta^{n+1,*} &= \zeta^n - i\alpha \cdot u^n, \\ u^{n+1,*} &= u^n - i\alpha \cdot \left[\beta\zeta^{n+1,*} + (1-\beta)\zeta^n\right], \end{aligned} \tag{2.6}$$

followed by a corrector substep,

$$\begin{aligned} \zeta^{n+1} &= \zeta^n - \frac{i\alpha}{2} \cdot \left(u^{n+1,*} + u^n\right), \\ u^{n+1} &= u^n - \frac{i\alpha}{2} \cdot \left[\epsilon\zeta^{n+1} + (1-\epsilon)\zeta^{n+1,*} + \zeta^n\right]. \end{aligned} \tag{2.7}$$

Setting $\beta = \epsilon = 0$ in the above reverts it to the standard RK2 time-stepping that is unstable for a non-dissipative system (purely real-valued α) since the eigenvalue magnitude is $|\lambda| = \sqrt{1+\alpha^4/4} \approx 1 + \alpha^4/8 > 1$, implying amplitude growth in time for any α. But, because in the limit $\alpha \to 0$ its growth rate asymptotes to unity faster than $1 + \mathcal{O}(\alpha^2)$, it is sufficient to add hyperdiffusivity rather than normal diffusivity to stabilize a forward-in-time, centered-in-space scheme. This behavior is called weak or asymptotic instability.

The presence of terms with β and ϵ brings FB feedback into the algorithm Eqs. (2.6) and (2.7), and both accuracy and stability can be improved by having them present. Using the r.h.s. of the predictor equations, we eliminate $\zeta^{n+1,*}$ and $u^{n+1,*}$ from the corrector and transform the algorithm into a single step written in matrix form as

$$\begin{pmatrix} \zeta \\ u \end{pmatrix}^{n+1} = \begin{pmatrix} 1 - \frac{\alpha^2}{2} & -i\alpha\left(1 - \frac{\alpha^2\beta}{2}\right) \\ -i\alpha\left(1 - \frac{\alpha^2\epsilon}{4}\right) & 1 - \frac{\alpha^2}{2} + \frac{\alpha^4\beta\epsilon}{4} \end{pmatrix} \begin{pmatrix} \zeta \\ u \end{pmatrix}^n. \tag{2.8}$$

This yields the characteristic equation for $\lambda(\alpha)$,

$$\lambda^2 - \left(2 - \alpha^2 + \frac{\alpha^4\beta\epsilon}{4}\right)\lambda + 1 + \frac{\alpha^4}{4}(1 - 2\beta - \epsilon + \beta\epsilon) = 0. \tag{2.9}$$

Since the exact solution of Eq. (2.4) has $\lambda = e^{\pm i\alpha}$, corresponding to right- and left-traveling waves in Eq. (2.2), we substitute the desired solution into Eq. (2.9) and expand it in a Taylor series for small α, seeking to approximate the ideal step-multiplier as accurately as possible by suppressing mismatch terms with successive powers of α:

$$\alpha^4\left(\frac{1}{3} - \frac{\beta}{2} - \frac{\epsilon}{4}\right) \pm i\alpha^5\left(\frac{1}{12} - \frac{\beta\epsilon}{4}\right) + \mathcal{O}(\alpha^6) = 0. \tag{2.10}$$

Choosing $\epsilon = 4/3 - 2\beta$ eliminates the $\mathcal{O}(\alpha^4)$ term, reducing the above to

$$\pm i\alpha^5\left[\frac{1}{36} + \frac{1}{2}\left(\beta - \frac{1}{3}\right)^2\right] + \mathcal{O}(\alpha^6) = 0. \tag{2.11}$$

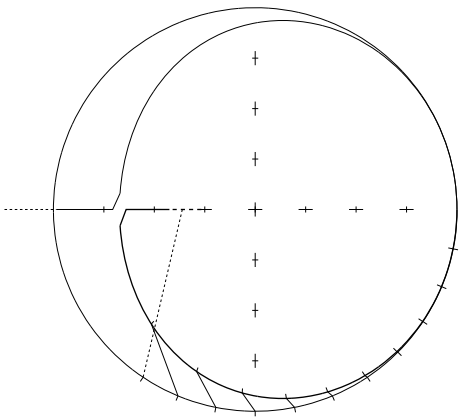


FIG. 2.1 Characteristic roots for the modified RK2 scheme (Eqs. (2.6) and (2.7)) with $\beta = 1/3$ and $\epsilon = 2/3$ relative to the unit circle. Tickmarks on the outer side of the unit circle point to the locations of "ideal" amplification factors $e^{-i\alpha}$ for $\alpha \in \{-\pi/16,\ -\pi/8,\ -3\pi/16, \dots\}$. Tickmarks on the inner side of bold solid curve indicate the actual roots corresponding to these values of α. The ideal and the actual root locations are connected by a thin straight line whose length and orientation show the magnitude and the nature (dispersive/dissipative) of numerical error. This algorithm has a third-order accurate step-multiplier $\lambda = \lambda(\alpha)$ and a stability limit $\alpha_{\max} = \sqrt{6(3-\sqrt{5})} = 2.14093$.

No real-valued β can eliminate the $\mathcal{O}(\alpha^5)$ term, one can only minimize the residual by setting $\beta = 1/3$, and correspondingly, $\epsilon = 2/3$. The position of characteristic roots relative to the unit circle (i.e., the exact solution) is shown in Fig. 2.1.

The stability range of this algorithm is limited by one of the modes leaving the unit circle through $\lambda = -1$. Substituting $\lambda = -1$ and $\epsilon = 4/3 - 2\beta$ into Eq. (2.9) yields

$$4 - \alpha^2 + \left[\frac{1}{36} - \left(\beta - \frac{1}{3}\right)^2\right]\alpha^4 = 0, \tag{2.12}$$

which is to be solved for $\alpha = \alpha(\beta)$ with β playing the role of an independent parameter. A simple analysis leads to the conclusion that $\beta = 1/3$ yields the maximum α ($= 2.14093$), hence the largest possible stability limit, and as shown in Eq. (2.11) and in the next paragraph, the same β value corresponds to the minimum possible truncation error among the whole subset of third-order schemes.

Overall, this modified RK2 algorithm is in line with two-time-level schemes of HALLBERG [1997] and HIGDON [2002], except that they do not contain any counterpart for the free parameter ϵ in Eq. (2.7) by always selecting $\epsilon = 1$ (hence their algorithms cannot be reverted back to classical RK2). Setting $\beta = 0$ and $\epsilon = 1$ in Eqs. (2.6) and (2.7) yields a non-dissipative scheme that makes Eq. (2.9) identical to the characteristic equation for classical FB. This leads to second-order accuracy and $\alpha_{\max} = 2$. In the absence of Coriolis force, the algorithm in Eqs. (4)–(6) of HIGDON [2002] has an identical characteristic equation, eigenvalues, accuracy, and stability limit. Setting $\beta = 1/2$ and $\epsilon = 1$ yields another second-order algorithm which is similar to Eq. (16) of HIGDON [2002]

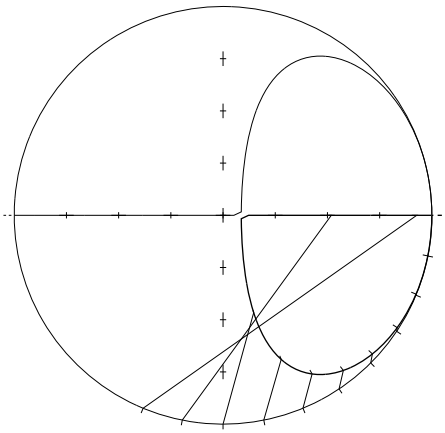


FIG. 2.2 Same as Fig. 2.1, but for $\beta = 1/2$ and $\epsilon = 1$. This setting is similar to HALLBERG [1997].

(cf., Eqs. (3.9), (3.10), (3.13), and (3.14) of HALLBERG [1997]). Again the stability limit is $\alpha_{\max} = 2$, but the scheme is highly dissipative (Fig. 2.2).[2]

RUEDA, SANMIGUEL-ROJAS and HODGES [2007] considered a family of RK2-type algorithms for the baroclinic mode of the tidal, residual, intertidal mudflat (TRIM) model.[3] They combine the predictor step (Eq. (2.6)) with[4]

$$\begin{aligned} \zeta^{n+1} &= \zeta^n - i\alpha \cdot \left[\gamma u^{n+1,*} + (1-\gamma)u^n\right], \\ u^{n+1} &= u^n - i\alpha \cdot \left[\theta \zeta^{n+1} + (1-\theta)\zeta^n\right], \end{aligned} \tag{2.13}$$

where again there is no ϵ-mixing between predicted and corrected ζ, but an extra degree of freedom is introduced by allowing γ and θ deviate from $\gamma = \theta = 1/2$. To be second-order accurate requires $\gamma + \theta = 1$. Once this is satisfied, an additional constraint, $\beta\gamma = 1/12$, makes this algorithm third-order accurate. RUEDA, SANMIGUEL-ROJAS and HODGES [2007] restricted their analysis to a set of discrete values with $\theta = 1/2$ or 1, and β and γ are various permutations of 0, 1/2, and 1, all of which result in either second- or first-order accuracy. They also showed that the choice of $\gamma = \theta = 1/2$ and $\beta = 1/6$ results in a third-order accurate algorithm. By making an analysis similar to Eq. (2.12), one can also show that this choice yields the largest possible stability limit, $\alpha_{\max} = 2$: any deviation of γ and θ from 1/2 while maintaining $\gamma + \theta = 1$ and $\beta\gamma = 1/12$ reduces $\alpha_{\max}$ relative to this value. Overall, it is comparable though slightly more dissipative than Eqs. (2.6) and (2.7) with $\beta = 1/3$ and $\epsilon = 2/3$ (Fig. 2.3).

[2]The algorithms of HIGDON [2002] and HALLBERG [1997] can be viewed as the two extreme members of the β-family of second-order schemes (Eqs. (2.6) and (2.7)) with $\epsilon = 1$ and $\beta \in [0, 1/2]$. All of them have a stability limit $\alpha_{\max} = 2$ independently of β, and they differ only by the dissipation rate that increases with β.

[3]TRIM is an ocean model, whose emphasis is on fine-scale coastal dynamics and coastal engineering (CASULLI and CHENG [1992]).

[4]Equations (2.6)–(2.13) can be remapped into Rueda's Eqs. (50), (51), (52), and (25) using the following our $\rightarrow$ their substitute of variables: $\zeta \rightarrow u$; $u \rightarrow \rho$, p; $\beta \rightarrow \theta_{\rm p}$; $\gamma \rightarrow \theta_{\rm b}$; $\theta \rightarrow \theta$.

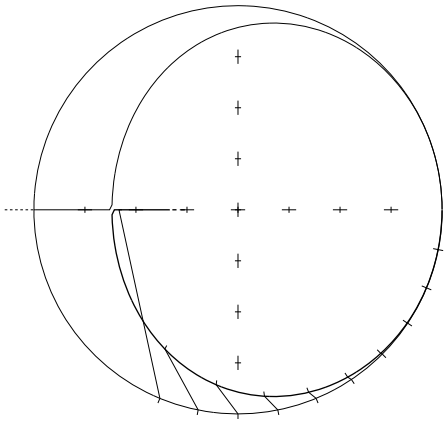


FIG. 2.3 RUEDA, SANMIGUEL-ROJAS and HODGES [2007] algorithm with their $\theta_p = 1/6$ and $\theta_b = \theta = 1/2$. This is equivalent to our Eqs. (2.6) and (2.13) with $\gamma = \theta = 1/2$ and $\beta = 1/6$.

Despite the fact that two-time-level algorithms for shallow-water equations[5] are perhaps the most studied (HALLBERG [1997], HIGDON [2002, 2005], RUEDA, SANMIGUEL-ROJAS and HODGES [2007], SHCHEPETKIN and MCWILLIAMS [2005]) and have an extensive history, none of the previous work has produced a scheme which is competitive with the classical FB step in terms of its stability limit relative to computational cost ($\alpha_{\max} = 2$ with the r.h.s. computed only once per time-step for each equation). An examination of the characteristic equations resulting from the two versions of a predictor-corrector algorithm – Eq. (2.6) in combination with either Eq. (2.7) or Eq. (2.13) – reveals that neither has sufficient degrees of freedom, despite the presence of three free coefficients in each. This can be remedied by combining ϵ- and θ-weightings for the corrector step, so it becomes

$$\begin{aligned} \zeta^{n+1} &= \zeta^n - i\alpha \cdot \left[(1-\theta)u^{n+1,*} + \theta u^n\right], \\ u^{n+1} &= u^n - i\alpha \cdot \left[\theta\left(\epsilon\zeta^{n+1} + (1-\epsilon)\zeta^{n+1,*}\right) + (1-\theta)\zeta^n\right], \end{aligned} \tag{2.14}$$

where we already replaced γ by $1-\theta$ in Eq. (2.13) to make it second-order accurate. As expected, the characteristic equation for (2.6) and (2.14) is

$$\lambda^2 - \lambda\left[2 - \alpha^2 + \alpha^4 A\right] + 1 - \alpha^4 B = 0, \quad \text{where} \quad \begin{cases} A = \beta\epsilon\theta(1-\theta) \\ B = (1-\theta)(\beta - \beta\epsilon\theta + \epsilon\theta - \theta) \end{cases}, \tag{2.15}$$

[5] In its classical sense, the term *shallow-water equations* refers to a single-layer of shallow, hydrostatically balanced homogeneous fluid. After CASULLI and CHENG [1992] and CASULLI and CATTANI [1994], it is frequently applied to hydrostatically balanced, stratified fluids, including ones admitting internal waves. Loosely, it is also applicable to governing equations for stratified, multilayer modeling in isopycnic coordinates.

and it reverts back to Eq. (2.9) if $\theta = 1/2$.[6] Substitution of $\lambda = e^{\pm i\alpha}$ and Taylor-series expansion leads to

$$\alpha^4\left(\frac{1}{12} - A - B\right) \pm i\alpha^5 B + \alpha^6\left(\frac{B}{2} - \frac{1}{360}\right) + \mathcal{O}\big(\alpha^7\big) = 0, \tag{2.16}$$

where the absence of a α^3-term guarantees second-order accuracy for any combination of β, θ, and ϵ. Obviously, one cannot eliminate both $\mathcal{O}\big(\alpha^5\big)$ and $\mathcal{O}\big(\alpha^6\big)$ terms simultaneously. To achieve third-order accuracy, one needs to satisfy $A + B = 1/12$, which leads to the condition

$$\epsilon = 1 + \frac{1}{12\theta(1-\theta)} - \frac{\beta}{\theta}\,, \tag{2.17}$$

and turns A and B in Eq. (2.15) into

$$\left.\begin{aligned} A &= (1-\theta)\left(C^2 - (\beta - C)^2\right) \\ B &= (1-\theta)\left((\beta - C)^2 - C^2 + \frac{1}{12(1-\theta)}\right) \end{aligned}\right\} \quad \text{where} \quad C = \frac{\theta}{2} + \frac{1}{24(1-\theta)}. \tag{2.18}$$

The expression for B can be made equal to zero to eliminate $\mathcal{O}\big(\alpha^5\big)$ term in Eq. (2.16) only when $\theta > 0.945$,[7] resulting in a non-dissipative fourth-order algorithm; however, it has unattractive properties: a significant portion of the α-range within the limit of stability yields a wrong phase speed without providing any damping at all, and the coefficients in Eq. (2.14) are no longer non-negative because values of (θ, β) which make $B = 0$ also result in $\epsilon > 1$ as follows from Eq. (2.17); e.g., $\theta > 0.945$ yields $\beta = 1.230$ and $\epsilon = 1.302$. For $0 \le \epsilon \le 1$, it is only possible to minimize the dissipation by selecting

$$\beta = \frac{\theta}{2} + \frac{1}{24(1-\theta)} \tag{2.19}$$

for any θ, which is still treated as a free parameter.

Algorithms of this kind become unstable when the two modes meet at some point on the real axis, after which one of them leaves the unit circle through either $\lambda = -1$ or $\lambda = +1$, whichever occurs earlier in α. Substituting $\lambda = \pm 1$ into Eq. (2.15) yields

$$\begin{aligned} &\lambda = -1: \qquad 4 - \alpha^2 + \alpha^4(A - B) = 0, \\ &\lambda = +1: \qquad \alpha^2\left[1 - \alpha^2(A + B)\right] = 0. \end{aligned} \tag{2.20}$$

[6] After setting $\epsilon = 1$ in Eq. (2.15), this also coincides with Eq. (53) from Rueda, Sanmiguel-Rojas and Hodges [2007] if θ_b is replaced with $1 - \theta$ there.

[7] The minimum possible value of θ which makes $B = 0$ in Eq. (2.18) occurs when $\beta = C$ (hence eliminating the first quartic term in the expression for B) and $C^2 = 1/[24(1-\theta)]$, which after substitution of the expression for C yields a quartic equation for θ alone. Its only solution within the range of interest, $0 < \theta < 1$, is $\theta = 0.9452697779$. Any change in β relative to $\beta = C$ results in a larger value of θ.

The first line results in $\alpha^2_{\max} = \left(1 \pm \sqrt{1 + 16(A - B)}\right) / [2(A - B)]$ where the sign $\pm$ must be chosen to be the same as the sign of $(A - B)$. The solution exists only if $A - B < 1/16$. As $A - B \to 1/16$, then $\alpha_{\max} \to \sqrt{8}$, which is the largest stability limit when this limitation applies. (Note that $\alpha_{\max} = 2$ in the case of $A - B = 0$ and changes continuously when $A - B$ changes sign.) The second line in Eq. (2.20) yields $\alpha^2_{\max} = 1/(A + B)$, which with Eq. (2.17) leads to a less restrictive $\alpha_{\max} = \sqrt{12}$ for the entire subset of third-order algorithms. Figure 2.4 summarizes this for the space of parameters θ, β, and ϵ within the domain to avoid negative coefficients in Eq. (2.14), $0 \le \theta$ and $\epsilon \le 1$. In contrast, $\beta > 0$ can, in principle, exceed 1 because no coefficient like $1 - \beta$ is present in Eq. (2.6). This figure reveals the existence of an area where the stability is limited only by the lower line in Eq. (2.20); i.e., none of the modes ever leaves the unit circle through $\lambda = -1$. We are therefore interested in (β, θ)-pairs from the portion of the shaded area in Fig. 2.4 just below the upper solid bold line that corresponds to $\epsilon = 1$. Furthermore, to minimize the $\mathcal{O}(\alpha^5)$ truncation term, we are interested in algorithms

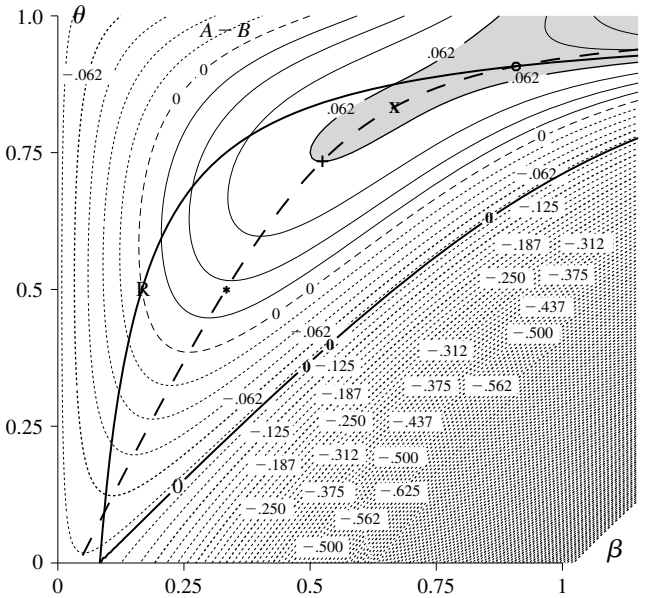


FIG. 2.4 Stability map for the two-parameter (β, θ)-family of third-order RK2 algorithms (Eqs. (2.6) and (2.14)) with ϵ set to satisfy Eq. (2.17). Thin contours show the difference of $A - B$ from Eq. (2.18) as a function of (β, θ), which controls the stability limit due to one of the modes leaving the unit circle at $\lambda = -1$. The shaded area corresponds to $A - B > 1/16$, where this no longer happens; hence, the stability range is limited only by the mode leaving through $\lambda = +1$ resulting in $\alpha_{\max} = \sqrt{12} \approx 3.4641$ for all settings within the shaded area. Superimposed bold solid curves correspond to $\epsilon = 0$ (lower) and $\epsilon = 1$ (upper); the values of (β, θ) must be chosen between these two curves in order for the algorithm to have all non-negative coefficients in Eq. (2.17). The bold-dashed curve corresponds to a minimal dissipation subset with $\beta = \beta(\theta)$ from Eq. (2.19). Specific settings shown on this map are **R** (RUEDA, SANMIGUEL-ROJAS and HODGES [2007]) and $*$ (SHCHEPETKIN and MCWILLIAMS [2005], also Fig. 2.1). The points $+$, **x**, and **o** refer to Fig. 2.5.

with β and θ related by Eq. (2.19) as represented by the bold-dashed line in Fig. 2.4. Remarkably, this line follows the maximum of $A - B$ for any given θ so that settings which minimize the truncation error are also optimal for stability.

Characteristic roots for three algorithms from the shaded area are shown in Fig. 2.5. The one with $\theta = 0.734$ corresponds to just after entrance into the shaded area along

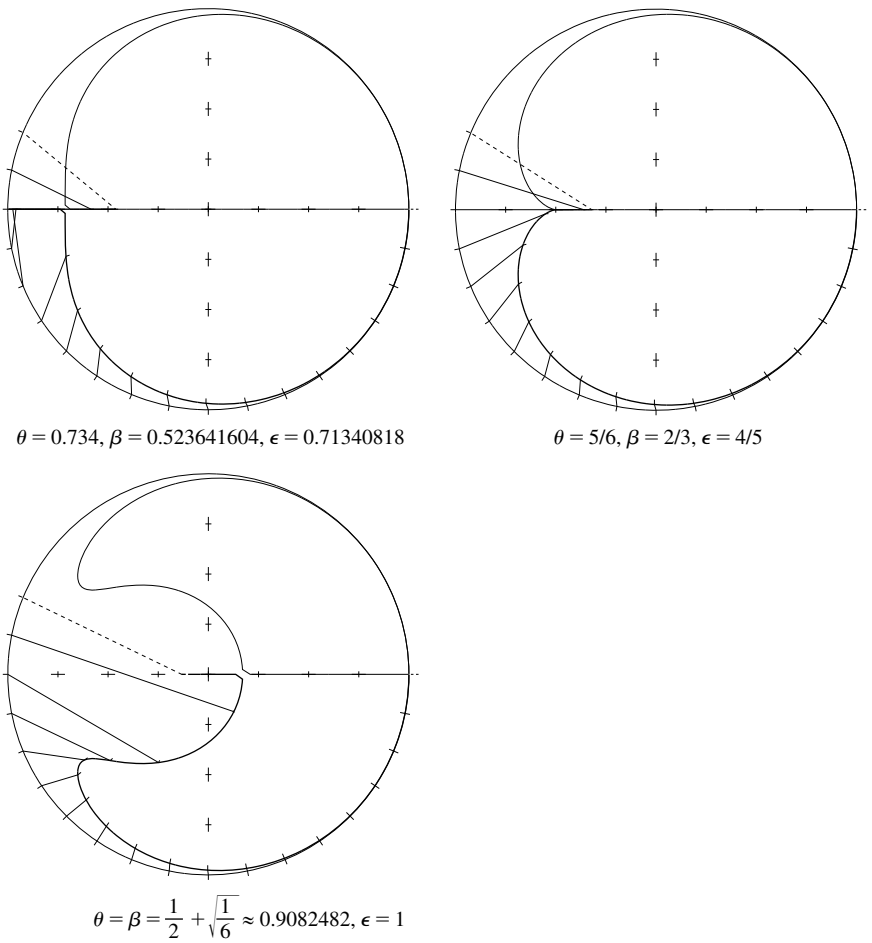


FIG. 2.5 Characteristic roots for the RK2 algorithm (Eqs. (2.6) and (2.14)) with coefficients chosen to yield third-order accuracy and minimal dissipation (i.e., both conditions (Eqs. (2.17) and (2.19)) are met) for three different values of θ. In the case of $\theta = 0.734$, the two arms meet each other at $\lambda \approx -0.7$ after which one of them proceeds along the real axis toward $\lambda = -1$ but stops when nearly reaching this point and reverses direction, continuing toward the center. (Note that roots corresponding to $\alpha = 7\pi/8$ and $\alpha = 15\pi/16$ are very close to each other, which indicates the existence of a stagnation point for $\lambda = \lambda(\alpha)$ in the vicinity (smaller values of θ result in one of the arms exiting the circle at $\lambda = -1$, as it occurs in Fig. 2.1, while larger θs move the reversal point closer to the center). Stability is limited by the other arm reaching $\lambda = +1$ at $\alpha = \alpha_{\max} = \sqrt{12}$, which is also the stability limit for the other two panels in this figure.

the dashed line[8] (denoted as "+" in Fig. 2.4). The overall behavior of the algorithm is similar to that in Fig. 2.1 except that it has slightly lower dissipation. More importantly, after the two arms meet each other at $\lambda \approx -0.7$, one of them continues toward the negative real axis, but instead of exiting at $\lambda = -1$, it stops there, reverses direction, and continues toward the center. Note the existence of the stagnation point discussed in the caption. Setting θ slightly smaller than 0.734 causes this mode to exit at $\lambda = -1$.

Increasing θ beyond 0.734 while following the dashed line in Fig. 2.4 moves the stagnation point toward the center of the unit circle, and subsequently it changes the behavior of the algorithm in the vicinity of the point where the two arms meet each other. For $\theta = 5/6$ (denoted as **x** in Fig. 2.4), they no longer approach the real axis at a 90-degree angle, but rather they bend inward and touch the real axis. The portion of the α-spectrum for which the roots λ are located on the real axis to the left of the merging point disappears when θ increases beyond $5/6$. This is beneficial for the algorithm because phase increments of α beyond π are within the aliasing range: wavenumber components corresponding to them cannot be propagated along the grid, so if the algorithm is used in this regime, these signals must be damped. Further increase of θ changes this behavior again. Instead of approaching the real axis, the arms bend inward, resulting in a highly dissipative algorithm for the upper portion of the spectrum, $13\pi/16 \le \alpha < \sqrt{12}$. Figure 2.5, lower left, shows the characteristic roots for an algorithm with maximum possible β and θ along the minimal dissipation curve with all-non-negative coefficients in Eqs. (2.6) and (2.14) (this is the point **o** in Fig. 2.4, located at intersection of the bold dashed and solid $\epsilon = 1$ lines).[9] Variation of θ within the range $0.734 \le \theta \le 0.91$ causes only minor effects on the behavior of this algorithm within the lower, physically accurate, portion of its spectrum, $|\alpha| < \pi/2$. All three examples on Fig. 2.5 demonstrate very small numerical dispersion and a dissipation-dominant truncation error outside this range.

This class of time-stepping algorithms is an attractive choice for isopycnic and high-resolution coastal engineering models because it is a two-time-level scheme that combines nicely with positive-definite advection algorithms as well as with wetting-and-drying schemes that also require the use of limiters. Having all non-negative coefficients in front of the r.h.s. terms in a time-stepping scheme is crucial (Stelling and Duinmeijer [2003]). Its accuracy, stability, and efficiency are superior to most of the known algorithms. It is somewhat less attractive for z- or σ-coordinate models in the context of long-term, large-scale simulation because it is incompatible with centered vertical advection needed to avoid long-term drift: although this requirement is mitigated relative to forward-in-time stepping, some degree of upstream-biasing of advection schemes is

[8]The exact value of θ for the point of entry into the shaded area in Fig. 2.4 comes from the equation, $\left(\theta^2 - \frac{1}{6}\right)(1-\theta)^2 - \frac{1-\theta}{8} + \frac{1}{144} = 0$, derived by substituting expressions for A, B, and C from Eq. (2.18) along with the condition $\beta = C$ from Eq. (2.19) into $A - B = 1/16$. This yields $\theta = 0.7332939955221$.

[9]Since this choice belongs to $\epsilon = 1$ – family, it can be used without modification in the TRIM code (Rueda, Sanmiguel-Rojas and Hodges [2007]), except for setting coefficients θ, $\theta_p = 1/2 + \sqrt{1/6}$ and $\theta_b = 1/2 - \sqrt{1/6}$ in their Eqs. (51), (52), and (25).

required for stability if RK2-type time-stepping is used.[10] The existence of two-time-level, predictor-corrector algorithms with a stability limit $\alpha_{\max}$ beyond 3 has been long overlooked, and, in fact, this makes it competitive with the FB-type algorithms considered later in this section in terms of computational efficiency (i.e., the ratio of the stability limit to the number of r.h.s. computations for each equation).

2.2. *LF–TR or LF–AM3 with FB feedback*

Another possibility is an algorithm comprised of a LF predictor substep followed by either a two-time TR or a three-time AM3 corrector:

$$\begin{aligned} \zeta^{n+1,*} &= \zeta^{n-1} - 2i\alpha\cdot u^n, \\ u^{n+1,*} &= u^{n-1} - 2i\alpha\cdot\left[(1-2\beta)\,\zeta^n + \beta\left(\zeta^{n+1,*}+\zeta^{n-1}\right)\right], \end{aligned} \tag{2.21}$$

and

$$\begin{aligned} \zeta^{n+1} &= \zeta^n - i\alpha\cdot\left\{\left(\frac{1}{2}-\gamma\right)u^{n+1,*} + \left(\frac{1}{2}+2\gamma\right)u^n - \gamma u^{n-1}\right\}, \\ u^{n+1} &= u^n - i\alpha\cdot\left\{\left(\frac{1}{2}-\gamma\right)\left[\epsilon\zeta^{n+1}+(1-\epsilon)\zeta^{n+1,*}\right] + \left(\frac{1}{2}+2\gamma\right)\zeta^n - \gamma\zeta^{n-1}\right\}, \end{aligned} \tag{2.22}$$

where the parameters β and ϵ introduce FB-feedback during both stages while γ controls the type of corrector scheme. Without FB-feedback, the standard algorithm is

$$\beta=\epsilon=0 \quad\Rightarrow\quad \begin{cases} \gamma = 0 & \Rightarrow \text{ LF–TR} & \alpha_{\max}=\sqrt{2} \\ \gamma = 1/12 & \Rightarrow \text{ LF–AM3} & \alpha_{\max}=1.5874 \\ \gamma = 0.0804 & \Rightarrow \text{ max stability} & \alpha_{\max}=1.5876\,, \end{cases}$$

which is one of the most efficient and attractive synchronous algorithms (cf., Fig. 20 in Shchepetkin and McWilliams [2005]).

Following exactly the same path as for RK2 above, we derive a set of constraints for coefficients β, γ, and ϵ to achieve the specified orders of accuracy (Shchepetkin and McWilliams [2005]):

$$\text{third-order:} \qquad \gamma = \frac{1}{12} \qquad \forall\ \beta,\epsilon, \tag{2.23}$$

$$\text{fourth-order:} \qquad \text{above and } \beta = \frac{7}{30} - \frac{\epsilon}{6} \qquad \forall\ \epsilon, \tag{2.24}$$

$$\text{fifth-order:} \qquad \text{both above and } -\frac{5}{6}\left(\epsilon-\frac{11}{20}\right)^2 - \frac{1603}{2400} = 0\,. \tag{2.25}$$

[10] For example, quick advection is asymptotically unstable in combination with forward-in-time stepping. (In contrast, quickest which explicitly contains the second-order, time-dependent terms is stable.) However, as discussed in Rueda, Sanmiguel-Rojas and Hodges [2007], quick is stable in combination with RK2 while centered scheme is asymptotically unstable.

No set of coefficients can satisfy the condition for fifth-order accuracy, so we can only minimize the leading-order truncation term by choosing $\epsilon = 11/20$, hence $\beta = 17/120$ and $\gamma = 1/12$. This yields a fourth-order scheme with extremely small numerical dispersion and dissipation within the whole range of its numerical stability, $\alpha_{\max} = 1.851640$ (Fig. 2.6 lower-left panel).

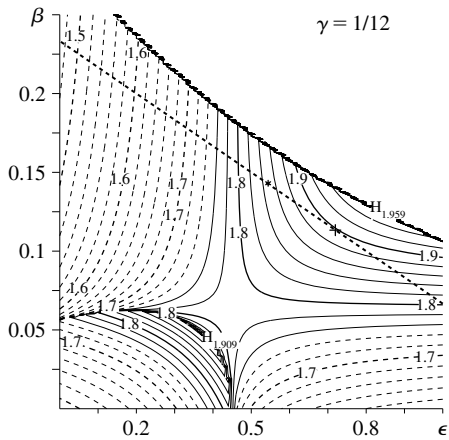
Since a primary goal is to extend the stability range, we progressively give up one order of accuracy at a time, which frees one or two parameters, ϵ, or (ϵ, β) be available for tuning. Figure 2.6 (upper-left) shows a map of the stability range $\alpha_{\max}$ in an ϵ, β-plane, for all third-order accurate schemes (hence $\gamma = 1/12$ is always respected). The subset of fourth-order schemes is represented by the diagonal line, $\beta = 7/30 - \epsilon/6$, that is nearly parallel to the edge of stability. Overall, there are two stability maxima in the ϵ, β-plane, and remarkably, the choices corresponding to maximum stability are not far away from the minimal truncation error within the fourth-order subset. As a result, $\epsilon = 0.83$, $\beta = 0.126$ corresponds to the largest possible $\alpha_{\max} = 1.958537$, and it is also very accurate within the whole stability range (Fig. 2.6, upper-right). It has a 25% larger stability limit than the 1.5874 of the original LF–AM3 scheme with $\beta = \epsilon = 0$. The secondary maximum (lower-left) is less attractive and, in fact, produces similar leading-order numerical dissipation and dispersion errors as does $\beta = \epsilon = 0$ LF–AM3, albeit with a wider stability range.

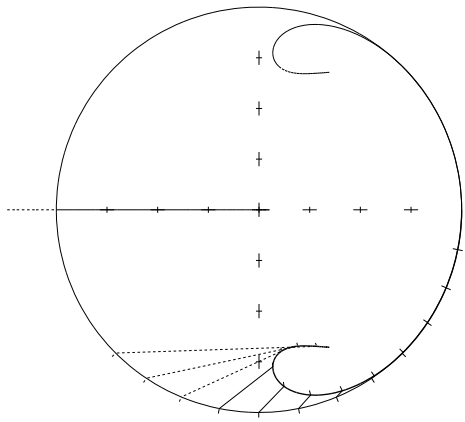
Searching for the maximum stability range in γ, β, ϵ-space while maintaining second-order accuracy (hence $\gamma \neq 1/12$ but is otherwise an adjustable parameter) requires essentially the same kind of analysis as in Figure 2.6 (upper-left) but repeated for different values of γ. This is summarized in Fig. 2.7 (upper-right), with the upper-left panel showing a particular example of $\alpha_{\max} = \alpha_{\max}(\epsilon, \beta)$ for $\gamma = 0$. It turns out that the stability range can be expanded significantly with a decrease of γ, however, at the expense of accuracy degradation. Given that these schemes are dissipative, this is acceptable and in fact desirable for the barotropic mode (since fast motions are fast-time-averaged anyway) and for applications where the wave propagation is not of primary interest. Thus, the introduction of FB-feedback into a LF–TR ($\gamma = 0$) scheme can achieve up to 70% gain in stability range relative to $\beta = \epsilon = 0$ (Fig. 2.7, lower-left). Going beyond $\gamma < 0$ is not desirable due to loss of accuracy. Still, none of these schemes can achieve an efficiency comparable to the classical FB scheme in terms of the ratio of $\alpha_{\max}$ and the number of r.h.s. computations.

2.3. Generalized FB with AB2–AM3

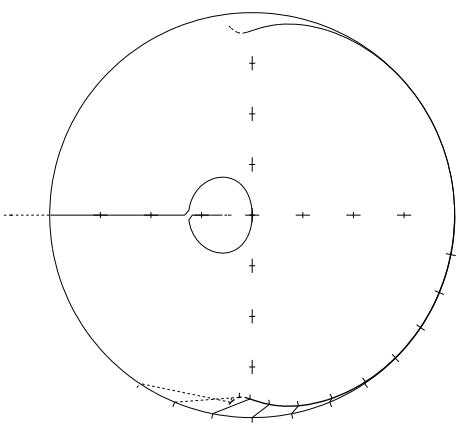
To approach the problem from the opposite direction – starting with a FB scheme and attempting to construct an algorithm compatible with both advection and wave propagation – we consider an explicit algorithm comprised of an AB2-like step for ζ followed by an AM3-like step for u:

$$\begin{aligned} \zeta^{n+1} &= \zeta^n - i\alpha\left[(1+\beta)u^n - \beta u^{n-1}\right], \\ u^{n+1} &= u^n - i\alpha\left[(1-\gamma-\epsilon)\zeta^{n+1} + \gamma\zeta^n + \epsilon\zeta^{n-1}\right]. \end{aligned} \tag{2.26}$$

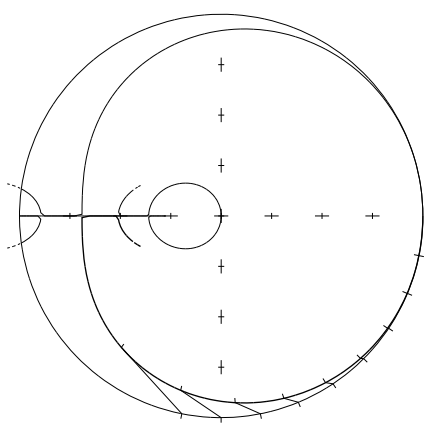




$\beta = 0.126$, $\epsilon = 0.83$: maximum possible stability range ($\alpha_{\max} = 1.958537$)



$\beta = 17/120$, $\epsilon = 11/20$: fourth-order accuracy with minimum possible truncation error ($\alpha_{\max} = 1.851640$)



$\beta = 0.044$, $\epsilon = 0.39$: secondary stability maximum ($\alpha_{\max} = 1.908525$)

FIG. 2.6 Upper left: stability limit $\alpha_{\max}$ as a function of ϵ and β with $\gamma = 1/12$ (i.e., among all third-order accurate schemes within the generalized LF-AM3 family). The empty area in the upper-right corner corresponds to schemes with an asymptotic instability for the physical modes. The straight dashed line $\beta = 7/30 - \epsilon/6$ approximately parallel to the edge corresponds to zero $\mathcal{O}(\alpha^5)$ truncation term (i.e., a fourth-order accurate subset). The asterisk (*) and cross (+) on this line denote locations of the minimal possible truncation error and maximum stability limit among the fourth-order algorithms, which are not far away from each other. Note the stability maxima at $(\epsilon, \beta) = (0.83, 0.126)$, just on the edge of asymptotic instability and $(0.39, 0.044)$. The three remaining panels show the characteristic roots for the β and ϵ choices yielding the indicated specific properties.

Obviously, it reverts to the classical FB scheme if $\beta = \gamma = \epsilon = 0$. Its characteristic equation is

$$\lambda^2 - \left[2 - \alpha^2 (1 - \gamma - \epsilon)(1 + \beta)\right]\lambda + 1 - \alpha^2 (\beta - \gamma - 2\beta\gamma - \beta\epsilon) + \alpha^2 (\epsilon + \beta\epsilon - \beta\gamma)\lambda^{-1} - \alpha^2 \beta\epsilon\lambda^{-2} = 0\,. \tag{2.27}$$

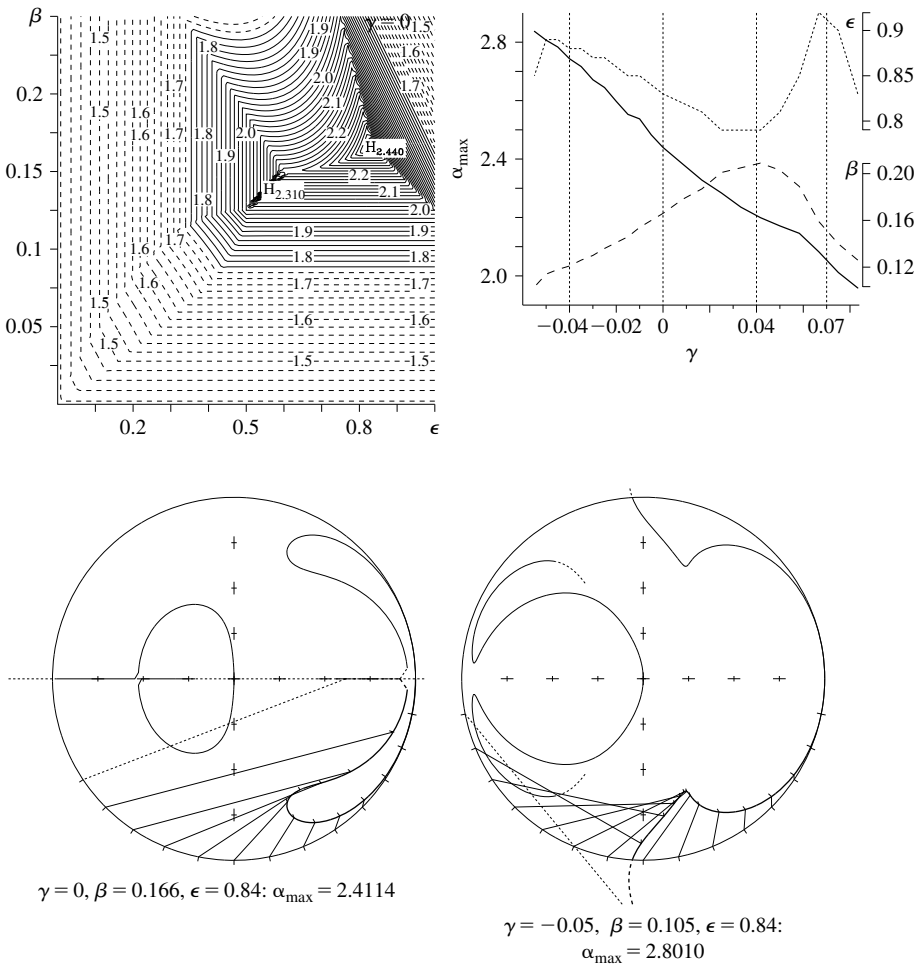


FIG. 2.7 Upper left: map of $\alpha_{\max} = \alpha_{\max}(\epsilon, \beta)$ for $\gamma = 0$. Upper right: $\alpha_{\max}$ and the corresponding ϵ and β as functions of γ. Lower panels show examples of β and ϵ choices that give the maximum stability range for a given γ.

After substitution of $\lambda = e^{\pm i\alpha}$ and Taylor-series expansion in α, a set of constraints arise for achieving progressive orders of accuracy,

$$\text{second-order:} \quad \gamma = \beta - 2\epsilon \qquad \forall\, \beta\,, \epsilon, \tag{2.28}$$

$$\text{third-order:} \quad \gamma = \beta - 2\beta^2 - \frac{1}{6} \quad \text{and} \quad \epsilon = \beta^2 + \frac{1}{12} \qquad \forall\, \beta, \tag{2.29}$$

$$\text{fourth-order:} \quad \gamma, \epsilon \quad \text{as above and} \quad \frac{1}{12} - \frac{\beta}{12} - \beta^3 = 0 \quad \Rightarrow \quad \beta = 0.3737076. \tag{2.30}$$

The second-order accuracy condition can be interpreted as a time-centering balance rule: once the r.h.s. for ζ is placed at $t_n + (1/2 - \delta)\,\Delta t$, the r.h.s. for u is centered at $t_n + (1/2 + \delta)\,\Delta t$ with the same offset $\delta \equiv 1/2 - \beta$ from the midway time $t_n + \Delta t/2$. The classical FB scheme obeys this rule, and it is also respected by the third- and fourth-order constraints. The third-order condition introduces a single-parameter family of schemes with a useful range of $0 < \beta \leq 1/2$ (Fig. 2.8).

The leading-order truncation term has a dissipative character, and it decreases with increasing β. It vanishes at $\beta = 0.3737076$ when the scheme becomes fourth-order, and it changes sign thereafter; this means that the physical modes become asymptotically

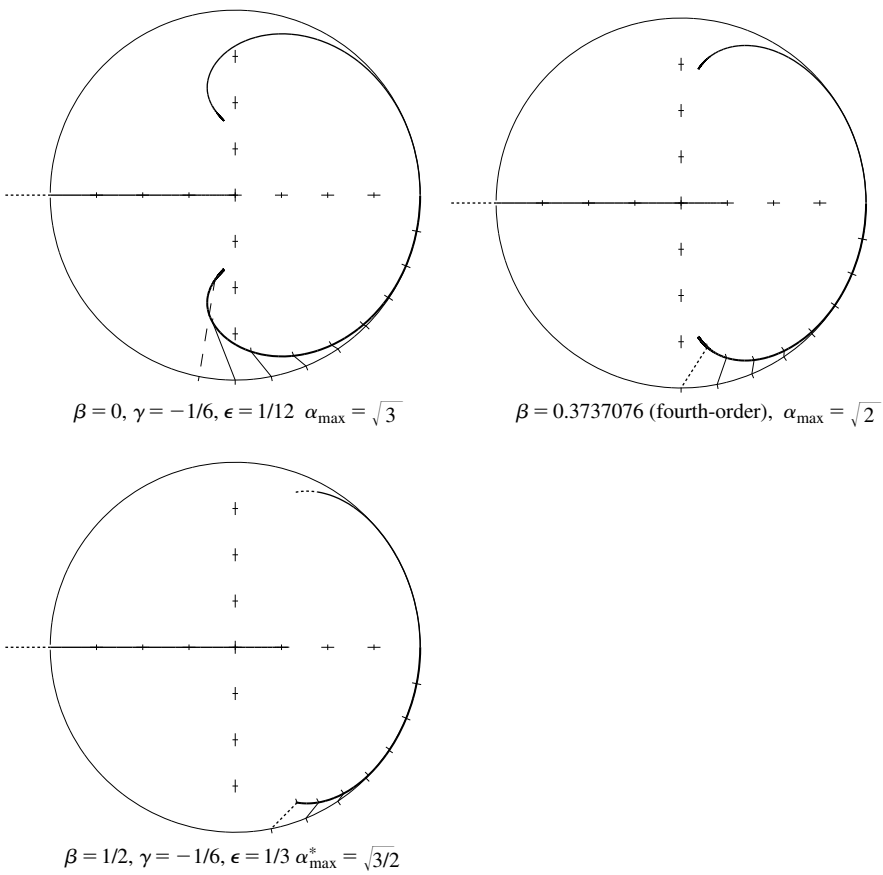


FIG. 2.8 Characteristic roots for the AB2–AM3 algorithm (Eq. (2.26)) with three different choices for β. In all three cases, the remaining parameters γ and ϵ satisfy the third-order accuracy condition. The leading third-order, dissipative truncation term changes sign at $\beta = 0.3737076$, resulting in fourth-order accuracy. The scheme becomes weakly unstable beyond this point (note that physical modes on the left-lower panel are slightly outside the unit circle, reaching $|\lambda| \approx 1.01$ for $\alpha \approx \pi/3$). The stability range decreases with increasing β, and for $\beta =< 1/2$, the AB2-type time-step is unconditionally unstable for an advection equation with centered spatial discretization.

unstable beyond this β value. Leaving the weak asymptotic instability aside, the overall stability range is limited by one of the computational modes that leaves the unit circle at $\lambda = -1$, hence

$$\alpha_{\max} = \sqrt{3} \Big/ \sqrt{1 + \frac{\beta}{2} + 6\beta^3}, \tag{2.31}$$

which decreases with β. Although potentially attractive and simple, this algorithm does not combine naturally with the other hyperbolic terms (advection, Coriolis) because there is no overlap in its β range: the AB2-like time-step is asymptotically unstable for the advection equation when $\beta \leq 1/2$ while the algorithm (Eq. (2.26)) for the wave system needs $\beta \leq 0.3737076$, and in fact, $\beta = 0$ is desirable to achieve the widest possible stability range.

2.4. Generalized FB with an AB3–AM4 step

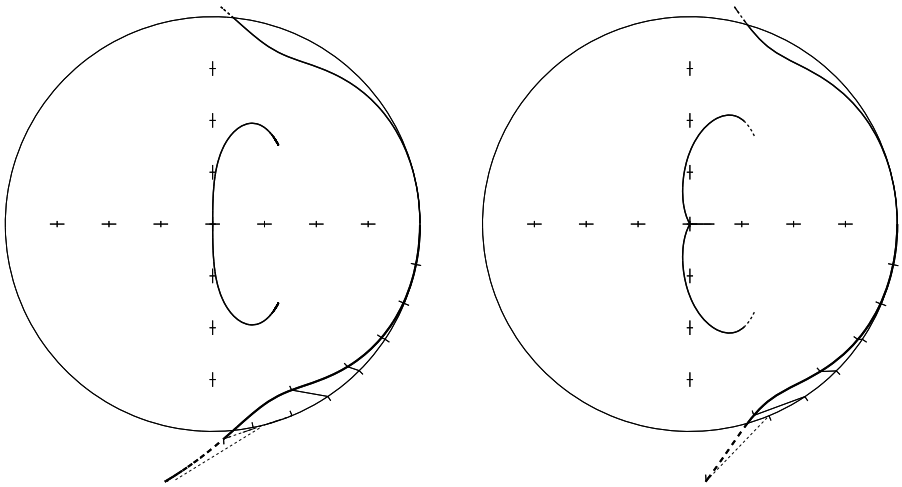
To overcome the limitation of Eq. (2.26), we explore the possibility of using a three-time, AB3-like step for ζ-equation followed by a four-time AM4-like step for u,

$$\begin{aligned} \zeta^{n+1} &= \zeta^n - i\alpha \left[\left(\frac{3}{2} + \beta \right) u^n - \left(\frac{1}{2} + 2\beta \right) u^{n-1} + \beta u^{n-2} \right] \\ u^{n+1} &= u^n - i\alpha \left[\left(\frac{1}{2} + \gamma + 2\epsilon \right) \zeta^{n+1} + \left(\frac{1}{2} - 2\gamma - 3\epsilon \right) \zeta^n + \gamma \zeta^{n-1} + \epsilon \zeta^{n-2} \right], \end{aligned} \tag{2.32}$$

where the r.h.s. for both equations are already time-centered at $t_n + \Delta t/2$ regardless of the values for β, γ, and ϵ (i.e., the r.h.s. time-centering rule (Eq. (2.28)) for the AB2–AM3 scheme is already respected). As a result, second-order accuracy is always guaranteed. Overall, the AB3-type (β-family) time-step for the advection equation is stable as long as $\beta > 1/6$ (otherwise, it is subject to an asymptotic instability of an AB2-type), and it is third-order accurate if $\beta = 5/12$ while a smaller value of $\beta = 0.281105$ yields the largest stability range. This time-step naturally combines with the Coriolis and advection terms (both centered and upstream-biased).

A viable choice would be a straightforward combination of third-order accurate AB3 (hence $\beta = 5/12$) with either a TR or a third-order accurate Adams-Moulton scheme ($\gamma = -1/12$, $\epsilon = 0$), resulting respectively in second- and third-order accuracy with a stability range $\alpha_{\max}$ slightly exceeding unity (Fig. 2.9). This is about 50% more efficient than a synchronous third-order AB3 scheme for both equations ($\alpha_{\max} = 0.71$) but has only half the efficiency of the classical FB scheme. In the remaining part of this section, we will show that the stability range of algorithm (Eq. (2.32)) can be significantly expanded by relaxing the condition $\beta = 5/12$, which is in fact the key to utilizing its full potential.

The analysis of the algorithm (Eq. (2.32)) follows the same path as for AB2–AM3 above. It again leads to a collection of conditions to achieve progressive



AB3–TR: $\beta = 5/12$, $\gamma = \epsilon = 0$
$\alpha_{\max} = 1.1441551$

AB3–AM3: $\beta = 5/12$, $\gamma = -1/12$, $\epsilon = 0$
$\alpha_{\max} = 1.003859$

FIG. 2.9 The algorithm (Eq. (3.1)) with a third-order accurate AB3 (hence $\beta = 5/12$) first step. These "naive" settings result in a stability limit of order of unity. The algorithm on the left was the original version for the main time-step in the ROMS family of codes, and it is still widely used.

orders of accuracy,

$$\text{third-order:} \quad \gamma = \frac{1}{3} - \beta - 3\epsilon \qquad \forall \beta, \epsilon, \tag{2.33}$$

$$\text{fourth-order:} \quad \beta = \frac{1}{12} - \epsilon \quad \text{and} \quad \gamma = \frac{1}{4} - 2\epsilon \qquad \forall \epsilon, \tag{2.34}$$

$$\text{fifth-order:} \quad \frac{7}{120} + \frac{2}{3}\epsilon + \epsilon^2 = 0 \quad \Rightarrow \quad \epsilon = -\frac{1}{3} \pm \frac{\sqrt{190}}{60}; \qquad \beta, \gamma \quad \text{from above.} \tag{2.35}$$

The fifth-order algorithm is asymptotically unstable and has $\alpha_{\max} = 1.0145$ limited by one of the computational modes leaving the unit circle at $\lambda = -1$ (Fig. 2.10, left). Overall, this is not an attractive choice due to both its modest stability range and the asymptotically instability of its physical modes.

Giving up, one order of accuracy allows us to treat ϵ as an adjustable parameter that can be tuned to achieve the maximum stability range. This search yields $\epsilon = 1/12$ and a stability limit of $\alpha_{\max} = \sqrt{3}$ (Fig. 2.10, right). (Here, one can substitute $\beta = 0$ and $\gamma = \epsilon = 1/12$ into the characteristic equation for Eq. (3.1) and verify that $\lambda = -1$ is a double root if $\alpha^2 = 3$.) An obvious drawback for this algorithm is that $\beta = 0$ means the time-stepping for the ζ-equation is only AB2, which is asymptotically unstable for advection and Coriolis force.

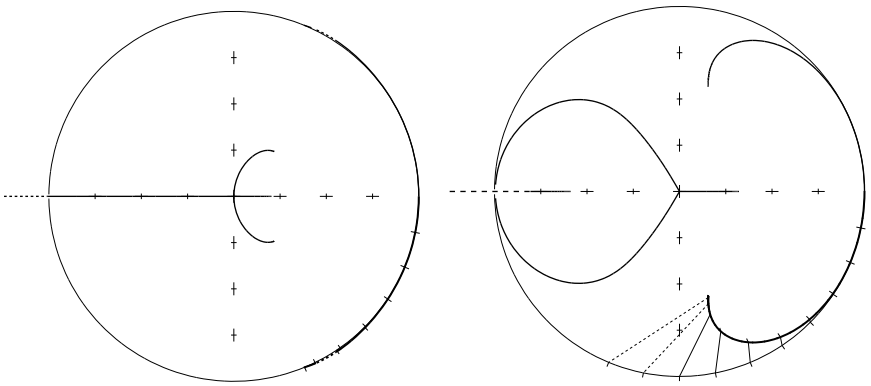


FIG. 2.10 Characteristic roots for the AB3–AM4 algorithm (Eq. (3.1)) with β, γ, and ϵ set to achieve either fifth-order accuracy (left panel) or the maximum possible stability limit while maintaining fourth-order accuracy (right panel). Note that at the optimum ϵ, two computational modes meet at $\lambda = -1$, after which one of them continues out of the unit circle along the negative real axis. If $\epsilon > 1/12$, the meeting occurs outside the circle (i.e., the computational modes leave the circle before they meet), while a smaller ϵ moves the meeting point inside, resulting in an earlier escape of one of the modes along the negative imaginary axis. Either way, $\alpha_{\max}$ ends up being smaller than $\sqrt{3}$ if $\epsilon \neq 1/12$.

The third-order, two-parameter (β,ϵ) family can reach up to $\alpha_{\max} = 1.939$ (Fig. 2.11, upper-left) that is now very close to that of the classical FB scheme. Its β value lies within the desirable range of $1/6 < \beta < 5/12$ (i.e., the range of stable choices for the advection equation with spatially centered schemes, as well as for Coriolis force). The only undesirable property of this algorithm is its nearly purely dispersive truncation error, resulting in weak damping of frequency components that are not accurately represented. This issue can be addressed by a slight bias of β away from the maximum stability (Fig. 2.11, upper-right), which leads to an insignificant decrease in $\alpha_{\max}$. Since this is achieved with a smaller value of β, the stability range for advection and Coriolis force is also decreased.

From a practical point of view, it is attractive to chose $\beta = 0.281105$, corresponding to the largest possible stability range for advection and Coriolis force within the β-family for AB3-like schemes, accompanied by $\gamma = 0.088$ and $\epsilon = 0.013$ that yield a sufficiently large stability range for waves (Fig. 2.11, lower-left) and a dissipation-dominant truncation error. This compromise gives second-order accuracy, and our experience is that it is robust even applied to the full nonlinear system (Section 4). Thus, it is the method of choice for the barotropic mode.

To summarize Eq. (3.1), we note that the crucial step to obtain an algorithm with a stability limit comparable to that of FB ($\alpha_{\max} = 2$) is to reduce the curvature parameter of the AB3-like step for ζ by setting $\beta < 5/12$. This brings a tension between the need to keep β relatively large to avoid an asymptotic instability for centered advection and Coriolis force, and the desire to expand the maximum stability range for the wave system that favors bringing β closer to zero. A simultaneous optimization of both stability ranges yields a useful range of $0.21 < \beta < 0.281105$. Remarkably, in this range, the AM4-like

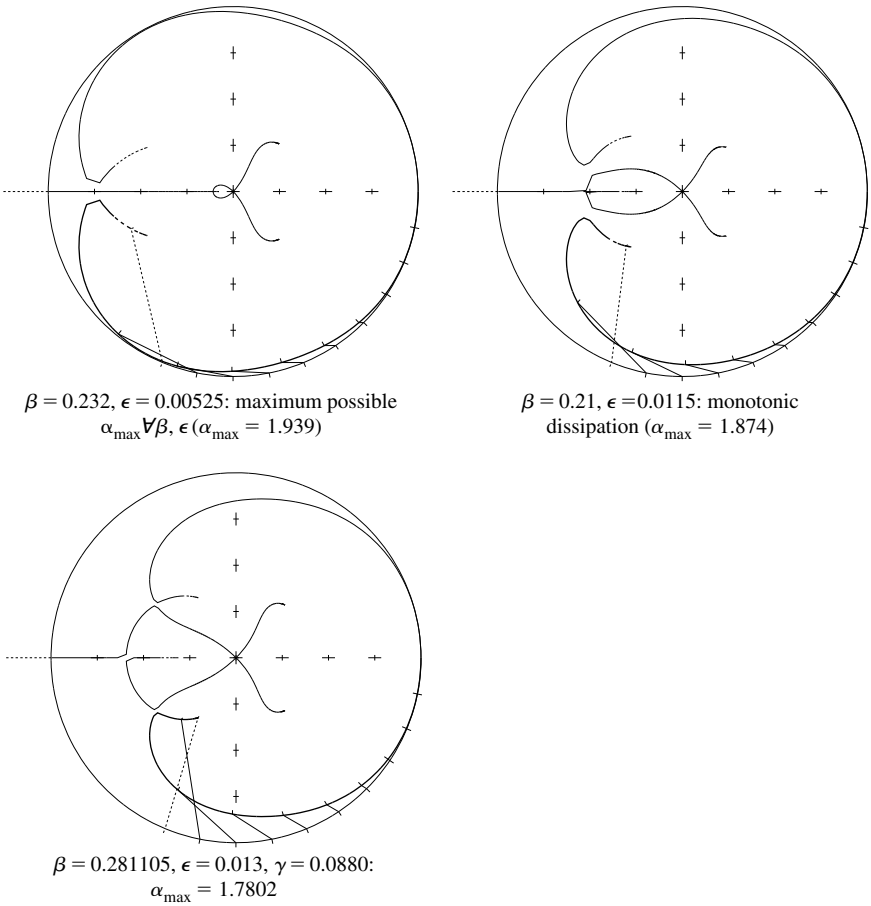


FIG. 2.11 Upper left: AB3–AM4 scheme (Eq. (3.1)) with the maximum possible stability range and third-order accuracy for $\lambda = \lambda(\alpha)$. The physical modes touch the unit circle at $\alpha \approx \pm 2\pi/3$. Larger values of β result in the physical modes going outside the circle near these α values (as in Fig. 2.9). Smaller β values cause an earlier escape of one of the computational modes along the negative real axis. Upper right: a third-order scheme with parameters slightly deviating from optimum stability to ensure that numerical dissipation increases monotonically with α. Lower left: a multipurpose compromise with β set to maximize the stability range for the advection equation, while ϵ and γ are set to yield a good stability range for the wave system while maintaining monotonically increasing dissipation.

coefficients in the second equation in (3.1) end up quite different from that in the classical AM4 weights, and one can verify that terms with ζ^{n+1}, ζ^n, ζ^{n-1}, and ζ^{n-2} in the r.h.s. all have positive coefficients for all of the cases shown in Fig. 2.11.

2.5. Summary for time-stepping algorithms

We have analyzed four different classes of algorithms for a wave system that use a degree of FB feedback to achieve better accuracy for modeling the phase speed for wave

motions and/or extend the stability limit $\alpha_{\max}$ over their previously known prototypes. Although in most oceanic applications external and internal waves are not of major interest, the wave properties of the discretized system are always a primary concern from a numerical viewpoint because they are likely to impose the most restrictive limit on the time-step size in flows with small Froude number. Unlike for a simple advection equation, where one can construct a stable algorithm using a simple forward-in-time stepping and upstream-biased, semi-Lagrangian discretization in space, the stability of a wave system cannot rely entirely on a specially designed spatial operator.[11] Thus, stability of an algorithm with respect to wave motions is always in consideration, even though the final selection of the time-stepping scheme among the ones described above depends on the choice of algorithms for spatial discretization that, in turn, depends on the physical application. The algorithms in Section 2.2 are fully compatible with centered advection for the tracer and momentum equations, and they naturally incorporate the treatment of Coriolis force using a synchronous LF–AM3 predictor-corrector step. The same remark applies to the algorithms in Section 2.4; however, compatibility with centered advection imposes some restriction on the choice of coefficients (formally $\beta > 1/6$ in Eq. (3.1) but in practice we use a greater value of β), which lead to a compromise in the stability limit $\alpha_{\max}$ for wave motions. In contrast, algorithms in Section 2.3 are incompatible with centered advection because of weak instability of second-order Adams-Bashforth step (one needs at least small viscosity to mitigate this, or introduce "a forward bias" into AB2 extrapolation coefficients, Campin, Adcroft, Hill and Marshall [2004], which is in essence setting $\beta > 1/2$ in Eq. (2.26) in it would be a single advection equation). Similarly, the RK2-type algorithms in Section 2.1 always require some degree of upstream bias in the advection scheme for stability because RK2 is asymptotically unstable in combination with centered advection. Monotonicity-preserving advection schemes typically require two-level time-stepping and have built-in compatibility with forward-in-time stepping but are incompatible with algorithms that have negative coefficients in their temporal interpolation. This makes RK2 preferable if monotonicity is desired (e.g., in modeling estuaries characterized by sharp fronts in temperature and salinity). The time-stepping algorithms described here are just linearizations of more general algorithms for the full nonlinear system (Section 4) which involve other considerations in their design (e.g., conservation properties), resulting in additional selection criteria.

3. Vertical mode-splitting

Although vertical mode-splitting has been used in oceanic modeling since the very beginning (Berntsen, Kowalik, Sælid and Sørli [1981], Bleck and Smith [1990], Blumberg and Mellor [1987], Bryan and Cox [1969], Killworth, Stainforth,

[11] In principle, one can separate signals propagating in different directions and construct an approximate Riemann solver (Roe [1981]), which essentially relies on upstream-biased algorithms for stability. However, this is not a viable option for oceanic modeling because of complexity (due to implied normal mode decomposition in vertical direction in the case of 3D mode (Shulman, Lewis and Mayer [1999]), computational cost, large numerical dissipation, and the implied directional splitting that is not desirable.

Webb and Paterson [1991]), a mature theoretical understanding of its stability and accuracy is relatively recent (Hallberg [1997], Higdon and Bennett [1996], Higdon and de Szoeke [1997], Higdon [1999, 2002], Nadiga, Hecht, Margolin and Smolarkiewicz [1997], Shchepetkin and McWilliams [2005], Skamarock and Klemp [1992]). The major issues to be resolved in this approach are (i) an inaccurate separation of fast- and slow-time (i.e., barotropic and baroclinic) components in the PGF that may cause "leakage" of fast-time signals into the slow evolution and numerical instability even for linearized systems (Higdon and Bennett [1996]); (ii) the time delay in calculating the vertically integrated r.h.s. terms of the slow component can, in effect, be a forward-in-time treatment of the barotropic mode, with associated loss of accuracy and numerical instability; (iii) an aliasing of fast barotropic signals due to sub-sampling in the baroclinic time-stepping; (iv) a loss of conservation and constancy preservation properties for tracers in both split-explicit (Griffies, Pacanowski, Schmidt and Balaji [2001]) and implicit free-surface models (Adcroft and Cadmin [2004]); (v) the compressibility effect in EOS complicates the definition of the barotropic PGF with the Boussinesq approximation; and (vi) the bottom stress must be known before the barotropic mode starts at every baroclinic time-step.

3.1. Tracer conservation and constancy preservation

In an incompressible fluid, the equation for material tracers q can be written in two forms, respectively, emphasizing the Lagrangian-parcel and volume-integral conservation properties:

$$\text{advection form:} \quad \frac{\partial q}{\partial t} + \big(\mathbf{u}\cdot\nabla\big)q = 0, \tag{3.1}$$

$$\text{conservation form:} \quad \frac{\partial q}{\partial t} + \nabla\cdot(\mathbf{u}q) = 0. \tag{3.2}$$

The continuity (nondivergence) equation $\nabla\cdot\mathbf{u} = 0$ plays the role of a compatibility condition making these two forms equivalent. If q is initially uniform in space, parcel conservation implies that it remains so: the property of constancy preservation.

Oceanic models always use the conservation form as the prototype for discrete equations,

$$\begin{aligned}\Delta\mathscr{V}^{n+1}_{i,j,k}\,q^{n+1}_{i,j,k} = \Delta\mathscr{V}^{n}_{i,j,k}\,q^{n}_{i,j,k} - \Delta t\Big[&\widetilde{q}_{i+\frac12,j,k}U_{i+\frac12,j,k} - \widetilde{q}_{i-\frac12,j,k}U_{i-\frac12,j,k} + \widetilde{q}_{i,j+\frac12,k}V_{i,j+\frac12,k}\\ &- \widetilde{q}_{i,j-\frac12,k}V_{i,j-\frac12,k} + \widetilde{q}_{i,j,k+\frac12}W_{i,j,k+\frac12} - \widetilde{q}_{i,j,k-\frac12}W_{i,j,k-\frac12}\Big],\end{aligned} \tag{3.3}$$

where discrete concentration values $q_{i,j,k}$ are understood as averages over the local control-volumes $\Delta\mathscr{V}_{i,j,k}$; i.e., $q_{i,j,k} = \dfrac{1}{\Delta\mathscr{V}_{i,j,k}}\displaystyle\int\limits_{\Delta\mathscr{V}^n_{i,j,k}} q(x,y,z)\,\mathrm{d}^3\mathscr{V}$. The tilde operator $\widetilde{q}_{i+\frac12,j,k}$ denotes an appropriate translation algorithm from grid-box averages to interface values, either as a simple spatial interpolation or as one involving both space and time in

a semi-Lagrangian approach. $U_{i+\frac{1}{2},j,k}$, $V_{i,j+\frac{1}{2},k}$, and $W_{i,j,k+\frac{1}{2}}$ are volume fluxes across grid-box interfaces. The discretized continuity equation,

$$\Delta\mathscr{V}_{i,j,k}^{n+1} = \Delta\mathscr{V}_{i,j,k}^{n} - \Delta t \cdot \left[U_{i+\frac{1}{2},j,k} - U_{i-\frac{1}{2},j,k} + V_{i,j+\frac{1}{2},k} - V_{i,j-\frac{1}{2},k} + W_{i,j,k+\frac{1}{2}} - W_{i,j,k-\frac{1}{2}}\right], \tag{3.4}$$

is formally consistent with Eq. (3.3) for $q_{i,j,k} \equiv 1$; therefore, as long as Eq. (3.4) holds, this time-stepping scheme has both conservation and constancy preservation.

The control volumes $\Delta\mathscr{V}_{i,j,k} = H_{i,j,k}\Delta\mathscr{A}_{i,j}$ in Eqs. (3.3) and (3.4) are time-dependent because grid-box heights $H_{i,j,k}$ depend on $\zeta(x, y, t)$,

$$H_{i,j,k} = z_{i,j,k+\frac{1}{2}} - z_{i,j,k-\frac{1}{2}} \text{ where } \begin{cases} z_{i,j,k+\frac{1}{2}} = z^{(0)}_{i,j,k+\frac{1}{2}} + \zeta_{i,j}\left(1 + \dfrac{z^{(0)}_{i,j,k+\frac{1}{2}}}{h_{i,j}}\right) \\ z^{(0)}_{i,j,k+\frac{1}{2}} \equiv z^{(0)}\left(\xi_{i,j}, \eta_{i,j}, s_{k+\frac{1}{2}}\right),\ k = 0, 1, \ldots, N. \end{cases} \tag{3.5}$$

The $z^{(0)}$ comprises a set of unperturbed (i.e., corresponding to $\zeta \equiv 0$) isosurfaces of a terrain-following vertical coordinate, $s_{k+\frac{1}{2}} \in [-1, 0]$. The lowest surface, $z_{i,j,\frac{1}{2}} \equiv z^{(0)}_{i,j,\frac{1}{2}} \equiv -h_{i,j}$, follows the bottom topography. Since $z^{(0)}_{i,j,N+\frac{1}{2}} \equiv 0$, the highest surface $z_{i,j,N+\frac{1}{2}} \equiv \zeta_{i,j}$ follows the oceanic top. Otherwise, the vertical coordinate transformation is a general one. In Eq. (3.5), the grid-box heights are proportionally stretched relative to their unperturbed values, $H^{(0)}_{i,j,k}$; i.e.,

$$H_{i,j,k} = H^{(0)}_{i,j,k} \cdot \left(1 + \frac{\zeta_{i,j}}{h_{i,j}}\right). \tag{3.6}$$

The loss of constancy preservation in Eq. (3.3) can occur if $\Delta\mathscr{V}_{i,j,k}^{n+1}$ does not come from Eq. (3.4), but rather is computed with a barotropic mode that uses a different time-step and time-stepping algorithm and, furthermore, is averaged in fast-time, replacing $\zeta \to \langle\zeta\rangle^{n+1} = \sum_{m=1}^{M^*} a_m \zeta^m$, to prevent aliasing of the barotropic frequencies unresolved by the baroclinic time-stepping. A vertical summation of Eq. (3.4) yields

$$\zeta_{i,j}^{n+1} = \zeta_{i,j}^{n} - \frac{\Delta t}{\Delta\mathscr{A}_{i,j}} \cdot \sum_{k=1}^{N}\left[U_{i+\frac{1}{2},j,k} - U_{i-\frac{1}{2},j,k} + V_{i,j+\frac{1}{2},k} - V_{i,j-\frac{1}{2},k}\right]. \tag{3.7}$$

This is not necessarily consistent with the fast-time-averaged free surface computed by the barotropic mode, implying that

$$\langle\zeta\rangle^{n+1} \neq \langle\zeta\rangle^{n} - \Delta t \cdot \mathrm{div}\langle\overline{\mathbf{U}}\rangle, \tag{3.8}$$

where indices n and $n+1$ correspond to the slow (baroclinic) time-step, and the overbar in $\overline{\mathbf{U}}$ means a vertically integrated volume flux.

Conversely, Eq. (3.4) is used for the computation of vertical velocity: start with $W_{i,j,\frac{1}{2}} = 0$ at the bottom and recursively proceed with

$$W_{i,j,k+\frac{1}{2}} = -\sum_{k'=1}^{k} \left\{ \frac{\Delta\mathscr{V}^{n+1}_{i,j,k'} - \Delta\mathscr{V}^{n}_{i,j,k'}}{\Delta t} + U_{i+\frac{1}{2},j,k'} - U_{i-\frac{1}{2},j,k'} + V_{i,j+\frac{1}{2},k'} - V_{i,j-\frac{1}{2},k'} \right\} \quad \text{for all} \quad k = 1, 2, \ldots, N. \tag{3.9}$$

This essentially defines $W_{i,j,k+\frac{1}{2}}$ as the finite-volume, finite-time-interval volume flux across the moving interface between vertically adjacent grid boxes, $\Delta\mathscr{V}_{i,j,k}$ and $\Delta\mathscr{V}_{i,j,k+1}$. This procedure does not automatically guarantee that the surface kinematic boundary condition,

$$W_{i,j,N+\frac{1}{2}} = 0, \tag{3.10}$$

is satisfied if $\Delta\mathscr{V}^{n+1}_{i,j,k}$ comes from the barotropic mode with a different time-stepping.

To ensure that slow-time continuity equation (Eq. (3.4)) is consistent with the barotropic mode, we must impose a constraint on the vertical integrals of the volume fluxes, $U_{i+\frac{1}{2},j,k}$ and $V_{i,j+\frac{1}{2},k}$,

$$\sum_{k=1}^{N} U_{i+\frac{1}{2},j,k} = \left\langle\!\left\langle \overline{U} \right\rangle\!\right\rangle^{n+\frac{1}{2}}_{i+\frac{1}{2},j} \qquad \text{and} \qquad \sum_{k=1}^{N} V_{i,j+\frac{1}{2},k} = \left\langle\!\left\langle \overline{V} \right\rangle\!\right\rangle^{n+\frac{1}{2}}_{i,j+\frac{1}{2}}, \tag{3.11}$$

so that

$$\langle\zeta\rangle^{n+1}_{i,j} = \langle\zeta\rangle^{n}_{i,j} - \frac{\Delta t}{\Delta\mathscr{A}_{i,j}} \left[\left\langle\!\left\langle \overline{U} \right\rangle\!\right\rangle^{n+\frac{1}{2}}_{i+\frac{1}{2},j} - \left\langle\!\left\langle \overline{U} \right\rangle\!\right\rangle^{n+\frac{1}{2}}_{i-\frac{1}{2},j} + \left\langle\!\left\langle \overline{V} \right\rangle\!\right\rangle^{n+\frac{1}{2}}_{i,j+\frac{1}{2}} - \left\langle\!\left\langle \overline{V} \right\rangle\!\right\rangle^{n+\frac{1}{2}}_{i,j-\frac{1}{2}} \right] \tag{3.12}$$

is consistent with the change in $\langle\zeta\rangle$ between two consecutive baroclinic time-steps. To define the second averaging operator $\langle\!\langle ... \rangle\!\rangle$, we note that a summation of consecutive barotropic time-steps yields

$$\zeta^{m+1} = \zeta^{m} - \frac{\Delta t}{M} \cdot \mathrm{div}\overline{\mathbf{U}}^{m+\frac{1}{2}}, \quad \text{hence} \quad \zeta^{m} = \zeta^{0} - \frac{\Delta t}{M} \sum_{m'=0}^{m-1} \mathrm{div}\overline{\mathbf{U}}^{m'+\frac{1}{2}}, \tag{3.13}$$

where m is the fast-time index and M is the integer mode-splitting ratio (i.e., ratio of commensurate baroclinic and barotropic time-step sizes). The $m = 0$ starting field ζ^0 corresponds to the baroclinic step n, and the barotropic mode restarts at the end of every

baroclinic time-step, $\langle \zeta, \overline{U}, \overline{V} \rangle^{n+1} \to \zeta^0, \overline{U}^0, \overline{V}^0$. After applying fast-time averaging $\langle ... \rangle$ to both sides of Eq. (3.13),

$$\langle \zeta \rangle^{n+1} \equiv \sum_{m=1}^{M^*} a_m \zeta^m = \zeta^0 - \frac{\Delta t}{M} \cdot \mathrm{div} \sum_{m=1}^{M^*} \left[a_m \sum_{m'=1}^{m} \overline{\mathbf{U}}^{m'-\frac{1}{2}} \right]. \tag{3.14}$$

This translates into

$$\langle \zeta \rangle^{n+1} = \langle \zeta \rangle^n - \Delta t \cdot \mathrm{div} \sum_{m'=1}^{M^*} b_{m'} \overline{\mathbf{U}}^{m'-\frac{1}{2}}, \quad \text{where} \quad b_{m'} = \frac{1}{M} \sum_{m=m'}^{M^*} a_m \tag{3.15}$$

$\forall\, m' = 1, \ldots, M^*$. The coefficients $\{a_m, m = 1, \ldots, M^*\}$ are the primary averaging weights (Fig. 3.1) that satisfy normalization and centroid conditions,

$$\sum_{m=1}^{M^*} a_m \equiv 1 \quad \text{and} \quad \sum_{m=1}^{M^*} \frac{m}{M} a_m \equiv 1, \tag{3.16}$$

but they are otherwise arbitrary thus far. $M^* \geq M$ is the fast-time index of the last non-zero a_m. We define

$$\left\langle\!\left\langle \overline{\mathbf{U}} \right\rangle\!\right\rangle^{n+\frac{1}{2}} \equiv \sum_{m=1}^{M^*} b_m \overline{\mathbf{U}}^{m-\frac{1}{2}}. \tag{3.17}$$

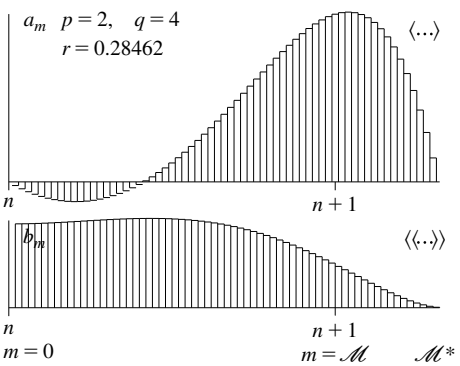


FIG. 3.1 Relationship between the primary, $\{a_m\}$, and secondary, $\{b_m\}$, fast-time-averaging weights. By definition, $\langle \zeta \rangle^{n+1} \equiv \sum_{m=1}^{M^*} a_m \zeta^m$ and $\left\langle\!\left\langle \overline{\mathbf{U}} \right\rangle\!\right\rangle^{n+\frac{1}{2}} \equiv \sum_{m=1}^{M^*} b_m \overline{\mathbf{U}}^m$. In order to satisfy normalization and centroid conditions (Eq. (3.16)), the integration of the barotropic mode must go beyond the $n+1$th baroclinic step, hence $M^* > M$. In this example, the a_m are negative at the beginning of their sequence (i.e., they have a S-shape).

The value of this negative lobe and the meaning of parameters p, q, and r are explained in Section 3.3.

Using this in the integral constraint, Eq. (3.11) with Eq. (3.15) guarantees that Eq. (3.4) holds exactly between baroclinic steps n and $n+1$ and, therefore, guarantees both conservation and constancy-preservation properties in Eq. (3.3). In practice, after completion of the barotropic time-stepping at every baroclinic time-step, five fields ($\langle\zeta\rangle^{n+1}$, $\langle\overline{U}\rangle^{n+1}$, $\langle\overline{V}\rangle^{n+1}$, $\left\langle\!\left\langle \overline{U} \right\rangle\!\right\rangle^{n+\frac{1}{2}}$, and $\left\langle\!\left\langle \overline{V} \right\rangle\!\right\rangle^{n+\frac{1}{2}}$) must be available for the baroclinic integration since $\langle ... \rangle$ fields cannot be expressed directly in terms of $\langle ... \rangle$ fields.

3.2. *Mode-splitting error in the PGF*

Vertical mode-splitting separates the vertically integrated, hydrostatic, horizontal PGF,

$$\mathscr{F} \equiv \mathscr{F}\left[\nabla_x\zeta,\ \zeta,\ \nabla_x\rho(z),\ \rho(z)\right] = -\frac{1}{\rho_0}\int_{-h}^{\zeta} \nabla_x P \,\mathrm{d}z = -\frac{g}{\rho_0}\int_{-h}^{\zeta}\left[\int_{z}^{\zeta} \nabla_x\rho\left(z'\right)\,\mathrm{d}z'\right]\mathrm{d}z, \tag{3.18}$$

into a "fast" term, $-gD\nabla_x\zeta$, and the remaining "slow" $\Big\{....\Big\}$ terms (these are also known as "coupling" or baroclinic-to-barotropic forcing terms),

$$\frac{\partial\left(D\overline{u}\right)}{\partial t} + \cdots = -gD\nabla_x\zeta + \Big\{gD\nabla_x\zeta + \mathscr{F}\Big\}. \tag{3.19}$$

The fast terms are recomputed at every barotropic step, while the slow terms are held constant since they change only once per baroclinic time-step. $D = h + \zeta$ is total depth of a vertical column. If the functional $\mathscr{F}$ contains nonlinear combinations of ζ and ρ (i.e., $\partial^2\mathscr{F}/\partial\zeta\,\partial\rho \neq 0$), freezing the slow terms can cause a mode-splitting error,

$$-gD\nabla_x\zeta' + \Big\{gD\nabla_x\zeta + \mathscr{F}\left[\nabla_x\zeta,\ \zeta,\ \nabla_x\rho(z),\ \rho(z)\right]\Big\} \neq \mathscr{F}\left[\nabla_x\zeta',\ \zeta',\ \nabla_x\rho(z),\ \rho(z)\right]; \tag{3.20}$$

i.e., at the end of barotropic time-stepping when $\zeta \to \zeta'$, the PGF seen by the barotropic mode no longer matches the vertical integral of the total PGF from the same ρ and the new ζ. Consequently, at the beginning of the new time-step when the full PGF is recomputed, its vertical integral is no longer in equilibrium with the state of the barotropic mode PGF even in the case when there is no change of the σ-level ρ values between consecutive baroclinic steps. The mismatch between the two contaminates the forcing terms computed and the new time-step and subsequently affects the state of barotropic mode one step later, thereby closing the feedback loop.

In isopycnic coordinates, Hallberg [1997], Higdon and Bennett [1996], Higdon and de Szoeke [1997] found an instability of the linearized mode-split system with non-dissipative time-stepping schemes (FB, LF). Their diagnosis and proposed remedies were that (i) mode-splitting can cause artificial mode-coupling; (ii) for some time-stepping schemes, the mode-coupling may cause a phase lag that induces a numerical instability similar to that of a forward time-step for a hyperbolic system; (iii) a perturbation analysis of weakly coupled linear system shows that the instability is a resonance of an

aliased barotropic mode subsampled at the baroclinic steps; (iv) the remedy is to redefine barotropic mode PGF to make it be equal to the vertical integral of the 3D PGF; and (v) a dissipative time-stepping scheme that filters the barotropic mode to prevent aliasing or a dissipative predictor-corrector scheme (HALLBERG [1997]) can be a useful way to achieve stability.

The common justification for Eq. (3.19) is $\zeta \ll D$ and $\rho' \equiv \rho - \rho_0 \ll \rho_0$, hence the magnitude of the mismatch in Eq. (3.20) is $\mathcal{O}\big(\max\big\{(\rho'/\rho_0)\nabla_x\zeta\ ,\ \zeta\nabla_x\rho'/\rho_0\big\}\big)$ relative to $\mathcal{O}\big(\nabla_x\zeta\big)$. Among other restrictions, this implies that ρ_0 must be chosen sufficiently close to the actual density ρ to avoid a "leakage" of barotropic signals into the baroclinic mode (HIGDON [2002]). Suppose that both modes are time-stepped within but close to their CFL limits of stability taken individually. This implies a choice of M in Eq. (3.13) as the ratio of the barotropic and first-baroclinic gravity-wave phase speeds adjusted by the ratio of the stability limits of their respective time-stepping algorithms. The coupled system may still be unstable if $M > \mathcal{O}\big(\rho_0/\left|\rho - \rho_0\right|\big)$. This is because in a Boussinesq model using splitting (Eq. (3.19)), the barotropic pressure gradient term arising from free surface gradient $\nabla_x\zeta$ creates an acceleration equal to $-g\nabla_x\zeta$ independently of the choice of ρ_0. On the other hand, the net vertically integrated PGF computed by full (baroclinic + barotropic) 3D scheme from a given density field and given state of free surface has slightly different sensitivity to $\nabla_x\zeta$; it creates acceleration more similar to $-g\nabla_x\left[\big(1+\rho^{*\prime}/\rho_0\big)\,\zeta\right]$ where $\rho^{*\prime}$ depends on the deviation of local density from ρ_0 in a manner quantified later in this section. This leads to the fact that phase speed of surface gravity waves as seen by the 3D part of the code is different from that seen by the barotropic mode. To avoid numerical instability, the difference in phase increment per one baroclinic time-step Δt between the two must be smaller that allowed by CFL criterion for the time-stepping scheme for the baroclinic mode. Since the density variation due to baroclinic effects can be estimated as large as 3% (i.e., comparable, and in some situations larger that the ratio phase speeds of barotropic and the first baroclinic modes) this potentially may force to chose a smaller Δt than required for stability of the baroclinic mode taken alone.

Furthermore, even if the mismatch in Eq. (3.20) is small in most cases, the primary concern here is that it still may cause a numerical instability even if ρ variations are small and ρ_0 is chosen so that the preceding M-criterion is respected. This is due to phase delays in computing the mismatch term associated with the organization of the coupled time-stepping algorithm. Another remedy to mitigate the consequences of this type of error is the use of a dissipative time filter for the barotropic mode (Section 3.3): however, this unavoidably degrades the numerical accuracy. Either way, it is always desirable to remove or minimize the mismatch.

Equation (3.20) suggests a general guideline for eliminating the mode-splitting PGF error by replacing $-gD\nabla_x\zeta$ in Eq. (3.19) with the variational derivative of $\mathscr{F} = \mathscr{F}\left[\nabla_x\zeta, \zeta, \ldots\right]$,

$$\delta\mathscr{F} = \frac{\partial\mathscr{F}}{\partial\left(\nabla_x\zeta\right)}\delta(\nabla_x\zeta) + \frac{\partial\mathscr{F}}{\partial\zeta}\delta\zeta. \tag{3.21}$$

ζ and $\nabla_x\zeta$ are treated as independent variables for the functional partial differentiation. In the discretized version, this corresponds to having ζ_i and ζ_{i+1} as independent degrees

of freedom that are alternatively expressible as their difference $\zeta_{i+1} - \zeta_i$ and average $(\zeta_{i+1} + \zeta_i)/2$. Substitution of Eq. (3.21) into Eq. (3.20) makes it into a Taylor-series expansion,

$$\mathscr{F}\left[\nabla_x\zeta,\ \zeta, \ldots\right] + \frac{\partial \mathscr{F}}{\partial\left(\nabla_x\zeta\right)}\nabla_x\left(\zeta' - \zeta\right) + \frac{\partial \mathscr{F}}{\partial\zeta}\left(\zeta' - \zeta\right) \approx \mathscr{F}\left[\nabla_x\zeta',\ \zeta', \ldots\right], \tag{3.22}$$

resulting in a cancellation of the dominant part of the mode-splitting error: recall that the mismatch between l.h.s. and r.h.s. of Eq. (3.22) can be estimated as $\mathcal{O}((\nabla_x(\zeta' - \zeta))^2) + \mathcal{O}((\zeta' - \zeta)^2)$.

Note that the net horizontal force applied to fluid element in Fig. 3.2 can be calculated as

$$\begin{aligned}\mathscr{F}_{i+\frac{1}{2}} &= \int_{-h_i}^{\zeta_i} P(x_i, z)\,\mathrm{d}z - \int_{-h_{i+1}}^{\zeta_{i+1}} P(x_{i+1}, z)\,\mathrm{d}z + \int_{x_i}^{x_{i+1}} P\big(x, -h(x)\big)\frac{\partial h(x)}{\partial x}\mathrm{d}x \\ &= \mathcal{I}_i - \mathcal{I}_{i+1} + \mathcal{I}_{i+\frac{1}{2}},\end{aligned} \tag{3.23}$$

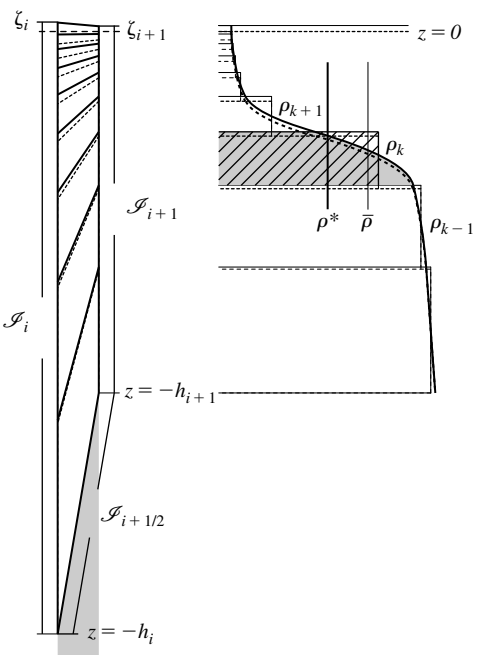


FIG. 3.2 Left: A segment of the vertical grid used in derivation of total vertically integrated PGF (Eq. (3.23)). Dashed lines correspond to the unperturbed ($\zeta = 0$) vertical coordinate and solid lines to the coordinate perturbed according to Eq. (3.6). Right: Computation of two-way vertically averaged densities (Eq. (3.25)) for a stratified water column. The ρ_k is interpreted as control-volume averages, hence the area of hatched rectangle is equal to the shaded area left from the continuous profile. Note that for a stably stratified profile, ρ^* is systematically smaller than $\overline{\rho}$ as illustrated here.

where $P_i(z)$ is hydrostatic pressure calculated separately in each vertical column,

$$P_i(z) = g \int_{z'_i}^{\zeta_i} \rho_i(z')\, \mathrm{d}z'. \tag{3.24}$$

By introducing

$$\overline{\rho}(x) = \frac{1}{D_i} \int_{-h_i}^{\zeta_i} \rho_i(z')\mathrm{d}z' \quad \text{and} \quad \rho_i^* = \frac{1}{\frac{1}{2}D_i^2} \int_{-h_i}^{\zeta_i} \left\{ \int_{z_i}^{\zeta_i} \rho_i(z')\mathrm{d}z' \right\} \mathrm{d}z, \tag{3.25}$$

where $D_i = \zeta_i + h_i$, the net force (Eq. (3.23)) can be expressed as

$$\mathscr{F}_{i+\frac{1}{2}} = g \left[\frac{\rho_i^* D_i^2}{2} - \frac{\rho_{i+1}^* D_{i+1}^2}{2} + \int_{x_i}^{x_{i+1}} \overline{\rho} D \frac{\partial h}{\partial x} \mathrm{d}x \right]. \tag{3.26}$$

This corresponds to the continuous form,

$$\frac{\partial}{\partial t}(D\overline{u}) + \ldots = -\frac{g}{\rho_0}\left[\nabla_x\left(\frac{\rho^* D^2}{2}\right) - \overline{\rho} D \nabla_x h\right] = -\frac{g}{\rho_0}\Bigg[\underbrace{\frac{h^2}{2}\nabla_x \rho^* + \left(\rho^* - \overline{\rho}\right)\nabla_x \frac{h^2}{2}}_{\mathscr{F}^{(0)}\ (\zeta=0 \text{ part})}$$

$$+ \underbrace{h\nabla_x\left(\rho^*\zeta\right) + \nabla_x\left(\frac{\rho^*\zeta^2}{2}\right) + \left(\rho^* - \overline{\rho}\right)\zeta\nabla_x h}_{\mathscr{F}'\ (\text{perturbation due to } \zeta\neq 0)}\Bigg], \tag{3.27}$$

where we have separated $\mathscr{F}^{(0)}$ which is independent of ζ and the remainder, ($\mathscr{F}'$). Since the mode-coupling algorithm already performs a vertical integration of the momentum r.h.s. terms, including the full PGF, $\mathscr{F}^{(0)}$ is not further required. However, $\mathscr{F}'$ satisfies Eq. (3.21) and is therefore a valid replacement for $-gD\nabla_x\zeta$ in Eq. (3.19) (as expected, one can easily verify that $\mathscr{F}'$ reverts back to $-gD\nabla_x\zeta$ if ρ is uniform, $\rho^* = \overline{\rho} = \rho_0$).

The accuracy of mode-splitting using the decomposition of $\mathscr{F} = \mathscr{F}^{(0)} + \mathscr{F}'$ fundamentally comes from the fact that changes in ζ from one time-step to the next do not modify the grid-box values of density $\rho_{i,j,k}$. In a purely barotropic motion, fluid parcels move up and down following changing free surface, and the grid-box locations move together with the parcels (Eq. (3.5)), resulting in no change in $\rho_{i,k}$. Hence, ρ^* and $\overline{\rho}$ in Eq. (3.25) are also nearly independent of ζ, which justifies keeping them constant during the fast time-stepping of barotropic mode. To ensure numerical stability and at least second-order accuracy, $\rho_{i,k}$, ρ^*, and $\overline{\rho}$ must be time-centered at $n + 1/2$ in baroclinic time.

An analogous discrete derivation (SHCHEPETKIN and MCWILLIAMS [2005]) yields

$$\mathscr{F}'_{i+\frac{1}{2}} = -\frac{g}{\rho_0}\left\{\frac{h_{i+1}+h_i}{2}\left(\rho^*_{i+1}\zeta_{i+1}-\rho^*_i\zeta_i\right)+\frac{\rho^*_{i+1}\zeta^2_{i+1}}{2}-\frac{\rho^*_i\zeta^2_i}{2}\right.$$
$$\left.+\,(h_{i+1}-h_i)\left[\frac{\left(\rho^*_{i+1}-\overline{\rho}_{i+1}\right)\zeta_{i+1}+\left(\rho^*_i-\overline{\rho}_i\right)\zeta_i}{2}+\frac{\left(\overline{\rho}_{i+1}-\overline{\rho}_i\right)\left(\zeta_{i+1}-\zeta_i\right)}{6}\right]\right\}. \tag{3.28}$$

The particular form of Eq. (3.28) depends on the discrete scheme for 3D PGF (Section 5). In principle, the splitting error can be eliminated entirely rather than just the leading-order term cancellation in Eq. (3.22). However, doing so imposes severe restrictions on the discretization choice for the 3D PGF that basically would then be limited to pressure Jacobian schemes (SHCHEPETKIN and MCWILLIAMS [2003]). This is undesirable because it raises the overall error in the PGF. For example, the scheme in LIN [1997] results in Eq. (3.28) without the last term inside [...] on the second line. Although this term is formally $\mathcal{O}\left(\Delta x^3\right)$-small (i.e., two orders higher than the preceding term), it is desirable to keep it since it makes Eq. (3.28) exact if ρ is a linear function of depth and horizontal coordinate, unlike the scheme in LIN [1997]. A density Jacobian scheme (as in BLUMBERG and MELLOR [1987]) does not allow separating ρ values belonging to different horizontal indices so that the vertical integral of $\mathscr{F}$ cannot be expressed in terms of ρ^* and $\overline{\rho}$ computed independently within each vertical column. The standard PGF scheme in ROMS (SHCHEPETKIN and MCWILLIAMS [2003]) uses a 4-point stencil in the horizontal and nonlinear interpolation of density to avoid spurious oscillations; both attributes make it impractical to derive an exactly consistent PGF scheme for the barotropic mode. Nevertheless, practical experience with Eq. (3.28) indicates that it is sufficiently accurate and stable.

For flat topography, ρ^* is the only relevant density for the barotropic mode. This choice is similar to Eq. (3.2) in HIGDON [1999], but it differs from BLECK and SMITH [1990] which uses the vertically averaged density (analogous here to $\overline{\rho}$) and from GRIFFIES, PACANOWSKI, SCHMIDT and BALAJI [2001] that uses the local density at the topmost grid cell instead of ρ^*. All other split-explicit models just use ρ_0. The terms proportional to $\nabla_x h$ in Eqs. (3.27) and (3.28) reflect the dynamical coupling between barotropic and baroclinic motions; it depends on the density difference, $\rho^* - \overline{\rho}$, and thus, it is part of what is sometimes referred to as the JEBAR effect (HOLLAND [1973]).

3.3. Design of the fast-time averaging filter

Averaging of the barotropic mode in a split-explicit model (i.e., choosing a_m in Eq. (3.14) distinct from a delta-function $\delta_{mM} = \{1,\ m = M;\quad 0,\ m \neq M\}$ is sometimes viewed as a "necessary evil" (GRIFFIES, PACANOWSKI, SCHMIDT and BALAJI [2001], HALLBERG [1997], HIGDON [1999]): while it yields a stable and robust numerical code, it undesirably degrades the temporal accuracy of the resolved barotropic motions and often introduces a numerical dissipation comparable to that of implicit backward-Euler time-stepping. We identify three reasons for averaging. First, although the effort is made to

remove mode-splitting error in PGF (Section 3.2), the split is never perfect in practice. If both the barotropic and baroclinic time-stepping algorithms are non-dissipative, barotropic aliasing may introduce numerical instability, HIGDON and BENNETT [1996], whereas fast-time averaging excludes the possibility of a coincidence of characteristic roots λ by placing the barotropic roots from the aliased range deep inside the unit circle (Section 3.2). Second, depending on the stage when the time-stepping algorithm computes the vertically integrated momentum advection terms that are kept constant during the barotropic time-stepping, they may incur a delay effectively like a forward time step for these terms. This leads to numerical instability of the same type as for a forward-in-time, centered-in-space advection equation. Fast-time averaging provides a mechanism to control this instability. This aspect puts an emphasis on damping at the low-frequency end, which is a very different requirement for the filter design compared with its anti-aliasing role. Third, depending on the algorithm for taking the first time-step (typically forward-in-time), the recurrent restart of the barotropic mode at each baroclinic time-step may introduce yet another numerical instability. Net dissipation in the barotropic time-stepping scheme and fast-time averaging can suppress this instability.

We now examine the design principles for the barotropic time filters. For simplicity of analysis, we assume $M \gg 1$, neglect the truncation error in the barotropic time-stepping, and replace the discrete summation over fast-time indices with a continuous time integration. $A(\tau)$ is defined as the continuous analog of $\{a_m | \, m = 1, \ldots, M^*\}$ with $\tau \sim m/M$ and $\tau_* \sim M^*/M$. A barotropic Fourier component ω_k gets a phase increment $\alpha = \omega_k \Delta t$ in one baroclinic time-step Δt. After fast-time averaging, its step-multiplier becomes

$$\lambda(\alpha) = \int_0^{\tau_*} \mathrm{e}^{-i\alpha\cdot\tau} A(\tau)\, \mathrm{d}\tau = \Re(\alpha)\mathrm{e}^{-i\alpha}, \tag{3.29}$$

where $\Re(\alpha)$ is the response function. Ideally, $\Re(\alpha) \approx 1$ for $\alpha \leq \alpha_0 \sim 1$ and $\Re(\alpha) \to 0$ rapidly in α once $\alpha > \alpha_0$. In the vicinity of $\alpha = 0$, $1 - \Re(\alpha) = \mathcal{O}(\alpha^n)$, where n is the temporal order of accuracy. Substitution of a Taylor-series expansion $\mathrm{e}^{-i\alpha\tau} = 1 - i\alpha\tau - \dfrac{\alpha^2\tau^2}{2} + \dfrac{i\alpha^3\tau^3}{6} + \ldots$ for $|\alpha| \ll 1$ in Eq. (3.29) leads to

$$\lambda(\alpha) = 1 - i\alpha - \frac{\alpha^2}{2}\mathfrak{I}_2 + \frac{i\alpha^3}{6}\mathfrak{I}_3 + \frac{\alpha^4}{24}\mathfrak{I}_4 + \ldots \quad \text{where} \quad \mathfrak{I}_n = \int_0^{\tau_*} \tau^n A(\tau)\, \mathrm{d}\tau, \tag{3.30}$$

with $\mathfrak{I}_0 \equiv \mathfrak{I}_1 \equiv 1$ due to the normalization and consistency conditions analogous to Eq. (3.16). Using the identity, $\tau^2 \equiv (\tau - 1)^2 + 2\tau - 1$, and the relation, $2\mathfrak{I}_1 - \mathfrak{I}_0 \equiv 1$, we find that

$$\mathfrak{I}_2 \equiv \int_0^{\tau_*} \tau^2 A(\tau)\, \mathrm{d}\tau = 1 + \int_0^{\tau_*} (\tau - 1)^2 A(\tau)\, \mathrm{d}\tau \equiv 1 + \epsilon. \tag{3.31}$$

If $A(\tau)$ is non-negative, the integrands are non-negative too; hence, $\epsilon \geq 0$ with equality reached only if $A(\tau)$ is a delta-function, $\delta(\tau - 1)$. Substitution of $\mathfrak{I}_2$ into Eq. (3.30) leads to the appearance of ϵ as a coefficient in the leading-order truncation term at second order. $\epsilon > 0$ corresponds to numerical dissipation. Therefore, any choice of a positive-definite $A(\tau)$ results in at most first-order accuracy for the fast-time-averaged barotropic mode (i.e., $\lambda(\alpha)$ agrees with $\mathrm{e}^{-i\alpha}$ only up to $\mathcal{O}(\alpha^2)$).

To achieve second-order accuracy, we introduce a shape function that allows some of the primary weights to be negative,

$$A(\tau) = A_0 \left\{ \left(\frac{\tau}{\tau_0}\right)^p \left[1 - \left(\frac{\tau}{\tau_0}\right)^q\right] - r\frac{\tau}{\tau_0} \right\}, \tag{3.32}$$

where p and q are independent parameters. A_0, τ_0, and r are then chosen to satisfy normalization, centroid, and second-order accuracy conditions in Eq. (3.30), viz., $\mathfrak{I}_n = 1$ for $n = 0, 1, 2$. In practice, we initially specify

$$r = 0 \quad \text{and} \quad \tau_0 = \frac{(p+2)\,(p+q+2)}{(p+1)\,(p+q+1)}, \tag{3.33}$$

(this choice of τ_0 centers $A(\tau)$ at $\tau = 1$; i.e., $\mathfrak{I}_1/\mathfrak{I}_0 = 1$), and then compute A_0 from the normalization condition. Using this initial $A(\tau)$, we adjust r, A_0, and τ_0 with an iterative procedure–adjust r to minimize $\epsilon = \mathfrak{I}_2 - 1$; recompute A_0 and τ_0 to restore $\mathfrak{I}_0 = \mathfrak{I}_1 = 1$; and repeat until $\epsilon \to 0$ – to satisfy the $\mathfrak{I}_n$ conditions. This yields a family $r = r(p, q)$ of second-order filters such as the following tabulated p, q, r-triplets.

$p = 2$	$q = 1$	$r = 0.1696907$	$p = 2$	$q = 4$	$r = 0.2846158$
2	2	0.2346283	2	6	0.2961888
2	3	0.2664452	3	8	0.1369941

The alternative choices, $p, q = 2, 4$ or $2, 2$, are the settings in ROMS for most applications; Fig. 3.1 is one of the corresponding shape functions.

Fig. 3.3 compares the step-multipliers for some fast-time-averaging algorithms with an S-shaped filter designed as described in this section. Ideally, $\lambda(\alpha) \approx 1$ for $\alpha \leq 2$ (the baroclinic time-stepping stability range), and $\lambda(\alpha) \ll 1$ thereafter. As expected, a flat averaging over $2\Delta t$ (left panel) results in very strong damping of the resolved frequencies (Griffies, Pacanowski, Schmidt and Balaji [2001]). A Hamming window (Oppenheim, Schafer and Buck [1999]) (middle panel) has much smaller dissipation for resolved frequencies and provides an efficient damping for the aliasing range. The $p, q = 2, 4$ filter (right panel) has virtually no damping for $|\alpha| \leq \pi/4$, and it is as efficient as the Hamming window in its anti-aliasing role. Another effect of having a negative lobe is that $A(\tau)$ makes the model more efficient by reducing the duration of the barotropic integration beyond t_{n+1} (i.e., $M^* - M$): the $p, q = 2, 4$ filter takes only 30% of the extra Δt step, while the Hamming window needs 50% and flat averaging needs 100%.

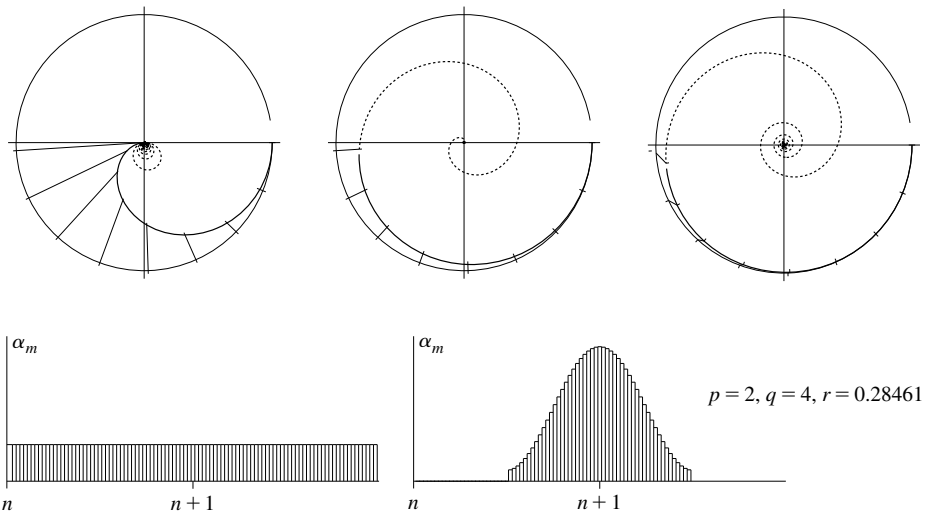


FIG. 3.3 Step-multiplier $\lambda(\alpha)$ for three different choices of the fast-time-averaging weights. Left: flat averaging over $2\Delta t$; middle: Hamming window; right: S-shaped weights from Fig. 3.1. The bold solid line on the diagrams turns dashed when entering the aliasing range.

3.4. Comparison with an implicit free-surface model

An implicit free-surface model entirely eliminates aliasing by simply restricting the phase increment of the barotropic mode. A particular scheme from the CFD community, the theta-method (CASULLI and CATTANI [1994]), is

$$\left.\begin{aligned} \zeta^{n+1} + i\alpha\theta u^{n+1} &= \zeta^n - i\alpha(1-\theta)u^n \\ u^{n+1} + i\alpha\theta\zeta^{n+1} &= u^n - i\alpha(1-\theta)\zeta^n \end{aligned}\right\} \quad \Rightarrow \quad \lambda(\alpha) = \frac{1 - \alpha^2\theta(1-\theta) \pm i\alpha}{1 + \alpha^2\theta^2}. \tag{3.34}$$

It is unconditionally stable if $1/2 \le \theta \le 1$ and is second-order accurate for $\theta = 1/2$. However, if used with $\alpha > 1$, the $\theta = 1/2$-scheme is prone to $2\Delta t$ oscillations, usually addressed by slightly biasing θ above 1/2, which makes it first-order accurate and dissipative. Setting $\theta = 2/3$ (Fig. 3.4, left) has a dissipation comparable to flat averaging over $2\Delta t$ (Fig. 3.3, left). A standard CFD practice is to use $\theta = 0.55$ (Fig. 3.4, right). Its damping is comparable (about twice as much) to the Hamming window. Since no third- or higher-order, unconditionally stable, implicit algorithm exists (note that an implicit AM3 scheme is asymptotically unstable for a purely hyperbolic problem), the theta-method is the only possibility for an implicit free-surface model, which constrains its accuracy to asymptotically approach second order when $\theta \to 1/2$. Therefore, a split-explicit model can be made inherently more accurate in representing even the relatively slow barotropic motions resolved by the baroclinic time-step (e.g., tides and topographic Rossby waves) than by an implicit model.

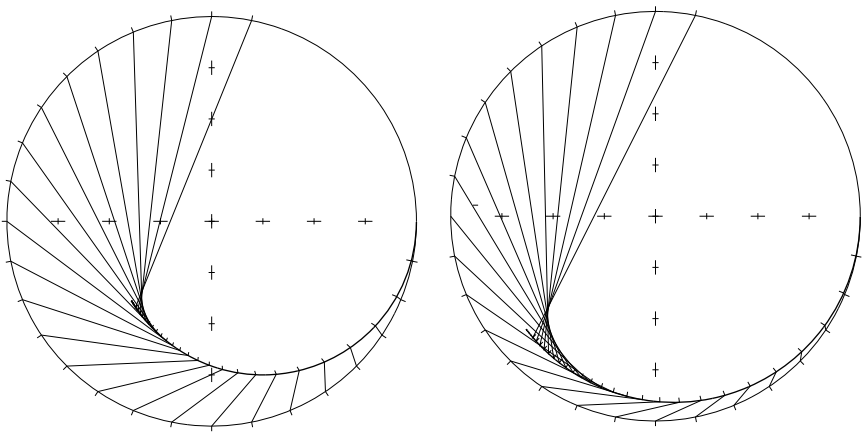


FIG. 3.4 $\lambda(\alpha)$ for the theta-method with two different θ values in the same format as Fig. 3.3. Comparing left and right panels shows that, while the dissipation increases with $\theta - 1/2|$, the phase error changes little with θ. The phase of $\lambda(\alpha)$ asymptotes to $-\pi$ when $\alpha \to \infty$, so $\lambda(\alpha)$ never enters the aliasing range.

4. Time-stepping the nonlinear system

4.1. Implementation of LF–AM3

The time-stepping algorithms in Section 2 are multi-time-level methods, relying on temporal interpolation or extrapolation of the r.h.s. terms computed at several consecutive steps to achieve the desired accuracy. This in principle can be applied to nonlinear systems as well (CANUTO, HUSSAINI, QUARTERONI and ZANG [1988]): compute and store the entire nonlinear r.h.s. at discrete time levels and interpolate it using these fields. On the other hand, mode-splitting in Section 3 restricts the choice of time-stepping algorithms to logically forward-in-time, two-time-level methods where the only available degree of freedom is the time placement of the tracer flux variables in Eq. (3.3) and similar quantities in momentum equations. Since tracer fluxes are products of volume fluxes and tracer values and volume fluxes are constrained by Eq. (3.11) to satisfy the finite-volume continuity equation (Eq. (3.4)), it is no longer possible to compute the complete tracer r.h.s. tendency terms at several consecutive time-steps and interpolate the result. Therefore, the algorithms from Section 2 must be adjusted for compatibility with mode-splitting.

The LF–AM3 scheme (Eqs. (2.21) and (2.22)) is rewritten as

$$\zeta^{n+\frac{1}{2}} = \left(\frac{1}{2} - 2\gamma\right)\zeta^{n-1} + \left(\frac{1}{2} + 2\gamma\right)\zeta^{n} - i\alpha\,(1-2\gamma)\,u^{n},$$

$$u^{n+\frac{1}{2}} = \left(\frac{1}{2} - 2\gamma\right)u^{n-1} + \left(\frac{1}{2} + 2\gamma\right)u^{n} - i\alpha\left[(1-2\gamma)\,\zeta^{n} + \beta\left(2\zeta^{n+\frac{1}{2}} - 3\zeta^{n} + \zeta^{n-1}\right)\right], \tag{4.1}$$

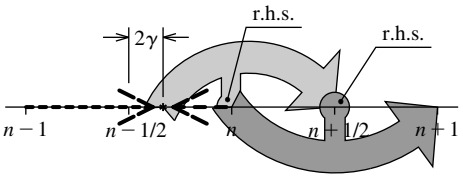


FIG. 4.1 Schematic diagram explaining the alternative LF–AM3 step: at first, $(n-1)$th and nth-step variables are interpolated linearly to $n-1/2+2\gamma$, which is used as the initial condition. It is advanced to $n+1/2$ using r.h.s. terms computed at nth step (predictor; $\gamma = 1/12$). Subsequently, the nth field is advanced to $n+1$ using the r.h.s. at $n+1/2$ (corrector).

followed by

$$\begin{aligned}
\zeta^{n+1} &= \zeta^n - i\alpha \cdot u^{n+\frac{1}{2}}, \\
u^{n+1} &= u^n - i\alpha \cdot \left\{(1-\epsilon)\,\zeta^{n+\frac{1}{2}} + \epsilon\left[\left(\frac{1}{2}-\gamma\right)\zeta^{n+1} + \left(\frac{1}{2}+2\gamma\right)\zeta^n - \gamma\zeta^{n-1}\right]\right\},
\end{aligned} \tag{4.2}$$

after which the provisional values $\zeta^{n+\frac{1}{2}}$ and $u^{n+\frac{1}{2}}$ are discarded. This alternative algorithm has a simple geometrical interpretation as a combination of interpolation and two consecutive LF-like steps (Fig. 4.1). It eliminates the need to store the full r.h.s. terms from one time-step to another, making the code more efficient. It is completely equivalent to the original algorithm if applied to a linear system (note that for the actual problem the symbolic operator $i\alpha[\ldots]$ here translates into a r.h.s. computation), while for a nonlinear system it differs by computing r.h.s. terms from the time-interpolated prognostic variables rather than an interpolation of the complete r.h.s. fields.

A comparison with LF-TR stepping, i.e., Eqs. (4.2) and (4.2) with $\gamma = 0$, offers another interpretation of Fig. 4.1: the 2γ bias relatively to $n-1/2$ in setting the initial condition introduces a pre-distortion that cancels the second-order truncation errors of the subsequent "logically LF" corrector stage, yielding an overall third-order accuracy of the algorithm as a whole.

Another difficulty with LF–AM3 is that the fluxes satisfying the discrete continuity equation (Eq. (3.4)) are available only during the corrector time-step not predictor step. Hence, it is impossible to achieve simultaneous conservation and constancy preservation for tracers during a predictor substep. Since the predicted values of the prognostic variables are used only to compute advective fluxes during the subsequent corrector step, the predictor substep does not necessarily have to be a conservative algorithm for the complete step to be conservative. A non-conservative, pseudo-compressible, predictor substep for tracers is

$$\begin{aligned}
q^{n+\frac{1}{2}}_{i,j,k} = \frac{1}{\Delta\mathscr{V}^+_{i,j,k}}\Bigg\{ & \Delta\mathscr{V}^-_{i,j,k}\left[\left(\frac{1}{2}+2\gamma\right)q^n_{i,j,k} + \left(\frac{1}{2}-2\gamma\right)q^{n-1}_{i,j,k}\right] \\
& - (1-2\gamma)\,\Delta t\Big[\widetilde{q}^n_{i+\frac{1}{2},j,k}U^n_{i+\frac{1}{2},j,k} - \widetilde{q}^n_{i-\frac{1}{2},j,k}U^n_{i-\frac{1}{2},j,k} + \widetilde{q}^n_{i,j+\frac{1}{2},k}V^n_{i,j+\frac{1}{2},k} \\
& - \widetilde{q}^n_{i,j-\frac{1}{2},k}V^n_{i,j-\frac{1}{2},k} + \widetilde{q}^n_{i,j,k+\frac{1}{2}}W^n_{i,j,k+\frac{1}{2}} - \widetilde{q}^n_{i,j,k-\frac{1}{2}}W^n_{i,j,k-\frac{1}{2}}\Big]\Bigg\},
\end{aligned} \tag{4.3}$$

where $\Delta\mathscr{V}^{\pm}_{i,j,k}$ is obtained from an artificial continuity equation,

$$\Delta\mathscr{V}^{\pm}_{i,j,k} = \Delta\mathscr{V}^{n}_{i,j,k} \mp \left(\tfrac{1}{2}-\gamma\right)\Delta t\Big[U^{n}_{i+\frac{1}{2},j,k} - U^{n}_{i-\frac{1}{2},j,k} + V^{n}_{i,j+\frac{1}{2},k} - V^{n}_{i,j-\frac{1}{2},k} + W^{n}_{i,j,k+\frac{1}{2}} - W^{n}_{i,j,k-\frac{1}{2}}\Big]. \tag{4.4}$$

The latter "absorbs" incompressibility errors in $U^{n}_{i+\frac{1}{2},j,k}$, $V^{n}_{i,j+\frac{1}{2},k}$, and $W^{n}_{i,j,k+\frac{1}{2}}$. The result is a conservative, constancy-preserving algorithm for $q^{n+\frac{1}{2}}_{i,j,k}$. Once the computation for $q^{n+\frac{1}{2}}_{i,j,k}$ is completed, $\Delta\mathscr{V}^{\pm}_{i,j,k}$ is discarded and recomputed during the next time-step. Because there is no guarantee that $\Delta\mathscr{V}^{+}_{i,j,k}$ is the same as $\Delta\mathscr{V}^{-}_{i,j,k}$ during the next time-step, Eq. (4.3) does not maintain the volume, $\sum_{i,j,k}\Delta\mathscr{V}_{i,j,k}q^{n+\frac{1}{2}}_{i,j,k}$. However, the complete algorithm – Eq. (4.3) in combination with corrector step via Eq. (3.3) – does.

4.2. Implementation of AB3–AM4

The AB3–AM4 FB scheme (Eq. (3.1)) is the method of choice for the barotropic mode because the time-step restriction imposed by the phase speed of barotropic waves dominates all other limitations (i.e., advection velocity and Coriolis frequency) by such a large degree that the other terms receive no consideration except for avoiding unconditionally unstable schemes. Its practical version consists of an AB3-extrapolation of prognostic variables,

$$\begin{pmatrix}\zeta\\ \overline{\mathbf{u}}\end{pmatrix}^{m+\frac{1}{2}} = \left(\frac{3}{2}+\beta\right)\begin{pmatrix}\zeta\\ \overline{\mathbf{u}}\end{pmatrix}^{m} - \left(\frac{1}{2}+2\beta\right)\begin{pmatrix}\zeta\\ \overline{\mathbf{u}}\end{pmatrix}^{m-1} + \beta\begin{pmatrix}\zeta\\ \overline{\mathbf{u}}\end{pmatrix}^{m-2}; \tag{4.5}$$

computation of finite-volume fluxes,

$$D^{m+\frac{1}{2}} = h + \zeta^{m+\frac{1}{2}}, \quad \overline{U}^{m+\frac{1}{2}} = D^{m+\frac{1}{2}}\Delta\eta\,\overline{u}^{m+\frac{1}{2}}, \quad \overline{V}^{m+\frac{1}{2}} = D^{m+\frac{1}{2}}\Delta\xi\,\overline{v}^{m+\frac{1}{2}}; \tag{4.6}$$

free-surface update,

$$\zeta^{m+1} = \zeta^{m} - \Delta t_{*}\cdot \mathrm{div}\overline{\mathbf{U}}^{m+\frac{1}{2}}; \tag{4.7}$$

computation of provisional ζ for the PGF,

$$\zeta' = \left(\frac{1}{2}+\gamma+2\epsilon\right)\zeta^{m+1} + \left(\frac{1}{2}-2\gamma-3\epsilon\right)\zeta^{m} + \gamma\zeta^{m-1} + \epsilon\zeta^{m-2}; \tag{4.8}$$

and the momentum step,

$$\overline{\mathbf{u}}^{m+1} = \frac{1}{D^{m+1}} \left\{ D^m \overline{\mathbf{u}}^m + \Delta t_* \cdot \left[\mathscr{F}\left(\zeta'\right) - D^{m+\frac{1}{2}} f\mathbf{k} \times \overline{\mathbf{u}}^{m+\frac{1}{2}} + \ldots \right] \right\}. \tag{4.9}$$

In the last step, the PGF $\mathscr{F}\left(\zeta'\right)$ is from Eq. (3.28), and the dots denote the other r.h.s. terms (advection, viscous diffusion, etc.,). This algorithm naturally accommodates advection (centered or upstream-biased) and the Coriolis force; it is stable without the need for viscosity or upstream-bias for the $\overline{\mathbf{U}}$ terms in the ζ equation; and it eliminates the need to store r.h.s. terms from one time-step to another.

A similar algorithm is applied for 3D mode in Kanarska, Shchepetkin and McWilliams [2007], except that unlike Eqs. (4.5)–(4.9), it starts with the update of momentum equation followed by the update of tracers. In that approach, the tracer fields were actually extrapolated toward $(n + 1/2)$th step twice using two different sets of AB3-like coefficients: the first time to compute density and then baroclinic pressure gradient (using coefficients optimized for stability of FB step), and the second time to compute advection terms for tracer equations (using coefficients chosen more close to the conventional AB3 set). This dual extrapolation removes the competitive requirements in setting of β in Eq. (2.32) as discussed in Section 2.4.

5. PGF

The discrete PGF error for a hydrostatic model in generalized vertical coordinates (including the σ family, e.g., ROMS) is widely recognized as a significant algorithmic problem (Arakawa and Suarez [1983], Blumberg and Mellor [1987], Chu and Fan [2003], Haney [1991], Kliem and Pietrzak [1999], Lin [1997], Mellor, Ezer and Oey [1994], Mesinger [1982], Mesinger and Janjic [1985], Shchepetkin and McWilliams [2003], Slordal [1997], Song [1998], Song and Wright [1998], Stelling and van Kester [1994]). It is often attributed to so-called hydrostatic inconsistency, i.e., a failure of the discretized PGF to vanish when isopycnic surfaces are horizontal. Because of deviation of quasi-horizontal coordinates from either geopotential-height (z) or isopycnic (ρ) surfaces, the PGF in the horizontal momentum equations appears in the form of two large terms that tend to cancel each other,

$$-\frac{1}{\rho_0} \left.\frac{\partial P}{\partial x}\right|_z = -\frac{1}{\rho_0} \left[\left.\frac{\partial P}{\partial x}\right|_s - \frac{\partial P}{\partial z} \cdot \left.\frac{\partial z}{\partial x}\right|_s \right]. \tag{5.1}$$

In the usual way, the partial-derivative subscript z means that it is computed with respect to a constant z surface, and the subscript s means that the differentiation is performed along the isosurface of the transformed vertical coordinate, $s = \text{const}$.

The most common focus has been on achieving accurate cancellation of the two terms in Eq. (5.1) in the special case of a horizontally uniform (i.e., flat) stratification, $\rho = \rho(z)$, where the correct answer is zero velocity (a state of rest). In this context Mellor, Oey and Ezer [1998] point out a Sigma-coordinate error of the second kind, which is the growth in time of a mainly barotropic flow with no mechanism of advective self-compensation

(in contrast to a baroclinic tendency to redistribute horizontal ρ surfaces by a flow generated by the PGF error to partially cancel the artificial flow). A small initial error does not guarantee that the error remain small at a later time. This experience brought attention to the integral properties such as material conservation and consistent conversion between potential and kinetic energy. Despite the vast published experience, there is not yet a consensus approach nor resolution of the problem. The approaches tend to fall into four major categories: (i) increase the order of accuracy in all coordinate directions (Beckmann and Haidvogel [1993], Chu and Fan [2003]); while this can be quite successful in idealized test cases, it has earned a reputation of being useless for realistic oceanic modeling (Kliem and Pietrzak [1999]); (ii) compute the PGF in z-coordinate space (Kliem and Pietrzak [1999] and its references); (iii) use a finite-volume, flux-form, pressure Jacobian formulation Chu and Fan [2003] and Lin [1997]; or (iv) use a density Jacobian discretization of an alternative form for PGF that computes the horizontal ρ gradient first then integrates it vertically (Blumberg and Mellor [1987], Song [1998]). The last two approaches rely on symmetric discretizations, mimicking the symmetries of the Jacobian operator, to reduce PGF error.

We found a successful technique to reduce PGF error. It is a generalization of the density Jacobian approach going to higher-order accuracy while retaining most of the symmetries of its original schemes (Shchepetkin and McWilliams [2003]). It can also be viewed as a polynomial reconstruction of the ρ field with subsequent analytical contour integration. A similar approach was applied to construct a high-order analog of the pressure Jacobian in Lin [1997]; however, the generalized density Jacobian is more attractive because of smaller truncation error and, more importantly, slower error growth in time for the flat stratification test cases. In this method, the PGF is formulated (similar to Blumberg and Mellor [1987]) as

$$-\frac{1}{\rho_0}\frac{\partial P}{\partial x}\bigg|_z = -\frac{1}{\rho_0}\frac{\partial P}{\partial x}\bigg|_{z=\zeta} - \frac{g}{\rho_0}\int_z^\zeta \frac{\partial \rho}{\partial x}\bigg|_z \mathrm{d}z' = -\frac{g}{\rho_0}\rho\bigg|_{z=\zeta}\frac{\partial \zeta}{\partial x} - \frac{g}{\rho_0}\int_z^\zeta \left[\frac{\partial \rho}{\partial x}\bigg|_s - \frac{\partial \rho}{\partial z'}\,\frac{\partial z'}{\partial x}\bigg|_s\right]\mathrm{d}z',$$

where the last term can be rewritten as

$$-\frac{g}{\rho_0}\int_s^0 \left[\frac{\partial \rho}{\partial x}\bigg|_s \frac{\partial z}{\partial s} - \frac{\partial \rho}{\partial s}\,\frac{\partial z}{\partial x}\bigg|_s\right] \mathrm{d}s' = -\frac{g}{\rho_0}\int_s^0 \mathscr{J}_{x,s}(\rho, z)\,\mathrm{d}s', \tag{5.2}$$

to justify the classification of the scheme as a density Jacobian type. The transformed vertical coordinate $s \in [-1, 0]$ is assumed in Eq. (3.5) to be neither isopycnic nor geopotential so that both terms inside the left-side integral in Eq. (5.2) are nontrivial. To discretize this, we introduce a control element $\mathscr{A}_{i+1/2,k+1/2}$ (the shaded area in Fig. 5.1) and apply Green's theorem,

$$\begin{aligned} -\Delta x\,\Delta z \cdot \mathscr{J}_{x,s}(\rho, z)\Big|_{i+\frac{1}{2},k+\frac{1}{2}} &\approx \iint_{\mathscr{A}} \mathscr{J}_{x,s}(\rho, z)\,\mathrm{d}z\,\mathrm{d}x = \oint_{\partial\mathscr{A}} \rho\, f \mathrm{d}z \\ &= FX_{i,k} - FX_{i+1,k} + FC_{i+\frac{1}{2},k+1} - FC_{i+\frac{1}{2},k}. \end{aligned} \tag{5.3}$$

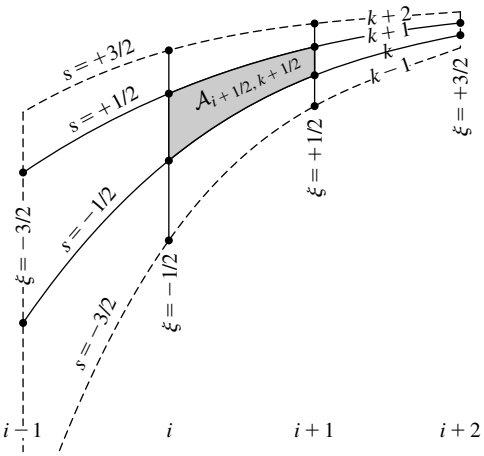


FIG. 5.1 Stencil in the x-z plane for computing the baroclinic PGF in the density Jacobian scheme. The Jacobian is approximated as a contour integral around the shaded area $\mathcal{A}_{i+1/2,k+1/2}$, $-\Delta x\,\Delta z\cdot \mathcal{J}_{x,s}(\rho, z) = \oint \rho(x, z)\,\mathrm{d}z$. The contour integral is approximated using one-dimensional cubic polynomial fits for both $\rho(x, z)$ and z as functions of the coordinates x and s along each of the four curvilinear facets bounding $\mathcal{A}_{i+1/2,k+1/2}$. Since a cubic fit requires a 4-point one-dimensional stencil, the whole Jacobian is evaluated using 12-points: a 4×4 grid without corners. Each of the line integrals *FX* and *FC* (Eq. (5.4)) participates in the computation of the density Jacobians for the two cell adjacent in either horizontally or vertically. The Jacobians are then integrated (via a simple summation) to compute PGF.

FX and *FC* are the line integrals along the vertical and quasi-horizontal curvilinear segments bounding $\mathcal{A}_{i+1/2,k+1/2}$,

$$FX_{i,k+\frac{1}{2}} = \int_{z_{i,k}}^{z_{i,k+1}} \rho\,\mathrm{d}z, \quad FC_{i+\frac{1}{2},k} = \int_{(x,s)_{i,k}}^{(x,s)_{i+1,k}} \rho\,\frac{\partial z}{\partial x}\bigg|_s\,\mathrm{d}x. \tag{5.4}$$

The problem thus reduces to interpolations for $\rho = \rho(x, s)$ and $z = z(x, s)$ along the integration contours. If linear interpolation is used for both ρ and z, the resultant scheme is equivalent to BLUMBERG and MELLOR [1987]. The natural extension is to use a cubic polynomial interpolation,

$$\rho(\xi) = \frac{\rho_j + \rho_{j+1}}{2} - \frac{d_{j+1} - d_j}{8} + \left[\frac{3}{2}\left(\rho_{j+1} - \rho_j\right) - \frac{d_j + d_{j+1}}{4}\right]\xi$$

$$+ \frac{d_{j+1} - d_j}{8}\xi^2 + \left[d_j + d_{j+1} - 2\left(\rho_{j+1} - \rho_j\right)\right]\xi^3, \tag{5.5}$$

where ξ defined for $-\frac{1}{2} \le \xi \le +\frac{1}{2}$ is either x or s; the index j is either i or k (see Fig. 5.1), and by construction,

$$\rho\,(\xi)\Big|_{\xi=-\frac{1}{2}} \equiv \rho_j \quad \rho\,(\xi)\Big|_{\xi=+\frac{1}{2}} \equiv \rho_{j+1} \quad \frac{\partial\rho}{\partial\xi}\Big|_{\xi=-\frac{1}{2}} \equiv d_j \quad \frac{\partial\rho}{\partial\xi}\Big|_{\xi=+\frac{1}{2}} \equiv d_{j+1}, \tag{5.6}$$

yields ρ values and derivatives at the side boundary of $\mathscr{A}_{i+1/2,k+1/2}$. Once Eq. (5.5) interpolates both ρ and z, the segment integrals (Eq. (5.4)) are evaluated analytically in terms of $z_{i,k}$, $\rho_{i,k}$, and their first spatial derivatives at the same location (see SHCHEPETKIN and MCWILLIAMS [2003] for full formulas).

The most important issue is the estimator for the derivative d_j, especially if ρ is not smooth on the grid. Using an algebraically averaged slope, the formula,

$$d_j = \frac{\Delta\rho_{j+\frac{1}{2}} + \Delta\rho_{j-\frac{1}{2}}}{2} \qquad \text{where} \qquad \Delta\rho_{j+\frac{1}{2}} \equiv \rho_{j+1} - \rho_j \qquad \forall j, \tag{5.7}$$

is sufficient to achieve the desired order of accuracy with a smooth ρ field and nearly uniform grid spacing. However, if ρ is not smooth, it admits spurious oscillations of the interpolant (Eq. (5.5)) that contaminate the PGF scheme as negative stratification patches, even when grid-point stratification values are positive everywhere, and this may result in numerical instability. In addition, Eq. (5.7) loses second-order accuracy if the grid spacing is not uniform. In contrast, a harmonic average,

$$d_j = \begin{cases} \dfrac{2\Delta\rho_{j+\frac{1}{2}}\Delta\rho_{j-\frac{1}{2}}}{\Delta\rho_{j+\frac{1}{2}} + \Delta\rho_{j-\frac{1}{2}}} & \text{if} \quad \Delta\rho_{j+\frac{1}{2}}\Delta\rho_{j-\frac{1}{2}} > 0 \\ 0 & \text{otherwise}\,, \end{cases} \tag{5.8}$$

has the property that if $\Delta\rho_{j+\frac{1}{2}}$ and $\Delta\rho_{j-\frac{1}{2}}$ have the same sign, d_j is no greater than twice the smaller of the two in magnitude; i.e.,

$$\left|d_j\right| < 2\left|\text{minmod}\left(\Delta\rho_{j+\frac{1}{2}}\,,\,\Delta\rho_{j-\frac{1}{2}}\right)\right|. \tag{5.9}$$

This guarantees that $\rho(\xi)$ in Eq. (5.5) is a monotonic, continuous function over the whole area of its definition.

Harmonic averaging (Eq. (5.8)) also escapes the loss of accuracy associated with non-uniformity of the vertical grid. This is extremely valuable for oceanic modeling since it is a common practice to choose only a moderate number of vertical levels with a grid spacing Δz that may change by as much as two orders of magnitude over the vertical column. Suppose that discretized values ρ_k are defined at locations z_k, such that

$\Delta z_{k+\frac{1}{2}} \equiv z_{k+1} - z_k \neq \Delta z_{k-\frac{1}{2}} \equiv z_k - z_{k-1}$. A Taylor series expansion around z_k gives

$$\rho_{k\pm1} = \rho_k \pm \rho' \Delta z_{k\pm\frac{1}{2}} + \frac{1}{2}\rho'' \Delta z^2_{k\pm\frac{1}{2}} \pm \frac{1}{6}\rho''' \Delta z^3_{k\pm\frac{1}{2}} + \ldots, \tag{5.10}$$

$$\Rightarrow \quad \begin{cases} \partial\rho_{k+\frac{1}{2}} \equiv \dfrac{\rho_{k+1} - \rho_k}{\Delta z_{k+\frac{1}{2}}} = \rho' + \dfrac{1}{2}\rho'' \Delta z_{k+\frac{1}{2}} + \dfrac{1}{6}\rho''' \Delta z^2_{k+\frac{1}{2}} + \ldots \\ \partial\rho_{k-\frac{1}{2}} \equiv \dfrac{\rho_k - \rho_{k-1}}{\Delta z_{k-\frac{1}{2}}} = \rho' - \dfrac{1}{2}\rho'' \Delta z_{k-\frac{1}{2}} + \dfrac{1}{6}\rho''' \Delta z^2_{k-\frac{1}{2}} + \ldots \end{cases} . \tag{5.11}$$

This leads to a second-order accurate approximation for $\partial\rho/\partial z$ at the location z_k,

$$\left.\frac{\partial\rho}{\partial z}\right|_{z=z_k} = \frac{\Delta z_{k-\frac{1}{2}} \partial\rho_{k+\frac{1}{2}} + \Delta z_{k+\frac{1}{2}} \partial\rho_{k-\frac{1}{2}}}{\Delta z_{k+\frac{1}{2}} + \Delta z_{k-\frac{1}{2}}} = \rho' + \frac{1}{6}\rho''' \Delta z_{k+\frac{1}{2}} \Delta z_{k-\frac{1}{2}} + \mathcal{O}\big(\Delta z^3\big), \tag{5.12}$$

which is just a linear interpolation of $\partial\rho$ on a non-uniform grid. On the other hand, since

$$\left.\frac{\partial\rho}{\partial z}\right|_{z=z_k} = \frac{\partial\rho/\partial s|_{s=s_k}}{\partial z/\partial s|_{s=s_k}}, \tag{5.13}$$

the use of Eq. (5.7) makes the estimator into

$$\left.\frac{\partial\rho}{\partial z}\right|_{z=z_k} = \frac{\Delta\rho_{k+\frac{1}{2}} + \Delta\rho_{k-\frac{1}{2}}}{\Delta z_{k+\frac{1}{2}} + \Delta z_{k-\frac{1}{2}}} = \rho' + \frac{1}{2}\rho'' \left(\Delta z_{k+\frac{1}{2}} - \Delta z_{k-\frac{1}{2}}\right) + \mathcal{O}\big(\Delta z^2\big). \tag{5.14}$$

This is only first-order accurate. It evaluates the derivative at the location $(z_{k+1} + z_{k-1})/2$ rather than the desired z_k. In contrast, Eqs. (5.13) and (5.8) applied to the elementary differences $\Delta\rho$ and Δz lead to

$$\left.\frac{\partial\rho}{\partial z}\right|_{z=z_k} = \frac{\Delta\rho_{k+\frac{1}{2}} \Delta\rho_{k-\frac{1}{2}} \left(\Delta z_{k+\frac{1}{2}} + \Delta z_{k-\frac{1}{2}}\right)}{\left(\Delta\rho_{k+\frac{1}{2}} + \Delta\rho_{k-\frac{1}{2}}\right) \Delta z_{k+\frac{1}{2}} \Delta z_{k-\frac{1}{2}}} = \frac{\partial\rho_{k+\frac{1}{2}} \partial\rho_{k-\frac{1}{2}} \left(\Delta z_{k+\frac{1}{2}} + \Delta z_{k-\frac{1}{2}}\right)}{\partial\rho_{k+\frac{1}{2}} \Delta z_{k+\frac{1}{2}} + \partial\rho_{k-\frac{1}{2}} \Delta z_{k-\frac{1}{2}}} . \tag{5.15}$$

We assume that $\Delta\rho_{k+\frac{1}{2}}$ and $\Delta\rho_{k-\frac{1}{2}}$ have the same sign and that $\rho = \rho(z)$ is sufficiently smooth on the grid scale to be accurately represented by a Taylor series. This essentially translates into the assumptions that

$$\left|\rho'' \cdot \Delta z\right| \ll |\rho'| \quad \text{and} \quad \left|\rho''' \cdot \Delta z^2\right| \ll |\rho'| \tag{5.16}$$

since the high-order derivatives in the Taylor series are presumed to be finite. Substitution of Eq. (5.11) into Eq. (5.15) yields

$$\left.\frac{\partial\rho}{\partial z}\right|_{z=z_k} = \rho' \frac{\left(1+\frac{1}{2}\frac{\rho''}{\rho'}\Delta z_{k+\frac{1}{2}} + \frac{1}{6}\frac{\rho'''}{\rho'}\Delta z^2_{k+\frac{1}{2}} + \ldots\right)\left(1-\frac{1}{2}\frac{\rho''}{\rho'}\Delta z_{k-\frac{1}{2}} + \frac{1}{6}\frac{\rho'''}{\rho'}\Delta z^2_{k-\frac{1}{2}} + \ldots\right)}{1+\frac{1}{2}\frac{\rho''}{\rho'}\frac{\Delta z^2_{k+\frac{1}{2}} - \Delta z^2_{k-\frac{1}{2}}}{\Delta z_{k-\frac{1}{2}} + \Delta z_{k-\frac{1}{2}}} + \frac{1}{6}\frac{\rho'''}{\rho'}\frac{\Delta z^3_{k+\frac{1}{2}} + \Delta z^3_{k-\frac{1}{2}}}{\Delta z_{k-\frac{1}{2}} + \Delta z_{k-\frac{1}{2}}} + \ldots}$$

$$= \rho' + \left(\frac{1}{6}\rho''' - \frac{1}{4}\frac{(\rho'')^2}{\rho'}\right)\Delta z_{k+\frac{1}{2}}\Delta z_{k-\frac{1}{2}} + \mathcal{O}(\Delta z^3), \tag{5.17}$$

indicating second-order accuracy of the estimator for $\partial\rho/\partial z|_{z=z_k}$. The leading order truncation term of Eq. (5.17) consists of two parts: the first one proportional to ρ''' is exactly the same as in Eq. (5.12), and the second is a nonlinear term,

$$-\frac{1}{4}\frac{(\rho'')^2}{\rho'}\Delta z_{k+\frac{1}{2}}\Delta z_{k-\frac{1}{2}} \approx -\rho'\left(\frac{\Delta\rho_{k+\frac{1}{2}} - \Delta\rho_{k-\frac{1}{2}}}{\Delta\rho_{k+\frac{1}{2}} + \Delta\rho_{k-\frac{1}{2}}}\right)^2.$$

The second formula always tends to reduce the estimated derivative and acts as a slope limiter if consecutive differences change abruptly on the grid scale. Because the same interpolation algorithm is applied to ρ and z, the discrete Jacobian guarantees the symmetry, $\mathscr{J}_{x,s}(\rho, z) = -\mathscr{J}_{x,s}(z, \rho)$. Although PGF cannot be eliminated entirely, it can be verified that for flat stratification, the cancellation of terms in Eq. (5.1) is fourth-order accurate, and the new scheme is robustly tolerant of "hydrostatically inconsistent" grids with $(\Delta x/\Delta z)\cdot\partial z/\partial x|_s > 1$ (Haney [1991]).

6. Impact of compressibility

The compressibility of seawater in the EOS raises two important design issues for oceanic models. The first is that the monotonicity constraint for $\rho(z)$ interpolation in Eqs. (5.5) and (5.8) no longer guarantees positive stratification for the interpolated profile if the constraint is applied to the in situ ρ, even if the point-wise stratification is strictly positive. This is because the grid-scale smoothness of ρ is judged by the ratio of consecutive differences, $\Delta\rho_{k+\frac{1}{2}}$ and $\Delta\rho_{k-\frac{1}{2}}$, both containing a component associated with bulk compressibility (i.e., a vertical change of in situ density that occurs even when potential temperature Θ and salinity S are spatially uniform). As a result, $\Delta\rho \approx -\Delta z\cdot g\rho_0/c_s^2 - \Delta z\cdot\rho_0 N^2/g$ (c_s is speed of sound and N is Brunt-Väisäla frequency), and the first term dominates under most oceanic conditions (Dukowicz [2001]). This obscures the detection of abrupt changes in stratification. The second issue is a consequence of the mode-splitting algorithm (Eqs. (3.21)–(3.28)) where $\overline{\rho}$ and ρ_* do not change in fast time, being kept constant at a time centered at $n+\frac{1}{2}$ to achieve second-order temporal accuracy during the barotropic time-stepping. When ρ is compressible, it depends on ζ though hydrostatic effects on pressure P. These changes are unaccounted for in the barotropic integration and thus are an additional source of mode-splitting error.

6.1. Compressibility and baroclinic PGF

The EOS for seawater expresses in situ ρ in terms of Θ, S, and P,

$$\rho = \rho\,(\Theta, S, P). \tag{6.1}$$

For oceanic modeling in situ, ρ is very interesting by itself, but it plays intermediate roles in several r.h.s. terms for prognostic variables, viz., horizontal PGF, stratification in vertical mixing parameterizations, and inclination of neutral surfaces along which lateral mixing occurs. The Boussinesq approximation replaces in situ ρ by a representative constant ρ_0 everywhere except in the gravitational force gravity; i.e., it retains the "gravitational" ρ in the gravitational force, but it approximates the "inertial" ρ in the Lagrangian acceleration by a constant ρ_0 that can be absorbed into PGF and otherwise disappear from the model. This approximation limits the EOS exclusively to the three purposes stated above, and the model is only sensitive to adiabatic gradients of ρ (defined in Eq. (6.6) below) but not to ρ itself. A consequence of the Boussinesq approximation is the replacement of mass conservation with an equivalent volume conservation based on a constant inertial ρ_0.

A common OGCM approximation is the replacement of in situ P in Eq. (6.1) with its bulk background value $P_0 = -g\rho_0 z$, viz.,

$$\rho = \rho\,(\Theta, S, |z|), \tag{6.2}$$

justified by $\rho - \rho_0 \ll \rho_0$. Free-surface, σ-coordinate models (Mellor [1991], Robinson, Padman and Levine [2001], Shchepetkin and McWilliams [2003]) often use an EOS in the form

$$\rho = \rho\,(\Theta, S, \zeta - z), \tag{6.3}$$

that selectively includes the barotropic contribution to the P used in the EOS but disregards the baroclinic part. The motivation for this choice comes not entirely from a physical consideration (i.e., $g\rho_0\zeta$ is often small compared with P) but more from a coding convenience where the vertical coordinate system is regenerated at every time-step from ζ and then used in the EOS routine. The use of standard EOS schemes, either as Eq. (6.1) or (6.2), implies a nonlinear dependence of ρ on z even if Θ and S are spatially uniform. For σ-coordinate models with coarse vertical resolution (often with a grid size as large as 500 m in the abyss), compressibility can cause significant PGF errors through hydrostatic non-cancellation in Eq. (5.1) (Shchepetkin and McWilliams [2003]). This type of error also exists in isopycnic models due to the non-equivalence of isopycnic and neutral surfaces caused by compressibility (Hallberg [2005]).

Consider for simplicity an EOS form within the approximation class of Eq. (6.2),

$$\rho\,(\Theta, S, z) = \rho^{(0)} + \rho'_1\,(\Theta, S) + \sum_{m=1}^{\infty} \left(q_m^{(0)} + q'_m\,(\Theta, S)\right) \cdot z^m. \tag{6.4}$$

$\rho^{(0)}$ and $q_m^{(0)}$, $m = 1, 2, \ldots$, are constant background values chosen so that $\rho^{(0)} \gg \rho'$, $q_1^{(0)} \gg q'_1$, etc. In practice, these are chosen by specifying representative constant values

for Θ and S and treating Eq. (6.4) as a series expansion around them. $q_1^{(0)}$ is the same as $g\rho^{(0)}/c_0^2$ (with c_0 a background value for c_s). With this EOS form, the density Jacobian (Eq. (5.2)) is

$$\mathscr{J}_{x,s}(\rho, z) = \mathscr{J}_{x,s}(\rho', z) + \sum_{m=1}^{\infty} \mathscr{J}_{x,s}(q'_m z) \cdot z^m. \tag{6.5}$$

Note that $\rho^{(0)}$ and $q_m^{(0)}$ contribute nothing.

Jackett and McDougall [1995] defined in situ adiabatic derivatives of ρ as differences of potential density with a local reference pressure (note that it is impossible to define potential density with any global reference pressure as a meaningful basis for determining stratification, unlike with the EOS for an ideal gas). In terms of Eq. (6.4), the adiabatic derivative with respect to the s coordinate is

$$\left.\frac{\partial \rho\,(\Theta,\, S,\, z)}{\partial s}\right|_{\mathrm{ad}} = \frac{\partial \rho'_1\,(\Theta,\, S)}{\partial s} + \sum_{m=1}^{\infty} z^m \frac{\partial q'_m\,(\Theta,\, S)}{\partial s}\,. \tag{6.6}$$

A similar expression applies to the horizontal (along $s = \mathrm{const}$) derivative $\partial \rho\,(\Theta,\, S,\, z)\,/\partial \xi|_{\mathrm{ad}}$. The baroclinic PGF (Eq. (6.5)) can be expressed entirely in terms of in situ adiabatic derivatives of ρ. For comparison, substituting the EOS (6.2) into Eq. (5.2) yields

$$\mathscr{J}_{x,s}(\rho, z) = -\hat{\alpha}\, \mathscr{J}_{x,s}(\Theta, z) + \hat{\beta}\, \mathscr{J}_{x,s}(S, z)\,. \tag{6.7}$$

Here, $\hat{\alpha} = -\,\partial\rho \big/ \partial\Theta|_{S,z}$ and $\hat{\beta} = \partial\rho \big/ \partial S|_{\Theta,z}$ are adiabatic thermal expansion and saline contraction factors (note that these differ from the conventional α and β coefficients by an added ρ multiplier). On the other hand, if the exact EOS Eq. (6.1) is used instead of Eq. (6.2), then the r.h.s. of Eq. (6.7) has an additional term, $-(1/c_s^2)\, \mathscr{J}_{x,s}(P, z)$ (i.e., $\propto \kappa$ in Eq. (6.17) below). This shows that the ability to express the baroclinic PGF entirely in terms of adiabatic ρ derivatives inherently relies on the EOS approximation $P \to z$ in Eq. (6.2). If the approximation in Eq. (6.2) is assumed valid (this aspect will be addressed in more detail in Section 6.3), then Eq. (6.7) indicates that the only requirement for accurately relating the gradients of Θ and S to the PGF is the correct computation of $\hat{\alpha}$ and $\hat{\beta}$, including their dependence on P or z (i.e., thermobaric effect). The in situ ρ by itself is irrelevant. This is also seen by the independence of Eq. (6.5) from the background terms $\rho^{(0)}$ and $q_m^{(0)}$.

Most of vertical change of ρ and much of the horizontal (along $s = \mathit{constant}$) change occur due to the bulk compressibility terms, i.e., $\partial \rho_{\text{in situ}}/\partial z \neq 0$ in Eq. (6.5). Consequently, a non-oscillatory profile of $\rho_{\text{in situ}}$ does not necessarily correspond to monotonic stratification. Therefore, it is meaningless to apply harmonic averaging (Eq. (5.8)) to consecutive differences of in situ ρ and to expect that monotonic positive stratification is guaranteed, even if the grid-box values of ρ are positively stratified. To achieve a monotonic stratification profile, we introduce elementary adiabatic differences, similar to Eq. (6.6) above; e.g., for $m = 1$,

$$\Delta\rho^{(\mathrm{ad})}_{i,k+\frac{1}{2}} = \rho'_{1\,i,k+1} - \rho'_{1\,i,k} + \left(q'_{1\,i,k+1} - q'_{1\,i,k}\right) \frac{z_{i,k+1} + z_{i,k}}{2}. \tag{6.8}$$

The averaged gradient (Eq. (5.8)) translates into

$$d_{i,k} \equiv \left.\frac{\partial \rho}{\partial s}\right|_{i,k} = \frac{2\Delta\rho^{(\mathrm{ad})}_{i,k+\frac{1}{2}} \cdot \Delta\rho^{(\mathrm{ad})}_{i,k-\frac{1}{2}}}{\Delta\rho^{(\mathrm{ad})}_{i,k+\frac{1}{2}} + \Delta\rho^{(\mathrm{ad})}_{i,k-\frac{1}{2}}} + q'_{1\,i,k} \left.\frac{\partial z}{\partial s}\right|_{i,k}, \tag{6.9}$$

where the adiabatic and compressible parts are separated at first, interpolated separately, and recombined at the end. This guarantees monotonicity of stratification for the interpolated profile. Because of the nonlinearity in Eq. (6.9), the resulting PGF scheme is incompatible with the common practice of subtracting a horizontally uniform background profile $\rho_{\mathrm{bak}} = \rho_{\mathrm{bak}}(z)$ in an attempt to reduce σ-coordinate PGF error. Similarly, the use of Eq. (6.7) as a basis for the PGF scheme is not desirable because separate computations of the Jacobians for Θ and S cannot ensure monotonicity of stratification if the Θ and S profiles are interpolated separately. For example, if there is a "spice" anomaly with large, smooth Θ and S gradients largely canceling each other to yield a ρ gradient that is small but non-smooth on the grid scale, then the monotonicity algorithm separately applied to Θ and S will fail to detect the sudden change in the ρ gradient.

6.2. Compressibility and barotropic mode-splitting

The mode-splitting algorithm described in Section 3.2 is derived using the assumption that ρ does not depend on ζ. This is no longer the case if the exact P dependence is included in the EOS (6.1) or even in its simplified version (Eq. (6.3)). Although the magnitude of the change is always small, a danger comes from the sensitivity of the EOS to ζ that implies a PGF contribution when ζ is computed at the previous time-step and kept constant during the barotropic time-stepping (i.e., effectively receiving a forward-in-time treatment). Consider a purely barotropic case with ρ changes due only to compressibility,

$$\rho = \rho(P) = \rho_1^{(0)} + \sum_m q_m^{(0)} P^m, \tag{6.10}$$

where $\rho_1^{(0)}$ and $q_m^{(0)}$ are spatially uniform. Without loss of generality, this can be replaced with

$$\rho = \rho_1^{(0)} + \sum_m q_m^{(0)} (\zeta - z)^m, \tag{6.11}$$

because the hydrostatic balance, $\partial P/\partial z = -g\rho$, makes it possible to remap Eq. (6.10) into Eq. (6.11) with an alternative set of coefficients q_m (e.g., $\rho = \rho_1 + q_1 P$ translates into $\rho = \rho_1 \exp\{g q_1 (\zeta - z)\}$).[12] A derivation similar to Eq. (3.23) yields the net PGF

[12]Another consequence of this $P \leftrightarrow z$ remapping is that it eliminates acoustic waves regardless whether or not the Boussinesq approximation is used. This makes it possible to build a hydrostatic, non-Boussinesq codes with relatively small additional effort. Non-hydrostatic, non-Boussinesq models must use other means to deal with unwanted acoustic waves (e.g., implicit time-stepping or the use of anelastic approximation) that may cause in a dramatic increase in code complexity.

applied to the fluid element (Fig. 3.2),

$$\mathscr{F}_{i+\frac{1}{2}} = g\big(\zeta_i - \zeta_{i+1}\big)\left[\rho_1^{(0)}\,\frac{D_i + D_{i+1}}{2} + \sum_m q_m^{(0)}\,\frac{D_i^{m+1} + D_i^m D_{i+1} + \cdots + D_i D_{i+1}^m + D_{i+1}^{m+1}}{(m+1)(m+2)}\right]. \tag{6.12}$$

This corresponds to the continuous form,

$$-g\left[\rho_1^{(0)} D + \sum_m q_m^{(0)}\,\frac{D^{m+1}}{m+1}\right]\nabla_x\zeta \equiv -g\overline{\rho}D\nabla_x\zeta, \tag{6.13}$$

where $\overline{\rho} = \rho_1^{(0)} + \sum_m q_m^{(0)}\,\dfrac{D^m}{m+1}$ is identified as the vertically averaged ρ. Therefore, we conclude that if ρ non-uniformity is caused exclusively by compressibility, then $\nabla_x\zeta$ generates exactly the same acceleration,

$$\frac{1}{\overline{\rho}D}\frac{\partial}{\partial t}\int_{-h}^{\zeta}\rho u\,\mathrm{d}z + \cdots = -g\nabla_x\zeta, \tag{6.14}$$

as in a uniform-density, shallow-water fluid. Furthermore, the acceleration by the full PGF, $-\dfrac{1}{\rho}\nabla_x P = -g\nabla_x\zeta$, is independent of depth throughout the vertical column even though both $P = P(z)$ and $\rho = \rho(z)$ are nonlinear functions of z; hence, a purely barotropic (i.e., vertically uniform) flow can remain barotropic.

Note that Eq. (6.11) is similar to Eq. (6.4), except that now it is expanded in powers of perturbed depth $\zeta - z$, rather than just z, and therefore, from Eq. (6.11), $\nabla_x\rho \neq 0$ as long as $\nabla_x\zeta \neq 0$. Still the $q_m^{(0)}$-terms in the EOS do not change the acceleration caused by the PGF. Here – unlike in the baroclinic case (Eqs. (6.4) and (6.5)) – the absence of spurious acceleration by the barotropic PGF is valid only in the non-Boussinesq case, with $\overline{u}$ and $\overline{v}$ defined as ρ-averaged rather than z-averaged velocity. The Boussinesq replacement of the inertial in situ ρ with ρ_0 creates a spurious multiplier ρ/ρ_0 in the PGF that destroys this property.[13] At a first glance, Eq. (6.14) suggests that taking into account ρ non-uniformity in the barotropic mode with ρ^* and $\overline{\rho}$ in Eq. (3.28) offers no benefit relative to the use of the shallow-water-like PGF term, $-gD\nabla_x\zeta$. However, Eqs. (3.28) and (6.14) are derived under two opposite assumptions about the ρ structure: Eq. (3.28) assumes that the ρ non-uniformity comes purely from baroclinic effects, and the flow is incompressible, hence ρ is conserved as Lagrangian tracer;

[13] This situation is similar to Case A of Dewar, Hsueh, McDougall and Yuan [1998] discussed in Section 6.3, but in reverse, the dependency $\rho = \rho(P)$ in Eq. (6.21) brings in PGF error when used within the modified Boussinesq model.

whereas Eq. (6.14) assumes that all non-uniformity comes exclusively from the bulk compressibility. Besides the spurious ρ/ρ_0 factor, we identify two types of dangerous error: (i) the mode-splitting error due to the $\rho = \rho(\ldots, \zeta - z)$ dependency since the computation of the 3D PGF in Eq. (3.21) is based on the previous-time ζ and thus receives a forward-in-time treatment; and (ii) an erroneous sensitivity of ρ_* and $\overline{\rho}$ to the vertical increase of in situ ρ by bulk compressibility that is mistaken for vertical stratification.

The magnitude of the mode-splitting error of type (i) is estimated from the vertical integral of the PGF due to ζ modulated by compressibility,

$$-g\nabla_x\zeta \cdot \int_{-h}^{\zeta} \exp\left\{\frac{g(\zeta - z')}{c_s^2}\right\} \mathrm{d}z' \approx -\underbrace{gD\nabla_x\zeta}_{\text{“fast”}} - \underbrace{\frac{1}{2}\cdot\frac{gD}{c_s^2}\cdot gD\nabla_x\zeta}_{\text{“slow”}}. \tag{6.15}$$

The “fast” term is treated within the barotropic mode using a small time-step. The “slow” term is never computed explicitly, but is rather an outcome of computing the vertical integral of 3D PGF based on ρ with the EOS using the ζ at the old time-step – the most recent available value before barotropic time-stepping begins. As a result, it remains unchanged during barotropic time-stepping even though it contains a gradient of ζ. $D = 5$ km and $c_s = 1500$ m/s yield an error estimate of $gD/(2c_s^2) = 0.01$, about 1% of the PGF due to the ζ perturbation. This is comparable with levels of other mode-splitting errors discussed in Section 3.2. It is expected to stay within the Courant-number limit of baroclinic (slow) time-stepping, leaving its forward-in-time treatment as the primary remaining concern. This type of splitting occurs whether or not the barotropic mode accounts for ρ non-uniformity, and furthermore, it occurs in non-Boussinesq models as well. For example, Robinson, Padman and Levine [2001] identified a similar error (although they do not classify it as mode-splitting error) and an associated instability in a model that uses a shallow-water form for the PGF in the barotropic mode. The instability is manifested as a tidal response with spuriously elevated amplitude. The source of instability is traced back to an inconsistency between ρ and horizontally averaged $\rho(z)$ profile (subtracted out in hopes of reducing PGF error); the former is computed using instantaneous ζ and the latter using $\zeta = 0$. Their proposed remedies include abandonment of the averaged $\rho(z)$ subtraction – a relatively minor effect – and total suppression of compressibility in the EOS – sufficient to suppress the instability but not acceptable in OGCMs because of loss of the thermobaric effect. Griffies, Pacanowski, Schmidt and Balaji [2001] and McDougall, Greatbatch and Lu [2002] advocate the use of the exact EOS Eq. (6.1) with a P that includes dynamic components due to both ζ and ρ taken from the previous time-step. However, we believe that this brings a similar mode-splitting error and potential instability that most likely is only controlled by their heavy barotropic-mode time-filtering by averaging over two baroclinic time-steps (Section 3.3).

A better treatment for both types of errors is presented in Section 6.3 after an analysis of alternative forms for the EOS.

6.3. Consistency of EOS and boussinesq approximation

The EOS form (Eq. (6.2)) as an approximation to Eq. (6.1) was challenged by DEWAR, HSUEH, MCDOUGALL and YUAN [1998].[14] Consider the response of a compressible barotropic fluid with uniform Θ and S to an imposed surface PGF $\nabla_x p_s$ (their Case A, Fig. 1.1). If Eq. (6.2) is used for the EOS, the PGF is constant and equal to its surface $\nabla_x p_s$ value throughout the vertical column. However, compressibility leads to changes in ρ, and if the EOS more correctly uses in situ P, the ρ changes depend on the PGF itself, and the true PGF will change with depth. Substituting their Eq. (2.3) into Eq. (2.2) yields

$$\nabla_x P = \nabla_x p_s + g \int_z^0 \frac{\nabla_x P}{c_s^2}\, \mathrm{d}z'. \tag{6.16}$$

This has the solution $\nabla_x P = \nabla_x p_s \cdot \mathrm{e}^{-gz/c_s^2}$; i.e., now the PGF has an exponential amplifier with depth. With typical abyssal values of $c_s = 1500\,\mathrm{m/s}$ and $z = -5000\,\mathrm{m}$, the amplification factor is about 1.022, which is comparable to a typical PGF error due to the Boussinesq approximation. However, the pressure gradient does not appear in the PGF by itself but in the combination, $(1/\rho)\nabla_x P$. Thus, by using the exact in situ ρ that has the same compressibility amplifying factor, instead of the Boussinesq ρ_0 without it, the depth-amplification effect is canceled in the PGF. For example, the balancing geostrophic velocity (cf., their Eq. (2.3)) does not change at all between a Boussinesq code with an approximate EOS and a non-Boussinesq code with the exact EOS. Their Cases B and C (Fig. 1.1) involve baroclinic variations of Θ and S. In contrast to the purely barotropic Case A, these cases do not have an exact cancellation of the compressibility errors. However, as shown by DUKOWICZ [2001], more than 90% of the error can be eliminated by a further modification of the EOS, so the danger identified by DEWAR, HSUEH, MCDOUGALL and YUAN [1998] is largely avoidable.

The approach of DUKOWICZ [2001] splits the compressibility coefficient κ into two parts,[15]

$$\kappa = \frac{1}{\rho}\left(\frac{\partial \rho}{\partial P}\right)_{\Theta,S}, \quad \kappa = \kappa^{(P)}(P) + \delta\kappa(\Theta, S, P), \tag{6.17}$$

[14] Although ROMS uses an intermediate approximation to EOS (6.3), this criticism is still a concern because of the absence of the $-(1/c_s^2)\mathscr{J}_{x,s}(P, z)$ term in Eq. (6.7) and its counterpart in Eq. (6.5). Sections 5.1 and 5.2 of SHCHEPETKIN and MCWILLIAMS [2003] introduce two PGF schemes. One computes the density Jacobian directly and then integrates it vertically (hence, entirely avoiding computation of P), and the other is a primitive form that first explicitly computes P. These two schemes are identical for an incompressible EOS, but the statement that the PGF can be expressed entirely in terms of adiabatic ρ differences applies only to the first scheme. Unless the EOS is modified to exclude bulk compressibility, the primitive form implicitly contains an equivalent of the $-(1/c_s^2)\mathscr{J}_{x,s}(P, z)$ component.

[15] To avoid confusion with ρ^* in the barotropic PGF in Sections 3.2 and 6.2, we modified the original notation of DUKOWICZ [2001] by $\rho^* \to \rho^\bullet$ and $P^* \to P^\bullet$.

where $\kappa^{(P)}$ is much larger than $\delta\kappa$. The exact EOS (6.1) can be rewritten with two ρ factors,

$$\rho = r(P) \cdot \rho^{\bullet}(\Theta, S, P). \tag{6.18}$$

Without any approximation, the PGF, hydrostatic balance, and EOS can be alternatively be expressed in terms of $\rho^{\bullet}$ and a related pressure quantity $P^{\bullet}$:

$$\frac{1}{\rho}\nabla_x P \quad \leftrightarrow \quad \frac{1}{\rho^{\bullet}}\nabla_x P^{\bullet}, \tag{6.19}$$

$$\frac{\partial P}{\partial z} = -g\rho \quad \leftrightarrow \quad \frac{\partial P^{\bullet}}{\partial z} = -g\rho^{\bullet}, \tag{6.20}$$

$$\rho = \rho(\Theta, S, P) \quad \leftrightarrow \quad \rho^{\bullet} = \frac{\rho\Big(\Theta, S, P(P^{\bullet})\Big)}{r(P^{\bullet})} = \rho^{\bullet}(\Theta, S, P^{\bullet}). \tag{6.21}$$

The relations in the right column have the same functional forms as the original ones in the left column, and the scaling factor $r(P)$ does not explicitly appear.

The practical value of the approximate EOS Eq. (6.4) or a fortiori the factored EOS Eq. (6.18) for oceanic simulations comes from a dramatic narrowing with depth of the range of realistic values for Θ and S (cf., Fig. 19 in Shchepetkin and McWilliams [2003] and Fig. 2 in McDougall, Jackett, Wright and Feistel [2003]). For the factored EOS form, $r(P)$ can be chosen so that $\kappa^{(P)}$ strongly dominates $\delta\kappa$ in Eq. (6.17) in the abyss; hence, the nonlinear dependence of ρ on P or z is mostly absorbed into $r(P)$, which is subsequently scaled out in the ρ, $P \rightarrow \rho^{\bullet}$, and $P^{\bullet}$ transformation (Eq. (6.21)). In the upper ocean, Θ and S are more variable, and factoring is not as effective in keeping $\delta\kappa$ small compared with $\kappa^{(P)}$; however, the nonlinear compressibility is not as important there, and useful approximations to the EOS can be made without sacrificing accuracy. We choose the definition,

$$r(P) = \rho_{\mathrm{JM95}}(\Theta_0, S_0, P) \,/\rho_{\mathrm{JM95}}(\Theta_0, S_0, 0), \tag{6.22}$$

where $\rho_{\mathrm{JM95}}(\Theta, S, P)$ refers to the particular form of the EOS in Jackett and McDougall [1995], and Θ_0 and S_0 are representative abyssal values (e.g., $\Theta_0 = 1.5$ and $S_0 = 34.74$ are good choices for global or basin-scale modeling). Then,

$$\rho^{\bullet} = \rho_{\mathrm{JM95}}(\Theta, S, P)/r(P) \tag{6.23}$$

has a substantially narrower dynamical range than the original $\rho = \rho_{\mathrm{JM95}}(\Theta, S, P)$, and, even more importantly, it does not grow with P or z. In the terminology of Dukowicz [2001], this procedure "stiffens" the EOS. In a Boussinesq model based on Eq. (6.21), $\rho^{\bullet}$ is replaced with the reference value ρ_0 (e.g., $\rho_0 = 1027.8\,\mathrm{kg/m^3}$ is consistent with the Θ_0 and S_0 choices above and is closer to the actual $\rho^{\bullet}$ than the more widely used ρ_0 values of 1000 or 1025 $\mathrm{kg/m^3}$). The in situ P used inside the EOS routine is approximated with a background $P_0(z)$ computed from

$$\frac{\mathrm{d}P_0}{\mathrm{d}z} = -g\rho_0 \cdot r(P_0). \tag{6.24}$$

This approximates the EOS in Eq. (6.23) as

$$\rho^{\bullet} = \rho^{\bullet}(\Theta, S, z). \tag{6.25}$$

This is the same functional form as Eq. (6.2), but it accounts for the main effect of ρ variation in computing P in the EOS; thus, it is closer to the exact EOS (6.1) in representing the changes of $\hat{\alpha}$ and $\hat{\beta}$ with depth. Finally, as in the PGF algorithm in SHCHEPETKIN and MCWILLIAMS [2003], the resulting EOS (6.25) is split as in Eq. (6.4), except that the expansion in powers of z is replaced with a $(\zeta - z)$ expansion that is truncated after the linear term. To minimize round-off errors, the EOS is expressed as a perturbation relative to ρ_0.

This form of the EOS in Eq. (6.25) allows computation of adiabatic ρ differences (Eq. (6.8)). These are averaged with a harmonic mean (Eq. (6.9)) that is subsequently needed to construct cubic interpolants (Eq. (5.5)), segment integrals (Eq. (5.4)), and discrete density Jacobian. The interpolant is guaranteed to maintain positive stratification as long as the discrete density field is positively stratified. Although removing the dominant part of the bulk compressibility, Eq. (6.21) makes point-wise differences of $\rho^{\bullet}$ much closer to adiabatic differences, one might be tempted to compute the baroclinic PGF directly from $\rho^{\bullet}$ without using adiabatic differencing. However, our experience has shown that this is neither sufficiently accurate in practice nor robust when there are sudden changes in stratification.

The transformation (Eq. (6.21)) offers a natural, simple remedy to reduce the mode-splitting errors of both types (i) and (ii) in Section 6.2; the elimination of bulk passive compressibility in the EOS effectively removes the second r.h.s. term in Eq. (6.15), but unlike the remedy of ROBINSON, PADMAN and LEVINE [2001], it retains a physically accurate representation of the thermobaric effect. Computing ρ^{*} and $\overline{\rho}$ from $\rho^{\bullet}$ is sufficient to eliminate their biases due to bulk compressibility, hence to avoid a type (ii) error.

Despite the multistage transformation described here, the functional forms of the EOS and PGF schemes in SHCHEPETKIN and MCWILLIAMS [2003] remain unchanged, requiring only a refitting of the polynomial coefficients in the EOS.[16]

6.4. Accuracy of the boussinesq approximation

The accuracy and utility of using the Boussinesq approximation for an OGCM is assessed in several papers GREATBATCH [2001], GREATBATCH and MCDOUGALL [2003], MCDOUGALL and GARRET [1992], MCDOUGALL, GREATBATCH and LU [2002], identifying, among other issues, an inherent conflict between the assumption of constancy of ρ (hence replacement of mass conservation with volume conservation) and the need to use

[16]Although more recent and supposedly more accurate versions of EOS have become available (JACKETT, MCDOUGALL, FEISTEL, WRIGHT and GRIFFIES [2006], MCDOUGALL, JACKETT, WRIGHT and FEISTEL [2003]), the EOS functional form in JACKETT and MCDOUGALL [1995], inherited from the UNESCO EOS, is preferable as the approximation standard because it is already close to the desired factored form, comprised of $\rho(\Theta, S)$ at 1 atm ($P = 0$ in our terms) multiplied by terms that account for compressibility effects. The rational functional form used in the newer EOS mixes P terms together with Θ and S terms and makes it harder to separate out P effects.

the fully compressible EOS that implies $\approx 5\%$ variation in ρ. This limits the accuracy of the Boussinesq approximation, and there has emerged a slow but steady advocacy for non-Boussinesq OGCMs (e.g., GRIFFIES, PACANOWSKI, SCHMIDT and BALAJI [2001] and the citations above).

In this situation, DUKOWICZ [2001] stands out because it constitutes a revision of the Boussinesq approximation as traditionally applied to OGCMs that include a compressible EOS in an ad hoc manner, breaking the internal consistency of the Boussinesq approximation. The revision restores consistency by bringing the properties of the EOS close to that for an incompressible fluid while still including the thermobaric effect. This approach stays within the spirit of the Boussinesq approximation by making the approximate PGF close to the full non-Boussinesq version without explicitly including any non-uniformity of the inertial ρ. The bulk compressibility ratio $r(P)$ is not used anywhere except in the $P_0 \leftrightarrow z$ remapping (Eqs. (6.24) and (6.25)) for the stiffened EOS, which brings a minor effect relative to the more traditional choice of replacing $P_0 = -\rho_0 g z$ in EOS.

This aspect of DUKOWICZ [2001] was criticized by MCDOUGALL, GREATBATCH and LU [2002] – in essence advocating discarding $r(P)$ – since it leaves "no choice but to interpret the horizontal velocity vector as the Eulerian-mean horizontal velocity, but not as the mass flux per unit area." This is viewed as a drawback because it prevents a reinterpretation of the prognostic variables in a Boussinesq model as density-weighted rather than Eulerian averages. When a solution reaches a stationary state, the difference between the reinterpreted Boussinesq model and a non-Boussinesq model disappears (cf., Sections 4 and 5 in MCDOUGALL, GREATBATCH and LU [2002], as well as similar approaches for including some non-Boussinesq effects in Boussinesq models GREATBATCH [2001], LU [2000]). This reinterpreted equivalence implies that the actual Boussinesq errors are less than the usual estimate of $\approx 5\%$ associated with the standard formulation. The use of $\rho^\bullet$ and $r(p)$ in a Boussinesq model prevents this reinterpretation.

This limitation of DUKOWICZ [2001] can be substantially mitigated in a finite-volume code by replacing $H_{i,j,k} = z_{i,j,k+\frac{1}{2}} - z_{i,j,k-\frac{1}{2}}$ in Eq. (3.5) with

$$H_{i,j,k} = \left(z_{i,j,k+\frac{1}{2}} - z_{i,j,k-\frac{1}{2}}\right) \cdot r\left[P_0\left(\zeta_{i,j} - \frac{z_{i,j,k+\frac{1}{2}} + z_{i,j,k-\frac{1}{2}}}{2}\right)\right]. \tag{6.26}$$

This replacement automatically, and without additional computational effort, implies a redefinition of the control volumes $\Delta\mathscr{V}_{i,j,k}$, interfacial contact surfaces, horizontal flux ($U_{i+\frac{1}{2},j,k}$, $V_{i,j+\frac{1}{2},j,k}$), and vertical flux $W_{i,j,k+\frac{1}{2}}$ (Eq. (3.9)) as mass-weighted by $\rho = r(P_0(z))$. This yields the major part of non-uniform inertial ρ in transforming volume conservation into approximate mass conservation with $\int\int\int r(P_0(z))\, d^3\mathscr{V}$.

Density Jacobian schemes use a contour integration to approximate $\Delta x \Delta z \cdot \mathscr{J}_{x,s}(\rho, z)$ which is then integrated vertically (via a simple summation) to compute point-wise pressure gradient. The latter one is subsequently multiplied by a horizontally averaged $H_{i,j,k}$ to convert it into the force applied to the control volume. This makes the PGF be invariant with respect to a change of definition for $H_{i,j,k}$ from the original

to Eq. (6.26) because the velocity component is also multiplied by the same $H_{i,j,k}$. The change in $H_{i,j,k}$ also leaves the transformation (Eq. (6.21)) unaffected. The analysis of DUKOWICZ [2001] only considers instantaneous errors associated with an inconsistent use of a fully compressible EOS in a Boussinesq model, but this is not a guarantee that the error will not grow in time. Recently, LOSCH, ADCROFT and CAMPIN [2004] and DE SZOEKE and SAMELSON [2002] pointed out that the hydrostatic, Boussinesq equations in z are isomorphic to the hydrostatic, non-Boussinesq equations in pressure coordinates. This implies that the solution differences between Boussinesq and non-Boussinesq models should stay bounded in time since P and z differences do so. Because Eq. (6.26) merely introduces a metric factor in the vertical coordinate while retaining the mathematical structure Boussinesq code, we expect that the Boussinesq errors using Eq. (6.26) also stay bounded.

The preceding discussion shows that the theoretical differences between Boussinesq and non-Boussinesq hydrostatic models are much less than the initial estimates of MCDOUGALL and GARRET [1992] and DEWAR, HSUEH, MCDOUGALL and YUAN [1998]. The differences can be further reduced by application of the transformation (Eq. (6.21)) in combination with the quasi-Boussinesq $r(P)$-remapping (Eq. (6.26)). The Boussinseq apprroximation offers an important advantage for a cleaner mode-spliting algorithm to avoid type (i) and (ii) errors (Section 6.2). Conversely, a more fundamental non-Boussinesq code does not escape the need to assure monotonic stratification profiles with higher-order Jacobian PGF schemes in generalized vertical coordinates and a compressible EOS that includes thermobaric effects. In summary, we do not presently see a strong case for preferring a non-Boussinesq OGCM.

7. Final remarks

In this chapter, we have discussed many of the central algorithmic elements – the computational kernel – in an OGCM designed for large computations of highly turbulent flows. Our currently preferred choices for these elements are summarized in Section 1 and discussed at length in the ensuing sections. A key aspect of OGCM design is the interplay among the kernel elements, with abundant possibilities for both destructive interference and constructive synergy. This perspective confounds any simple expectation that better code modularity is the principal software step toward better OGCMs: while a modular structure may facilitate code adaptation, the most important design consideration is the overall model performance in physical and numerical accuracy and computational efficiency.

The use of oceanic models has historically followed a path downward in scale, from basins and global domains to flows with smaller space and time scales and more turbulent dynamics. At the present time, the simulation battle front is at mesoscale-eddy resolution, but we can anticipate continuing scale refinements through a combination of larger computers, further algorithmic advances, multiscale (nested-grid) methods, and, of course, improved dynamical understanding of the simulated phenomena. We intend to participate in these developmental directions and mention, in closing,

a newly constructed, non-hydrostatic version of ROMS KANARSKA, SHCHEPETKIN and MCWILLIAMS [2007].

8. Acknowledgments

The authors appreciate the sustained support for model development research provided by the Office of Naval Research through its grants N00014-98-1-0165, N00014-00-1-0249, N00014-02-1-0236, and N00014-05-10293.

References

ADCROFT, A., CAMPIN, J.-M. (2004). Rescaled height coordinates for accurate free-surface flows in ocean circullation models. *Ocean Model.* **7**, 269–284.

ARAKAWA, A., LAMB, V.R. (1977). Computational design of the basic dynamical processes of the UCLA general circulation model. *Methods Comput. Phys.* **17**, 174–267.

ARAKAWA, A., SUAREZ, M.J. (1983). Vertical differencing of the primitive equations in sigma coordinates. *Mon. Wea. Rev.* **111**, 34–45.

BECKMANN, A., HAIDVOGEL, D.B. (1993). Numerical simulation of flow around a tall isolated seamount. Part. 1: problem formulation and model accuracy. *J. Phys. Ocean.* **23**, 1736–1753.

BECKMANN, A., HAIDVOGEL, D.B. (1997). A numerical simulation of flow at Fieberling Guyot. *J. Geophys. Res.* **102**, 5595–5613.

BERNTSEN, H., KOWALIK, Z., SÆLID, S., SØRLI, K. (1981). Efficient numerical simulation of ocean hydrodynamics by a splitting procedure. *Model. Identif. Control*. **2** (4), 181–199.

BLECK, R., SMITH, L.T. (1990) A wind-driven isopycnic coordinate model of the north and equatorial Atlantic Ocean: 1. Model development and supporting experiments. *J. Geophys. Res.* **95C**, 3273–3285.

BLUMBERG, A.F., MELLOR, G.L. (1987). A description of a three-dimensional coastal ocean circulation model. In: Heaps, N. (ed.), *Three-Dimensional Coastal Ocean Models* (AGU), pp. 1–16.

BRYAN, K., COX, M. (1969). A numerical method for the study of the circulation of the world ocean. *J. Comp. Phys.* **4**, 347–376.

CAMPIN, J.-M., ADCROFT, A., HILL, C., MARSHALL, J. (2004). Conservation of properties in a free-surface model. *Ocean Model.* **6**, 221–244.

CANUTO, C., HUSSAINI, M.Y., QUARTERONI, A., ZANG, T.A. (1988). *Spectral Methods in Fluid Mechanics* (Springer-Verlag), pp. 567.

CASULLI, V., CHENG, R.T. (1992). Semi-implicit finite-difference methods for 3-dimensional shallowwater flow. *Int. J. Numer. Methods Fluids* **15**, 629–648.

CASULLI, V., CATTANI, E. (1994). Stability, accuracy and efficiency of a semi-implicit method for 3-dimensional shallow-water flow. *Comput. Math. Al.* **27**, 99–112.

CASULLI, V., STELLING, G. (1998). Numerical simulation of 3D quasi-hydrostatic free-surface flows. *J. Hydr. Eng.* **124**, 678–686.

CASULLI, V. (1999). A semi-implicit finite-difference method for non-hydrostatic free-surface flows. *Int. J. Numer. Methods Fluids*. **30**, 425–440.

CHENG, R.T., CASULLI, V. (2001). Evaluation of the UnTRIM model for 3-D tidal circulation. In: *Proceedings of the 7-th Intern. Conf. on Estuarine and Coastal Modeling* (St. Petersburg, FL), pp. 628–642.

CHU, P.C., FAN, C.-W. (1997). Sixth-order difference scheme for sigma-coordinate ocean models. *J. Phys. Ocean.* **27**, 2064–2071.

CHU, P.C., FAN, C.-W. (2003). Hydrostatic correction for sigma coordinate ocean models. *J. Geophys. Res.* **106** (C6), 3206, doi: 10.1029/2002JC001668.

DEWAR, W.K., HSUEH, Y., MCDOUGALL, T.J., YUAN, D.I. (1998). Calculation of pressure in ocean simulations. *J. Phys. Ocean.* **28**, 577–588.

DIETRICH, D.E., LIN, C.A., MESTAS-NUNEZ, A., KO, D.-S. (1997). A high resolution numerical study of Gulf of Mexico fronts and eddies. *Meteorol. Atmos. Phys.* **64**, 187–201.

DIETRICH, D.E., MARIETTA, M.G., ROACHE, P.J. (1987). An ocean modeling system with turbulent boundary layers and topography, part 1: numerical description. *Int. J. Numer. Methods Fluids*. **7**, 833–855.

DUKOWITZ, J.K., SMITH, R.D. (1994). Implicit free-surface method for the Bryan-Cox-Semtner ocean model. *J. Geophys. Res.* **99**, 7991–8014.

DUKOWICZ, J.K. (2001). Reduction of density and pressure gradient errors in ocean simulations. *J. Phys. Ocean.* **31**, 1915–1921.

DUKOWICZ, J.K. (2006). Structure of the barotropic mode in layered ocean models. *Ocean Model.* **11**, 49–68.

DURRAN, D.R. (1991). The third order Adams-Bashforth method: an attractive alternative to leapfrog time differencing. *Mon. Wea. Rev.* **119**, 702–720.

DURRAN, D.R. (1998). *Numerical Methods for Wave Equations in Geophysical Fluid Dynamics* (Springer-Verlag), pp. 465.

GENT, P.R., MCWILLIAMS, J.C. (1990). Isopycnal mixing in ocean circulation models, *J. Phys. Ocean.* **20**, 150–155.

GOLDBERG, M., TADMOR, E. (1988). Simple stability criteria for difference approximations of hyperbolic initial-boundary value problems. In: Ballmann, J., Jeltsch, R. (eds.), *Nonlinear Hyperbolic Equations - Theory, Computation Methods, and Applications*, Proceedings of the Second International Conference on Nonlinear Hyperbolic Problems, Notes on Numerical Fluid Mechanics, vol. 24 (Vieweg Verlag), pp. 179–185.

GREATBATCH, R.J., LU, Y., CAI, Y. (2001). Relaxing the Boussinesq approximation in ocean circulation models. *J. Atmos. Oceanic Technol.* **18**, 1911–1923.

GREATBATCH, R.J., MCDOUGALL, T.J. (2003). The non-Boussinesq temporal residual mean. *J. Phys. Ocean.* **33**, 1231–1239.

GRIFFIES, S.M., GNANADESIKAN, A., PACANOWSKI, R.C., LARICHEV, V.D., DUKOWICZ, J.K., SMITH, R.D. (1998). Isoneutral diffusion in a z-coordinate ocean model. *J. Phys. Ocean.* **28**, 805–830.

GRIFFIES, S.M., BÖNING, C., BRYAN, F.O., CHASSIGNET, E.P., GERDES, R., HASUMI, H., HIRST, A., TREGUIER, A.-M., WEBB, D. (2000). Development in ocean climate modeling. *Ocean Model.* **2**, 123–192.

GRIFFIES, S.M., PACANOWSKI, R.C., SCHMIDT, M., BALAJI, V. (2001). Tracer conservation with an explicit free surface method for z-coordinate ocean models. *Mon. Wea. Rev.* **5**, 1081–1098.

GROPP, W., LUSK, E., SKJELLUM, A. (1999). Using MPI, 2nd Edition: portable Parallel Programming with the Message Passing Interface. MIT Press In Scientific And Engineering Computation Series, Cambridge, MA, USA. pp. 395.

HAIDVOGEL, D.B., ARANGO, H., HEDSTROM, K., BECKMANN, A., RIZZOLI, P., SHCHEPETKIN, A.F. (2000). Model evaluation experiments in the North Atlantic Basin: simulations in non-linear terrain-following coordinates. *Dyn. Atmos. Oceans*. **32**, 239–281.

HALLBERG, R. (1997). Stable split time stepping schemes for large-scale ocean modeling. *J. Comp. Phys.* **135**, 54–65.

HALLBERG, R. (2005). A thermobaric instability of Lagrangian vertical coordinate ocean models. *Ocean Model.* **8**, 279–300.

HANEY, R.L. (1991). On the pressure gradient force over steep topography in sigma coordinate ocean models. *J. Phys. Ocean.* **21**, 610–618.

HIGDON, R.L., BENNETT, A.F. (1996). Stability analysis of operator splitting for large-scale ocean modeling. *J. Comp. Phys.* **123**, 311–329.

HIGDON, R.L., DE SZOEKE, R.A. (1997). Barotropic-baroclinic time splitting for ocean circullation modeling. *J. Comp. Phys.* **135**, 31–53.

HIGDON, R.L. (1999). Implementation of barotropic-baroclinic time splitting for isopycnic-coordinate ocean modeling. *J. Comp. Phys.* **148**, 579–604.

HIGDON, R.L. (2002). A two-level time-stepping method for layered ocean circulation models. *J. Comp. Phys.* **177**, 59–94.

HIGDON, R.L. (2005). A two-level time-stepping method for layered ocean circulation models: further development and testing. *J. Comp. Phys.* **206**, 463–504.

HOLLAND, W.R. (1973). Baroclinic and topographic influences on the transport in western boundary currents. *Geophys. Fluid Dyn.* **4**, 187–210.

JACKETT, D.R., MCDOUGALL, T.J. (1995). Minimal adjustment of hydrostatic profiles to achieve static stability. *J. Atmos. Oceanic Technol.* **12**, 381–389.

JACKETT, D.R., MCDOUGALL, T.J., FEISTEL, R., WRIGHT, D.G., GRIFFIES, S.M. (2006). Algorithms for density, potential temperature, conservative temperature and freezing temperature of seawater. *J. Atmos. Oceanic Technol.* Submitted.

Kanarska, Y., Shchepetkin, A.F., McWilliams, J.C. (2007). Algorithm for non-hydrostatic dynamics in ROMS. *Ocean Model.* **18**, 143–174, doi: 10.1016/j.ocemod.2007.04.01.

Killworth, P.D., Stainforth, D., Webb, D.J., Paterson, S.M. (1991). The development of a free-surface Bryan-Cox-Semtner ocean model. *J. Phys. Ocean.* **21**, 1333–1348.

Kliem, N., Pietrzak, J.D. (1999). On the pressure gradient errors in sigma coordinate ocean models: a comparison with laboratory experiments. *J. Geophys. Res.* **104**, 29781–29799.

Leonard, B.P., Lock, A.P., McVean, M.K. (1996). Conservative explicit unrestricted-time-step constancy preserving advection schemes. *Mon. Wea. Rev.* **124**, 2588–2606.

Lilly, D.K. (1965). On the computational stability of time-dependent non-linear geophysical fluid dynamics problem. *Mon. Wea. Rev.* **93**, 11–26.

Lin, S.J. (1997). A finite volume integration method for computing pressure gradient force in general vertical coordinates. *Q. J. R. Meteorol. Soc.* **123**, 1749–1762.

Liu, H., Tadmor, E. (2003). Critical thresholds and conditional stability for Euler equations and related models. In: Hou, T., Tadmor, E. (eds.), *"Hyperbolic Problems: Theory, Numerics, Applications", Proceedings of the 9th International Conference on Hyperbolic Problems* (Springer-Verlag, Pasadena, CA) pp. 227–240.

Losch, M., Adcroft, A., Campin, J.-M. (2004). How sensitive are coarse general circulation models to fundamental approximations in the equations of motion? *J. Phys. Ocean.* **34**, 306–319.

Lu, Y. (2000). Including non-Boussinesq effect in Boussinesq ocean circulation models. *J. Phys. Ocean.* **31**, 1616–1622.

Marchesiello, P., McWilliams, J.C., Shchepetkin, A. (2003). Equilibrium structure and dynamics of the California Current System. *J. Phys. Ocean.* **33**, 753–783.

Marsaleix, P., Auclair, F., Herrmann, M.-J., Estournel, C., Pairaud, I., Ulses, C. (2008). Energy conservation in sigma-coordinate free-surface ocean models. *Ocean Model.* **20**, 61–89, doi: 10.101/j.ocemod.2007.07.005.

Marshall, J., Adcroft, A., Hill, C., Perelman, L., Heisey, C. (1997). A finite-volume, incompressible Navier Stokes model for studies of the ocean on parallel computers. *J. Geophys. Res.* **102**, 5753–5766.

McDougall, T.J., Garett, C.J.R. (1992). Scalar conservation equations in a turbulent ocean. *Deep-Sea Res.* **39**, 1953–1966.

McDougall, T.J., Greatbatch, R.J., Lu, Y. (2002). On conservation equations in oceanography: how accurate are Boussinesq ocean models? *J. Phys. Ocean.* **32**, 1574–1584.

McDougall, T.J., Jackett, D.J., Wright, D.G., Feistel, R. (2003). Accurate and computationally efficient algorithms for potential temperature and density and of seawater. *J. Atmos. Ocean. Technol.* **20**, 730–741.

McWilliams, J.C. (1996). Modeling the oceanic general circulation. *Annu. Rev. Fluid Mech.* **28**, 1–34.

Mellor, G.L. (1991). An equation of state for numerical modeling of oceans and estuaries. *J. Atmos. Ocean. Technol.* **1991**, 609–611.

Mellor, G.L., Ezer, T., Oey, L.-Y. (1994). The pressure gradient conundrum of sigma coordinate ocean models. *J. Atmos. Ocean. Technol.* **11**, 1126–1134.

Mellor, G.L., Oey, L.-Y., Ezer, T. (1998). Sigma coordinate pressure gradient errors and the seamount problem. *J. Atmos. Ocean. Technol.* **15**, 1122–1131.

Mesinger, F. (1982). On the convergence and error problems of the calculation of the pressure gradient force in sigma coordinate ocean models. *Geophys. Astrophys. Fluid Dyn.* **19**, 105–117.

Mesinger, F., Janjic, Z.I. (1985). Problems and numerical methods of the incorporation of mountains in atmospheric models. *Lect. Appl. Math.* **22**, 81–120.

Nadiga, B.T., Hecht, M.W., Margolin, L.G., Smolarkiewicz, P.K. (1997). On simulating flows with multiple time scales using a method of averages. *Theor. Comput. Fluid Dyn.* **9**, 281–292.

Oppenheim, A.V., Schafer, R.W., Buck, J.R. (1999). *Discrete-Time Signal Processing*, 2nd Edition, Prentice-Hall, pp. 870.

Orszag, S.A. (1971). Numerical simulation of incompressible flows with simple boundaries: accuracy. *J. Fluid Mech.* **49**, 76–112.

Robinson, R., Padman, L., Levine, M.D. (2001). A correction to the baroclinic pressure gradient term in the Princeton ocean model. *J. Atmos. Ocean. Technol.* **18**, 1068–1075.

Roe, P.L. (1981). Approximate Riemann solvers, parameter vectors and difference schemes. *J. Comp. Phys.* **43**, 357–372.

Rueda, F.J., Sanmiguel-Rojas, E., Hodges, B.R. (2007). Baroclinic stability for a family of two-level, semi-implicit numerical methods for the 3D shallow water equations. *Int. J. Numer. Meth. Fluids* **54**, 237–268, doi: 10.1002/fld.1391.

Shchepetkin, A.F., McWilliams, J.C. (1998). Quasi-monotone advection schemes based on explicit locally adaptive dissipation. *Mon. Wea. Rev.* **126**, 1541–1580.

Shchepetkin, A.F., McWilliams, J.C. (2003). A method for computing horizontal pressure-gradient force in an oceanic model with a non-aligned vertical coordinate. *J. Geophys. Res.* **108** (C3), 3090, doi: 10.1029/2001JC001047 35.1-35.34.

Shchepetkin, A.F., McWilliams, J.C. (2005). The regional ocean modeling system: a splitexplicit, free-surface, topography-following-coordinate oceanic model. *Ocean Model.* **9**, 347–404, doi: 10.106/j.oceanmod.2004.08.002.

Shulman, I., Lewis, J.K., Mayer, J.G. (1999). Local data assimilation in the estimation of barotropic and baroclinic open boundary conditions. *J. Geophys. Res.* **104** (C6), 13667–13680.

Slordal, L.H. (1997). The pressure gradient force in sigma coordinate ocean models. *Int. J. Numer. Methods Fluids*. **24**, 987–1017.

Skamarock, W.C., Klemp, J.B. (1992). The stability of time-split numerical methods for the hydrostatic and the nonhydrostatic elastic equations. *Mon. Wea. Rev.* **120**, 2109–2197.

Song, Y.T. (1998). A general pressure gradient formulation for ocean models. Part I: scheme design and diagnostic analysis. *Mon. Wea. Rev.* **126**, 3213–3230.

Song, Y.T., Wright, D.G. (1998). A general pressure gradient formulation for ocean models. Part II: energy, momentum and bottom torque consistency. *Mon. Wea. Rev.* **126**, 3213–3230.

Stelling, G.S., Duinmeijer, S.P.A. (2003). A staggered conservative scheme for every Froude number in rapidly varied shallow water flows. *Int. J. Numer. Meth. Fluids*. **43**, 1329–1354, doi: 10.1002/fld.537.

Stelling, G.S., van Kester, J.A.Th. (1994). On the approximation of horizontal gradients in sigma coordinates for bathymetry with steep bottom slopes. *Int. J. Numer. Meth. Fluids*. **18**, 915–935.

de Szoeke, R.A., Samelson, R.M. (2002). The duality between the Boussinesq and non-Boussinesq hydrostatic equations of motion. *J. Phys. Ocean.* **32**, 2194–2203.

Willebrand, J., Barnier, B., Böning, C., Dieterich, C., Killworth, P.D., LeProvost, C., Jia, Y., Molines, J.M., New, A.L. (2001). Circulation characteristics in three eddy-permitting models of the North Atlantic. *Prog. Oceanogr.* **48**, 123–161.

Bifurcation Analysis of Ocean, Atmosphere, and Climate Models

Eric Simonnet, Henk A. Dijkstra, Michael Ghil

Institut Non Linéaire de Nice, UMR6612, CNRS, France. Institute for Marine and Atmospheric research Utrecht, Utrecht, The Netherlands. Département Terre-Atmosphère-Océan, Ecole Normale Supérieure, and Laboratoire de Météorologie Dynamique (CNRS and IPSL), Paris, France, and Atmospheric and Oceanic Sciences Department and Institute of Geophysics and Planetary Physics, University of California, Los Angeles, USA

Abstract

We provide an overview of numerical bifurcation methods and how they can help one understand the variability of large-scale oceanic and atmospheric flows, as well of the climate system as a whole. As a particular example, we consider the problem of how low-frequency variability of the wind-driven ocean circulation arises due to internal dynamics. The methods are illustrated with the help of a hierarchy of models of so-called double-gyre, wind-driven flows.

1. Introduction

The climate system is highly complex. It is composed of many subsystems, each with its own characteristic time scales, spatial scales, and physical complexities. To better describe, understand, and predict how the climate is evolving requires an in-depth understanding of these subsystems and how they interact with each other. Up until today, even the most sophisticated, high-resolution, global earth system models are unable to provide accurate and reliable simulations of climate change, subject to both natural and anthropogenic forcing. Although it is now possible to predict some general tendencies of atmospheric and oceanic fields over several decades, many quantitative as well as qualitative aspects of climate change are still far from being well understood. The main difficulties are related not only to the turbulent behavior of the smaller scales of motion,

Computational Methods for the Atmosphere and the Oceans
Special Volume (Roger M. Temam and Joseph J. Tribbia, Guest Editors) of
HANDBOOK OF NUMERICAL ANALYSIS, VOL. XIV
P.G. Ciarlet (Editor)
Copyright © 2009 Elsevier B.V.
All rights reserved
ISSN 1570-8659
DOI 10.1016/S1570-8659(08)00203-2

and their effects on the large scales, but also to a lack of understanding of the physics driving the whole system.

A key prerequisite to greater scientific insight into the climate system as a whole is the availability of a hierarchy of climate models, from the highly simplified to the very detailed. Relying on such a model hierarchy is probably the most reasonable strategy for understanding fundamental aspects of climate change (Dijkstra and Ghil [2005], Ghil and Robertson [2000], Held [2005], Schneider and Dickinson [1974]). This strategy aims at increasing our knowledge of the system from its most basic to its highly applied aspects; it does so by studying, on the one hand, each subsystem in isolation, from the bottom up, while on the other, it considers increasingly complex versions of the whole system in combination.

The case of the oceans is particularly illuminating. The simplest zero-dimensional ocean models are represented by a small number of ordinary differential equations (ODEs), typically O(10) at most, which are referred to as box models. They are, for example, used to study the stability of the oceans' thermohaline circulation (Stommel [1961]). One-dimensional partial differential equation (PDE) models can also be used to study the vertical structure of the upper ocean alone or the oceanic mixed layer.

Two-dimensional PDE models of the oceans correspond to models that resolve either the two horizontal coordinates (x, y), like in traditional studies of the wind-driven ocean circulation (Jiang, Jin and Ghil [1995], McCalpin and Haidvogel [1996]) (see also Section 4.1, Dijkstra [2005], Dijkstra and Ghil [2005], pp. 532) or a vertical versus an horizontal coordinate (y, z), like in studies of the thermohaline circulation (Cessi and Young [1992], Quon and Ghil [1992, 1995], Thual and McWilliams [1992]). Finally, general circulation models (GCMs) or primitive equation models (PEs) are essentially 3D models that are composed of several subsystems coupled with each other. These models, such as coupled ocean–atmosphere models, include both dynamical and thermodynamical processes. Results from GCMs are, in general, difficult to interpret without the knowledge accumulated from the study of lower dimensional models, but are essential for understanding climate change.

The hierarchy of models outlined above must not hide the fact that complexity increases not only in the dimensionality of the models but in several directions at once. This is especially true when one considers PDE models. Apart from the physical processes, which are already included, one may consider two other directions of increasing complexity. First, as one increases the spatial resolution of a PDE model, the effects of the unresolved, turbulent small scales on the resolved, large scales become crucial; sophisticated techniques to parametrize these subgrid-scale effects are part of the success of any GCM. Indeed, as one increases the spatial resolution, new phenomena may appear which were completely ignored or neglected before. A recent, and striking example can be given regarding the influence of oceanic fronts on the atmosphere above, where new instabilities emerge in eddy-resolving 2D models (Feliks, Ghil and Simonnet [2004, 2007]). These instabilities are not found in coarse-resolution GCMs (see also Chelton and Wentz [2005]). Second, the geometry of the problem–e.g., the shape of the domain and the spatial structure of the forcing–is also an important factor that

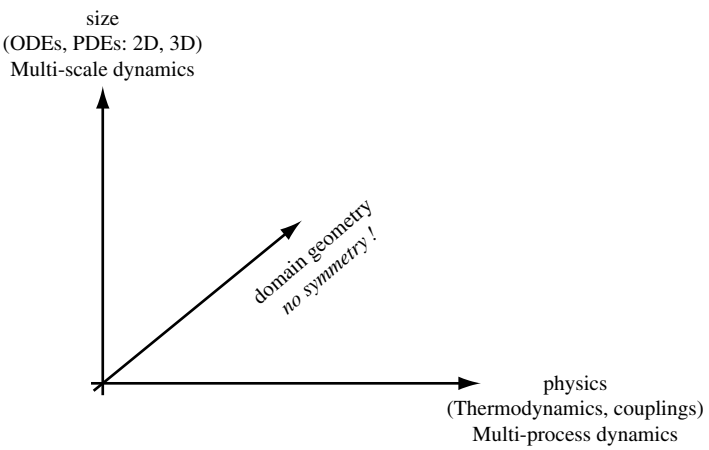


FIG. 1.1 Directions of increasing complexity in a hierarchy of models.

cannot be neglected and plays a key role in bringing the model problem closer to reality. Figure 1.1 illustrates the three axes of increasing complexity described herein.

In confronting modeling results with observations, one has to realize that it is the largest scales that are best and most reliably captured. The variability of the large scales arises from two sources: (i) the competition among the finite-amplitude instabilities and (ii) the net effects of the smaller scales. In many situations, in the atmosphere as well as in the ocean, the analysis of climate data shows that the large-scale flows exhibit quasi-periodic behavior, with strongly localized spectral peaks emerging above the background red-noise spectrum. This feature might indicate that the weakly turbulent large scales are dominated by low-dimensional dynamics; in other words, at specific scales, some form of approximate closure might exist. A typical example is given by the low-frequency dynamics of the wind-driven ocean circulation, where certain large-scale instabilities are almost unaffected by the small-scale eddies, due to a significant spectral gap between these scales. These instabilities are related to the existence of homoclinic cycles and strange attractors of very low dimension (SIMONNET, GHIL and DIJKSTRA [2005, 2007]). A more detailed description is given in Section 4.

Dynamical systems theory thus appears to provide a good set of tools to study the large-scale dynamics and in particular to address the first source of variability above. This theory is particularly efficient at explaining how chaotic behavior of small-dimensional systems emerges and the phase transitions and bifurcations that might lead from simple to more complex behavior. The theory is most fully developed for systems with a finite number of degrees of freedom. However, the remarks above might give the reader some hint of why it appears very promising in the context of climate studies. We will also comment in the last section on the second source of variability, i.e., the effects of the small scales on the large ones.

Most of the PDE systems that govern climate dynamics are dissipative and possess finite-dimensional attractors as well as nice regularity properties. A specific solution

of such a system – for given initial and boundary data and a fixed set of parameter values – may be considered as a single point in a suitably defined function space. Such a solution evolves along a deterministic trajectory and belongs to an associated semigroup of solutions for different initial data. In general, these solutions asymptotically converge to a global attractor of finite but still fairly large dimension (TEMAM [1997], pp. 643). An important aspect of climate studies is to probe these attractors so as to extract their topological and metric properties. This approach can provide some predictive answers to the following questions:

1. What are the generic transition routes to chaos in these systems?
2. How well do these routes explain the climate system's observed behavior?
3. How sensitive is this behavior to natural or anthropogenic perturbations?

We cannot present, in the allotted space, a global method to answer all these questions but will concentrate instead on some numerical techniques to succeed in a less ambitious although highly pertinent quest: how to extract local properties of the PDEs attractors? This presentation will mostly focus on bifurcation theory and its numerical applications.

Our aim is to describe numerical algorithms for computing in a rather simple way the branches of steady states and periodic orbits of geophysical flows that are governed by systems of nonlinear PDES. A by-product of these algorithms is to help detect the bifurcations of these steady and periodic solutions. Bifurcation theory is outlined in Section 2. Its applications to the numerical computation of branches of steady states and periodic orbits are given in Section 3. In Section 4, we describe a typical example of the application of these techniques to a hierarchy of models of the wind-driven ocean circulation. An outlook on future developments is provided in Section 5.

2. Bifurcation theory

Bifurcation theory studies changes in the qualitative behavior of a dynamical system as one or several of its parameters vary. The results of this theory permit one to follow systematically climatic behavior from the simplest kind of model solutions to the most complex, from single to multiple equilibria, periodic, chaotic to fully turbulent solutions. Through our approach, we mainly aim to detect local bifurcations, i.e., in situations where the only knowledge of the leading eigenvalues of the linearized equations is enough to conclude on the phase-space dynamics. We will also give some examples of global bifurcations such as homoclinic ones that are surprisingly common in climate models (GHIL and CHILDRESS [1987], pp. 485, CROMMELIN [2002], MEACHAM [2000], NADIGA and LUCE [2001]).

Here, we sketch the basic concepts of bifurcation theory for a general system of autonomous ODEs of dimension n that can be written as

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{f}(\mathbf{x}, \mathbf{p}). \tag{2.1}$$

Here $\mathbf{x}$ is the state vector in the phase space $\mathbf{R}^n$ which typically corresponds to the discretized solutions of the PDE models using either finite differences, finite elements or

spectral methods. Although we do not assume here that the number of degrees of freedom n is large, it must be kept in mind that it is of the order of $O(10^4)$ for 2D PDEs and $O(10^{5-6})$ for 3D models. The vector $\mathbf{p}$ is the parameter vector in $\mathbf{R}^p$, where p is usually much smaller than n and for now on we choose $p = 1$, i.e., we consider bifurcations that appear through variation of one parameter only; these bifurcations are usually referred to as codimension-1 bifurcations. The right-hand side $\mathbf{f}$ contains the model dynamics including the forcing terms, and it depends on $\mathbf{x}$ in a nonlinear fashion, at time t. We also restrict here the discussion to autonomous systems, i.e., $\mathbf{f}$ does not depend explicitly on time. A trajectory of the dynamical system, starting at the initial state $\mathbf{x}(t_0) = \mathbf{x}_0$, is a curve $\Gamma = \{\mathbf{x}(t); -\infty < t < \infty\}$ in $\mathbf{R}^n$ which satisfies Eq. (2.1). The solution at time t with initial condition $\mathbf{x}_0$ will be indicated later on $\phi_t(\mathbf{x}_0)$. In the following, we will refer to the solution $\bar{\mathbf{x}}$ of

$$\mathbf{f}(\bar{\mathbf{x}}, \bar{\mathbf{p}}) = 0, \tag{2.2}$$

as a fixed point or equilibrium for which $\mathbf{x}(t) = \bar{\mathbf{x}}$ for all t, or also a steady state in the context of PDEs.

Linear stability analysis of a particular fixed point $(\bar{\mathbf{x}}, \bar{\mathbf{p}})$ amounts to consider infinitesimally small perturbations $\mathbf{y}$, i.e.,

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{y}; \tag{2.3}$$

linearization of Eq. (2.1) around $\bar{\mathbf{x}}$ then gives

$$\frac{\mathrm{d}\mathbf{y}}{\mathrm{d}t} = \mathbf{J}(\bar{\mathbf{x}}, \bar{\mathbf{p}})\mathbf{y}, \tag{2.4}$$

where $\mathbf{J}$ is the Jacobian matrix

$$\mathbf{J} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \cdots & \cdots & \cdots \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_n} \end{pmatrix}. \tag{2.5}$$

The linear autonomous ODE system (Eq. (2.4)) has solutions of the form $\mathbf{y}(t) = \mathrm{e}^{\sigma t}\hat{\mathbf{y}}$. Substituting such a solution into Eq. (2.4) leads to an eigenvalue problem for the complex growth factor $\sigma = \sigma_r + i\sigma_c$, where $i^2 = -1$, namely

$$\mathbf{J}(\bar{\mathbf{x}}, \bar{\mathbf{p}})\hat{\mathbf{y}} = \sigma\hat{\mathbf{y}}. \tag{2.6}$$

Fixed points for which there are eigenvalues with $\sigma_r > 0$ are unstable since the associated perturbations are exponentially growing, whereas fixed points for which $\sigma_r < 0$ are linearly stable. In the situation where $\sigma_c \neq 0$, the associated eigenmodes will be oscillatory with frequency σ_c, i.e., with a characteristic period of $2\pi/\sigma_c$. The eigenspaces associated with eigenvalues $\sigma_r > 0$ (resp. $\sigma_r < 0$) will be denoted E^u (resp. E^s), whereas the eigenspace associated with eigenvalues $\sigma_r = 0$ will be denoted E^c.

Before enumerating the various possible codimension-1 bifurcations, we briefly recall some basics and essential notions of dynamical systems theory.

THEOREM 2.1. (Hartman–Grobman). *If* $\mathbf{J}(\bar{\mathbf{x}}, \bar{\mathbf{p}})$ *has no purely imaginary eigenvalues, the number of eigenvalues with positive and negative real parts determine the topological equivalence (homeomorphism) of the flow near* $\bar{\mathbf{x}}$.

This result roughly says that the local structure of the flow near the fixed point is structurally stable. In such a case, $\bar{\mathbf{x}}$ is called an *hyperbolic* fixed point. Bifurcations correspond precisely to situations where this theorem does not apply, i.e., where the flow is not structurally stable. The *stable* (W^s) and *unstable* (W^u) *manifolds* are defined as

$$\begin{aligned} W^s(\bar{\mathbf{x}}) &= \{\mathbf{x} \mid \lim_{t\to+\infty} \phi_t(\mathbf{x}) = \bar{\mathbf{x}}\}, \\ W^u(\bar{\mathbf{x}}) &= \{\mathbf{x} \mid \lim_{t\to-\infty} \phi_t(\mathbf{x}) = \bar{\mathbf{x}}\}. \end{aligned} \tag{2.7}$$

These manifolds are unique and are tangent to the corresponding eigenspaces E^u and E^s of the linearized system (Eq. (2.6)) at $\bar{\mathbf{x}}$. The situation where bifurcations occur, i.e., there are n_0 eigenvalues with vanishing reals part, $\sigma_r = 0$, leads to the center manifold theorem.

THEOREM 2.2. *There exist unique stable and unstable manifolds* W^u *and* W^s *tangent to* E^u *and* E^s *at* $\bar{\mathbf{x}}$ *and a (nonunique) center manifold* W^c *tangent to* E^c *at* $\bar{\mathbf{x}}$. *The three manifolds are all invariant by the flow* ϕ_t.

The center manifold is in general associated with a loss of regularity contrary to the unstable and stable manifolds. Theorem 2.2 indeed implies that it is possible to reduce (locally) the dynamics on the center manifold, typically, taking $\bar{\mathbf{x}} = 0$ for simplicity, one has

$$\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t} = \mathbf{L}_0\mathbf{u} + \mathbf{N}(\mathbf{u}, \mathbf{p}), \tag{2.8}$$

where $\mathbf{N}$, which depends on the vector parameter $\mathbf{p}$, has a Taylor expansion starting with at least quadratic terms, $\mathbf{u}$ lives in $\mathbf{R}^{n_0}$ and $\mathbf{L}_0$ has n_0 eigenvalues with zero real part. Another well-known approach is the Lyapunov–Schmidt reduction that we do not present here (CHOSSAT and LAUTERBACH [2000], GOLUBITSKY, STEWART and SCHAEFFER [2000], pp. 533).

Having reduced the system in Eq. (2.1), into the system in Eq. (2.8), it is possible to find a change of coordinates (e.g., $\mathbf{u} = \mathbf{v} + H(\mathbf{v})$ for the system in Eq. (2.8) so that the system becomes as simple as possible. The resulting vector field thus obtained is called the *normal form*. It is an extension of the reduction to Jordan form for matrices to the nonlinear case. Normal form theory provides a way to classify the different kinds of bifurcations that may occur with only knowledge of the eigenvalues that lie on the imaginary axis. We do not present normal form theory in this short presentation, the reader may refer to GUCKENHEIMER and HOLMES [1990], pp. 453, IOOSS and JOSEPH [1999], pp. 356 for further details.

2.1. Local codimension-1 bifurcations of steady states

We are now ready to present the most important bifurcation cases, starting with the situation of a single zero eigenvalue ($n_0 = 1$). In this case, there are three important normal forms:

1. Saddle-node bifurcation: it corresponds to the case where the system (Eq. (2.8)) (when reduced to its normal form) is

$$\dot{u} = p \pm u^2. \tag{2.9}$$

   The sign characterizes supercriticality ($p - u^2$) or subcriticality ($p + u^2$). In the supercritical case, it is straightforward to check that the branch of solutions $u = \sqrt{p}$ is stable and the branch $u = -\sqrt{p}$ is unstable (see Fig. 2.1).
2. Transcritical bifurcation (see Fig. 2.2): the normal form corresponds in this case to

$$\dot{u} = pu \pm u^2. \tag{2.10}$$

3. Pitchfork bifurcation (symmetry breaking): the normal form is

$$\dot{u} = pu \pm u^3. \tag{2.11}$$

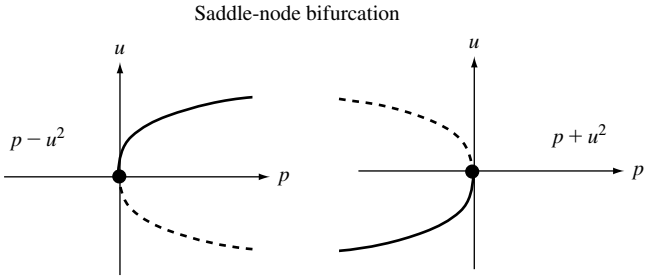


FIG. 2.1 Supercritical (left) and subcritical (right) saddle-node bifurcation. The solid (dotted) branches indicate stable (unstable) solutions.

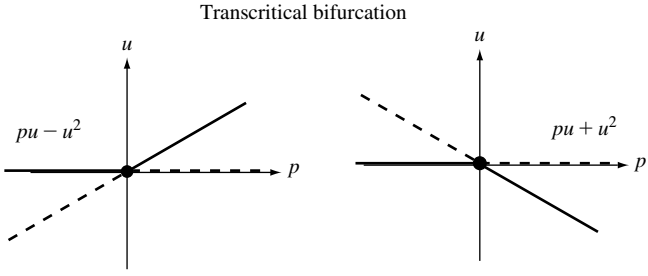


FIG. 2.2 Supercritical (left) and subcritical (right) transcritical bifurcation.

In the supercritical situation, there is a transfer of stability from the symmetric solution $u = 0$ to the pair of conjugated solutions $u = \pm\sqrt{p}$ (see Fig. 2.3). In the supercritical case, it must be noted that the system remains in a neighborhood of the equilibrium so that one observes a *soft* or *noncatastrophic* loss of stability. In the subcritical case, the situation is very different as can be seen in the left panel of Fig. 2.3. The domain of attraction of the fixed point is bounded by the unstable ones and shrinks as the parameter p approaches zero to disappear. The system is thus pushed out from the neighborhood of the now unstable fixed point leading to a *sharp* or *catastrophic* loss of stability. Decreasing again the parameter to negative values will not return the system to the previously stable equibrium since it may have already left its domain of attraction.

The importance of the saddle-node case is that all codimension-1 bifurcations with a zero eigenvalue can be perturbed to saddle-node bifurcations (see Fig. 2.4). These perturbations arise through imperfections of the physical system, domain geometry, boundary conditions, or forcing terms. The last two cases (pitchfork and transcritical) indeed illustrate situations where there is something special about the formulation of the problem: the pitchfork bifurcation is the rule for systems constrained by some (e.g., reflection) symmetry.

Whereas in the previous cases, the number of fixed points changed as the parameter was varied, it is also possible that a steady solution transfers its stability to a limit cycle. This kind of transition is called a Hopf bifurcation. It corresponds to the special case of a simple conjugate pair of pure imaginary eigenvalues $\sigma = \pm\omega i$ ($n_0 = 2$) crossing the imaginary axis. The normal form can be written in polar coordinate as

$$\begin{aligned} \dot{r} &= pr \pm r^3, \\ \dot{\theta} &= \omega. \end{aligned} \tag{2.12}$$

Again the sign determines whether the Hopf bifurcation is supercritical or subcritical, and the discussion is similar to the pitchfork bifurcation case. Figure 2.5 shows the phase portrait behavior for the supercritical case. For more complete details on Hopf bifurcations, see, e.g., MARSDEN and MCCRACKEN [1976].

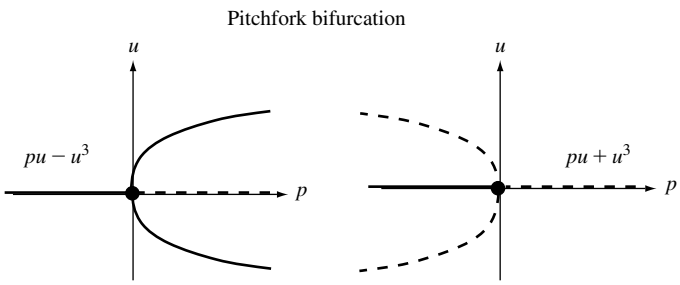


FIG. 2.3 Supercritical (left) and subcritical (right) pitchfork bifurcation.

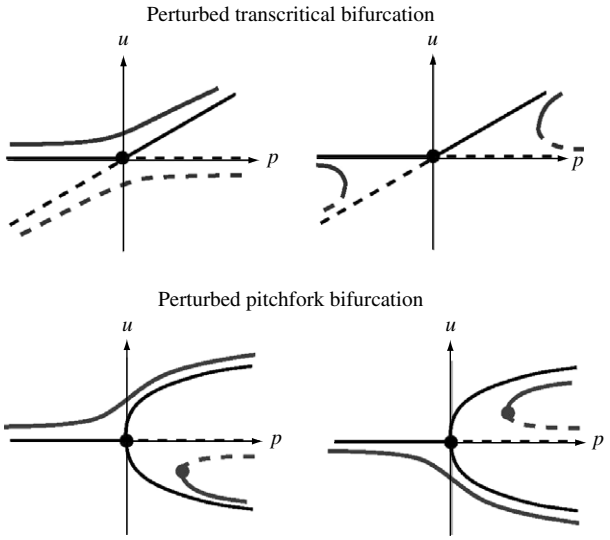


FIG. 2.4 Perturbed transcritical bifurcation (upper panel) and perturbed pitchfork bifurcation (lower panel). The left and right panels for each case are different possibilities arising from a perturbation of each normal form.

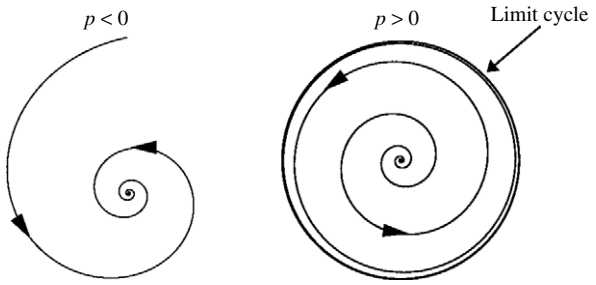


FIG. 2.5 Phase space trajectories associated with a supercritical Hopf bifurcation at $p = 0$. For $p < 0$, there is only one stable fixed point (left panel), whereas a stable limit cycle appears for $p > 0$ (right panel).

2.2. *Local codimension-1 bifurcations of limit cycles*

One may ask whether it is possible to apply the procedures described above to more complex (limit) sets. A very similar discussion applies for bifurcation of limit cycles although there are some additional complications. Let us assume that one has a limit cycle γ of the original system (Eq. (2.1)) for a parameter $\bar{p}$ that we omit in the notations for simplicity and whose corresponding solution is $\bar{\mathbf{x}}(t) = \bar{\mathbf{x}}(t + T)$. We consider an infinitesimal perturbation $\xi(t)$ of γ, i.e., we let $\mathbf{x}(t) = \bar{\mathbf{x}}(t) + \xi(t)$ in Eq. (2.1), and neglecting quadratic terms, one obtains

$$\dot{\xi} = \mathbf{J}(\bar{\mathbf{x}}(t))\xi, \tag{2.13}$$

and $\mathbf{J}(\bar{\mathbf{x}}(t))$ is now a T-periodic matrix. It can be shown that the fundamental solution matrix X of the system (Eq. (2.13)) can be written as $X(t) = Y(t)\mathrm{e}^{t\mathcal{R}}$, where $Y(t+T) = Y(t)$. One thus obtains that

$$X(t+T) = \mathbf{M}X(t),\ \mathbf{M} = \mathrm{e}^{T\mathcal{R}}. \tag{2.14}$$

The matrix $\mathbf{M}$ is called the *monodromy* matrix, and its eigenvalues $\sigma_1, \ldots, \sigma_n$ are called the *Floquet multipliers*. The monodromy matrix is not uniquely determined by the solutions of Eq. (2.13) but its eigenvalues are uniquely determined. Since the perturbation $\xi(t) = \bar{\mathbf{x}}(t+\epsilon) - \bar{\mathbf{x}}(t)$, ϵ small, is T-periodic, it immediately implies that $\mathbf{M}$ has an unit eigenvalue, i.e., perturbations along γ neither diverge or converge. The linear stability of γ is thus determined by the remaining $n-1$ eigenvalues.

Let Σ be a (fixed) local cross-section of dimension $n-1$ of the limit cycle γ such that the periodic orbit is not tangent to this hypersurface and denote $\mathbf{x}^\star$ the intersection of Σ with γ. There is a nice geometrical interpretation of the monodromy matrix in term of the Poincaré map defined as $\mathcal{P}(\mathbf{x}) = \phi_\tau(\mathbf{x})$, where $\mathbf{x}$ is assumed to be in a neighborhood of $\mathbf{x}^\star$ and τ is the time taken for the orbit $\phi_t(\mathbf{x})$ to first return to Σ (as $\mathbf{x}$ approaches $\mathbf{x}^\star$, τ will tend to T). After a change of basis such that the matrix $\mathbf{M}$ has a column $(0, \ldots 0, 1)^T$ corresponding to the unit eigenvalue, the remaining block $(n-1)\times(n-1)$ matrix corresponds to the linearized Poincaré map.

These remarks show that the bifurcations of limit cycles are related to the behavior of a discrete dynamical system (the Poincaré map)

$$x_{n+1} = \mathcal{P}x_n, \tag{2.15}$$

rather than a continuous dynamical system like in the case of fixed points. The bifurcation theory for fixed points of the iterative map with eigenvalue having unit norm is completely analogous to the bifurcation theory for equilibria with an eigenvalue on the imaginary axis. Periodic orbits become unstable when Floquet multipliers σ_i cross the unit circle as the parameter p is changed (remember that the Floquet multipliers depend on the parameter p). There are three important cases.

1. A real Floquet multiplier is crossing the unit circle $\sigma(\bar{p}) = 1$ (saddle-node). This situation can be shown to be topologically equivalent to the one-dimensional discrete dynamical system

   $$x_{n+1} = \mathcal{P}(x_n),\ \text{with } \mathcal{P}(x) = p + x \pm x^2. \tag{2.16}$$

   Let us consider the supercritical case $\mathcal{P}(x) = p + x - x^2$ and assume that $\bar{p} = 0$ for simplicity. As p becomes positive, two fixed points $x_1^\star$ and $x_2^\star$ of the iterative map (Eq. (2.16)) appear which are solutions of $\mathcal{P}(x) = x$. These two fixed points correspond to the appearance of two new families of periodic orbits. One family is stable ($\mathcal{P}'(x_1^\star) < 1$) while the other is unstable ($\mathcal{P}'(x_2^\star) > 1$). Like in the case of equilibria, particular constraints may lead to *transcritical* or *pitchfork* bifurcations (see Fig. 2.6).
2. A real Floquet multiplier is crossing the unit circle with $\sigma(\bar{p}) = -1$. This situation is called *flip* or *period-doubling* bifurcation and has no equivalent for equilibria.

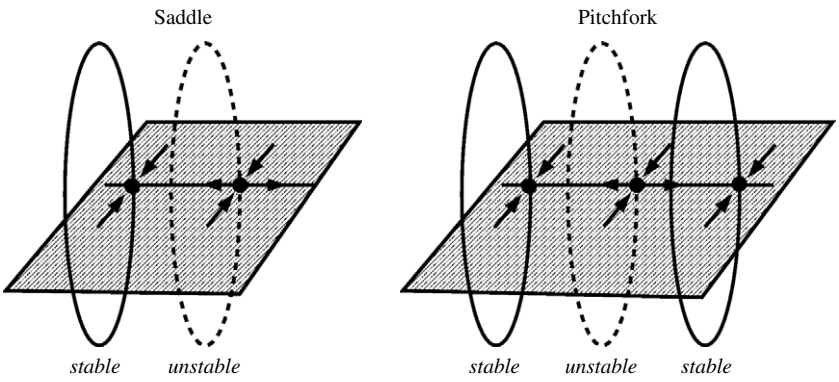


FIG. 2.6 Phase space view associated with the saddle-node (left panel) and pitchfork (right panel) bifurcations of periodic orbits.

The system is topologically equivalent to

$$x_{n+1} = \mathcal{P}(x_n), \text{ with } \mathcal{P}(x) = -(1+p)x \pm x^3. \tag{2.17}$$

This situation corresponds to the pitchfork case for the second iterate $\mathcal{P}^2$ map. Again consider (with $\bar{p} = 0$) the supercritical case $\mathcal{P}(x) = -(1+p)x + x^3$. As p becomes positive, two fixed points of the second iterate $\mathcal{P}^2$ appear which are not fixed points of the first iterate. This means that another stable periodic orbit of period $2T$ arises, whereas the original periodic orbit γ becomes unstable (see Fig. 2.7). The corresponding trajectories alternate from one side of γ to the other along the direction of the eigenvector associated with the eigenvalue $\sigma = -1$. The periodic orbit is twisted around the original periodic orbit like a Möbius band. The consequence is that this bifurcation cannot occur in a two-dimensional system since one cannot embed a Möbius band in a two-dimensional manifold.

3. The final example corresponds to the case of a pair of complex conjugate eigenvalues $\sigma, \bar{\sigma}$ crossing the unit circle such that $|\sigma(\bar{p})| = |e^{i\varphi}| = 1$. This bifurcation is called *Neimark–Sacker* or torus bifurcation. If one assumes after reduction on a two-dimensional invariant manifold that $d\sigma(p)/dp \neq 0$ at $p = \bar{p}$ and $\sigma^j(\bar{p}) \neq 1$ for $j = 1, 2, 3, 4$, then there is a change of coordinates such that the (Poincaré) map takes the following form in polar coordinates

$$\begin{aligned} \mathcal{P}_r(r,\theta) &= r + d(p-\bar{p})r + ar^3, \\ \mathcal{P}_\theta(r,\theta) &= \theta + \varphi + br^2, \end{aligned} \tag{2.18}$$

where a, b, and d are parameters. Provided $a \neq 0$, this normal form indicates, that a close curve generically bifurcates from the fixed point, this closed curve corresponds to a two-dimensional invariant torus. Note that the strong resonance cases $\sigma^j(\bar{p}) \neq 1$ for $j = 1$ and $j = 2$ correspond to the saddle-node and period-doubling bifurcations, and the two other cases $j = 3, 4$ may lead to the absence of a closed curve or even several invariant curves (see KUZNETSOV [1995], pp. 515).

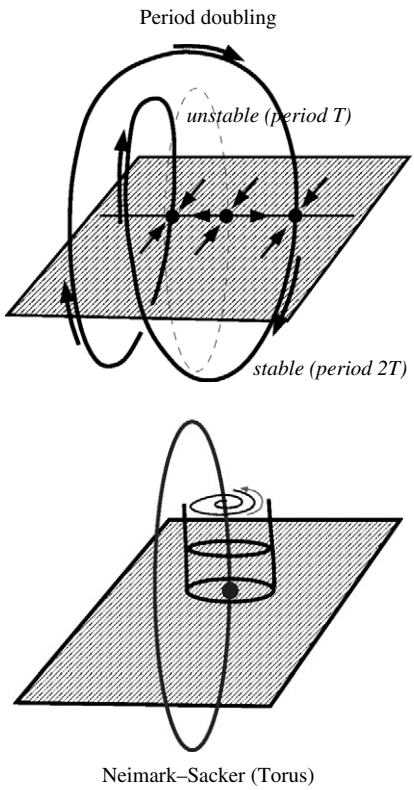


FIG. 2.7 Phase space view of a period-doubling or flip bifurcation of a periodic orbit (upper panel) and a Neimark–Sacker or torus bifurcation (lower panel).

2.3. *Homoclinic bifurcations of equilibria*

We now discuss an important situation where there is a global change of phase space as one parameter is varied which cannot be detected by local analysis of the dynamics in the neighborhood of equilibria (or limit cycles). This situation corresponds to a *global* bifurcation, and we focus here on the case of homoclinic orbits. This is a central issue in dynamical systems theory since global bifurcations in dynamical systems with $n > 2$ are in general responsible for the emergence of chaos and strange attractors. RUELLE and TAKENS [1971] introduced the notion of a strange attractor, in referring to a chaotic attractor characterized by sensitivity to the initial state and whose dimension is fractional, rather than integer.

Homoclinic orbits correspond to the interaction between an unstable fixed point and a periodic orbit. This is exemplified by Fig. 2.8 which corresponds to a planar homoclinic orbit to a saddle-node fixed point. The corresponding system is

$$\begin{aligned} \dot{u} &= -u + 2v + u^2, \\ \dot{v} &= (2-p)u - v - 3u^2 + \tfrac{3}{2}uv. \end{aligned} \tag{2.19}$$

Indeed, it can be shown that homoclinic bifurcations on the plane are determined by the saddle quantity $\kappa = \sigma_1 + \sigma_2$, where σ_i are the eigenvalues at the (saddle-type) fixed point at the homoclinic bifurcation ($p = \bar{p}$). The system (Eq. (2.19)) corresponds to $\kappa = -2 < 0$, with $\bar{p} = 0$. As both objects connect to each other, the period of the orbit becomes longer as p approaches zero to becomes infinite at $p = 0$. The trajectory along the unstable manifold of the fixed point approaches in infinite time the same fixed point along its stable manifold. The homoclinic orbit is then attracting the trajectories inside. For p positive, a stable periodic orbit (L in Fig. 2.8) is created.

Homoclinic bifurcations in systems with $n > 2$ are considerably more complicated. They most often induce stretching and folding in some bounded set of phase space and lead to the appearance of (Smale) horseshoes and possibly chaos. Typical horseshoes are given by the celebrated map sketched in Fig. 2.9. At the limit where this map is iterated an infinite number of time, a fractal Cantor set is obtained. In the neighborhood of the saddle fixed point where the homoclinic orbit is connected, the leading eigenvalues determine how much stretching and folding occurs in phase space. In (un)favorable cases, Poincaré

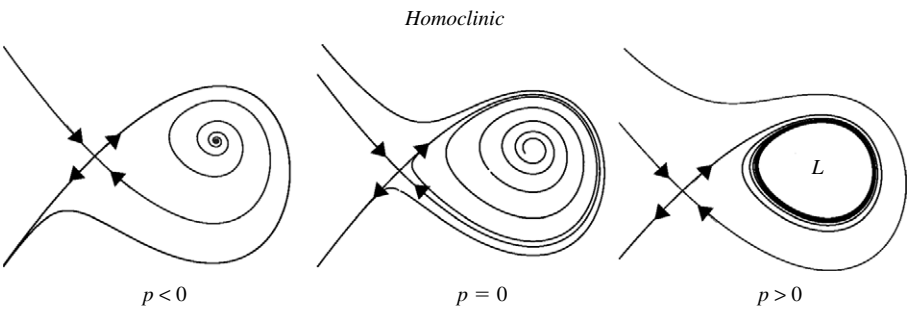


FIG. 2.8 Phase trajectories of the system (Eq. (2.19)) in the (u, v) plane.

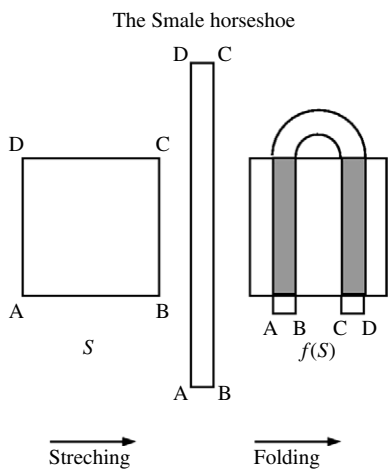


FIG. 2.9 One iteration of the Smale map f on the square S.

maps exhibit a similar iterative behavior than in the transformation (Eq. (2.10)) enabling the existence of horseshoes. In the following, we mention two important situations. Consider a three-dimensional system with two eigenvalues $\sigma_1, \sigma_2 < 0$ and one with $\sigma_3 > 0$ at the hyperbolic fixed point (saddle) and assume that homoclinic orbits exist for $p = \bar{p}$.

1. $\sigma_i \in \mathbf{R}$. In situations where there is only one homoclinic orbit, the system behaves more or less similarly to the planar case (Eq. (2.19)). In particular, a periodic orbit is created for $p \neq \bar{p}$ (WIGGINS [1988], pp. 494).

   The interesting case corresponds to the situation when two homoclinic orbits connect to the saddle point. In particular, when there is some $\mathbb{Z}_2$ symmetry in the system, it amounts to have only one parameter controlling the existence of the two homoclinic orbits at the same time. The symmetric case is of particular historical interest since this is precisely the situation that arises in the much studied Lorenz system (LORENZ [1963a], SPARROW [1983], pp. 269). Under particular conditions on the eigenvalues (e.g., $-\sigma_2 < \sigma_3 < -\sigma_1$) and invariance of the system by symmetry ($(x_1, x_2, x_3) \to (-x_1, x_2, -x_3)$), one observes a *one-sided homoclinic explosion*. This means that for, say, $p \leq \bar{p}$, there is nothing spectacular associated with the dynamics near the broken homoclinic orbit, but for $p > \bar{p}$, horseshoes appear seemingly out of nowhere. There exists a value of p, say p_0, such that for all p such that $\bar{p} < p < p_0$ the Poincaré map exhibits an invariant Cantor set (for p fixed, there is only one horseshoe). An infinite number of unstable periodic orbits of all possible periods are created. Although no strange attractor actually exists, these horseshoes are considered as being the chaotic heart of numerically observed strange attractors.

   We briefly illustrate the situation of the Lorenz system. This is a three-mode truncation of the Rayleigh–Bénard convection problem and corresponds to

$$\begin{aligned} \dot{x} &= s(y - x), \\ \dot{y} &= -xz + rx - y, \\ \dot{z} &= xz - bz, \end{aligned} \tag{2.20}$$

   with fixed parameters $s = 10$, $b = 10/3$, and control parameter r. As r increases, one observes the following bifurcations: for $0 < r < 1$, the origin is globally stable; at $r = 1$, there is a supercritical pitchfork bifurcation; at $r = 470/19$, there are subcritical asymmetric Hopf bifurcations, and the first homoclinic explosion is observed at $r = 13.93$. It must be noted that an important bifurcation occurs at $r = 24.06\ldots$, where the invariant Cantor set is destroyed through successive homoclinic explosions which yields to the existence of a genuine strange attractor. Note that a proof of its existence has only recently been given (TUCKER [2002]).
2. $\sigma_1 = \bar{\sigma}_2 = \rho \pm i\omega$ with $\rho < 0$ and $\sigma_3 > 0$ (saddle-focus). Let $\delta = -\rho/\sigma_3$. This case was first studied by SHILNIKOV [1965], and the dynamics is known as the *Shilnikov phenomenon*. It can be shown that, provided $\delta < 1$, there is a finite number of fixed points (and thus periodic orbits) for $p < \bar{p}$ and $p > \bar{p}$ and a countable infinity of fixed points for $p = \bar{p}$. Thus, contrary to the Lorenz bifurcation, the Shilnikov case is *two-sided*. As a matter of fact, at $p = \bar{p}$, the situation is

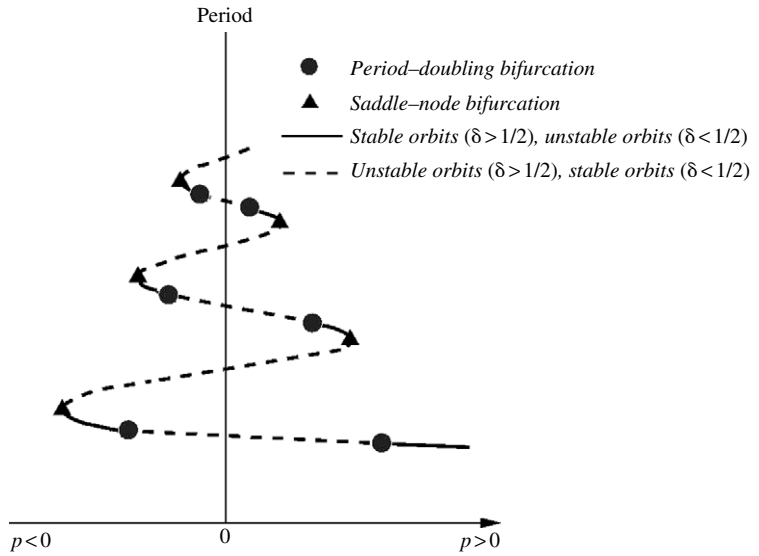


FIG. 2.10 Dependence of the period and stability of the bifurcated periodic orbits for the case $\delta < 1$. Here the critical parameter value $\bar{p}$ is set to zero for convenience.

considerably far more complicated than in the Lorenz case since there is a countable infinity of horseshoes in the neighborhood of the fixed point, each giving an infinite number of saddle fixed points. It must be noted that no symmetry is required contrary to the Lorenz case. To render the picture even more complicated, for $p \neq \bar{p}$, other homoclinic orbits appear and the Shilnikov phenomena are repeated for each of these new homoclinic orbits. We give in Fig. 2.10 the bifurcation behavior of the periodic orbits for the principal homoclinic orbit; such behavior is often referred to as *Shilnikov wiggles*. The reader may find detailed proofs in GLENDINNING and SPARROW [1984], GUCKENHEIMER and HOLMES [1990], pp. 453 and, WIGGINS [1988], pp. 494. We do not present here the case when the system possesses some symmetry; it is studied in GLENDINNING [1984].

For $\delta > 1$, the situation is simple since there are no horseshoes. Only one periodic orbit exists at one side of the critical parameter value (say $p > \bar{p}$) which becomes homoclinic at $p = \bar{p}$. When a reflection symmetry is present, this case is referred to as a *gluing bifurcation*.

3. Continuation methods

The idea of continuation methods is to compute branches of steady states as one parameter is varied. It enables one to obtain meaningful and generic information on the local dynamics of a PDE model for a large range of parameter values. Although time integrations of the model are ultimately needed to detect global bifurcations as well as statistical properties of the flow, continuation methods are able to determine the first

bifurcations off the branches of steady states in a rather cheap way. When interested in steady states, it naturally avoids to perform long-time integrations for several values of the parameter. It is fair to say that in the context of PDEs, continuation algorithms are the same as in the context of small-dimensional dynamical systems. The difficulty rather lies on an accurate estimation of the Jacobian, the solution of the nonlinear systems of equations, and its leading eigenvalues.

The basic technique involves two main steps. The first one allows one to advance along a branch of steady states as a parameter is varied one step at a time. The second one involves a linear stability analysis of the previously computed steady state using a generalized eigenvalue solver at each step. Before going to the details, we reformulated the system (Eq. (2.1)) in the context of PDEs, namely

$$\mathcal{M}\mathbf{u}_t = -\mathcal{L}\mathbf{u} - \mathcal{N}(\mathbf{u}) + \mathcal{F}, \tag{3.1}$$

where $\mathbf{u}(\mathbf{x}, t) \in \mathbf{R}^n$ corresponds to a discretized solution of the original PDE, $\mathcal{M}$ is the mass matrix, $\mathcal{L}$ the discretized linear operator, $\mathcal{N}(\mathbf{u}) : \mathbf{R}^n \to \mathbf{R}^n$ the (discretized) nonlinear operator, and $\mathcal{F} \in \mathbf{R}^n$ the discretized forcing term. Note that $\mathcal{M}$ is not invertible in general since we implicitly include the boundary conditions in the formulation (Eq. (3.1)). For instance, Dirichlet conditions correspond to $\mathcal{M}, \mathcal{N}(\mathbf{u})$ and $\mathcal{F}$ being zero on the boundary of the domain and $\mathcal{L}$ being the identity. Formulation (Eq. (3.1)) is thus rather general.

3.1. Pseudo-arclength methods

Finding steady states of the system (Eq. (3.1)) amounts to solve

$$\Phi(\mathbf{u}, p) = 0, \text{ where } \Phi(\mathbf{u}, p) = \mathcal{L}\mathbf{u} + \mathcal{N}(\mathbf{u}) - \mathcal{F}. \tag{3.2}$$

Pseudo-arclength methods (KELLER [1977], SEYDEL [1994]) are very often used. The idea is to parametrize branches of solutions $\Gamma(s) \equiv (\mathbf{u}(s), p(s))$ with an arclength parameter s (or an approximation of it, thus the term 'pseudo') and choose this parameter as the continuation parameter instead of p. An additional equation is obtained by approximating the normalization condition of the tangent $(\dot{\mathbf{u}}(s), \dot{p}(s)) \equiv \dot{\Gamma}(s)$ to the branch $\Gamma(s)$, where the dot refers to the derivative with respect to s, namely $|\dot{\Gamma}|^2 = 1$. This condition is approximated by

$$\dot{\mathbf{u}}_0^T(\mathbf{u} - \mathbf{u_0}) + \dot{p}_0(p - p_0) - \Delta s = 0, \tag{3.3}$$

where $(\mathbf{u}_0, p_0)$ is a known starting solution (or a previously computed point) on a particular branch, and Δs is the (small) steplength. In order to compute the tangent $\dot{\Gamma}_0(s)$, one differentiates $\Phi(\mathbf{u}(\mathbf{s}), p(s)) = 0$ with respect to s at $(\mathbf{u}_0, p_0)$ to find

$$\left[\Phi_{\mathbf{u}}^0, \Phi_p^0\right]\dot{\Gamma}_0(s) = \begin{pmatrix} \frac{\partial \Phi_1^0}{\partial u_1} & \cdots & \frac{\partial \Phi_1^0}{\partial u_n} & \frac{\partial \Phi_1^0}{\partial p} \\ & & & \\ \frac{\partial \Phi_n^0}{\partial u_1} & \cdots & \frac{\partial \Phi_n^0}{\partial u_n} & \frac{\partial \Phi_n^0}{\partial p} \end{pmatrix} \dot{\Gamma}_0(s) = 0, \tag{3.4}$$

where $\Phi^0_{\mathbf{u}}$ and Φ^0_p correspond to the derivatives evaluated at $(\mathbf{u}_0, p_0)$. If $(\mathbf{u}_0, p_0)$ is not a bifurcation point, then dim ker $\left[\Phi^0_{\mathbf{u}}, \Phi^0_p\right] = 1$ and therefore $\left[\Phi^0_{\mathbf{u}}, \Phi^0_p\right]$ has rank n. We may thus determine $\dot{\Gamma}_0(s)$ as the null vector of the $n \times (n+1)$ matrix $\left[\Phi_{\mathbf{u}}, \Phi_p\right]$. In practice, this can be done by upper triangulating the matrix $\left[\Phi_{\mathbf{u}}, \Phi_p\right]$ and solving a $(n+1) \times (n+1)$ extended system with a one in the right-lower corner and right-hand side. Then, the normalization condition $|\dot{\Gamma}_0|^2 = 1$ is used. A *predictor* solution of Eq. (3.3) is given by

$$\begin{aligned} \mathbf{u} &= \mathbf{u_0} + \Delta s \quad \dot{\mathbf{u}}_0, \\ p &= p_0 + \Delta s \quad \dot{p}_0. \end{aligned} \tag{3.5}$$

The next step is then to project the predictor solution $(\mathbf{u}, p)$ back onto the branch in a direction orthogonal to the tangent $\dot{\Gamma}_0$. This is called the *corrector* algorithm. It should rely on a robust nonlinear solver for the system (Eq. (3.2)). The one of common use is the Newton–Raphson method. This method has a quadratic convergence provided the initial starting solution is close enough to the solution. The *predictor* step just does that, all the more so as Δs is small. The initial guess will thus be the predictor solution (Eq. (3.5)). Newton–Raphson iterations with iteration index k can then be written as

$$\begin{pmatrix} \Phi_{\mathbf{u}}(\mathbf{u}^k, p^k) & \Phi_p(\mathbf{u}^k, p^k) \\ \dot{\mathbf{u}}_0^T & \dot{p}_0 \end{pmatrix} \begin{pmatrix} \Delta\mathbf{u}^{k+1} \\ \Delta p^{k+1} \end{pmatrix} = \begin{pmatrix} -\Phi(\mathbf{u}^k, p^k) \\ r_{n+1} \end{pmatrix}, \tag{3.6}$$

where $r_{n+1} = \Delta s - \dot{\mathbf{u}}_0^T(\mathbf{u}^k - \mathbf{u_0}) - \dot{p}_0(p^k - p_0)$.

Once $(\Delta\mathbf{u}^{k+1}, \Delta p^{k+1})$ is found, one sets

$$\begin{aligned} \mathbf{u}^{k+1} &= \mathbf{u}^k + \Delta\mathbf{u}^{k+1}, \\ p^{k+1} &= p^k + \Delta p^{k+1}. \end{aligned} \tag{3.7}$$

In practical situations, it is better to solve two $n \times n$ linear systems instead of solving directly Eq. (3.6), namely

$$\begin{aligned} \Phi_{\mathbf{u}}(\mathbf{u}^k, p^k)\mathbf{z}_1 &= -\Phi(\mathbf{u}^k, p^k), \\ \Phi_{\mathbf{u}}(\mathbf{u}^k, p^k)\mathbf{z}_2 &= \Phi_p(\mathbf{u}^k, p^k). \end{aligned} \tag{3.8}$$

Then, the solution of Eq. (3.6) is

$$\Delta p^{k+1} = \frac{r_{n+1} - \dot{\mathbf{u}}_0^T\mathbf{z}_1}{\dot{p}_0 - \dot{\mathbf{u}}_0^T\mathbf{z}_2}, \text{ and } \Delta\mathbf{u}^{k+1} = \mathbf{z}_1 - \Delta p^{k+1}\mathbf{z}_2. \tag{3.9}$$

The method is illustrated geometrically by Fig. 3.1. It must be noted that pseudo-arclength methods rely essentially on a correct estimation of the Jacobian matrix $\left[\Phi_{\mathbf{u}}, \Phi_p\right]$, and this can be done in at least two ways: explicitly 'by hand' or numerically using the approximation $\Phi_x \simeq (\Phi(x+\epsilon) - \Phi(x))/\epsilon$ for small ϵ. It is usually faster to do it 'by hand' provided one is cautious in treating the boundary conditions. Failure to do so in general prevents the Newton–Raphson method to converge quadratically.

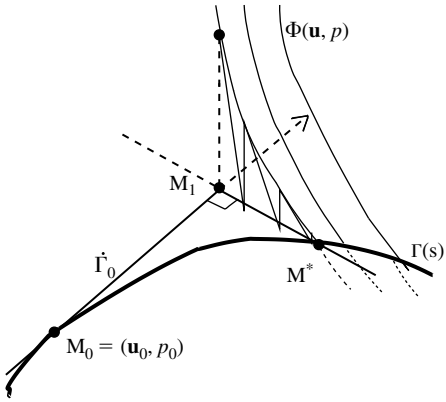


FIG. 3.1 Sketch of the pseudo-arclength method.

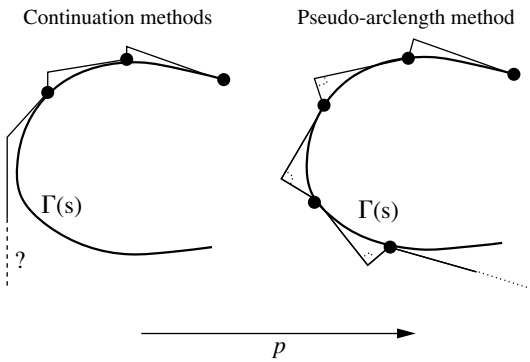


FIG. 3.2 Pseudo-arclength methods can go through saddle-node bifurcation. The direction of continuation is to the left, i.e., toward decreasing values of p.

The advantage of pseudo-arclength methods over traditional continuation methods is that the Jacobian of the extended system $\left[\Phi_{\mathbf{u}}, \Phi_p\right]$ has rank n even at saddle-node points where $\Phi_{\mathbf{u}}$ becomes singular. Hence, one can easily continue around saddle-node bifurcation as illustrated in Fig. 3.2. One other advantage is that it allows to compute branches of unstable as well as stable solutions.

3.2. *Linear stability problem*

In order to detect bifurcations of the steady states $(\bar{\mathbf{u}}, \bar{p})$ which have been previously computed by a pseudo-arclength method, one sets first $\mathbf{u} = \bar{\mathbf{u}} + \mathbf{v}$ in Eq. (3.1) to yield the linearized (discretized) PDE:

$$\mathcal{M}\mathbf{v}_t = -\mathcal{L}\mathbf{v} - \mathcal{N}_{\mathbf{u}}(\bar{\mathbf{u}})\mathbf{v}. \tag{3.10}$$

We look for solutions in the form $\mathbf{v} = \hat{\mathbf{v}}e^{\sigma t}$ so that one obtains the generalized eigenvalue problem:

$$\mathcal{A}\hat{\mathbf{v}} = \sigma\mathcal{B}\hat{\mathbf{v}}, \text{ with } \mathcal{A} = \mathcal{L} + \mathcal{N}_{\mathbf{u}}(\bar{\mathbf{u}}) \text{ and } \mathcal{B} = -\mathcal{M}. \tag{3.11}$$

The matrix $\mathcal{B}$ is usually singular due to the presence of (time-independent) boundary conditions and indeed whenever time derivatives are absent from the original PDE system. Like in the traditional approach presented in the Section 2, one is mostly interested in eigenvalues σ which cross the imaginary axis. The eigenvalues closest to the imaginary axis are often called *leading* eigenvalues. When n is very large, it is not possible to compute all the eigenvalues of the system (Eq. (3.11)), and one must rely on more sophisticated methods which are able to compute the leading ones. We present here a very efficient method called the simultaneous iteration technique (SIT) (Stewart and Jennings [1981]).

The idea is to apply a transformation mapping of the left complex half plane into the unit disk so that complex eigenvalues crossing the imaginary axis are transformed into complex eigenvalues leaving the unit disk. The complex transformation is

$$z \to z' = \frac{a - b + z}{a + b - z}, \; b \in \mathbf{R}, a > 0. \tag{3.12}$$

It is easy to check that for $\mathcal{R}e\, z \leq b$ one has $|z'| \leq 1$. The parameter b shifts the spectrum along the real axis, whereas the parameter a stretches it. One then applies the transformation (Eq. (3.12)) to the generalized eigenvalue problem (Eq. (3.11)) to obtain

$$\mathcal{D}^{-1}\mathcal{C}\hat{\mathbf{v}} = \lambda\hat{\mathbf{v}}, \text{ where } \lambda = \frac{a - b + \sigma}{a + b - \sigma}, \text{ and } \begin{array}{rl} \mathcal{C} = & \mathcal{A} + (a - b)\mathcal{B} \\ \mathcal{D} = & -\mathcal{A} + (a + b)\mathcal{B} \end{array}. \tag{3.13}$$

Although the matrix $\mathcal{B}$ is in general singular, the matrices $\mathcal{D}$ and $\mathcal{C}$ are generically not singular. The new eigenvalue problem (Eq. (3.13)) can then be treated using generalized power methods which compute the eigenvalues with the largest absolute values (Golub and Van Loan [1996]). Another technique used to solve the generalized eigenvalue problems is the Jacobi–Davidson QZ method (Sleijpen and Van der Vorst [1996]). In this method, a Krylov subspace of $\mathbf{R}^n$ is constructed, and special projection techniques are used to find the eigenvalues and eigenvectors.

3.3. Branch switching and homotopy methods

Although a linear stability analysis is aimed at detecting the bifurcations, it does not provide a way to compute the new branches of solutions which appear at the bifurcations. There are traditionally at least two ways to switch on branches of steady states when $\det \Phi_u$ in Eq. (3.6) changes sign but $\dot{p}$ does not, namely for transcritical and pitchfork bifurcations. Let the tangent to the original branch be denoted by $(\dot{\mathbf{u}}_0, \dot{p}_0)$ and the solution at the bifurcation $(\mathbf{u}^\star, p^\star)$.

The first method consists in computing a vector $(\hat{\mathbf{u}}, \hat{p})$ which belongs to the plane spanned by $(\dot{\mathbf{u}}_0, \dot{p}_0)$ and the tangent to the new branch, and which is orthogonal to $(\dot{\mathbf{u}}_0, \dot{p}_0)$. It exactly amounts to solve the system

$$\begin{pmatrix} \Phi_{\mathbf{u}}(\mathbf{u}^\star, p^\star) & \Phi_p(\mathbf{u}^\star, p^\star) \\ \dot{\mathbf{u}}_0{}^T & \dot{p}_0 \end{pmatrix} \begin{pmatrix} \hat{\mathbf{u}} \\ \hat{p} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ 0 \end{pmatrix}. \tag{3.14}$$

The first equation $\Phi_{\mathbf{u}}\hat{\mathbf{u}} + \Phi_p \hat{p} = 0$ indeed insures that $(\hat{\mathbf{u}}, \hat{p})$ belongs to the plane spanned by the two tangent vectors since $\Phi_{\mathbf{u}}$ has a nontrivial kernel at the bifurcation point. The solution to Eq. (3.14) is easily determined to be

$$\hat{p} = \frac{-\mathbf{u}_0^T \phi}{\dot{p}_0 - \dot{\mathbf{u}}_0^T \mathbf{z}}, \ \hat{\mathbf{u}} = \phi - \hat{p}\mathbf{z}, \tag{3.15}$$

where ϕ is a null vector of $\Phi_{\mathbf{u}}(\mathbf{u}^\star, p^\star)$ and $\mathbf{z}$ is the solution of $\Phi_{\mathbf{u}}(\mathbf{u}^\star, p^\star)\mathbf{z} = \Phi_p(\mathbf{u}^\star, p^\star)$. In practice, the null vector is calculated by inverse iteration (ATKINSON [1976]) or a good approximation of it may be given by the null eigenvector obtained by the linear stability analysis. To determine a point on the new branch, the Newton process is started with the *predictor*

$$\mathbf{u} = \mathbf{u}^\star \pm \Delta s \, \hat{\mathbf{u}}; \ p = p^\star \pm \Delta s \, \hat{p}, \tag{3.16}$$

where the $\pm$ indicates that solutions can be found on either side of the known branch.

The other method consists in using an homotopy parameter θ that perturbs the original problem to a problem where only saddle-node bifurcations are observed. As stated in Section 2, pitchfork and transcritical bifurcations can always be perturbed to give saddle-node bifurcations. The method is illustrated for the case of the pitchfork bifurcation in Fig. 3.3. One is starting from the solution a and wants to reach either the solution b or c.

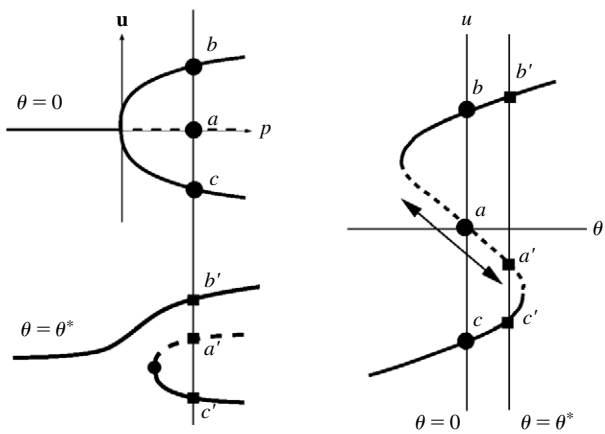


FIG. 3.3 Use of homotopy methods to compute branch switching. The left panels correspond to the bifurcation diagrams when varying p for two different values of the perturbation homotopy parameter θ, the right panel is a cross-section of the right panels at constant p corresponding to the thin red line. One initially starts from the fixed point a and follows the branch along the homotopy parameter θ (left panel).

The idea is simply to compute the branch of solutions along the direction θ which corresponds here to a parameter which perturbs the problem symmetry (for instance, the forcing term in Eq. (3.1)). The bifurcation diagram for $\theta = \theta^\star \neq 0$ is given to have an idea of the two-dimensional set of solutions. Homotopy methods are very powerful and can even solve more complicated problems like finding isolated branches of solutions in nontrivial systems.

3.4. Continuation of periodic orbits

We now turn to the numerical computation of stable and unstable periodic orbits versus a parameter p of autonomous problems. There are basically two methods to compute periodic orbits: one method directly solves a boundary value problem, while in the other method, fixed points of the Poincaré map are computed.

3.4.1. Boundary value problem approach

As the period T has an a priori unknown dependence on p, time is usually rescaled as $\tilde{t} = t/T$, such that the original problem (Eq. (2.1)) transforms into

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = T\mathbf{f}(\mathbf{x}, \mathbf{p}), \tag{3.17}$$

where $t \in [0, 1]$. The computation of a periodic orbit of Eq. (3.17) can be considered as a boundary value problem in time, since $\mathbf{x}(0) = \mathbf{x}(1)$. For autonomous systems, these boundary conditions do not fully determine the orbit since its phase can be freely shifted. Hence, in a pseudo-arclength method, we need, in addition to the parametrizing equation, also a normalization of the phase, i.e., a phase condition.

Suppose we have determined a periodic solution $(\mathbf{x}^0, T^0)$ at $p = p^0$, then many methods solve the problem (Eq. (3.17)) with

$$\int_0^1 (\mathbf{x}(t) - \mathbf{x}^0(t))^T \dot{\mathbf{x}}^0(t)\mathrm{d}t + (T - T^0)\dot{T}^0 + (p - p^0)\dot{p}^0 = \Delta s, \tag{3.18a}$$

$$\int_0^1 \mathbf{x}^T(t)\dot{\mathbf{x}}^0(t)\mathrm{d}t = 0, \tag{3.18b}$$

$$\mathbf{x}(0) = \mathbf{x}(1), \tag{3.18c}$$

where Δs is again the step length. The phase condition (Eq. (3.18b)) is obtained by requiring that the distance between the new and the old periodic solution is minimized with respect to the phase. The problem (Eq. (3.18)) is solved either by so-called single or multiple shooting techniques or by collocation techniques. A starting point of the periodic orbit can usually be obtained by a perturbation analysis near a Hopf bifurcation.

3.4.2. Fixed points of the Poincaré map

In this approach, we first have to define a Poincaré section which is transversal to the flow. This can be done by defining a hyperplane Σ as

$$g(\mathbf{x}) = \mathbf{n}.(\mathbf{x} - \mathbf{x}_*), \tag{3.19}$$

where $\mathbf{n}$ is the normal to the hyperplane and $\mathbf{x}_*$ a certain point in Σ. In practice, one can often take intersections with one of the coordinate planes. The Poincaré map on Σ is defined as

$$\mathcal{P}(\mathbf{x}, p) = \phi(t(\mathbf{x}), \mathbf{x}, p), \tag{3.20}$$

where ϕ is a solution of Eq. (2.1). Here, $t(\mathbf{x})$ is the return time and $sign(\mathbf{n}.\phi(0, \mathbf{x}, p)) = sign(\mathbf{n}.\phi(t(\mathbf{x}), \mathbf{x}, p))$ such that the orbit intersects Σ with the same orientation at both times.

Periodic orbits are computed from

$$\mathbf{x} - \mathcal{P}(\mathbf{x}, p) = 0, \tag{3.21}$$

with $\mathbf{x} \in \Sigma$. In practice, the Poincaré section is parametrized, and, leaving out this detail, the solutions of Eq. (3.21) versus the parameter p are again calculated by the pseudo-arclength method. This leads to systems of equations (Eq. (3.6)), where the Jacobian matrix is now given by $I - \mathcal{P}_x(\mathbf{x}^k, p^k)$. The additional difficulty is now that we do not want to construct the Poincaré map explicitly.

If we solve the linear systems Eq. (3.8) with an iterative method, however, we need only the result of a product of the Jacobian with a vector, say $\mathbf{v}$. Thereto, we need to perform two transient integrations, one with the full system (Eq. (2.1)) and one with the first variational equation (Eq. (2.13)). If $\mathbf{x}^k$ is again the k^{th} iterate of the Newton–Raphson process, then the initial conditions of the full system are $\mathbf{x}^k + \mathbf{v}$, while for the variational equation, the initial conditions are $\mathbf{v}$. Integration of Eq. (2.8) until an intersection with Σ provides the vector $\mathbf{x}^*$, while integration of Eq. (2.13) provides $\mathbf{y}^*$. It can be shown that in this case

$$P_x\mathbf{v} = \mathbf{y}^* - \mathbf{z}^* \frac{\mathbf{n}.\mathbf{y}^*}{\mathbf{n}.\mathbf{z}^*}, \tag{3.22}$$

where $\mathbf{z}^* = \mathbf{f}(\mathbf{x}^*, p)$. A more detailed description of this algorithm in the context of PDEs can be found in Sanchez, Net, Garcia-Archilla and Simo [2004].

3.5. Application potential

The early and best-known applications of continuation methods to atmospheric and ocean dynamics involved a very small number of degrees of freedom (Lorenz [1963a,b], Veronis [1963, 1966]). For instance, the Lorenz model (Eq. (2.20)) is probably one of the most studied dynamical systems of the past 40 years. As dynamical systems theory is being rapidly extended to infinite-dimensional systems governed by PDEs (Temam [1997], pp. 643), the applications to atmospheric, oceanic, and climate dynamics are becoming more and more sophisticated (Dijkstra [2005], pp. 532, Ghil and Childress [1987], pp. 485). Indeed, the computer power available to study bifurcation sequences is increasing rapidly so that more sophisticated numerical methods can be applied.

In the mid-1970s, it was possible to compute the first one or two bifurcations for spatially 1D energy balance models with (Held and Suarez [1974], North [1975], North, Mengel and Short [1983]) or without (Ghil [1976]) spectral truncation.

In the mid-1980s, truncations to tens of degrees of freedom of 2D geophysical flow problems could be approached in this manner (GHIL and CHILDRESS [1987], pp. 485). In particular, continuation methods were first seen in the seminal paper of Legras and Ghil on atmospheric flow regimes (LEGRAS and GHIL [1985]). In the last few years, there has been a spectacular increase of bifurcation studies involving a large number of degrees of freedom. For instance, bifurcation sequences have been computed for 2D oceanic flows (CESSI and YOUNG [1992], PRIMEAU [1998], QUON and GHIL [1992, 1995], SIMONNET, GHIL, IDE, TEMAM and WANG [2003a,b], SIMONNET, TEMAM, WANG, GHIL and IDE [1998], SPEICH, DIJKSTRA and GHIL [1995]) as well as 3D flows (CHEN and GHIL [1996], DIJKSTRA, OKSUZOGLU, WUBS and BOTTA [2001], GHIL and ROBERTSON [2000], WEIJER, DIJKSTRA, OKSUZOGLU, WUBS and NIET [2003]). All these studies involved tens of thousands of degrees of freedom with the noticeable exception of WEIJER, DIJKSTRA, OKSUZOGLU, WUBS and NIET [2003] which involved about 300,000 degrees of freedom. Simplified atmospheric, oceanic, or coupled GCMs have thus become amenable to a systematic study of their large-scale variability.

The applicability strongly depends on the availability of efficient solvers for the linear systems arising from the Newton–Raphson method. Recently, the development of targeted solvers for ocean models (DE NIET, WUBS, VAN SCHELTINGA and DIJKSTRA [2007], WUBS, DE NIET and DIJKSTRA [2006]) has opened the way to tackle problems with up to 10^6 degrees of freedom. The latest solver is based on a block Gauss–Seidel preconditioner, which uses the special structure and properties of the hydrostatic and geostrophic balances.

4. Application to the wind-driven ocean circulation

One of the central problems of physical oceanography is to understand the physics of the near-surface ocean circulation at mid-latitudes. The North Atlantic circulation is composed of two large-scale gyres, namely a subpolar cyclonic and a subtropical anticyclonic gyre associated with a well-known eastward jet, the Gulf Stream. These two cyclonic and anticyclonic gyres and the associated intense zonal jet are essentially driven by the mid-latitude easterlies and trade winds in the northern hemisphere. This system is referred to as the *wind-driven double-gyre circulation*. Figure 4.1 gives a schematic representation of this circulation in the Gulf Stream region.

The Kuroshio extension is another example of a double-gyre system in the North Pacific. Similar systems exist in the Southern hemisphere as well such as the Brazil and Malvinas currents in the South Atlantic. These intense oceanic currents exhibit highly complex, multi-scale spatio-temporal structures. Spatial scales from several kilometers (eddies, rings) up to hundred kilometers (meanders, recirculation gyres) are associated with time scales of months to decades.

The fundamental question is to understand the origin of the low-frequency variability of the double-gyre wind-driven circulation. This low-frequency variability involves timescales of several years to several decades. Although the atmosphere above the mid-latitude oceans is highly variable, some low-frequency phenomena may be intrinsically caused by nonlinear oceanic dynamical processes. In the following, we describe how the

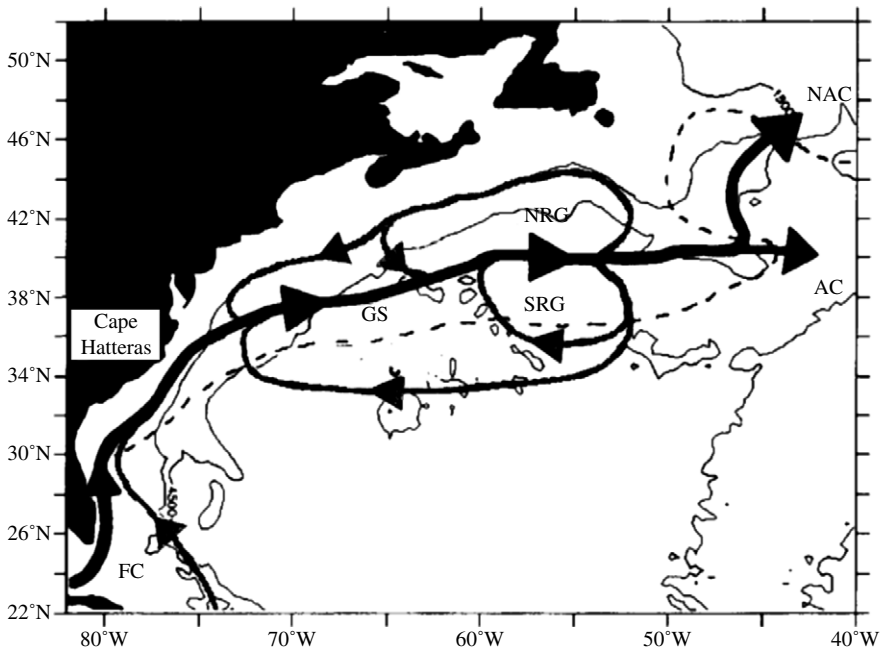


FIG. 4.1 Sketch of the near-surface circulation in the Gulf Stream region. Bold lines: Florida current (FC) and Gulf Stream (GS), branching into the North Atlantic Current (NAC) and Azores Current (AC). The abbreviations NRG and SRG indicate Northern (cyclonic) and Southern (anticyclonic) recirculation gyre, respectively (from DENGG, BECKMANN and GERDES [1996]).

use of dynamical systems and bifurcation theory enables one to achieve a rather deep understanding in this particular problem. In particular, we will demonstrate that low-frequency variability can have a sole internal origin and that it is related to instabilities of the mean flow.

4.1. The double-gyre quasi-geostrophic model

Quasi-geostrophic (QG) equations are very often used in ocean and atmosphere studies. They provide a rather simple 2D model which describes qualitatively the behavior of geophysical flows at mid-latitudes. In the context of the double-gyre wind-driven circulation, these equations read

$$\begin{aligned} &q_t + \left[\psi_x q_y - \psi_y q_x\right] + \beta\psi_x = -\mu\Delta\psi + \nu\Delta^2\psi - \alpha\sin(2\pi y/L_y), \\ &q = \Delta\psi - \lambda_R^{-2}\psi, \end{aligned} \tag{4.1}$$

where x denotes the longitudinal direction and y the latitudinal one on a rectangular domain $\Omega = (0, L_x) \times (0, L_y)$. These equations are indeed very similar to the incompressible 2D Navier–Stokes equations expressed in vorticity form, and in particular the velocities are related to the streamfunction ψ by $u = -\psi_y$ and $v = \psi_x$ and q is called

here the potential vorticity. The RHS of Eq. (4.1) comes from the vertical integration of the fluid layer from bottom to surface. It is composed of $-\mu\Delta\psi$, the stress at the bottom which is proportional to the vorticity, and the wind stress at the surface $\alpha\sin(2\pi y/L_y)$ which is prescribed and time-independent. The eddy viscosity operator $\nu\Delta^2\psi$ plays a rather different role and is needed here as a turbulent closure in order to stop the enstrophy cascade toward the small scales. The term $\beta\psi_x$ is of paramount importance since it represents the variation of the Coriolis force on a β plane. Indeed, on such a plane, the Coriolis parameter f is assumed to be equal to $f = f_0 + \beta y$. Note that it is rather the variations of f with latitude which are important and permit the propagation of the so-called Rossby waves (PEDLOSKY [1987], pp. 710).

The equations (Eq. (4.1)) describe the dynamics of flows which are in geostrophic equilibrium, i.e., when there is a balance between Coriolis and pressure forces so that the ocean or atmospheric fluid parcels flow along the direction of lines of constant pressure. The boundary conditions can either be no-slip ($\psi = \partial\psi/\partial\mathbf{n} = 0$) or free-slip ($\psi = \Delta\psi = 0$) although more accurate boundary conditions can also be used (MCWILLIAMS [1977]). The sinusoidal forcing term corresponds to the curl of the wind stress field, namely easterlies in the northern and southern part of the basin and westerlies in the middle. It is assumed here that the forcing term obeys some symmetry and in particular that the amounts of negative and positive vorticity due to the winds are equal. This choice implies that the model (Eq. (4.1)) is invariant with respect to the $\mathbb{Z}_2$ (reflection) symmetry

$$\mathcal{S}\psi(x, y) = -\psi(x, L_y - y). \tag{4.2}$$

This property is important since one wants to start from the most idealized picture where the system has the largest number of symmetries and then perturb it when proceeding along the axis of complexity (see Fig. 1.1).

Figure 4.2 shows a typical bifurcation diagram of the double-gyre circulation in small oceanic basins (SIMONNET and DIJKSTRA [2002]). This diagram is obtained using a pseudo-arclength method with linear stability analysis as described in Section 3. Three distinct branches of steady solutions are obtained. The first branch is characterized by perfectly antisymmetric solutions ψ for which $\mathcal{S}\psi = \psi$. After the first saddle-node bifurcation point L, these solutions become inertially dominated, the flows become very energetic, and the recirculation gyres eventually fill the entire basin. Two pitchfork bifurcations (see Fig. 2.3 in Section 2), P_1 and P_2, occur on the antisymmetric branch. They are responsible for the appearance of asymmetric solutions, say ψ_1 and ψ_2, which are conjugated to each other, i.e., $\mathcal{S}\psi_1 = \psi_2$. The second pitchfork bifurcation at P_2 leads to a branch of asymmetric flows with their jet aligned in the west–east direction and a confluence point either shifted to the south or to the north of the mid-axis of the basin. The flows on the two branches starting from P_1 exhibit meandering of the jet downstream of the recirculation dipole (see details in CHANG, IDE, GHIL and LAI [2001]), with either a stronger subtropical gyre (upper branch) or a stronger subpolar gyre (lower branch). The weaker gyre is more affected by the asymmetry and wraps around the stronger one. Hopf bifurcations H_1, H_{gyre}, H_3 are detected along these asymmetric branches indicating that the steady states are destabilized by successive oscillatory instabilities.

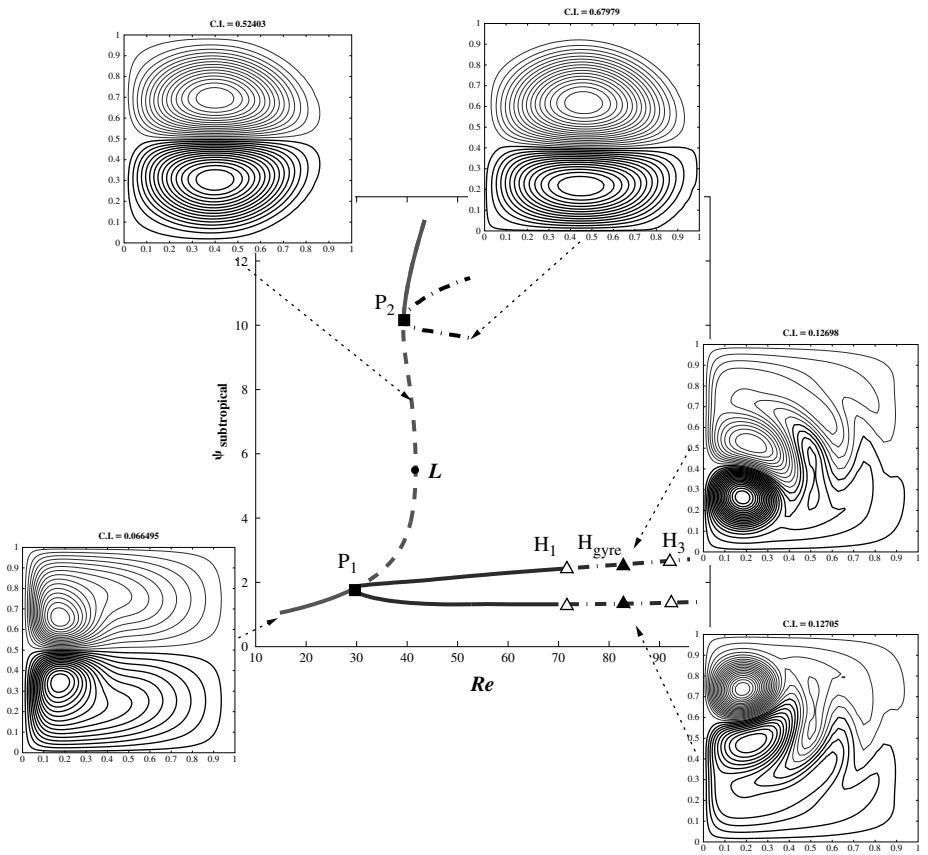


FIG. 4.2 Bifurcation diagram for the QG model (Eq. (4.1)). Here, the maximum value of the streamfunction of the subtropical gyre ($\Psi_{\text{subtropical}}$) is plotted versus the Reynold number $Re = UL_X/\nu$, where U is a typical characteristic horizontal velocity. The bifurcations P_1, P_2 correspond to the pitchfork bifurcations (squares), whereas the triangles indicate the location of the Hopf bifurcations. The saddle-node bifurcation L is indicated by a dot. The panels correspond to the upper layer streamfunctions, positive contours are represented by thick lines, negative contours by thin lines together with contour intervals (from SIMONNET and DIJKSTRA [2002]). (See also color insert).

The path of the real part of the leading eigenvalues and associated eigenvectors of Eq. (3.11) is shown in Fig. 4.3. The connection between this figure and Fig. 4.2 is simple: instead of showing the quantity $\Psi_{\text{subtropical}}$ along the various branches of steady states, one plots instead the real and imaginary parts of the leading eigenvalue of the linearized equation. These paths are found by solving Eq. (3.11) at each step of the continuation algorithm. One thus observes the crossing of the imaginary part in the complex plane by real eigenvalues (thick curves) giving the saddle-node and the two pitchfork bifurcations as seen previously in Fig. 4.2. As expected, the streamfunction patterns are symmetric for the case of the pitchfork bifurcation and antisymmetric for the case of the saddle-node bifurcation. The important thing about Fig. 4.3, however, is the

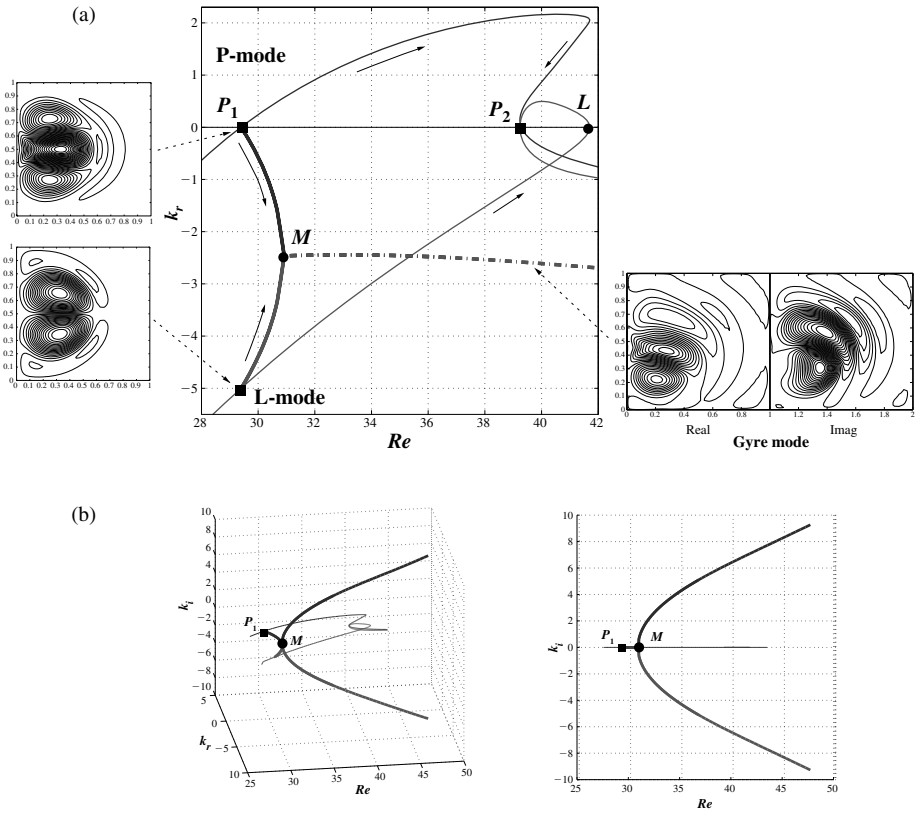


FIG. 4.3 Spectral behavior of the eigenmodes involved into the various bifurcations off the antisymmetric branch for the barotropic QG model shown in red and blue in Fig. 4.2. The real part, panel (a), and imaginary part, panel (b), of the P-mode and L-mode eigenvalues are plotted versus *Re*. Thick lines, beginning at P_1 in Fig. 4.2 and ending at M, correspond to the path of these two modes on the asymmetric branch, thin lines to the path on the antisymmetric branch. The dash-dot and thick line indicates a non-zero imaginary part. At M, the merging between the P-mode and L-mode occurs. The various panels show the streamfunction patterns at the locations indicated by the arrows (from SIMONNET and DIJKSTRA [2002]). (See also color insert).

remarkable mode-merging observed at M, where the two real eigenvalues become complex conjugated. Although it corresponds to a topological change of behavior in phase space, it is not rigorously speaking a bifurcation since, at M, the oscillatory behavior is damped. Nevertheless, this oscillatory eigenmode is eventually destabilized at H_{gyre} in Fig. 4.2 through a Hopf bifurcation. It should be clear from the lower panel of Fig. 4.3 that the imaginary part of the eigenvalue is growing quadratically and thus leads to an oscillatory instability of rather low frequency compared to other complex eigenvalues in the spectrum. This merging phenomenon is as generic as the pitchfork bifurcation P_1 and is seen in much more complex situations and models (SIMONNET [2005], SIMONNET, GHIL, IDE, TEMAM and WANG [2003b]).

This remarkable result is indeed one of the starting points of a much deeper understanding of the low-frequency dynamics of the double-gyre circulation in the recent years. To start with, the mode referred to as *gyre mode* in Fig. 4.3 was already known since the work of JIANG, JIN and GHIL [1995] on a shallow-water version of the model (Eq. (4.1)). Its particular spatial structure induces a relaxation oscillation of the gyres associated with periods of several years. These relaxation oscillations moreover become chaotic in strongly nonlinear regimes. For a long time, the origin of such modes was unexplained. Indeed, they appeared seemingly from nowhere, and contrary to most of the other instabilities found (e.g., Hopf bifurcations H_1 and H_3 in Fig. 4.2), linear theories of the double-gyre flow cannot explain their existence, as it appears that this phenomenon is essentially nonlinear. This result illustrates the power of dynamical systems and bifurcation theory.

At the same time, some authors (CHANG, IDE, GHIL and LAI [2001], MEACHAM [2000], SIMONNET, TEMAM, WANG, GHIL and IDE [1998]) argued that the chaos observed in more nonlinear regimes could not be the result of several competing instabilities interacting according to the Ruelle–Takens route to chaos (Takens [1981]). They suggested the presence of a global homoclinic (or possibly heteroclinic) bifurcation in the model (Eq. (4.1)). It is only recently that such a bifurcation has been explicitly detected (NADIGA and LUCE [2001], SIMONNET, GHIL and DIJKSTRA [2005], SIMONNET, GHIL, IDE, TEMAM and WANG [2003b]). The connection with the gyre modes is also clearly established in SIMONNET, GHIL, IDE, TEMAM and WANG [2003b] and SIMONNET, GHIL and DIJKSTRA [2005]: these modes span larger and larger regions of phase space as the nonlinearity of the model increases and form a symmetric homoclinic orbit later on. This is illustrated by Fig. 4.4, where the unfolding of the limit cycles a, b, c, and d for a parameter approaching the critical value at the symmetric homoclinic bifurcation is clearly seen.

Note that in this case, a poor man's continuation method of periodic orbits was used, i.e., several time integrations for particularly well-chosen parameters near the critical value were achieved. It is also possible using continuation techniques to have some ideas about the nature of the homoclinic bifurcation by computing the leading eigenvalues at the hyperbolic fixed point to which the homoclinic cycle is connected. This is illustrated by Fig. 4.5, which provides a loci of pitchfork, homoclinic, and saddle-node bifurcations in the two-parameter plane spanned by the wind stress intensity α and lateral viscosity ν (see Eq. (4.1)). Figure 4.6 corresponds to the linear stability analysis at the hyperbolic point where the homoclinic orbit is connected. For small ν, it appears that the homoclinic orbit is of saddle-focus type similar to the Shilnikov case (SHILNIKOV [1965]) presented in Section 2 although there is now an additional symmetry.

Complex dynamics is found on both sides of the homoclinic orbit. A sketch of the bifurcations of periodic orbits in the vicinity of the principal homoclinic orbit is shown for values of the wind stress (ν fixed) near the critical value $\bar{\alpha}$ in Fig. 4.7. One thus observes period-doubling bifurcations of the periodic orbits related to the gyre modes for $\alpha < \bar{\alpha}$ similar to the ones observed in JIANG, JIN and GHIL [1995] and SIMONNET, GHIL, IDE, TEMAM and WANG [2003b] as well as symmetry-breaking bifurcations of the now symmetric limit cycles 'after' the homoclinic bifurcation. Note that each symmetry-breaking bifurcation provides a new family of asymmetric periodic orbits which undergo sequences of period-doubling bifurcations. This wiggling behavior is indeed confirmed

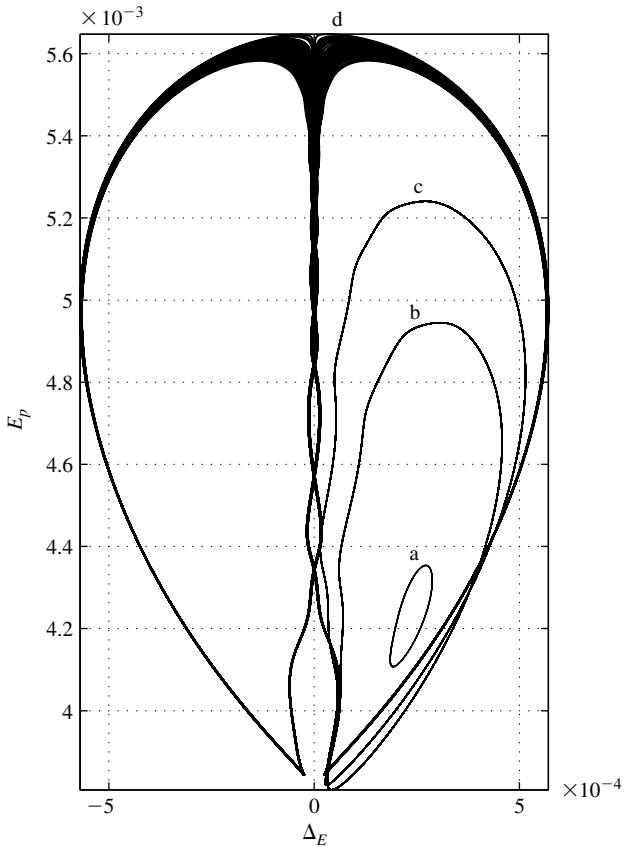


FIG. 4.4 Unfolding of the relaxation oscillations induced by the gyre modes in the plane spanned by the total potential energy E_p versus $\Delta_E = E_p(\psi > 0) - E_p(\psi < 0)$ (from SIMONNET, GHIL and DIJKSTRA [2005]).

directly by integrating the model (Eq. (4.1)) for many values of α near $\bar{\alpha}$ in SIMONNET, GHIL and DIJKSTRA [2007] (not shown).

It is clear from bifurcation theory that one is limited to secondary bifurcations and that as further instabilities emerge, the dimensions of the chaotic attractors as well as their topological complexities increase. One may thus legitimately ask about the nature of chaos in the double-gyre circulation when the nonlinearity is increased even further. A surprising answer is given in SIMONNET [2005] which justifies even more the use of those dynamical systems concepts when the flows approach more turbulent states. As far as large oceanic basins like the North Atlantic and Pacific are concerned, a new phenomenon emerges which has been referred to as *the quantization of the low-frequency dynamics* in SIMONNET [2005]. Roughly speaking the successive bifurcations culminating with the symmetric homoclinic bifurcation as described previously are repeated but now involve larger wavenumbers. Figure 4.8 is indeed a generalization of Fig. 4.5. The first colored band corresponds exactly to the situation shown in Fig. 4.5 except

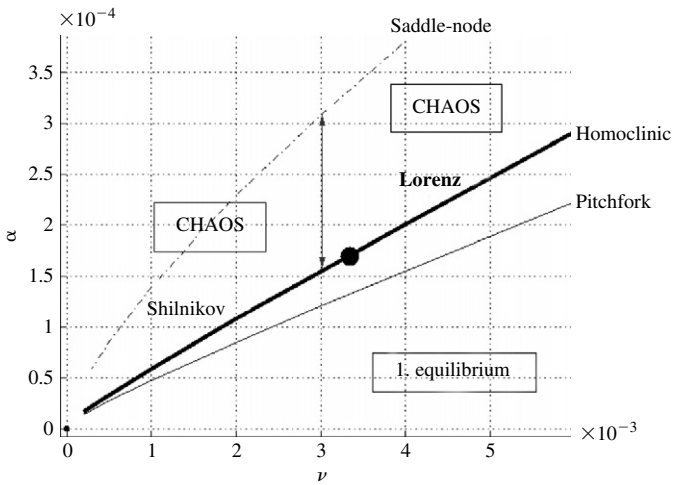


FIG. 4.5 Loci of pitchfork (light solid curve), homoclinic (heavy solid curve), and saddle-node bifurcations (dash-dotted curve) in the two-parameter plane spanned by the nondimensional viscosity ν and wind stress intensity α. The filled circle marks the transition between Lorenz chaos and Shilnikov chaos. (from SIMONNET, GHIL and DIJKSTRA [2005]).

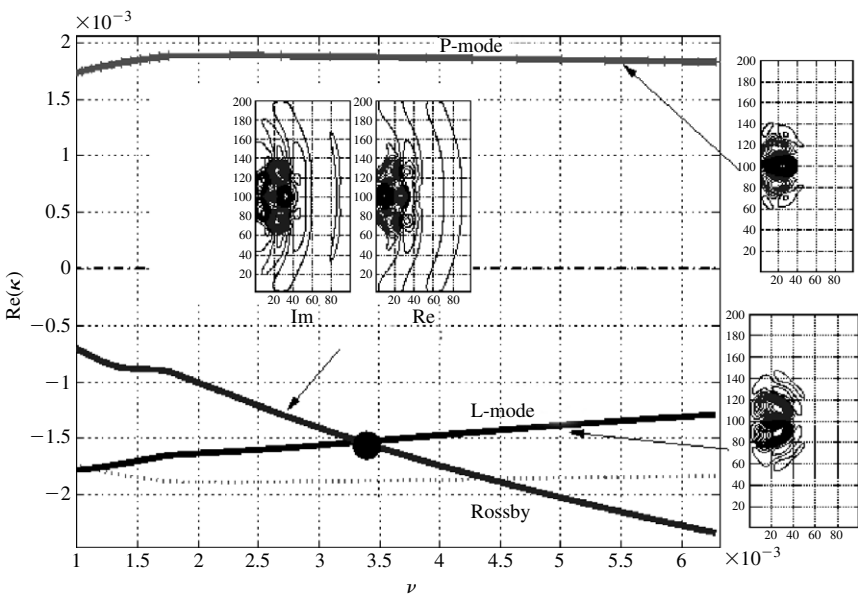


FIG. 4.6 Leading instabilities that occur along the homoclinic locus of Fig. 4.5, as a function of the viscosity ν. The curves plotted in the large panel are the real parts of the leading eigenvalues (the ones closest to the imaginary axis). The spatial patterns of the associated streamfunction fields (real for the P- and L-mode and complex conjugate for the Rossby basin mode) are plotted in the small panels (from SIMONNET, GHIL and DIJKSTRA [2005]).

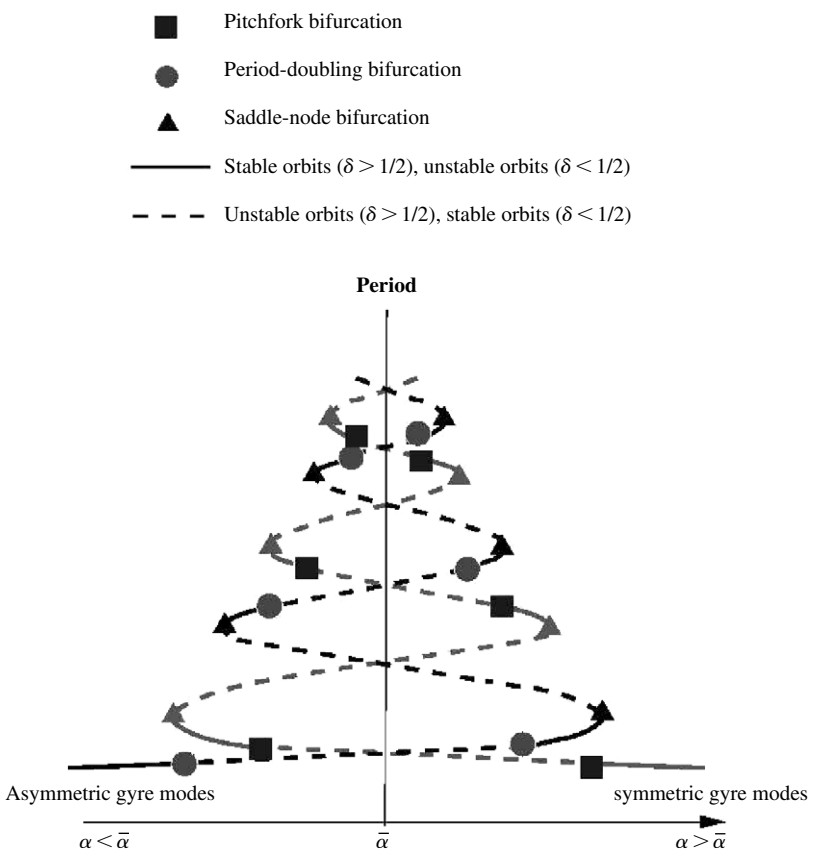


FIG. 4.7 Shilnikov wiggles in the $\mathbb{Z}_2$-symmetric case and for $\delta < 1$. The period of the asymmetric periodic orbits (left curve) has been multiplied by 2.

that the saddle-node curve is now replaced by a locus of subcritical pitchfork birfurcations. Between the two curves, Shilnikov phenomena are observed (not shown). The next colored bands exhibit the same successive bifurcations namely supercritical pitchfork bifurcations, real eigenmode merging off the asymmetric branches of solutions like in Fig. 4.3 followed by Hopf bifurcations of gyre modes and finally subcritical pitchfork bifurcation. The unfolding of the relaxation oscillations (see Fig. 4.4) into a symmetric homoclinic orbit is not explicitly detected however. The reason is related to the appearance of high-frequency instabilities (through Hopf bifurcations) along the antisymmetric branch. These instabilities interact nonlinearly with the low-frequency structures (the gyre modes) so that the poor man's continuation approach is inefficient here.

There are at least two ways to detect the presence of a global bifurcation in this situation: the first is through statistical methods and spectral analysis like multichannel singular spectrum analysis (M-SSA) that we do not present here (see, e.g., GHIL,

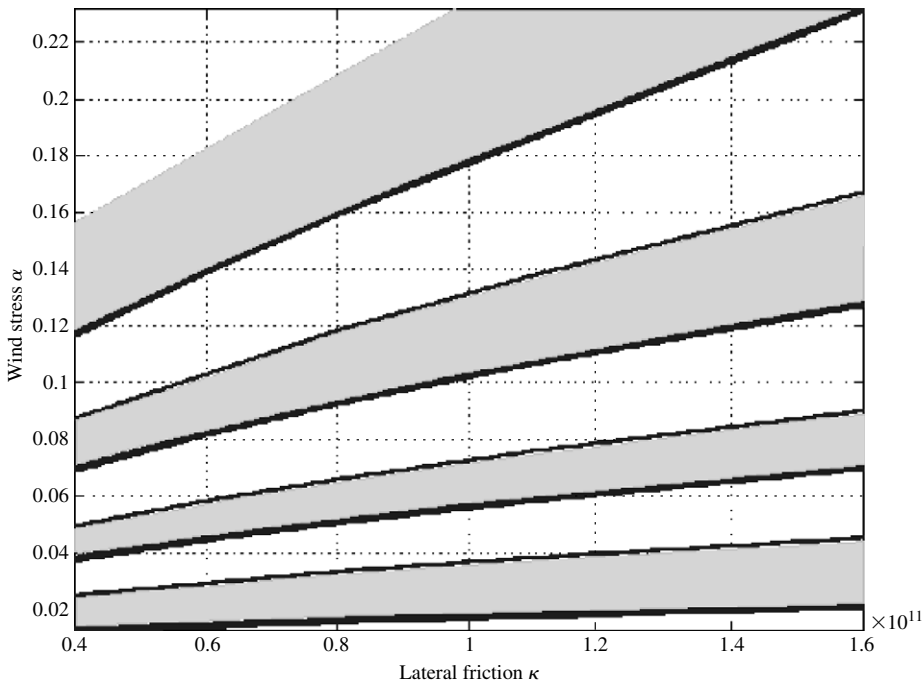


FIG. 4.8 Loci of supercritical and subcritical pitchfork bifurcations The two-parameter plane is spanned by the wind stress intensity α and the lateral hyperviscosity κ. The shaded areas correspond to the n-bump supercritical regimes and the areas in-between to the n-bump subcritical regimes ($n = 0, 1, 2, 3$) (From SIMONNET [2005]).

ALLEN, DETTINGER, IDE, KONDRASHOV, MANN, ROBERTSON, SAUNDERS, TIAN, VARADI and YIOU [2002], PLAUT and VAUTARD [1994]). The second is through continuation methods of periodic orbits which, as far as we know, have been used only in the work of SAMELSON [2001], WOLFE and SAMELSON [2006] and ZOLDI and GREENSIDE [1998] in a different context and for rather small to medium dimensional dynamical systems. The conclusion from SIMONNET [2005] is that the low-frequency skeleton of the attractors in more nonlinear regimes is likely to be self-similar to the so-called ground regime described at length above and that such a property can rather be easily detected using mere continuation techniques.

4.2. The shallow-water model

We discuss here an original example (unpublished) which uses the great strength and flexibility of continuation and homotopy methods. This example is directly connected to the problem of having explicit numerical control on the hierarchy of models available in geophysics. As seen previously, the QG equations (Eq. (4.1)) satisfy the exact symmetry (Eq. (4.2)). One of the consequence is that the branches of steady states are all connected. In more complex models, such a property disappears and one is confronted to situations where it is, most of the time, difficult to detect the isolated branches. A typical situation

is the case of the shallow-water (SW) equations that are given by the following system

$$\begin{aligned} U_t + \nabla \cdot (\mathbf{v}U) - fV &= -g'hh_x + \nu\Delta U - \mu U + \alpha\frac{\tau^x}{\rho}, \\ V_t + \nabla \cdot (\mathbf{v}V) + fU &= -g'hh_y + \nu\Delta V - \mu V + \alpha\frac{\tau^x}{\rho}, \\ h_t &= -U_x - V_y, \end{aligned} \tag{4.3}$$

where the eastward and northward velocities are given by $\mathbf{v} = (u, v)$ and h is the thickness of the upper-layer. The upper layer mass flux vector is represented by $\mathbf{V} = (U, V) = (uh, vh)$. The reduced gravity g' is given by $g' = g\Delta\rho/\rho$. The dissipative terms are given by the lateral friction coefficient ν and the bottom friction coefficient μ. We consider these equations, often referred to as *1.5-layer* or *equivalent-barotropic or reduced-gravity model*, in a closed domain Ω as in (Eq. (4.1)) together with either no-slip or free-slip boundary conditions. The double-gyre forcing term is given by

$$\tau^x = -\cos(\frac{2\pi y}{L_y}),\ \tau^y = 0.$$

It is straightforward to check that the system of equations (Eq. (4.3)) is no longer invariant with respect to the symmetry (Eq. (4.2)). Now, a close inspection shows that it is possible to connect the system Eq. (4.1) to Eq. (4.3) through a so-called homotopy parameter η, namely

$$\begin{aligned} U_t + \eta\left[\nabla \cdot (\mathbf{v}U) + g'hh_x\right] + (1-\eta)\left[\nabla \cdot (\mathbf{V}U) + g'h_x\right] - fV \\ = \nu\Delta U - \mu U + \alpha\frac{\tau^x}{\rho}, \\ V_t + \eta\left[\nabla \cdot (\mathbf{v}V) + g'hh_y\right] + (1-\eta)\left[\nabla \cdot (\mathbf{V}U) + g'h_y\right] + fU \\ = \nu\Delta V - \mu V + \alpha\frac{\tau^y}{\rho}, \\ h_t + U_x + V_y = 0. \end{aligned} \tag{4.4}$$

When $\eta = 0$, one recovers the steady-state version of Eq. (4.1) by taking the curl of (4.4) with $U = -\psi_y$, $V = \psi_x$, whereas $\eta = 1$ yields Eq. (4.3). It thus becomes possible to compute the branches of solutions for $\eta = 0$ and then smoothly going to the case $\eta = 1$ by pseudo-arclength continuation. The result is given in Fig. 4.9 for a small oceanic domain. Previous studies of SW models of the double-gyre circulation (more Simonnet, Ghil, Ide, Temam and Wang [2003a,b], Speich, Dijkstra and Ghil [1995]) were able to compute branches *a* and *b* in Fig. 4.9 but not branch *c*. The methodology is even more interesting in large basins since one observes successive pitchfork bifurcations in QG models which have never been computed in SW models.

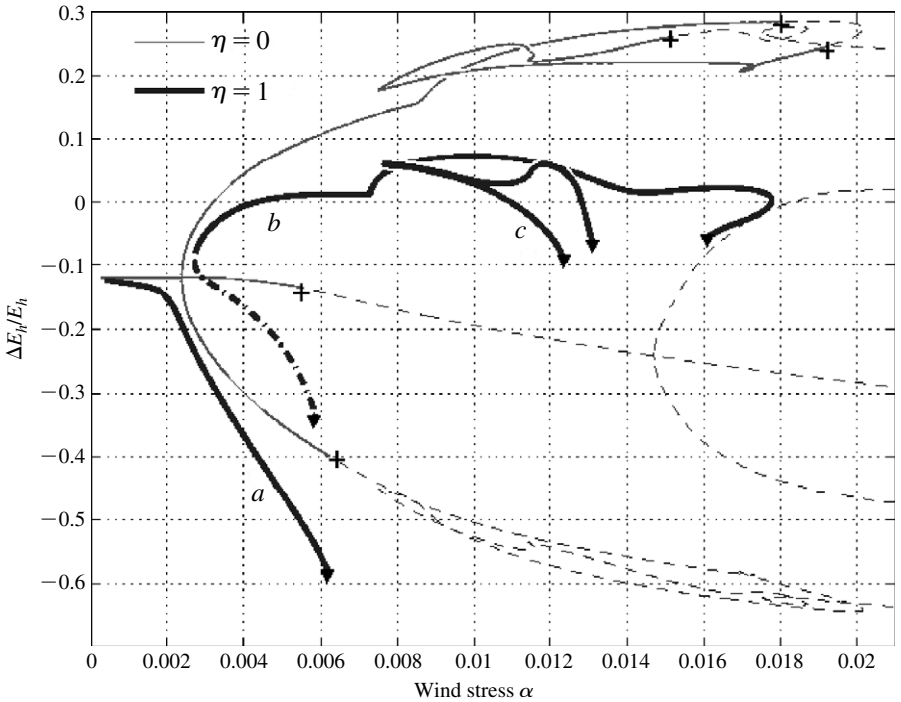


FIG. 4.9 Branches of steady states for the system (Eq. (4.4)) for $\eta = 0$ (light curves) and $\eta = 1$ (thick curves). The thin-dotted curves correspond to the case $\eta = 0$ which disappear when going to $\eta = 1$.

4.3. Bifurcations in primitive-equation models

Very recently, the double-gyre problem was considered in a 3D primitive-equation model (DE NIET, WUBS, VAN SCHELTINGA and DIJKSTRA [2007]). Similar as in the studies with quasi-geostrophic and shallow-water models, a basin of 10° length and 10° width centered around 45°N was considered. In longitude ϕ and latitude θ, the boundaries of the domain are given by $\phi_w = 270°\text{E}$, $\phi_e = 280°\text{E}$, $\theta_s = 40°\text{N}$ and $\theta_n = 50°\text{N}$; the basin has a constant depth $D = 2400\,\text{m}$. The flows in this domain are forced by a zonal wind stress given by

$$\tau^{\phi}(\phi, \theta) = -\tau_0 \cos\left(2\pi \frac{\theta - \theta_s}{\theta_n - \theta_s}\right), \tag{4.5a}$$

$$\tau^{\theta}(\phi, \theta) = 0, \tag{4.5b}$$

where $\tau_0 = 0.1$ Pa is a typical amplitude. This wind forcing is distributed as a body forcing over the first (upper) layer of the ocean having a depth $H_m = 200$ m.

With $r_0 = 6.4 \cdot 10^6$ m and $\Omega = 7.5 \cdot 10^{-5}\,\text{s}^{-1}$ being the radius and angular velocity of the Earth, the governing equations for the zonal, meridional, and vertical velocity u, v,

and w and the pressure p become

$$\frac{Du}{\mathrm{d}t} - uv\tan\theta - 2\Omega\, v\sin\theta + \frac{1}{\rho_0 r_0\cos\theta}\frac{\partial p}{\partial\phi} = A_V\frac{\partial^2 u}{\partial z^2} + A_H\, L_u(u,v) + \frac{\tau_0}{\rho_0 H_m}\tau^{\phi}\mathcal{G}(z), \tag{4.6a}$$

$$\frac{Dv}{\mathrm{d}t} + u^2\tan\theta + 2\Omega\, u\sin\theta + \frac{1}{\rho_0 r_0}\frac{\partial p}{\partial\theta} = A_V\frac{\partial^2 v}{\partial z^2} + A_H\, L_v(u,v) + \frac{\tau_0}{\rho_0 H_m}\tau^{\theta}\mathcal{G}(z), \tag{4.6b}$$

$$\frac{\partial p}{\partial z} = \rho_0 g, \tag{4.6c}$$

$$\frac{\partial w}{\partial z} + \frac{1}{r_0\cos\theta}\left(\frac{\partial u}{\partial\phi} + \frac{\partial(v\cos\theta)}{\partial\theta}\right) = 0, \tag{4.6d}$$

where $\mathcal{G}(z) = \mathcal{H}(z/H_m + 1)$, $\mathcal{H}$ is a continuous approximation of the Heaviside function, $\rho_0 = 10^3\,\mathrm{kgm}^{-3}$ is the constant density of the ocean water, and $g = 9.8\ \mathrm{ms}^{-2}$ is the gravitational acceleration. In these equations, A_H and A_V are the horizontal and vertical momentum (eddy) viscosity, respectively. We fix $A_V = 10^{-3}\,\mathrm{m}^2\,\mathrm{s}^{-1}$ and use A_H as a control parameter. In addition,

$$\frac{D}{\mathrm{d}t} = \frac{\partial}{\partial t} + \frac{u}{r_0\cos\theta}\frac{\partial}{\partial\phi} + \frac{v}{r_0}\frac{\partial}{\partial\theta} + w\frac{\partial}{\partial z},$$

$$L_u(u,v) = \nabla_H^2 u + \frac{u}{r_0^2\cos^2\theta} - \frac{2\sin\theta}{r_0^2\cos^2\theta}\frac{\partial v}{\partial\phi},$$

$$L_v(u,v) = \nabla_H^2 v + \frac{v}{r_0^2\cos^2\theta} + \frac{2\sin\theta}{r_0^2\cos^2\theta}\frac{\partial u}{\partial\phi},$$

where ∇_H^2 is the horizontal Laplace operator. Slip conditions are assumed at the bottom boundary, while at all lateral boundaries no-slip conditions are applied. As the forcing is represented as a body force over the first layer, slip conditions also apply at the ocean surface.

The equations are discretized in space using a second-order accurate control volume discretization method on a staggered Arakawa B-grid with $i = 0, \ldots, N$, $j = 0, \ldots, M$, $k = 0, \ldots, L$. The solution vector $\mathbf{u}$ contains the unknowns (u, v, w, p) at each grid point and hence has dimension $d = 4 \times N \times M \times L$. A typical bifurcation diagram, with the lateral friction coefficient A_H as a control parameter, for a resolution of $100 \times 100 \times 12$ points is plotted in Fig. 4.10. In agreement to what is found in shallow-water models, we here also find an imperfect pitchfork bifurcation, which leads to the existence of multiple flow patterns below a value of $A_H = 320\,\mathrm{m}^2\,\mathrm{s}^{-1}$. The isolated branch in Fig. 4.10 was found by the residue continuation algorithm as described in Dijkstra [2005, pp 532] and analyzed in Gruais, Rittemard and Dijkstra [2005].

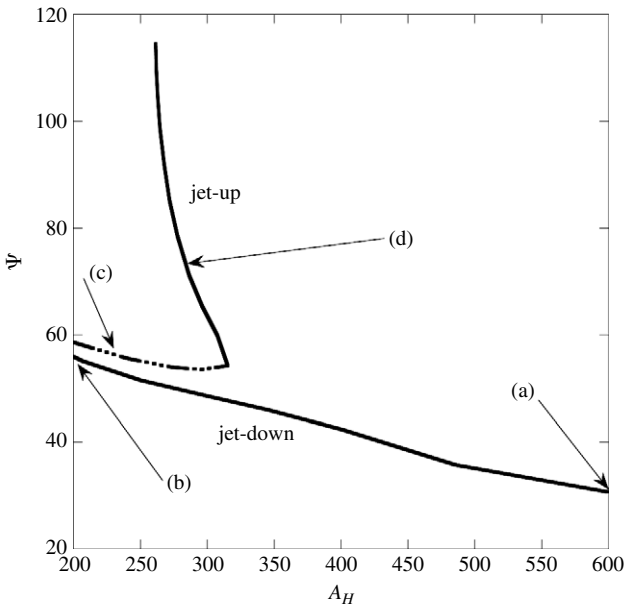


FIG. 4.10 Bifurcation diagram of the constant density circulation and control parameter A_H.

The barotropic streamfunction of patterns at specific locations in Fig. 4.10 is plotted in Fig. 4.11. The solution for large A_H on the connected branch is the near antisymmetric double-gyre flow (Fig. 4.11a). When A_H decreases along this branch, a so-called jet-down solution appears (Fig. 4.11b), a solution very well known from QG and SW models. Along the isolated branch, a jet-up solution exists (Fig. 4.11c) along the lower branch, and after the saddle-node bifurcation, the flow becomes inertially controlled (Fig. 4.11d). Although these results were expected, this is the first time that such a multiple equilibria structure is computed for a 3D primitive-equation model (a dynamical system having 480,000 degrees of freedom).

5. Outlook

Dynamical systems theory is a qualitative mathematical theory that deals with the spatio-temporal behavior of general systems of evolution equations. The theory analyzes systematically the changes in system behavior when parameters are varied. The brief overview in Section 2 described the most elementary types of transitions, those associated with codimension-1 bifurcations. In GUCKENHEIMER and HOLMES [1990, pp. 453], and KUZNETSOV [1995, pp. 515], the theory is further developed to deal with more complex transition scenarios.

The methodology of bifurcation theory was first applied in the 1960s and 1970s to highly simplified models of the ocean, atmosphere, and climate, governed by systems

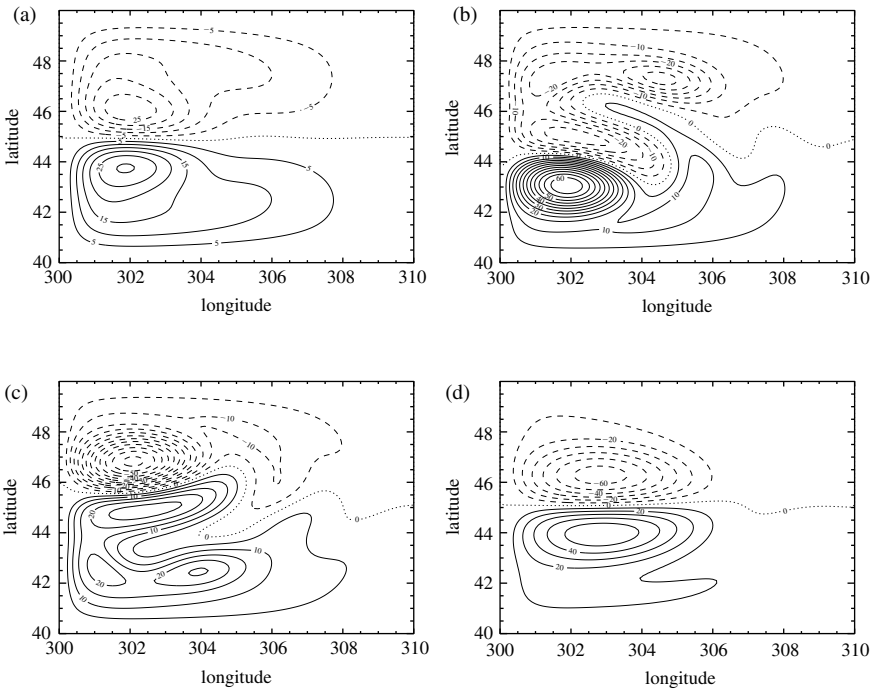


FIG. 4.11 Patterns of the barotropic streamfunction (the streamfunction of the vertically integrated flow) at labelled solutions marked in Fig. 4.10. Contour values are in Sv (1 Sv = 10^6 $m^3 s^{-1}$).

of a few coupled ODEs (LORENZ [1963a], STOMMEL [1961], VERONIS [1963, 1966]), or by one-dimensional PDEs, GHIL [1976], HELD and SUAREZ [1974], NORTH [1975]. The methodology of continuation methods and numerical bifurcation theory, briefly described in Section 3, is now certainly capable of handling rather sophisticated models governed by 2D and 3D systems of PDEs. The systematic use of pseudo-arclength continuation algorithms coupled with linear stability analysis allows one to infer generic and meaningful information on the dynamics of large-scale flows. Local continuation techniques together with well-chosen time integrations enable one to detect global bifurcations as well. The use of a hierarchy of models is necessary to eliminate model-dependent dynamical processes and to focus on the essential dynamics.

The solutions of certain members of this hierarchy, such as the quasi-geostrophic (QG) equations (Eq. (4.1)), seem to be dominated by low-dimensional dynamics, embedded into an a priori infinite-dimensional phase space. This feature may be due to a significant separation of scales of motion; in the QG models, for example, the fast gravity waves are filtered out and only the dynamics related to low-frequency Rossby waves is represented. As far as the large scales are concerned, we might be able to view small-scale, fully developed turbulence as a background soup that can be filtered out in the first approximation.

As an example of this approach, we considered a hierarchy of models of the constant-density, wind-driven double-gyre ocean circulation. It is possible to analyze the QG, shallow-water (SW), and primitive-equation (PE) models of this circulation systematically and to determine robust elements of the dynamics and of the transition behavior of these flows. We showed how to handle the corresponding PDEs as dynamical systems with a phase space of very large dimension and to detect and compute the successive bifurcations and branches of steady states. For example, the perfect pitchfork bifurcation associated with symmetry-breaking in the QG model is transformed into a perturbed pitchfork in the SW model, and this imperfect pitchfork is still present in the PE model. Within the QG model, low-frequency variability originates from an oscillatory mode, the gyre mode. This mode, in turn, arises through the merging of two purely exponential modes, and it subsequently causes global bifurcations associated with homoclinic orbits. The longest time scales involved in this low-frequency variability are of a few decades at most.

The effect of the small scales on the large ones is a difficult issue closely related to the system's spectral behavior. New, efficient coarse-graining bifurcation techniques have been recently developed by Kevrekidis and colleagues (e.g., Gear, Kevrekidis and Theodoropoulos [2002], Kevrekidis, Gear, Hyman, Kevrekidis, Runborg and Theodoropoulos) in the context of chemistry and reaction-diffusion systems. These techniques enable models at a "fine," microscopic level of description, like the 3D PE model presented in Section 4.3, to perform modeling tasks at a "coarse," macroscopic level, like our QG models. This feat is achieved by extracting from the microscopic description the information that traditional numerical procedures would obtain through function evaluation based on the macroscopic evolution equation, had this equation been available.

A characteristic feature of complex systems is the emergence of macroscopic, coherent types of behavior from the interactions of small-scale, microscopic elements. Macroscopic rules can therefore often be deduced from microscopic ones. Such is the case for the Navier–Stokes equations, which represent the (coarse) evolution of certain moments of the Boltzmann equation. In many situations, however, the macroscopic equations are not necessarily available. For instance, one may not have closed equations at the system level of interest; that is, the macroscopic velocity field may not be available as a function of just concentrations, or the form of the viscous terms may not be just a known Newtonian expression. In these cases, one often has a correct microscopic description of the true physics, at a molecular level or at a very fine resolution, but system-level tasks are not feasible anymore, directly at the macroscopic level. Still, it turns out to be possible to use information from short-time integrations of the micromodels to obtain information on the macromodel's Jacobian and thus detect (macro-)bifurcations. It is only recently that these coarse-graining techniques are being tested in the context of geophysical fluid dynamics.

The methodology of dynamical systems is now being applied to many problems in physical oceanography and climate dynamics. Examples are (i) the stability and variability of the thermohaline circulation (Te Raa and Dijkstra [2002]); (ii) transitions in the El-Niño ocean–atmosphere system (Dijkstra [2005, pp 532], Neelin, Battisti, Hirst, Jin, Wakata, Yamagata, and Zebiak [1998], Neelin, Latif and Jin [1994],

DER VAART DIJKSTRA and JIN [2000]), including those due to plate motion (OMTA and DIJKSTRA [2003]); and (iii) variability of the mid-latitude atmosphere–ocean system (VAN DER AVOIRD, DIJKSTRA, NAUW and SCHUURMANS [2002]). This methodology is likely to provide a major avenue to understanding the processes that cause large-scale, low-frequency variability in the ocean, atmosphere, and the whole climate system.

6. Acknowledgments

MG's and ES's work was supported by Grants ATM00-82131 from the National Science Foundation's Programs on Climate Dynamics, Physical Oceanography, and Applied and Computational Mathematics, and DE-FG02-04ER63881 from the Department of Energy's Climate Change Prediction Program.

References

ATKINSON, K.E. (1976). *An Introduction to Numerical Analysis* (Jon Wiley and Sons).

CESSI, P., YOUNG, W.R. (1992). Multiple equilibria in two-dimensional thermohaline circulation. *J. Fluid Mech.* **241**, 291–309.

CHANG, K.I., IDE, K., GHIL, M., LAI, C.-C.A. (2001). Transition to aperiodic variability in a wind-driven double-gyre circulation model. *J. Phys. Oceanogr.* **31**, 1260–1286.

CHELTON, D.B., WENTZ, F.J. (2005). Global microwave satellite observations of sea surface temperature for numerical weather prediction and climate research. *Bull. Am. Met. Soc.* **86**, 1097–1115.

CHEN, F., GHIL, M. (1996). Interdecadal variability in an hybrid coupled ocean-atmosphere model. *J. Phys. Oceanogr.* **26**, 1561–1578.

CHOSSAT, P., LAUTERBACH, R. (2000). *Methods in Equivariant Bifurcation and Dynamical Systems. Advanced Series in Nonlinear Dynamics, vol. 15* (World Scientific, Singapore).

CROMMELIN, D.T. (2002). Homoclinic dynamics: a scenario for atmospheric ultralow-frequency variability. *J. Atmos. Sci.* **59**, 1533–1549.

DE NIET, A.C., WUBS, F.W., TERWISSCHA VAN SCHELTINGA, A., DIJKSTRA, H.A. (2007). A tailored solver for bifurcation studies of high-resolution ocean models. *J. Comp. Phys.* **227**, 654–679.

DENGG, J., BECKMANN, A., GERDES, R. (1996). The gulf stream separation problem. In: Kraus,W.A. (ed.), *The Warmwatersphere of the North Atlantic Ocean* (Gebr. Borntraeger, Berlin, Germany), pp. 253–290.

DIJKSTRA, H.A. (2005). *Nonlinear Physical Oceanography, A Dynamical Systems Approach to the Large Scale Ocean Circulation and El Ni˜no*, second ed. (Kluwer, Dordrecht/Boston, MA/London), pp. 532.

DIJKSTRA, H.A., GHIL, M. (2005). Low-frequency variability of the large-scale ocean circulation: a dynamical systems approach. *Rev. Geophys.* **43**, RG3002, doi: 10.1029/2002RG000122.

DIJKSTRA, H.A., OKSUZOGLU, H., WUBS, F.W., BOTTA, E.F.F. (2001). A fully implicit model of the three-dimensional thermohaline ocean circulation. *J. Comp. Phys.* **173**, 685–715.

FELIKS, Y., GHIL, M., SIMONNET, E. (2004). Low-frequency variability in the mid-latitude atmosphere induced by an oceanic thermal front. *J. Atmos. Sci.* **61**, 961–981.

FELIKS, Y., GHIL, M., SIMONNET, E. (2007). Low-frequency variability in the mid-latitude baroclinic atmosphere induced by an oceanic thermal front. *J. Atmos. Sci.* **64**, 97–116.

GEAR, C.W., KEVREKIDIS, I.G., THEODOROPOULOS, C. (2002). Coarse integration/bifurcation analysis via microscopic simulators: micro-galerkin method. *Comp. Chem. Engrg.* **26**, 941–963.

GHIL, M. (1976). Climate stability for a Sellers-type model. *J. Atmos. Sci.* **33**, 6–20.

GHIL, M., ALLEN, R.M., DETTINGER, M.D., IDE, K., KONDRASHOV, D., MANN, M.E., ROBERTSON, A., SAUNDERS, A., TIAN, Y., VARADI, F., YIOU, P. (2002). Advanced spectral methods for climatic time series. *Rev. Geophys.* **40** (1), 3.1–3.41, doi: 10.1029/2000GR000092.

GHIL, M., CHILDRESS, S. (1987). *Topics in Geophysical Fluid Dynamics: Atmospheric Dynamics, Dynamo Theory and Climate Dynamics* (Springer-Verlag, New York/Heidelberg, Germany/Berlin, Germany), pp. 485.

GHIL, M., ROBERTSON, A.W. (2000). Solving problems with GCMs: general circulation models and their role in the climate modeling hierarchy. In: *General Circulation Model Developments: Past, Present and Future*, pp. 285–325.

GLENDINNING, P. (1984). Bifurcations near homoclinic orbits with symmetry. *Phys. Lett.* **103A**, 163–166.

GLENDINNING, P., SPARROW, C. (1984). Local and global behavior near homoclinic orbits. *J. Stat. Phys.* **35**, 645–696.

GOLUB, G., VAN LOAN, C. (1996). *Matrix Computations*, third ed. (The Johns Hopkins University Press, London).

GOLUBITSKY, M.I., STEWART, I., SCHAEFFER, D.G. (2000). *Singularities and Group in Bifurcation Theory, vol. 2* (Springer-Verlag, New York), pp. 533.

GRUAIS, I., RITTEMARD, N., DIJKSTRA, H.A. (2005). A priori estimations of a global homotopy residue continuation method. *Nonlinear Funct. Analysis Optim.* **26** (4-5), 507–521.

GUCKENHEIMER, J., HOLMES, P. (1990). *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields*, second ed. (Springer-Verlag, New York/Berlin, Germany), pp. 453.

HELD, I.M. (2005). The gap between simulation and understanding in climate modeling. *Bull. Am. Met. Soc.* **86**, 1609–1614.

HELD, I.M., SUAREZ, M.J. (1974). Simple albedo feedback models of the ice caps. *Tellus* **26**, 613–629.

IOOSS, G., JOSEPH, D.D. (1999). *Elementary Stability and Bifurcation Theory*, second ed. (Springer-Verlag), pp. 356.

JIANG, S., JIN, F.F., GHIL, M. (1995). Multiple equilibria, periodic, and aperiodic solutions in a wind-driven, double-gyre, shallow-water model. *J. Phys. Oceanogr.* **25**, 764–786.

KELLER, H.B. (1977). Numerical solution of bifurcation and nonlinear eigenvalue problems. In: *Applications of Bifurcation Theory* (Academic Press), pp. 359–384.

KEVREKIDIS, I.G., GEAR, C.W., HYMAN, J.M., KEVREKIDIS, P.G., RUNBORG, O., THEODOROPOULOS, C. Equation-free multiscale computation: enabling microscopic simulators to perform system-level tasks. Physics/0209043 at arXiv.org.

KUZNETSOV, Y.A. (1995). *Elements of Applied Bifurcation Theory* (Springer Verlag), pp. 515.

LEGRAS, B., GHIL, M. (1985). Persistent anomalies, blocking and variations in atmospheric predictability. *J. Atmos. Sci.* **42**, 433–471.

LORENZ, E.N. (1963a). Deterministic non-periodic flow. *J. Atmos. Sci.* **20**, 130–141.

LORENZ, E.N. (1963b). The mechanics of vacillation. *J. Atmos. Sci.* **20**, 448–464.

MARSDEN, J.E., MCCRACKEN, M. (1976). *The Hopf Bifurcation and its Applications* (Springer-Verlag, New-York).

MCCALPIN, J., HAIDVOGEL, D. (1996). Phenomenology of the low-frequency variability in a reduced-gravity, quasi-geostrophic double-gyre model. *J. Phys. Oceanogr.* **26**, 739–752.

MCWILLIAMS, J.C. (1977). A note on a consistent quasigeostrophic model in a multiply connected domain. *Dyn. Atmos. Ocean.* **1**, 427–441.

MEACHAM, S.P. (2000). Low-frequency variability in the wind-driven circulation. *J. Phys. Oceanogr.* **30**, 269–293.

NADIGA, B.T., LUCE, B.P. (2001). Global bifurcation of Shilnikov type in a double-gyre ocean model. *J. Phys. Oceanogr.* **31**, 2669–2690.

NEELIN, J.D., BATTISTI, D.S., HIRST, A.C., JIN, F.-F., WAKATA, Y., YAMAGATA, T., ZEBIAK, S. (1998). ENSO theory. *J. Geophys. Res.* **103**, 14261–14290.

NEELIN, J.D., LATIF, M., JIN, F.-F. (1994). Dynamics of coupled ocean-atmosphere models: the tropical problem. *Annu. Rev. Fluid Mech.* **26**, 617–659.

NORTH, G. (1975). Analytical solution to a simple climate model with diffusive heat transport. *J. Atmos. Sci.* **32**, 1301–1307.

NORTH, G.R., MENGEL, J.G., SHORT, D.A. (1983). Simple energy-balance model resolving the seasons and the continents - application to the astronomical theory of the ice ages. *J. Geophys. Res.* **88**, 6576–6586.

OMTA, A.W., DIJKSTRA, H.A. (2003). A physical mechanism for the Atlantic-Pacific flow reversal in the early Miocene. *Glob. Planet. Change* **36**, 265–276.

PEDLOSKY, J. (1987). *Geophysical Fluid Dynamics*, second ed. (Springer-Verlag, New York/Heidelberg, Germany/Berlin, Germany), pp. 710.

PLAUT, G., VAUTARD, R. (1994). Spells of low-frequency oscillations and weather regimes in the northern hemisphere. *J. Atmos. Sci.* **51**, 210–236.

PRIMEAU, F.W. (1998). Multiple equilibria of a double-gyre ocean model with super-slip boundary conditions. *J. Phys. Oceanogr.* **28**, 2130–2147.

QUON, C., GHIL, M. (1992). Multiple equilibria in thermosolutal convection due to salt-flux boundary conditions. *J. Fluid Mech.* **245**, 449–484.

QUON, C., GHIL, M. (1995). Multiple equilibria and stable oscillations in thermosolutal convection at small aspect ratio. *J. Fluid Mech.* **291**, 33–56.

RUELLE, D., TAKENS, F. (1971). On the nature of turbulence. *Comm. Math. Phys.* **20**, 167–192.

SAMELSON, R.M. (2001). Periodic orbits and disturbance growth for baroclinic waves. *J. Atmos. Sci.* **58**, 436–450.

SANCHEZ, J., NET, M., GARCIA-ARCHILLA, B., SIMO, C. (2004). Newton-Krylov continuation of periodic orbits for Navier-Stokes flows. *J. Comput. Phys.* **201**, 13–33.

SCHNEIDER, S.H., DICKINSON, R.E. (1974). Climate modeling. *Rev. Geophys. Space Phys.* **12**, 447–493.

SEYDEL, R. (1994). *Practical Bifurcation and Stability Analysis: From Equilibrium to Chaos* (Springer-Verlag, New York).

SHILNIKOV, L.P. (1965). A case of the existence of a denumerable set of periodic motions. *Sov. Math. Dokl.* **6**, 163–166.

SIMONNET, E. (2005). Quantization of the low-frequency variability of the double-gyre circulation. *J. Phys. Oceanogr.* **35**, 2268–2290.

SIMONNET, E., DIJKSTRA, H.A. (2002). Spontaneous generation of low-frequency modes of variability in the wind-driven ocean circulation. *J. Phys. Oceanogr.* **32**, 1747–1762.

SIMONNET, E., GHIL, M., DIJKSTRA, H.A. (2005). Homoclinic bifurcations in the barotropic quasi-geostrophic double-gyre circulation. *J. Mar. Research* **63**, 931–956.

SIMONNET, E., GHIL, M., DIJKSTRA, H.A. (2007). Quasi-homoclinic behavior of the barotropic quasi-geostrophic double-gyre circulation. *Chaos* in preparation.

SIMONNET, E., GHIL, M., IDE, K., TEMAM, R., WANG, S. (2003a). Low-frequency variability in shallow-water models of the wind-driven ocean circulation. Part I: steady-state solutions. *J. Phys. Oceanogr.* **33**, 712–728.

SIMONNET, E., GHIL, M., IDE, K., TEMAM, R., WANG, S. (2003b). Low-frequency variability in shallow-water models of the wind-driven ocean circulation. Part II: time-dependent solutions. *J. Phys. Oceanogr.* **33**, 729–752.

SIMONNET, E., TEMAM, R., WANG, S., GHIL, M., IDE, K. (1998). Successive bifurcations in a shallowwater ocean model. In: Lecture Notes in Physics **515** (Springer-Verlag), pp. 225–230.

SLEIJPEN, G.L.G., VAN DER VORST, H.A. (1996). A Jacobi-Davidson iteration method for linear eigenvalue problems. *SIAM J. Matrix Anal. Appl.* **17**, 410–425.

SPARROW, C.T. (1983). *Lorenz Equations: Bifurcations, Chaos, and Strange Attractors* (Springer-Verlag, [Applied Mathematical Sciences]), pp. 269.

SPEICH, S., DIJKSTRA, H.A., GHIL, M. (1995). Successive bifurcations in a shallow-water model applied to the wind-driven ocean circulation. *Nonlinear Processes Geophys.* **2**, 241–268.

STEWART, W.J., JENNINGS, A. (1981). A simultaneous iteration algorithm for real matrices. *ACM Trans. Math. Softw.* **7**, 184–198.

STOMMEL, H. (1961). Thermohaline convection with two stable regimes of flow. *Tellus* **2**, 230–244.

TAKENS, F. (1981). Detecting strange attractors in turbulence. In: Rand, D.A., Young, L.-S. (eds.), *Lecture Notes in Mathematics* **898** (Springer), pp. 366–381.

TE RAA, L.A., DIJKSTRA, H.A. (2002). Instability of the thermohaline ocean circulation on interdecadal time scales. *J. Phys. Oceanogr.* **32**, 138–160.

TEMAM, R. (1997). *Infinite-Dimensional Dynamical Systems in Mechanics and Physics, vol. 68*, Applied Mathematical Science (Springer-Verlag, New York), pp. 643.

THUAL, O., MCWILLIAMS, J.C. (1992). The catastrophic structure of thermohaline convection in a two-dimensional fluid model and a comparison with low-order box models. *Geophys. Astrophys. Fluid Dyn.* **64**, 67–95.

TUCKER, W. (2002). A rigorous ODE solver and Smale's 14th problem. *Found. Comput. Math.* **2**, 53–117.

VAN DER AVOIRD, F., DIJKSTRA, H.A., NAUW, J.J., SCHUURMANS, C.J.E. (2002). Nonlinearly induced low-frequency variability in a midlatitude coupled ocean-atmosphere model of intermediate complexity. *Clim. Dyn.* **19**, 303–320.

VAN DER VAART, P.C.F., DIJKSTRA, H.A., JIN, F.-F. (2000). The Pacific cold tongue and the ENSO mode: unified theory within the Zebiak-Cane model. *J. Atmos. Sci.* **57**, 967–988.

VERONIS, G. (1963). An analysis of wind-driven ocean circulation with a limited number of Fourier components. *J. Atmos. Sci.* **20**, 577–593.

VERONIS, G. (1966). Wind-driven ocean circulation. ii: Numerical solution of the nonlinear problem. *Deep Sea Res.* **13**, 31–55.

WEIJER, W., DIJKSTRA, H.A., OKSUZOGLU, H., WUBS, F.W., DE NIET, A.C. (2003). A fully-implicit model of the global ocean circulation. *J. Comp. Phys.* **192**, 452–470.

WIGGINS, S. (1988). *Global Bifurcations and Chaos: Analytical Methods* (Springer-Verlag [Applied Mathematical Sciences]), pp. 494.

WOLFE, C.L., SAMELSON, R.M. (2006). Normal-mode analysis of a baroclinic wave-mean oscillation. *J. Atmos. Sci.* **63**, 2795–2812.

WUBS, F., DE NIET, A.C., DIJKSTRA, H.A. (2006). The performance of implicit ocean models on b- and c-grids. *J. Comp. Phys.* **211**, 210–228.

ZOLDI, S.M., GREENSIDE, H.S. (1998). Spatially localized unstable periodic orbits of a highdimensional chaotic system. *J. Atmos. Sci.* **57**, 2511–2514.

Time-Periodic Flows in Geophysical and Classical Fluid Dynamics

R. M. Samelson

College of Oceanic and Atmospheric Sciences, 104 COAS Administration Building, Oregon State University, Corvallis, OR 97331-5503, USA

Abstract

Time-periodic flows form an important special class of fluid motions. This class includes, for example, the many well-known families of linear propagating waves. Recently, nonlinear time-periodic flows have received increased attention in several contexts in geophysical and classical fluid mechanics. Despite the relative simplicity of their Eulerian representation, time-periodic velocity fields can give rise to complex, aperiodic Lagrangian motion; models of this type can be useful for understanding fluid transport processes in the ocean and atmosphere. One approach to the analysis of chaotic dynamical systems, including certain models of geophysical flows, is based on the identification of a large set of unstable periodic cycles, each of which corresponds to an independent time-periodic flow. Time-periodic flows provide accessible examples in which to explore mechanisms of disturbance growth in general time-dependent flows, a problem of practical interest in numerical weather and ocean prediction. Recent advances in the theory of the transition to turbulence in classical pipe flow involve bifurcation analysis for a special class of time-periodic flows.

1. Introduction

The motion of a fluid in two- or three-dimensional space can be represented by the time-dependent map of the fluid elements from their initial positions $\mathbf{a}$ to their final positions $\mathbf{X}(\tau; \mathbf{a})$,

$$\mathbf{x} = \mathbf{X}(\tau; \mathbf{a}), \qquad \mathbf{X}(0; \mathbf{a}) = \mathbf{a}, \tag{1.1}$$

Computational Methods for the Atmosphere and the Oceans
Special Volume (Roger M. Temam and Joseph J. Tribbia, Guest Editors) of
HANDBOOK OF NUMERICAL ANALYSIS, VOL. XIV
P.G. Ciarlet (Editor)
Copyright © 2009 Elsevier B.V.
All rights reserved
ISSN 1570-8659
DOI 10.1016/S1570-8659(08)00204-4

where time τ appears as a parameter. The fluid velocity $\mathbf{v}(\mathbf{x}, t)$ is the rate of change of position of a fluid element,

$$\frac{\partial \mathbf{x}(\tau, \mathbf{a})}{\partial \tau}|_{\tau=t} = \mathbf{v}[\mathbf{X}(t, \mathbf{a}), t] = \mathbf{v}(\mathbf{x}, t). \tag{1.2}$$

In Eq. (1.2), the partial derivatives are taken with the initial positions $\mathbf{a}$ held constant, i.e., with respect to τ in Eq. (1.1) and following the fluid motion. In many situations, it is possible to obtain dynamical equations that determine the evolution of $\mathbf{v}(\mathbf{x}, t)$ directly, without solving for the trajectory map $\mathbf{X}$. More generally, the dynamical equations can frequently be posed and solved in the Eulerian spatial coordinates $\mathbf{x}$, without reference to the Lagrangian coordinates $\mathbf{a}$. In that case, the flow can be considered time-periodic if $\mathbf{v}(\mathbf{x}, t)$, and any other variables needed to complete the dynamical description are periodic in t:

$$\mathbf{v}(\mathbf{x}, t + T) = \mathbf{v}(\mathbf{x}, t) \tag{1.3}$$

for all t and some fixed constant $T > 0$, where T is the period of the flow. Note that the trajectory map $\mathbf{X}$ need not in general be periodic, even if the flow is periodic by the criterion (Eq. (1.3)); this apparent contradiction figures in one of the topics considered below.

Time-periodic flows that satisfy Eq. (1.3) arise in many situations in geophysical and classical fluid dynamics. The simplest of these are linear traveling waves. An example of these are the Rossby waves of quasi-geostrophic theory (e.g., Gill [1982], Pedlosky [1987]). Rossby waves satisfy the linearized forms of various potential vorticity equations, such as

$$q_t + \psi_x q_y - \psi_y q_x = 0, \quad q = \beta y + \nabla\psi - F\psi, \tag{1.4}$$

in which β and F are constants, and the streamfunction ψ and potential vorticity q depend on the horizontal coordinates x and y and on the time t. For small amplitude disturbances, Eq. (1.4) may be approximated by

$$(\nabla\psi - F\psi)_t + \beta\psi_x = 0, \tag{1.5}$$

which has plane wave solutions

$$\psi = \psi_0 \,\mathrm{Re}\{\exp[-i(kx + ly - \omega t + \phi_0)]\}, \quad \omega = -\frac{\beta k}{k^2 + l^2 + F}, \tag{1.6}$$

where Re denotes the real part, and k, l, ω, ϕ_0, and ψ_0 are real constants. These are dispersive, propagating waves, with phase speed $c = \omega/k$. By the criterion of Eq. (1.3), they are also time-periodic flows, with velocity field $\mathbf{v}(x, y, t) = (-\psi_y, \psi_x)$ and period $T = 2\pi/\omega$.

Linear time-periodic propagating-wave flows are fundamental to many fields of geophysical and classical fluid mechanics (Gill [1982], Lighthill [1978], Pedlosky [2003]). From a mathematical point of view, they arise naturally through Fourier analysis of linear wave equations, and their properties are generally well understood. The present review will instead briefly summarize a few areas of recent interest involving nonlinear time-periodic flows.

2. Lagrangian motion

An interesting property of time-periodic flows, alluded to in the Introduction, is that the motion of fluid elements in time-periodic velocity fields need not themselves be periodic. Instead, they can be complex and aperiodic. An illustrative geophysical example of this is the model of a two-dimensional meandering jet introduced by PIERREHUMBERT [1991] and discussed at length in a recent pedagogical text (SAMELSON and WIGGINS [2006]).

In this model, the meandering jet is represented by the streamfunction $\psi(x, y, t)$, where the zonal coordinate x is moving with the phase speed of the primary translating meander, and ψ is the sum of terms representing the primary meander and a disturbance,

$$\psi(x, y, t) = \psi_0(x, y) + \varepsilon\psi_1(x, y, t). \tag{2.1}$$

Note that the disturbance amplitude ε need not be small, though some elements of the analysis are simplified if it is. The streamfunction ψ_0 for the steady, unperturbed ($\varepsilon = 0$) flow is given by,

$$\psi_0(x, y) = -cy + A \sin kx \sin y, \tag{2.2}$$

and the perturbation ψ_1 is periodic in t,

$$\psi_1(x, y, t + T) = \psi_1(x, y, t). \tag{2.3}$$

The constant c is the phase speed of the primary meander represented by ψ_0, A is a constant meander amplitude, and the entire flow is confined in a periodic or infinite channel in x, with rigid walls at $y = 0$ and $y = \pi$. The corresponding equations for the fluid trajectories are

$$\frac{\mathrm{d}x}{\mathrm{d}t} = c - A \sin kx \cos y - \varepsilon\psi_{1y}(x, y, t), \quad \frac{\mathrm{d}y}{\mathrm{d}t} = kA \cos kx \sin y + \varepsilon\psi_{1x}(x, y, t). \tag{2.4}$$

Note that the streamfunction (Eq. (2.1)) is a kinematic model; the streamfunction for the velocity field is specified directly to represent the general structure of a meandering jet rather than being obtained as the solution to a given set of fluid dynamical equations.

Since $\psi_0(x, 0) = \psi_0(x, \pi) = 0$, the normal velocity of the unperturbed flow vanishes on the rigid channel walls. For $|c/A| < 1$, the tangential velocity will also vanish at the two points on $y = 0$ where $c = A \sin kx$ and the two points on $y = \pi$ where $c = -A \sin kx$. These four points are thus stagnation points of the unperturbed flow. In this case, the level set $\psi_0(x, y) = 0$ includes interior curves that connect pairs of these points and divide the unperturbed flow into distinct flow regimes. The flow regimes consist of either recirculations adjoining the channel walls, with no mean velocity, or a contiguous, meandering jet with a mean velocity directed along the channel. In the unperturbed flow, a fluid element that is initially located within a recirculation regime is trapped there forever, while an element in the jet regime can never join a recirculation cell. From the point of view of classical dynamics, the bounding curves that connect the pairs of stagnation points are separatrices.

When $\varepsilon \neq 0$ and ψ_1 has the form of a propagating wave, such as

$$\psi_1(x, y, t) = \sin(k_1 x - \omega_1 t) \sin y, \tag{2.5}$$

the velocity field $\mathbf{v}(\mathbf{x}, t) = (-\psi_y, \psi_x)$ in Eq. (2.4) is time-periodic, but the solutions $\mathbf{x} = \mathbf{X}(x_0, y_0, t)$ of Eq. (2.4) may include trajectories that are aperiodic. In this case, the boundaries between recirculation and jet regimes break down, and fluid may be exchanged between the regimes: fluid elements that are initially located within a recirculation regime may escape into the jet and be carried arbitrarily far along the channel, while fluid elements initially in the jet may be captured by recirculation and remain trapped there ever after. From the point of view of classical dynamics, this situation corresponds to splitting of the separatrices under the perturbation. Although the breakdown of regime boundaries and the quantification of fluid exchange between regimes are probably of greater physical interest, note that, in this situation, the fluid motion may be chaotic in the rigorous sense of the word, with a specific collection of fluid trajectories forming a chaotic invariant set: a set in a bounded region containing a countably infinite number of periodic trajectories, an uncountably infinite number of a aperiodic trajectories, and a trajectory that approaches every point in the set arbitrarily closely (a "dense" trajectory). Such sets typically have a fractal structure similar to that of Cantor sets.

The mathematical framework of dynamical systems theory can be used to analyze fluid exchange processes of this type in extensive detail. Time-periodic flows are of special interest in this connection because the periodicity allows regime boundaries, with respect to which fluid exchange is computed, to be constructed that are fixed in space. For this reason, time-periodic flows provide particularly appealing illustrative and didactic examples, which have stimulated much of the original research in this area (e.g., Aref [1984], Knobloch and Weiss [1987], Pierrehumbert [1991], Samelson [1992]). More generally, recent work, some of which has been summarized in a geophysically based pedagogical text (Samelson and Wiggins [2006]), has shown that this approach can be extended well beyond the special class of time-periodic flows.

3. Cycle expansions

The possibility that chaotic motions associated with the strange attractors of deterministic dynamical systems could be useful quantitative models of the observed irregular, unpredictable evolution of large-scale flows in the atmosphere and ocean has fascinated geophysical fluid dynamicists since it was first suggested by Lorenz [1963], in his pioneering analysis of aperiodic, deterministic motion. In this case, the invariant set appears in a multidimensional phase space that describes the state of the fluid system rather than in physical space as in the Lagrangian motion problem. Different trajectories, or orbits, in phase space describe different possible histories of the system, depending on initial conditions. The orbit on a strange attractor is aperiodic and unstable: the state of the system never repeats exactly, and almost any arbitrarily small disturbance to the state of the system will grow exponentially, leading to exponentially sensitive dependence of the evolution of the system on its initial conditions. Qualitatively, the unstable and quasi-oscillatory but non-repeating character of these orbits is consistent with observations

of the large-scale atmosphere and ocean. Quantitatively, the relevance of this model is more difficult to establish because the atmosphere and ocean are not isolated systems with fixed external forcing and because the statistical recurrence times for the global atmosphere and ocean are so long that near recurrences have not been documented in the historical record, despite the recognizably oscillatory nature of many aspects of observed atmospheric and oceanic variability.

Nonetheless, chaotic motion on strange attractors remains at least a powerful quantitative metaphor for the observed irregularity of the atmosphere and ocean and an object of continuing geophysical interest. A beautiful and elegant approach to the analysis of dynamics on such sets is the periodic orbit theory (Artuso, Aurell and Cvitanović [1990a,b], Cvitanović, Artuso, Mainieri, Tanner and Vattay [2005], Gutzwiller [1990]). This approach uses the set of unstable periodic orbits associated with the attractor as a basis for analysis of the aperiodic motion. Essentially, the countable set of unstable periodic trajectories of a chaotic invariant set is used to approximate an associated dense orbit that describes the attractor. The mathematical theory for this approach is still relatively incomplete. From a physical point of view, however, the approach is appealing for its combination of decomposition into orbits with finite – and thus arguably physical – cycle lengths and an intuitive stability-weighted averaging procedure. This combination has allowed physical progress to be made using variants of the approach even in the absence of a comprehensive and rigorous mathematical theory.

The periodic orbit, or cycle, expansion procedure and the supporting theory are described in detail by Cvitanović, Artuso, Mainieri, Tanner and Vattay [2005]. A brief sketch of the approach is given here, with notation closely following that of Cvitanović, Artuso, Mainieri, Tanner and Vattay [2005]. The present summary resembles that of McNamara [1994]. The cycle expansion allows the computation of averages over the chaotic attractor through the dynamical ζ function,

$$\frac{1}{\zeta(s)} = \prod_p (1 - t_p), \tag{3.1}$$

where the product is taken over all the distinct prime cycles, indexed by p, and t_p is a weight associated with the p-th cycle (Cvitanović, Artuso, Mainieri, Tanner and Vattay [2005]). The average is computed by finding the zero of the dynamical zeta function, i.e., the point $s = s_0$ such that

$$\frac{1}{\zeta(s_0)} = 0. \tag{3.2}$$

The deceptively simple Eq. (3.1) hides an infinite amount of detail. The weight t_p is defined by

$$t_p = t_p(s, \beta; z) = \frac{1}{|\Lambda_p|} \exp(\beta \Phi_p - s T_p)\, z^{n_p}, \tag{3.3}$$

where Λ_p is the leading Floquet multiplier of the p-th cycle, β is a parameter, Φ_p is the average of the property of interest over the p-th cycle, and s is the dependent variable.

The last term is inserted as a counter that is useful for ordering the cycles during the necessary truncation of the infinite product (Eq. (3.1)); in it, z is a dummy variable associated with the index n (which measures the order of a cycle in terms of the number of times it traverses a Poincaré section), n_p is the order of the p-th cycle, and eventually z is set to 1. For scalar functions ϕ of the dynamical variables $\mathbf{x}$, the cycle average Φ_p is defined by

$$\Phi_p = \int_0^{T_p} \phi[\mathbf{x}_p(t)]\,\mathrm{d}t, \tag{3.4}$$

where $\mathbf{x}_p(t)$ is the solution along the p-th cycle.

The desired time and phase-space volume average $\langle\phi\rangle$ is defined by

$$\langle\phi\rangle = \lim_{t\to\infty} \frac{1}{Vt} \int_V \Phi^t(\mathbf{a})\,\mathrm{d}a, \tag{3.5}$$

where $\Phi^t(\mathbf{a})$ is the integral of the function ϕ along the trajectory $\mathbf{x}(t;\mathbf{a})$ passing through $\mathbf{a}$ at $t = 0$,

$$\Phi^t(\mathbf{a}) = \int_0^t \phi[\mathbf{x}(t;\mathbf{a})]\,\mathrm{d}t, \tag{3.6}$$

$\mathrm{d}a$ is the phase-space volume element, and $V = \int_V \mathrm{d}a$ is the bounded volume of phase space in which the solution remains. Remarkably, the average $\langle\phi\rangle$ is related to the variable s in Eq. (3.3) by

$$\langle\phi\rangle = \left(\frac{\partial s_0(\beta)}{\partial\beta}\right)_{\beta=0}, \tag{3.7}$$

where the zero $s = s_0$ is regarded as a function of the parameter β.

The relation expressed in Eq. (3.7) is not obvious. The starting point for the argument that establishes it is the definition of the generating function $Q(\beta)$, where

$$Q(\beta) = \frac{1}{t} \ln\left(\frac{1}{V}\int_V\int_V \exp[\beta\Phi^t(\mathbf{a})]L^t(\mathbf{a}',\mathbf{a})\,\mathrm{d}a\,\mathrm{d}a'\right). \tag{3.8}$$

In Eq. (3.8), the Perron-Frobenius operator $L^t(\mathbf{a}',\mathbf{a}) = \delta[\mathbf{a}' - \mathbf{x}(\mathbf{t};\mathbf{a})]$ translates densities $\rho(0;\mathbf{a})$ to densities $\rho(t;\mathbf{a})$ following the motion $\mathbf{x}(t;\mathbf{a})$. Then,

$$\left(\frac{\partial Q(\beta)}{\partial\beta}\right)_{\beta=0} = \frac{1}{Vt}\int_V \Phi^t(\mathbf{a})\,\mathrm{d}a \to \langle\phi\rangle \ \text{ as } \ t\to\infty, \tag{3.9}$$

which will agree with Eq. (3.7) if also $Q(\beta) \to s_0(\beta)$ as $t\to\infty$. The integrand in Eq. (3.8) is a continuous transformation of $\mathbf{a}$-space into itself, and one can interpret the longtime asymptotics in terms of an eigenvalue problem, for which the leading eigenvalue determines the average. The expansion of the associated determinant leads to the polynomial product (Eq. (3.1)). Moreover, since the Perron-Frobenius operator

vanishes except at exact recurrences, it is only the periodic cycles that contribute to the expansion. CVITANOVIĆ, ARTUSO, MAINIERI, TANNER and VATTAY [2005] show that the eventual consequence of these considerations is the expressions in Eq. (3.1) – (3.3) and (3.7) for the average (Eq. (3.5)).

To evaluate Eq. (3.7), it is necessary to compute a large number of cycles, expand the infinite product in Eq. (3.1) into an infinite series, and truncate the series at the corresponding point. The convergence of the expansion depends on the approximate cancellation in the series of contributions from longer-period cycles by products of contributions from related, shorter-period cycles. In order to optimize this cancellation in the truncated series, and simply to locate the cycles, it is generally necessary to construct a set of symbol sequences, each of which corresponds to a particular cycle in the truncation. The symbols may, for example, identify the distinct subdomains on a Poincaré section through which the orbits pass.

Unfortunately, the requirement that this explicit symbolic dynamics be constructed for the cycles severely restricts the complexity of the systems to which the approach can be applied in its full power. In practice, it seems to mean essentially that the system dynamics must be approximately reducible to a one-dimensional or, perhaps, a two-dimensional map. Geophysically relevant examples of such systems do exist, however. MCNAMARA [1994] has carried out the cycle expansion analysis for the LORENZ [1963] equations, and SAMELSON [2001b] developed a symbolic dynamics and identified the first 157 cycles for the weakly nonlinear baroclinic wave model of PEDLOSKY [1971]. From a practical point of view, MCNAMARA [1994] finds that, if averages of only a small number of quantities are needed, direct integration of the equations over a long time interval is more efficient than the cycle expansion, but also that once the cycle expansion is obtained, many different averages can be computed quickly with little additional effort. In either case, the cycle expansion offers conceptual insight through its approximate representation of the chaotic attractor orbit in terms of the finite-period cycles.

In some cases, useful physical results can still be extracted from a periodic orbit analysis, even if the symbolic dynamics construction is unavailable. It may be possible to obtain a qualitative, and even partial quantitative, description of the basic structure of a sufficiently simple attractor from a few periodic orbits or even a single one. The weighting factor t_p in Eq. (3.1) is inversely proportional to the stability of the cycle as Eq. (3.3) shows. An empirical expansion based on sets of cycles located by direct numerical computation, and inversely weighted by stability in an analogous manner, can be used in the absence of information on the symbolic dynamics. Such an approach has been taken, for example, by ZOLDI and GREENSIDE [1998] in a study of the Kuramoto-Sivashinsky equation, a partial differential equation of advection-diffusion form that has been studied in part with geophysical motivations.

The periodic orbit theory offers an elegant and suggestive conceptual framework for the analysis of the physical mechanisms that give rise to irregular, aperiodic variability in deterministic systems. While systematic computations of the cycle expansion have thus far been carried out only for relatively simple, low-order dynamical models, with a limited number of unstable modes, CHRISTIANSEN, CVITANOVIĆ and PUTKARADZE [1997] suggest that the periodic orbit theory may relate directly to the problem of fluid

turbulence. Although the analysis and computations that would rigorously test this suggestion appear still out of reach, the hypothesis can be naturally extended to the broader phenomena of irregular, aperiodic large-scale motions in the atmosphere and ocean, many of which have a quasi-oscillatory character. This fascinating possibility is, at the very least, sure to stimulate new views and analyses of atmospheric and oceanic variability.

4. Linear disturbance growth

The tendency for small disturbances to amplify in midlatitude storm tracks and other regions of atmospheric instability is the main source of the loss of medium-range predictability in operational numerical weather forecasting. Related dynamical processes in the ocean, and in coupled ocean-atmosphere models, present analogous obstacles to prediction, which will become increasingly apparent as ocean forecasting systems based on ocean dynamical models become operational over the next decade. A fundamental understanding of the mechanisms of disturbance growth in geophysical fluid flows is thus of considerable practical, as well as scientific, interest.

The stability properties of steady flows have been studied extensively in a wide variety of geophysical contexts. Baroclinic instability, the fundamental disturbance growth mechanism associated with cyclogenesis in the midlatitude atmosphere, was discovered by Charney [1947] and Eady [1949] through normal mode analysis of steady parallel baroclinic shear flows representing the midlatitude westerlies. These classical analyses provide the fundamental mechanistic models that inform our conceptual understanding of the physical processes of disturbance growth in midlatitude storm tracks.

In numerical weather prediction, ensemble forecast methods are now routinely used to provide probabilistic information to supplement a deterministic "control" forecast that is based on the best estimate of the initial atmospheric state. Similar methods will be in increasing use in ocean prediction where the frequently sparse data will generally lead to substantial initial uncertainty. In these methods, a set of numerical forecasts based on slightly different initial conditions is used to provide information on the reliability of the control forecast and possible alternative outcomes. The size of the ensemble is limited by computational considerations, so the ensemble members must be chosen carefully.

The ensemble is generated by perturbing the best estimate of the initial state in various ways. Clearly, an ensemble member that is generated by adding a growing disturbance will be of greater interest than one generated by adding a decaying disturbance since in the latter case the ensemble member will essentially duplicate the control forecast. The classical theory of baroclinic instability of a steady zonal flow remains the primary source of our conceptual understanding of the role of instabilities in disturbance growth. However, the atmosphere and ocean are not steady flows, but are continuously evolving. Relatively little theoretical attention has been paid to the instability of time-dependent flows. Thus, the dynamical element of ensemble generation theory remains fundamentally incomplete.

In this context, time-periodic basic flows provide an accessible framework in which to investigate the mechanisms of disturbance growth on time-dependent flows. Such flows contain many of the same technical and conceptual elements as flows with more

general time-dependence, but admit analysis in terms of normal modes. In this case, the linear eigenmodes are also themselves time-dependent. A number of recent studies have examined periodic solutions of geophysical fluid models and their stability (FANTINI and DAVOLIO [2001], ITOH and KIMOTO [1996], JIANG and GHIL [1997], JIANG, JIN and GHIL [1995], KAZANTSEV [1998, 2001]). The present review will briefly summarize two examples: a sequence of studies of unstable baroclinic wave-mean oscillations in a two-layer quasi-geostrophic model (SAMELSON [2001a,b], SAMELSON and WOLFE [2003], WOLFE and SAMELSON [2006]) and an analysis of a model of the El Niño-Southern Oscillation (ENSO) (SAMELSON and TZIPERMAN [2001]).

The baroclinic wave model is based on the finite amplitude generalization of the PHILLIPS [1954] model of baroclinic instability. It consists of a two-layer, f-plane, quasi-geostrophic fluid in a periodic channel with a rigid lid at the upper boundary and Ekman dissipation at both upper and lower boundaries. The mean layer depths are equal, and the standard no-normal-flow boundary conditions are imposed at the channel walls along $y = \{0, 1\}$. The equations are

$$Q_{jt} + J(\Psi_j, Q_j) = -r\nabla^2\Psi_j, \qquad j = 1, 2, \tag{4.1}$$

where Ψ_j and Q_j are the streamfunction and potential vorticity, respectively, in layer j, $j = 1, 2$, and $\nabla^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2$. The upper- and lower-layer ($j = 1$ and $j = 2$, respectively) streamfunctions each consist of a constant mean shear plus an arbitrary disturbance,

$$\Psi_1(x, y, t) = -\frac{1}{2}U_s y + \psi_1(x, y, t), \tag{4.2}$$

$$\Psi_2(x, y, t) = +\frac{1}{2}U_s y + \psi_2(x, y, t). \tag{4.3}$$

The potential vorticities are

$$Q_1(x, y, t) = \nabla^2\Psi_1 - F(\Psi_1 - \Psi_2), \tag{4.4}$$

$$Q_2(x, y, t) = \nabla^2\Psi_2 - F(\Psi_2 - \Psi_1). \tag{4.5}$$

Here U_s is the mean shear of the basic flow, and r is the Ekman damping coefficient.

Periodic orbits of this model, and their normal-mode stability, have been analyzed in two distinct parameter ranges: a weakly nonlinear limit, in which the mean shear U_s is marginally supercritical, and a strongly nonlinear – but not turbulent – regime, in which the mean shear U_s is of order one. In the weakly nonlinear limit (SAMELSON [2001a,b]), the dynamical system reduces to the evolution of a single zonal wave, with independent upper- and lower-layer amplitudes, and its interaction with a zonal mean-flow correction that depends continuously on the cross-channel coordinate y but can be accurately represented by a small number of cross-channel Fourier components. In the strongly nonlinear regime (SAMELSON and WOLFE [2003], WOLFE and SAMELSON [2006, 2008]), the system is high-dimensional, and the analysis is conducted using a spectral model with several thousand spectral components.

Most of the periodic wave-mean oscillations identified in these studies were linearly unstable so that special numerical techniques based on Newton's method were necessary to locate them. For the weakly nonlinear, low-dimensional case, a direct Newton solver in the full eight-dimensional phase space was used, with a first guess for the Newton iteration obtained from a stable periodic orbit that exists for a nearby value of the friction parameter. For the strongly nonlinear, high-dimensional case, a Newton-Picard method with subspace iteration (LUST, ROOSE, SPENCE and CHAMPNEYS [1998]) was used, with a first guess from a near recurrence in a long aperiodic time series obtained by direct numerical integration at the same, fixed parameter values. The idea of the Newton-Picard iteration is to use Newton's method in a small subspace that contains the growing modes, and direct forward (Picard) integration in the much larger complement subspace that contains only the decaying modes. The combined Newton-Picard iteration is much more efficient than the direct Newton's method on the full space because of the great reduction in the dimension of the space in which the Newton's method is used. The subspace iteration is necessary because the subspace of growing modes is not known until after the solution is obtained.

In either case, small disturbances to the basic wave-mean oscillation satisfy a linear version of Eq. (4.1), in which the equations are linearized about the periodic wave-mean oscillation. After the solution for the basic, nonlinear oscillation has been obtained by the Newton or Newton-Picard methods outlined above, the solutions of these linearized equations may be analyzed following standard methods for linear differential systems with periodic coefficients (e.g., CODDINGTON and LEVINSON [1955]), often referred to as Floquet theory. For the finite-dimensional case relevant to the numerical solutions considered here, the standard Floquet theory is sufficient. For the infinite-dimensional case relevant to the underlying partial differential equations, such as Eq. (4.1), some results on Floquet theory are available (e.g., BREVDO and BRIDGES [1997], KUCHMENT [1993]), but the mathematical description is considerably less mature.

For a time-periodic solution with period T in an N-dimensional phase space, any linear disturbance to the solution may be written as a sum of the Floquet eigenvectors $\{\phi^j(t),\, j = 1, 2, \ldots, N\}$. The Floquet vectors $\phi^j(t)$, which are themselves time-dependent, are the normal modes of the time-periodic solution. Each Floquet vector has the form $\phi^j(x, y, t) = \Phi^j(x, y, t) \exp(\lambda_j t)$, for a time-periodic function $\Phi^j(x, y, t)$, with $\Phi^j(x, y, t + T) = \Phi^j(x, y, t)$ (or possibly $\Phi^j(x, y, t + T) = -\Phi^j(x, y, t)$), where in the two-layer model the notation Φ^j represents both the upper- and lower-layer components of the disturbance, and λ_j is a constant. In general, the functions ϕ^j and Φ^j and the characteristic exponent λ_j may be complex. If the magnitude $|\Lambda_j| = |\exp(\lambda_j T)|$ of the j-th Floquet multiplier is greater than one, the j-th mode is growing; if it is equal to one, the j-mode is neutral; if it is less than one, it is decaying. An equivalent statement can be made based on the sign of the real part of λ_j. Since the time-derivative of the basic cycle is itself a linear mode, there is always at least one neutral mode. If the dynamics admit a translation symmetry, there will be at least one additional neutral mode (WOLFE and SAMELSON [2006]). The Floquet vectors may be obtained numerically by integrating a complete set of independent linear disturbances over one cycle of the basic oscillation and then solving the resulting eigenvalue problem. If only a few leading vectors are

required, iterative techniques (LUST, ROOSE, SPENCE and CHAMPNEYS [1998]) may be used instead.

Despite the differences in the amplitudes and dynamics of the basic baroclinic wave-mean oscillations in the weakly (SAMELSON [2001a,b]) and strongly (SAMELSON and WOLFE [2003], WOLFE and SAMELSON [2006]) nonlinear cases, the qualitative structure of the corresponding Floquet spectra are similar. In each case, there are a small number of growing or relatively weakly decaying modes with spatial scales similar to those of the basic oscillation, a corresponding number of rapidly decaying, inviscidly damped modes, and a larger number of modes that decay at an intermediate rate that is very close to the imposed Ekman-friction damping rate. The inviscidly damped modes are analogous to the complex conjugates of the growing modes in classical steady-flow instability theory, and each generally resembles one of the growing or weakly decaying modes but with phase differences that support inviscid decay rather than growth. These sets of modes may be collectively referred to as "wave-dynamical" modes as their dynamics are dominated by large-scale wave processes. The large number of frictionally decaying modes, on the other hand, have dynamics that are dominated by mean-flow advection and frictional dissipation and may be collectively referred to as "damped-advective" modes. They are evidently related to the singular neutral modes that arise in classical baroclinic instability theory for disturbances to steady zonal shear flows, which represent purely advective solutions to the inviscid linearized equations.

Direct numerical integration and solution to the resulting several-thousand-square matrix eigenvalue problem were used by WOLFE and SAMELSON [2006] to obtain a complete set of Floquet vectors for the strongly nonlinear baroclinic wave-mean oscillation. Different, empirically based methods were used by SAMELSON and TZIPERMAN [2001] to solve for Floquet vectors in a geophysical model of even greater computational complexity: an idealized coupled ocean-atmosphere model with roughly 34,000 degrees of freedom, which had been developed for the study of ENSO dynamics in the tropical ocean and atmosphere. In the SAMELSON and TZIPERMAN [2001] analysis, long time series of multiple fields were first decomposed into empirical orthogonal functions (EOFs), and an analysis of the distribution of states in a low-order phase space consisting of the EOF amplitudes was then conducted. This analysis yielded an approximate periodic cycle in the reduced phase space.

The stability of this model ENSO cycle was then estimated by constructing linear maps over subintervals of the cycle from least-square fits to the distributions of states near the cycle at the beginning and end of each subinterval. The Floquet vectors were obtained by composing these maps over all the subintervals and solving the resulting eigenvalue problem. The time-dependent structure of the vectors was subsequently reconstructed using the sequence of subinterval maps. SAMELSON and TZIPERMAN [2001] used the results to show that the instability of the model ENSO cycle represented by the approximate periodic cycle arose from dynamical processes very similar to those that sustained the cycle itself, and that the time-dependent structure of the growing mode was consistent with the observed loss of predictability during the growth-phase of El Niño conditions. Thus, as in the baroclinic wave problem, the Floquet vectors of the time-periodic flow proved to be interesting dynamical structures that warranted close examination.

This general result should not be surprising. As the normal modes for linear disturbances on the time-dependent basic flow, the Floquet modes are intrinsic dynamical objects. Their structure and characteristic exponents are norm-independent, unlike other frequently used measures of disturbance growth, including optimal disturbances (Farrell [1989]), whose structure and growth rates depend directly on the choice of norm and on the optimization interval over which they are computed. If an asymptotically stable cycle becomes asymptotically unstable as a parameter of the basic flow is varied, this must occur through a change in sign of the real part of one of the Floquet exponents. Thus, from the point of view of bifurcation theory, the Floquet modes are the fundamental objects determining the stability of the basic oscillation, and it follows that they must therefore also be of fundamental importance for a basic understanding of the ensemble generation problem.

Moreover, if the unstable periodic cycles are associated with a chaotic attractor that describes the long-term, aperiodic behavior of the flow, as in the weakly nonlinear baroclinic wave model, these normal modes can be used to define a "tangent" direction that describes the local structure of the attractor (Samelson [2001b]). This structure determines the distribution of accessible states that should be spanned by an appropriate ensemble system, in the case in which uncertainty in initial conditions is the dominant error source. The periodic orbit theory discussed in the previous section makes explicit the connection between the stability properties of the cycles and the stability properties of the chaotic attractor, and thus between the Floquet modes of the cycles and the Lyapunov characteristic vectors of the attractor: the former may be considered a special case of the latter, for periodic basic flows, or the latter as a generalization of the former, for aperiodic basic flows. In either case, the results of normal-mode analyses of time-periodic basic flows provide, at minimum, an accessible and concrete illustration of the relevance of the normal-mode framework to the study of disturbance growth on time-dependent basic flows.

5. Transition to turbulence

Reynolds [1883] presented the first careful experimental studies of the the transition to turbulence in pressure-driven fluid flow through a pipe. These studies showed that, at sufficiently large values of the dimensionless ratio $\mathrm{Re} = UD/\nu$ – where U is the mean speed of the flow through the pipe, D is the pipe diameter, and ν is the kinematic viscosity – steady laminar flow spontaneously gives way to irregular, fluctuating, turbulent flow. Despite over one hundred years of study, this problem of transition to turbulence in pipe flow remains incompletely understood (Hof, van Doorne, Westerweel, Nieuwstadt, Faisst, Eckhardt, Wedin, Kerswell and Waleffe [2004], Wedin and Kerswell [2004]).

The laminar pipe flow continues to exist as a mathematical solution to the fluid equations at arbitrarily large Re. It might be anticipated, by analogy with other classical fluid problems such as Rayleigh-Benard convection, that the transition to time-dependent motion should be associated with the bifurcation to instability of the laminar solution at a critical value Re_c of the Reynolds number Re. However, the linear stability properties

of the laminar pipe flow solution are inconsistent with established experimental results. Although the stability analysis is complicated by the three-dimensional geometry of the pipe, the laminar flow is thought to be linearly stable (WEDIN and KERSWELL [2004]), and no critical Reynolds numbers Re_c have been found at which the laminar flow becomes linearly unstable. Experimental results, on the other hand, indicate that the laminar flow always becomes unstable at sufficiently large Re, but also that the value of Re at which this happens depends upon the amplitude of the disturbances to which the laminar flow is subjected (HOF, VAN DOORNE, WESTERWEEL, NIEUWSTADT, FAISST, ECKHARDT, WEDIN, KERSWELL and WALEFFE [2004]). The transition to turbulence often occurs near $\mathrm{Re} = 2000$, but in some experiments with carefully controlled conditions, laminar flow has been observed for values of Re as large as 100,000 (HOF, VAN DOORNE, WESTERWEEL, NIEUWSTADT, FAISST, ECKHARDT, WEDIN, KERSWELL and WALEFFE [2004], PFENNIGER [1961]).

The dependence of the transition point on the disturbance amplitude suggests that the transition occurs as the result of a finite amplitude instability of the laminar flow, which cannot be revealed by linear stability analysis. Recent progress toward understanding of this problem, and related classical flows of fluid between parallel plates, includes as an important element the discovery of new, periodic, traveling-wave solutions that play an essential role in a complex chain of events that is believed to support the amplification of small disturbances and thereby lead to a turbulent transition in response to finite amplitude perturbation (FAISST and ECKHARDT [2003], WALEFFE [1997], WEDIN and KERSWELL [2004]). In brief, this essentially nonlinear disturbance growth process is believed to occur through the self-sustaining interaction of several different flow structures: streamwise rolls, streaks, and wavelike disturbances (WALEFFE [1997], WEDIN and KERSWELL [2004]).

Since these interacting flow structures do not appear through bifurcations of the laminar flow solution, they cannot be located by stability analysis of that solution. Instead, a novel continuation procedure has been devised to locate new exact solutions associated with these dynamical processes (WALEFFE [1998, 2001], WEDIN and KERSWELL [2004]). Based on detailed physical reasoning, a body force of carefully chosen form is introduced into the equations of motion, representing posited nonlinear interactions in an hypothesized self-sustained turbulent state at a particular, relatively low value of Re. If the artifice is successful, the new solution branch generated in this way can be continued back to an exact solution by smoothly reducing the body force to zero.

For pipe flow, FAISST and ECKHARDT [2003] and WEDIN and KERSWELL [2004] have successfully used this approach to compute many new traveling wave solutions. The new solutions have m-fold degrees of rotational symmetry, with $m = 1, 2, 3, 4, 5$, or 6. These solutions arise through saddle-node bifurcations at relatively small Reynolds number, with the smallest at $\mathrm{Re} = 1251$ (WEDIN and KERSWELL [2004]). They appear typically to be unstable, as anticipated, for all or at least most parameter values for which they exist. Recently, transient traveling wave motions resembling the theoretical traveling wave solutions have been observed in laboratory experiments (HOF, VAN DOORNE, WESTERWEEL, NIEUWSTADT, FAISST, ECKHARDT, WEDIN, KERSWELL and WALEFFE [2004]).

The perspective that arises from this work is of the appearance near turbulent transition of many unstable traveling wave solutions of differing symmetries and structures; a finite amplitude perturbation may then be sufficient to force the flow into an irregular oscillation between these unstable oscillatory solutions (Hof, van Doorne, Westerweel, Nieuwstadt, Faisst, Eckhardt, Wedin, Kerswell and Waleffe [2004]). At sufficiently large Re, a chaotic attractor representing the turbulent state may develop from these unstable cycles, reducing the basin of attraction of the stable laminar flow solution and the corresponding critical amplitude of the disturbance necessary to cause the transition from laminar to turbulent flow (Hof, van Doorne, Westerweel, Nieuwstadt, Faisst, Eckhardt, Wedin, Kerswell and Waleffe [2004]). Under some circumstances, it may be possible to analyze these transitions by considering the attractor bifurcations directly (Ma and Wang [2007]).

At fixed points in space, the traveling wave solutions are periodic in time. Thus, they are specific and particularly interesting examples of the general class of time-periodic flows discussed in the Introduction. There are intriguing parallels between the view of turbulent transition suggested by the these analyses, and the cycle-expansion approach to chaotic motion outlined above in Section 3 (Kerswell [2005]). Whether a rigorous connection may exist between these remains to be seen.

6. Summary

Time-periodic flows are of special interest in many areas of geophysical and classical fluid dynamics. In addition to the many varieties of linear wave solutions supported by the equations of fluid motion, there are also many varieties of nonlinear time-periodic solutions. Recent research in several different areas illustrates the richness of these nonlinear solutions and the insight that they offer into the structure and dynamics of more general flows. Despite their apparently unnatural, exactly repetitive behavior, they form useful models of flows with more complicated time-dependence, and the competition and interaction between them may ultimately prove to be an essential source of the irregular, quasi-oscillatory, fluctuating motions that are characteristic of fluid flows in nature.

7. Acknowledgments

Preparation of this review was supported in part by grants from the Office of Naval Research and the National Science Foundation.

References

Aref, H. (1984). Stirring by chaotic advection. *J. Fluid Mech.* **143**, 1–21.

Artuso, R., Aurell, E., Cvitanović, P. (1990a). Recycling of strange sets I: cycle expansions. *Nonlinearity*. **3**, 325–359.

Artuso, R., Aurell, E., Cvitanović, P. (1990b). Recycling of strange sets II: applications. *Nonlinearity*. **3**, 361–386.

Brevdo, L., Bridges, T.J. (1997). Local and global instabilities of spatially developing flows: cautionary examples. *Proc. R. Soc. Lond. A*. **453**, 145–1364.

Charney, J. (1947). The dynamics of long waves in a baroclinic westerly current. *J. Meteorol.* **4**, 135–162.

Christiansen, F., Cvitanović, P., Putkaradze, V. (1997). Spatio-temporal chaos in terms of unstable recurrent patterns. *Nonlinearity*. **10**, 55–70.

Coddington, E., Levinson, N. (1955). *Theory of Ordinary Differential Equations* (McGraw-Hill, New York), pp. 429.

Cvitanović, P., Artuso, R., Mainieri, R., Tanner, G., Vattay, G. (2005). *Chaos: Classical and Quantum* (ChaosBook.org. Niels Bohr Institute, Copenhagen, Denmark).

Eady, E. (1949). Long waves and cyclone waves. *Tellus*. **1**, 33–52.

Fantini, M., Davolio, S. (2001). Instability of neutral Eady waves and orography. *J. Atmos. Sci.* **58**, 1146–1154.

Faisst, H., Eckhardt, B. (2003). Traveling waves in pipe flow. *Phys. Rev. Lett.* **91**, 224502.

Farrell, B. (1989). Optimal excitation of baroclinic waves. *J. Atmos. Sci.* **46**, 1193–1206.

Gill, A. (1982). *Atmosphere-Ocean Dynamics* (Academic Press, New York).

Gutzwiller, M. (1990). *Chaos in classical and quantum mechanics*, Springer, New York, pp. 432.

Hof, B., van Doorne, C., Westerweel, J., Nieuwstadt, F., Faisst, H., Eckhardt, B., Wedin, H., Kerswell, R., Waleffe, F. (2004). Experimental observations of nonlinear traveling waves in turbulent pipe flow. *Science*. **305**, 1594–1598.

Itoh, H., Kimoto, M. (1996). Multiple attractors and chaotic itinerancy in a quasi-geostrophic model with realistic topography: implications for weather regimes and low-frequency variability. *J. Atmos. Sci.* **53**, 2217–2231.

Jiang, S., Ghil, M. (1997). Tracking nonlinear solutions with simulated altimetric data in a shallow-water model. *J. Phys. Oceanogr.* **27**, 72–95.

Jiang, S., Jin, F., Ghil, M. (1995). Multiple equilibria, periodic and aperiodic solutions in a wind-driven double-gyre shallow-water model. *J. Phys. Oceanogr.* **25**, 764–785.

Kazantsev, E. (1998). Unstable periodic orbits and attractor of the barotropic ocean model. *Nonlinear Proc. Geophys.* **5**, 193–208.

Kazantsev, E. (2001). Sensitivity of the attractor of the barotropic ocean model to external influences: approach by unstable periodic orbits. *Nonlinear Proc. Geophys.* **8**, 281–300.

Kerswell, R. (2005). Recent progress in understanding the transition to turbulence in a pipe. *Nonlinearity*. **18**, R17–R44.

Knobloch, E., Weiss, J.B. (1987). Chaotic advection by modulated traveling waves. *Phys. Rev. A*. **36**, 1522–1524.

Kuchment, P. (1993). *Floquet Theory for Partial Differential Equations* (Birkhäuser Verlag, Basle, Switzerland).

Lighthill, J. (1978). *Waves in Fluids*. (Cambridge University Press, Cambridge, UK).

Lorenz, E. (1963). Deterministic nonperiodic flow. *J. Atmos. Sci.* **12**, 130–141.

Lust, K., Roose, D., Spence, A., Champneys, A. (1998). An adaptive Newton-Picard algorithm with subspace iteration for computing periodic solutions. *SIAM J. Sci. Comp.* **19**, 1188–1209.

Ma, T., Wang, S. (2007). *Phase Transition Dynamics in Nonlinear Sciences*, in preparation.

McNamara, S. (1994). A periodic orbit expansion of the lorenz system. In: Salmon, R., Ewing-Deremer, B. (eds.), *Geometrical Methods in Fluid Dynamics* (Woods Hole Oceanographic Institution Technical Report, WHOI-94-12).

Pedlosky, J. (1971). Finite-amplitude baroclinic waves with small dissipation. *J. Atmos. Sci.* **28**, 587–597.

Pedlosky, J. (1987). *Geophysical Fluid Dynamics* (Springer-Verlag, New York), pp. 710.

Pedlosky, J. (2003). *Waves in the Ocean and the Atmosphere* (Springer-Verlag, New York).

Pfenniger, W. (1961). Transition in the inlet length of tubes at high Reynolds numbers. In: Lachman, G. (ed.), *Boundary Layer and Flow Control* (Pergamon, Oxford, UK).

Phillips, N. (1954). Energy transformations and meridional circulations associated with simple baroclinic waves in a two-level, quasi-geostrophic model. *Tellus*. **6**, 273–286.

Pierrehumbert, R. (1991). Chaotic mixing of tracer and vorticity by modulated traveling Rossby waves. *Geophys. Astrophys. Fluid Dyn.* **58**, 285–319.

Reynolds, O. (1883). An experimental investigation of the circumstances which determine whether the motion of water shall be direct or sinuous, and of the law of resistance in parallel channels. *Philos. Trans. R. Soc. Lond. A* **174**, 935–982.

Samelson, R.M. (1992). Fluid exchange across a meandering jet. *J. Phys. Oceanogr.* **22**, 431–440.

Samelson, R.M. (2001a). Periodic orbits and disturbance growth for baroclinic waves. *J. Atmos. Sci* **58**, 436–450.

Samelson, R.M. (2001b). Lyapunov, Floquet, and singular vectors for baroclinic waves. *Nonlinear Proc. Geophys.* **8**, 439–448.

Samelson, R.M., Tziperman, E. (2001). Instability of the chaotic ENSO: the growth-phase predictability barrier. *J. Atmos. Sci.* **58**, 3613–3625.

Samelson, R.M., Wiggins, S. (2006). *Lagrangian Transport in Geophysical Jets and Waves* (Springer-Verlag, New York).

Samelson, R. M., Wolfe, C. (2003). A nonlinear baroclinic wave-mean oscillation with multiple normal-mode instabilities. *J. Atmos. Sci.*, **60**, 1186–1199.

Waleffe, F. (1997). On a self-sustaining process in shear flows. *Phys. Fluids*. **9**, 883–900.

Waleffe, F. (1998). Three-dimensional coherent states in plane in shear flows. *Phys. Rev. Lett.* **81**, 4140–4143.

Waleffe, F. (2001). Exact coherent structures in channel flow. *J. Fluid Mech.* **435**, 93–102.

Wedin, H., Kerswell, R. (2004). Exact coherent structures in pipe flow: travelling wave solutions. *J. Fluid Mech.* **508**, 333.

Wolfe, C.L., Samelson, R.M. (2006). Normal-mode analysis of a baroclinic wave-mean oscillation. *J. Atmos. Sci.* **63**, 2795–2812.

Wolfe, C.L., Samelson, R.M. (2008). Singular vectors and time-dependent normal modes of a baroclinic wave-mean oscillation. *J. Atmos. Sci*. **65**, 875–894.

Zoldi, S., Greenside, H. (1998). Spatially localized unstable periodic orbits of a high-dimensional chaotic system. *Phys. Rev. E* **57** (3), R2511–R2514.

Momentum Maps for Lattice EPDiff

Colin J. Cotter

Department of Mathematics, Imperial College London, SW7 2AZ, UK

Darryl D. Holm

Computer and Computational Science,
Los Alamos National Laboratory, Los Alamos, NM 87545, USA

1. Introduction

1.1. Transverse internal wave interactions

Synthetic Aperture Radar (SAR) observations from the Space Shuttle often show nonlinear internal wave trains that propagate for many hundreds of kilometers across large basins such as the South China Sea (SCS) shown in Fig. 1.1.

These wave trains are characterized as *Great Lines on the Sea* in YODER, ANKLESON, BARBER, FLAMENT and BALCH [1994]. Both lines and spirals on the sea arise as flow phenomena rather than wave phenomena per se. The flow phenomenon detected in the the SAR imagery is associated with nonlinear internal waves, whose crests may be as much as 200 km long. The amplitude of these internal waves results in about 150 m of deflection in the thermocline over a distance of about 1 km. Thus, their aspect ratio satisfies the first criterion to be nonlinear shallow water waves. Their amplitude is also considerably less than the typical thickness of the thermocline, but it is not actually infinitesimal compared with the thermocline thickness. The flow along the crests of these waves also indicates they are not precisely the same as usual shallow water waves.

The particular nonlinear internal waves found in the SCS are generated by the tides flowing east to west through the Luzon Strait over submerged ridges between Taiwan and the Philippines. The SAR images in Fig. 1.1 show that the momentum of the tides flowing westward over these ridges concentrates into internal waves on the thermocline that emerge into the SCS basin as thin wave fronts which may extend in length for hundreds of kilometers (much larger than the Straits in which they were created) and may propagate for thousands of kilometers. Perhaps because of the complex topography,

Computational Methods for the Atmosphere and the Oceans
Special Volume (Roger M. Temam and Joseph J. Tribbia, Guest Editors) of
HANDBOOK OF NUMERICAL ANALYSIS, VOL. XIV
P.G. Ciarlet (Editor)
Copyright © 2009 Elsevier B.V.
All rights reserved
ISSN 1570-8659
DOI 10.1016/S1570-8659(08)00205-6

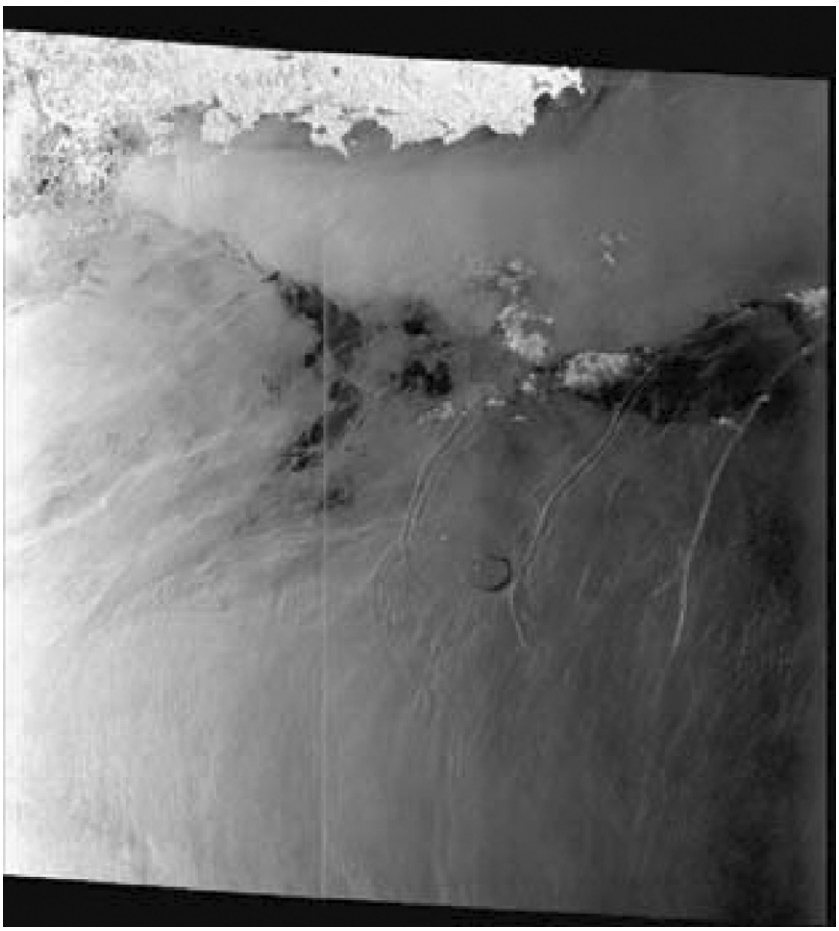


FIG. 1.1 Image from the Space Shuttle of long, tidally-excited waves in the South China Sea near the Dongsha Atoll. The waves are propagating from east to west and are produced with every tide (about 12 hours). The waves interact with the Atoll and then undergo nonlinear reconnections. (Picture courtesy of A. Liu.)

the tides flowing over the mouth of the Luzon Strait do not produce internal waves propagating in both directions. The significant wave trains propagate Westward.

Propagating wave trains may intersect transversely with other wave trains. Sometimes these wave trains merely pass through each other as linear waves. However, in nonlinear wave encounters such as those captured by SAR imaging of the region of the SCS West of Dongsha Island in Fig. 1.2, two wave fronts may intersect transversely, merge together, and produce a single wave front. This merger of the wave fronts is the hallmark of a nonlinear process. These particular wave interactions possess strong transverse dynamics (flow along the crests) and momentum exchange in the direction of propagation, which allow the wave fronts to merge and reconnect, rather than merely passing through each other, as weaker waves do when they intersect in an interference pattern.

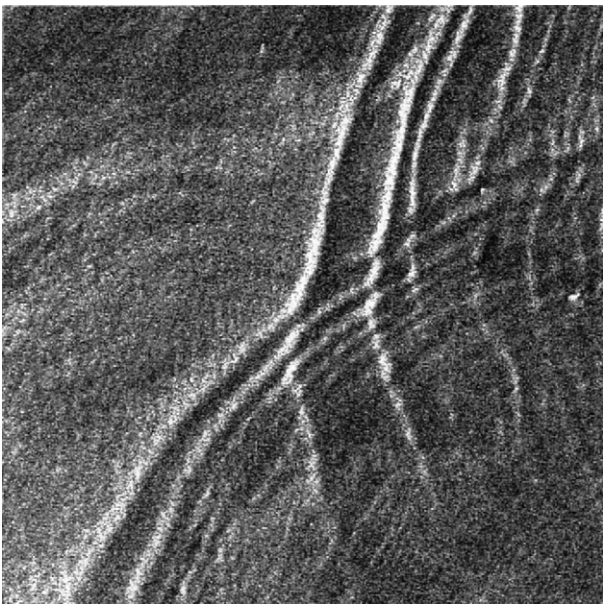


FIG. 1.2 Enlargement of part of Fig. 1.1 showing reconnecting long waves. (Picture courtesy of A. Liu.)

Nonlinear internal wave interactions have been well studied in one dimension (1D), often by using the weakly nonlinear Boussinesq approximation. These studies have usually resulted in a variant of the Korteweg-de Vries (KdV) equation, which has soliton solutions that interact by exchange of momentum in unidirectional elastic collisions (SELIGER and WHITHAM [1968]). However, the complex wave front interactions shown in Fig. 1.2 are plainly at least two-dimensional (2D). We shall pursue the qualitative description of these higher-dimensional wave interactions by using a simple 2D model equation called EPDiff.[1] EPDiff may be derived in 1D from the asymptotic expansion for shallow water wave motion of the Euler equations for the unidirectional flow of an incompressible fluid with a free surface moving under gravity. In 1D, the result is the Camassa-Holm (CH) equation, which arises at quadratic order in this expansion. That is, CH is one order of accuracy in the asymptotic expansion beyond KdV, which arises at linear order. Just as for KdV, the CH equation is completely integrable, so CH also has soliton solutions that interact by elastic collisions in 1D. Moreover, in the limit of zero linear dispersion, the CH solitons develop a sharp peak at which their profile has a jump in derivative that forms a sharp peak. In this limit, the CH solitons are called "peakons." The CH peakons are weak solutions, in the sense that their momentum is concentrated on delta functions that move with the velocity of the fluid flow.

In its zero-dispersion limit, CH has a geometric property that allows it to be immediately generalized to higher dimensions, in which it is called EPDiff. The term "EPDiff" distinguishes CH, which is a 1D shallow water wave equation with physical wave

[1]EPDiff is the "Euler-Poincaré equation on the diffeomorphisms."

dispersion, from its dispersionless limit which belongs to a larger class of equations. This larger class of equations – the Euler-Poincaré (EP) equations HOLM, MARSDEN and RATIU [1998] – describes geodesic motion with respect to any metric defining a norm on the vector space of the Lie algebra of a Lie group. In the geometric theory of fluid mechanics, the fluid velocity belongs to the tangent space of the group of smooth invertible maps called "diffeomorphisms" (or diffeos, for short). The EP equation on the diffeomorphisms is called EPDiff. EPDiff is a larger class of equations than CH also because it is defined for geodesic motion on the diffeos with respect to any metric, not just for the H^1 norm of the velocity, which appears as the kinetic energy norm in the derivation of CH. (The gradient part of the H^1 norm for CH corresponds to the vertically averaged kinetic energy associated with vertical motion.) Thus, among the EPDiff equations, the dispersionless limit of CH is 1D EPDiff(H^1). In 1D, the momentum of the EPDiff(H^1) peakons is concentrated at points moving along with the flow; but in higher dimensions, their momentum is distributed on embedded subspaces moving with the flow. In particular, EPDiff(H^1) in 2D has singular solutions whose momentum is distributed along curves in the plane. As solutions of the 2D version of a unidirectional shallow water wave equation in its limit of zero linear dispersion, these moving curves in the plane evolving under the dynamics of EPDiff(H^1) are prototypes for studying the interactions of the Great Lines on the Sea.

To jump ahead, the singular (or weak) solutions of the EPDiff equation that emerge in finite time from any confined smooth initial conditions and are supported on embedded subspaces moving with the flow velocity, just as seen in the Great Lines on the Sea captured in Fig. 1.1. We developed a numerical method for simulating the singular solutions of EPDiff in the framework of its geometric definition, which is natural for the Variational Particle Mesh (VPM) method. Our numerical results using VPM show that

- Singular solutions for EPDiff may be simulated by VPM as curve-segments move with the 2D flow velocity that possesses no internal degrees of freedom.
- In collisions between any two of these curve-segment solutions for EPDiff, the momentum of the one that overtakes from behind is imparted to the one ahead. Thus, overtaking collisions between two finite-length wave packets are elastic.
- The transverse collision of two curve-segment solutions for EPDiff may result in merger (or reconnection) of the curve segments due to a combination of exchange of momentum between the wave trains and flow along their wave crests. In 2D, the reconnection or merger of singular wave fronts under numerical EPDiff dynamics using VPM is evident in Fig. 1.3.

1.1.1. Plan of the chapter

In this chapter, we introduce the VPM method for EPDiff, and discuss some of the properties that arise from the variational structure, in the following sections:

- The particle-mesh calculus is set out in Section 2.
- We give a variational principle associated with the method in Section 3.

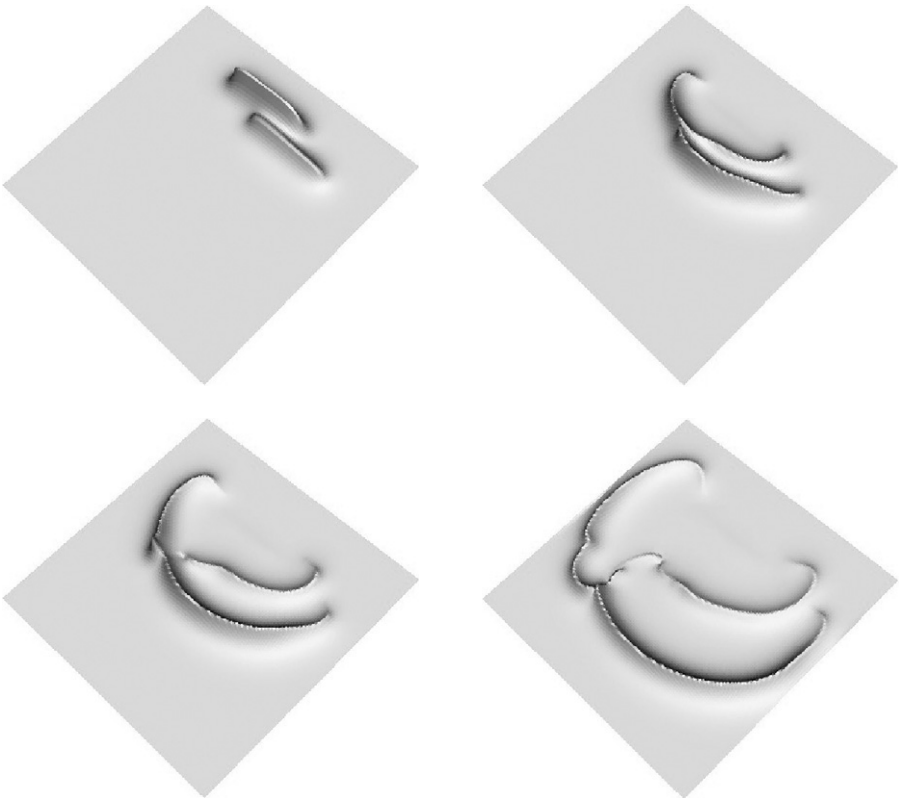


FIG. 1.3 Plots showing surfaces of velocity magnitude $|u|$ at times $t = 0$, 1.55, 2.7, and 4.75 showing an "overtaking" collision between two singular momentum filaments. The filament which is initially behind has greater momentum, and so it catches up with the filament in front, transferring momentum to the front filament and causing a reconnection to occur. This is the 2D version of the process illustrated in Fig. 8.3. This is the nonlinear reconnection process which is illustrated in the Space Shuttle image in Figs. 1.1 and 1.2. These results were obtained using the same method as Fig. 8.5.

- Section 4 shows that the Eulerian grid quantities satisfy an approximation of the EPDiff equation in EP form.
- Section 5 defines a left-action of D_{VPM} on Ω and provides the corresponding momentum map.
- Sections 6 defines a right-action in an extended space which can be interpreted as a discrete form of relabeling of Lagrangian particles. The Hamiltonian for the continuous time evolution of discretized EPDiff solutions is invariant under the action, and so from Noether's theorem, we obtain a conserved momentum.
- Section 7 shows how this conserved momentum can be interpreted as a discrete form of Kelvin's circulation theorem.
- Section 8 gives some numerical examples, as well as convergence tests for the method.

1.2. Theoretical development

Much of the theoretical development in this chapter is inspired by the following theorem (ARNOLD [1966]).

THEOREM 1.1 (ARNOLD [1966]). *The solutions to Euler's equations for the incompressible motion of an ideal fluid describe coadjoint geodesic motion on the volume preserving diffeomorphisms, with respect to the L^2 norm of the fluid velocity (the kinetic energy).*

The Euler equations for incompressible motion of an ideal fluid may be written in the material frame as

$$\mathcal{P}\Big(\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t}\Big) = 0 \quad \text{along} \quad \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{u} \quad \text{with} \quad \nabla\cdot\mathbf{u} = 0,$$

where $\mathcal{P}$ is the Leray projection onto the incompressible vector fields. These equations may also be written in the spatial frame as

$$\mathcal{P}\Big(\partial_t\mathbf{u} + \mathrm{ad}^*_{\mathbf{u}}\mathbf{u}\Big) = 0,$$

where ad* is the dual of the ad-action among incompressible vector fields under the L^2 pairing. That is, ad* is defined by $\langle\mu,\ \mathrm{ad}_{\mathbf{u}}\mathbf{w}\rangle = -\langle\mathrm{ad}^*_{\mathbf{u}}\mu,\ \mathbf{w}\rangle$. Here, $\mathrm{ad}_u\, w = [u, w]$ is the Lie-algebra commutator between vector fields $\mathbf{u}$, $\mathbf{w}$, and $\langle\,\cdot\,,\,\cdot\,\rangle$ denotes the L^2 pairing between such vector fields and one-form densities such as μ.

1.2.1. EPDiff

The EPDiff equation describes the corresponding coadjoint geodesic motion on the full diffeomorphism group, allowing for compressibility and an arbitrary norm, $\|\cdot\|$,

$$\text{EPDiff is} \quad \partial_t\mu + \mathrm{ad}^*_u\mu = 0, \quad \text{with} \quad \mu = \frac{\delta\ell}{\delta u}, \quad \text{where} \quad \ell = \frac{1}{2}\|u\|^2.$$

The momentum density μ is a one-form density, and the EPDiff equation describes coadjoint dynamics under the action of the corresponding velocity vector field. In EPDiff, ad^*_u is the coadjoint action of a vector field u acting on a one-form density $\mu = \delta\ell/\delta u$ for a Lagrangian $\ell[u]$ in Hamilton's principle $\delta S = 0$ for $S = \int \ell[u]\mathrm{d}t$. In components,

$$\mu = \mathbf{m}\cdot\mathrm{d}\mathbf{x}\otimes\mathrm{d\,Vol},$$

and EPDiff may be written as the invariance condition,

$$\frac{\mathrm{d}\mu}{\mathrm{d}t} = 0 \quad \text{along} \quad \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{u} = G * \mathbf{m},$$

where $G*$ denotes convolution with the Green's function relating the components of $\mathbf{m}$ and $\mathbf{u}$. In particular, for the H^1 norm $\|u\|^2 \equiv \int |\mathbf{u}|^2 + \alpha^2|\nabla\mathbf{u}|^2\ \mathrm{d\,Vol}$, we have the component relation

$$\mathbf{m} = \mathbf{u} - \alpha^2\Delta\mathbf{u}, \tag{1.1}$$

and G is the Green's function for the Helmholtz operator, $Id - \alpha^2\Delta$, Δ is the Laplacian, and α is a length scale. Thus, EPDiff for the H^s norm with $s > 0$ is an integro-partial differential equation.

Originally derived (HOLM, MARSDEN and RATIU [1998]) as an n-dimensional generalization of the CH equation for shallow-water dynamics in 1D (CAMASSA and HOLM [1993]), EPDiff arises in several other applications. For example, EPDiff for the H^1 norm is the pressureless version of the Lagrangian-averaged Navier-Stokes-alpha model of turbulence (FOIAS, HOLM and TITI [2001]). EPDiff for H^1 also emerges in the limit in which one ignores variations in height of the Green-Nagdhi equation for shallow water dynamics (CAMASSA, HOLM and LEVERMORE [1996]). In 1D, this is the dispersionless limit of the CH equation (CAMASSA and HOLM [1993]). In general, EPDiff is the equation for coadjoint geodesic motion on the diffeomorphisms with respect to any given norm on the Eulerian particle velocity (kinetic energy). Finally, EPDiff also describes the process of template matching in computational anatomy (MILLER, TROUVÉ and YOUNES [2002]). In this application, EPDiff has recently become a conduit for technology transfer from soliton theory to computational anatomy (HOLM, RANANATHER, TROUVÉ and YOUNES [2004]). Thus, EPDiff turns out to be a prototype equation for a number of applications.

The present chapter describes the underlying principles for using the VPM method in numerically integrating EPDiff in the study of its nonlinear wave interactions.

1.2.2. VPM method

The VPM method introduced in COTTER [2005] produces Hamiltonian spatial discretizations of fluid equations which may then be integrated in discrete time by using a variational integrator. VPM may be regarded as a descendant of the Hamiltonian Particle-Mesh method (FRANK, GOTTWALD and REICH [2002]), which is a Hamiltonian discretization of the rotating shallow-water equations. The difference is that HPM combines an Eulerian representation of the potential energy (which gives rise to the pressure term) with a Lagrangian representation of the kinetic energy while VPM uses an Eulerian representation of the entire Lagrangian. This means that the VPM method is much more general than HPM and may be applied to many different fluid partial differential equations (PDEs) (e.g., shallow-water, Green-Nagdhi, incompressible Euler, etc.). In this chapter, we focus on EPDiff, which is an equation for fluid velocity only. Consequently, symmetries of the discretized fluid velocity will be symmetries of the equations. In future, we will extend this work to include advected quantities such as density, scalars etc. Our ultimate aim is to use geometric properties in constructing general numerical methods for PDEs describing the continuum dynamics of fluids, complex fluids, and plasmas.

The conservative properties of variational integrators are well understood (LEW, MARSDEN, ORTIZ and WEST [2003]). In this chapter, we will discuss preservation under VPM spatial discretization of the geometric properties of the well-known EPDiff equation for coadjoint motion under the diffeomorphisms (HOLM and MARSDEN [2004]),

$$\partial_t \mu + \mathrm{ad}^*_u \mu = 0, \quad \text{with } \mu = \frac{\delta l}{\delta u}.$$

In particular, we shall discuss discrete VPM analogs of the momentum maps for the left- and right-actions of the diffeomorphisms on embedded subspaces of $\mathbb{R}^n$ (HOLM and

MARSDEN [2004]). The Lagrangian we shall choose is the H^1 norm, $\ell[u] = \frac{1}{2}\|u\|^2_{H^1}$, so the components of velocity u and momentum density $\mu = \delta\ell/\delta u$ will be related by the Helmholtz operator as in Eq. (1.1). In this case, velocity $u \in H^1$ implies that its dual momentum density $\mu \in H^{-1}$; so the solutions of EPDiff may be measure valued in μ. That is, weak solutions of EPDiff are allowed in this case, which are expressed in terms of delta functions supported on embedded subspaces of $\mathbb{R}^n$ (HOLM and MARSDEN [2004]). The left-action of the diffeomorphisms on these embedded subspaces of $\mathbb{R}^n$ generates the motion of spatially discrete EPDiff (lattice EPDiff), while the right-action is a symmetry and generates the conservation law for circulation according to the Kelvin-Noether theorem (HOLM, MARSDEN and RATIU [1998]). Thus, we seek the spatially discrete version of the corresponding theorem for continuum solutions in HOLM and MARSDEN [2004]. All of these properties will then be preserved by an appropriate variational time integrator.

2. Particle-mesh calculus

This section describes the particle-mesh calculus that will be used in discretising EPDiff. We shall describe its discretization in space with continuous time, and later we shall describe how to construct variational time integrators to assemble a fully discrete space-time integration scheme.

A finite dimensional subspace of $\mathfrak{X}(\Omega)$: The infinite-dimensional space of smooth vector fields $\mathfrak{X}(\Omega)$ generates the diffeomorphisms (smooth invertible maps with smooth inverse) of the domain Ω onto itself. To make a numerical algorithm that can be calculated on a finite computer, we first need to choose a finite-dimensional subspace $\mathfrak{X}_0$ of $\mathfrak{X}(\Omega)$ that will generate our diffeomorphisms. We begin with a fixed grid consisting of n_g points in the domain Ω with vector coordinates $\{x_k\}_{k=1}^{n_g} \in \mathbb{R}^{d\times n_g}$ in d dimensions. At each grid point x_k, we shall associate a velocity vector u_k. The finite-dimensional space of possible sets of velocity vectors $u = (u_1, \ldots, u_{n_g}) \in \mathbb{R}^{d\times n_g}$ will then represent the required subspace $\mathfrak{X}_0$. We call a set of values $\{u_k\}_{k=1}^{n_g}$ the *grid representative* of the corresponding vector field.

To obtain the element of $\mathfrak{X}(\Omega)$ corresponding to u, we use a set of basis functions with $\psi_k(x)$ representing a distribution centred around x_k. These basis functions are taken to have compact support and to satisfy the *Partition-of-Unity (PoU) property*

$$\sum_{k=1}^{n_g} \psi_k(x) = 1, \qquad \forall\, x \in \Omega\,.$$

The vector field X_u is then defined as follows:

DEFINITION 2.1. The vector field X_u on Ω whose grid representative is u takes the coordinate form

$$X_u(x) = \sum_k u_k \psi_k(x) \cdot \frac{\partial}{\partial x}\,.$$

A plot of a typical basis function in 1D is given in Fig. 2.1.

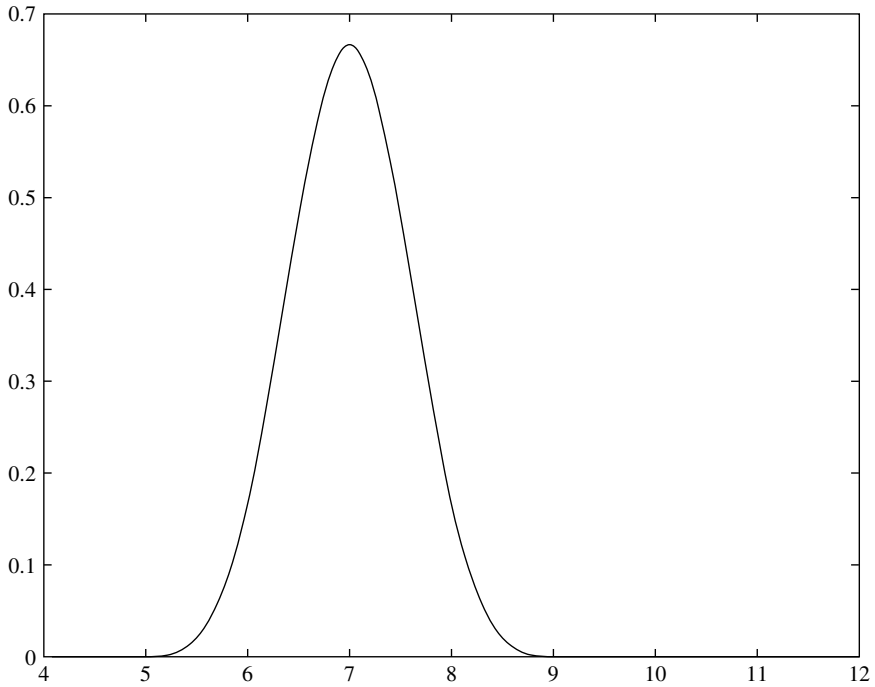


FIG. 2.1 Plot of a B-spline basis function ψ_k centred on $x_k = 7$ with a grid width of 1. These basis functions satisfy the partition-of-unity property.

REMARK 2.1. In general, these vector fields do not commute among themselves in the Lie bracket, so they do not form a Lie subalgebra of $\mathfrak{X}(\Omega)$. The consequence of this is that we cannot obtain a reduced form for the Eulerian variables but must solve for the Lagrangian particles as state variables. Also in general, the value of $X_u(x_k)$ is not exactly equal to u_k but is convergent to it in the continuum limit.

Dynamics of a finite set of Lagrangian particles: We shall proceed in describing our numerical method by introducing a finite set of n_p *Lagrangian fluid particles* $\{Q_\beta\}_{\beta=1}^{n_p}$, whose velocities $\{\dot{Q}_\beta\}_{\beta=1}^{n_p}$ are entirely determined by the grid velocity representation $\{u_k\}_{k=1}^{n_g}$ *via* the vector field X_u as follows:

DEFINITION 2.2. The ***PoU vector field*** $X_{u;n_p} \in T\,\Omega^{n_p}$ associated with a velocity grid representative u is defined as

$$X_{u;n_p}(Q) = \sum_\beta \sum_k u_k \psi_k(Q) \cdot \frac{\partial}{\partial Q}, \quad Q \in (Q_1, \ldots, Q_{n_p}) = \mathbb{R}^{d \times n_p}. \tag{2.1}$$

We shall constrain the dynamics of the particles so that a tangent vector $\dot{Q}$ may be represented as a PoU vector field evaluated at the point Q. That is, $(Q, \dot{Q})$ lies in a distribution D^{VPM} defined as follows:

DEFINITION 2.3. [The distribution D^{VPM}] Let $D^{\mathrm{VPM}} \subset T\,\Omega^{n_p}$ be the distribution defined by

$$D^{\mathrm{VPM}} = \left\{ \left(\dot{Q}, Q\right) : \dot{Q}_\beta = \sum_k u_k \psi_k(Q_\beta) \text{ for some } u \in \mathbb{R}^{d\times n_g} \text{ and } \forall\, k = 1, \ldots, n_g \right\}.$$

DEFINITION 2.4. A time series $Q(t) = (Q_1(t), \ldots, Q_{n_p}(t))$ with $(\dot{Q}(t), Q(t)) \in D^{\mathrm{VPM}}\ \forall\, t_0 \le t \le t_1$ is called a ***VPM trajectory***. Each VPM trajectory defines a time series $u_k(t) \in \mathbb{R}^{d\times n_g} \times [t_0, t_1]$ such that

$$\dot{Q}_\beta(t) = \sum_k u_k(t)\psi_k(Q_\beta), \tag{2.2}$$

for $\beta = 1, \ldots, n_p$. This is the ***VPM tangent vector relation***, which we will enforce as a constraint for the variational principle resulting in the VPM method.

Gradient and divergence: In this section, we describe how the operations of gradient, divergence, and curl may be approximated using the particle-mesh discretization. These approximations apply the *two dual purposes of the basis functions* ψ_k:

1. The ψ_k interpolate functions from the grid to the particles.
2. The ψ_k also construct densities on the grid from weights stored on the particles.

Notation: Square brackets $[\,\cdot\,]^G$ and $[\,\cdot\,]^P$ will denote these two maps from particles to grid and vice versa. Superscripts distinguish whether the quantity is evaluated on the grid or on the particles. That is, $[\,\cdot\,]^P$ indicates mapping from grid to particles, and $[\,\cdot\,]^G$ indicates mapping from particles to grid.

DEFINITION 2.5. Let $\{f_k\}_{k=1}^{n_g}$ be a scalar quantity stored at the Eulerian grid points. Then,

$$[f]^P_\beta = \sum_k f_k \psi_k(Q_\beta)$$

is an approximation of f evaluated at the particle locations. Furthermore,

$$[\nabla f]^P_\beta = \sum_k f_k \frac{\partial}{\partial Q_\beta}\psi_k(Q_\beta)$$

is an approximation of the gradient of the scalar f evaluated at the particle locations.

DEFINITION 2.6. Given a discrete approximation to integration on the grid, we define an inner product

$$\langle u, v\rangle = \sum_{kl} u_k \cdot M_{kl} v_{kl},$$

where M is a positive-definite symmetric matrix (often referred to as the mass matrix).

DEFINITION 2.7. Let $\{g_\beta\}_{\beta=1}^{n_p}$ be a distribution of values stored at particle locations. We construct a density on the Eulerian grid as

$$[g]_k^G = \sum_l (M^{-1})_{kl} \sum_\beta g_\beta \psi_l(Q_\beta).$$

Furthermore, if the distribution is vector-valued g, then

$$[\nabla \cdot g]_k^G = -\sum_l (M^{-1})_{kl} \sum_\beta g_\beta \cdot \frac{\partial}{\partial Q_\beta} \psi_l(Q_\beta)$$

is an approximation to the divergence of g on the grid.

Discretized continuity equation: Given a set of constant weights $\tilde{D}_\beta$ on the particles $\beta = 1, \ldots, n_p$ to construct a density

$$D_k = \sum_l (M^{-1})_{kl} \sum_\beta \tilde{D}_\beta \psi_k(Q_\beta),$$

one computes

$$\begin{aligned}\frac{\mathrm{d}D_k}{\mathrm{d}t} &= \sum_l (M^{-1})_{kl} \sum_\beta \tilde{D}_\beta \frac{\partial \psi_l}{\partial Q_\beta}(Q_\beta) \cdot \dot{Q}_\beta, \\ &= \sum_l (M^{-1})_{kl} \sum_\beta \tilde{D}_\beta \frac{\partial \psi_l}{\partial Q_\beta} \cdot \sum_k u_k \psi_k(Q_\beta),\end{aligned}$$

and so

$$\frac{\mathrm{d}}{\mathrm{d}t}[D]^G = -\left[\nabla \cdot \left([u]^P D\right)\right]^G,$$

so the corresponding grid representative $[D]^G$ satisfies a discretized continuity equation.

2.2. Lagrangian for semidiscrete EPDiff

Next, we form the Lagrangian for semidiscrete EPDiff. Since the discrete velocity field is defined on the grid, any grid-based method (finite difference, finite element, discrete Fourier transform etc.) may be used. In the examples in this chapter, we use the continuous

EPDiff Lagrangian

$$L_C = \frac{1}{2}\int_\Omega (||u||^2 + \alpha^2||\nabla u||^2)\ \mathrm{d\,Vol},$$

in which the constant α has dimensions of length, with the integral approximated using standard piecewise-linear finite elements to approximate this integral, leading to the discrete Lagrangian

$$L(u) = \frac{1}{2}\sum_{k,l} u_k \cdot H_{kl} u_l \equiv \frac{1}{2} u \cdot Hu, \tag{2.3}$$

where H_{kl} is a positive-definite symmetric matrix.

Remark 2.2. As in the continuous case, this Lagrangian is written entirely in terms of the Eulerian velocity (in this case, the velocity grid representation). In the continuum case, this form of the Lagrangian admits EP reduction (as Eulerian velocity is invariant under the right-action of the diffeomorphism group Diff (Ω)). This reduction results in the EP equation

$$\mu_t + \mathrm{ad}^*_u \mu = 0\,,$$

where $\mu = \delta L/\delta u$ and ad* is the *dual* of the ad-action (Lie algebra commutator) of vector fields on the domain. In the VPM discretization of EPDiff, an analogous equation will emerge, written on the Eulerian grid.

3. Variational principle for discrete EPDiff

In this section we shall derive the equations for $Q(t)$ from a variational principle applied to the Lagrangian (Eq. (2.3)) and required to satisfy the VPM tangent vector constraint. Namely, the variational principle is constrained to restrict the solutions so that $\dot{Q} \in D_Q^{\mathrm{VPM}}$ (defined as the subspace $\{\alpha : (\alpha, Q) \in D^{\mathrm{VPM}}\} \subset T_Q\Omega^{n_p}$). This constraint on VPM trajectories is the discrete analog of the *Lin constraints* in the *Clebsch variational approach* to continuum ideal fluid dynamics as discussed in Holm and Kupershmidt [1983]; Holm, Marsden and Ratiu [1998]. At the end of this section, we shall give a fully discrete variational principle which produces the numerical scheme.

3.1. Constrained action principle for semidiscrete EPDiff

We begin by defining the *grid momentum* as follows.

Definition 3.1. [Grid momentum] The ***grid momentum*** $m = (m_1, \ldots, m_{n_g}) \in \mathbb{R}^{d\times n_g}$ is defined from the grid representative Lagrangian $L(u)$ in Eq. (2.3) *via* its derivative

$$\sum_l M_{kl} m_l = \frac{\partial L}{\partial u_k}, \tag{3.1}$$

where M_{kl} is the mass matrix given in Definition (2.6).

This expression for the grid momentum is an approximation to $\delta L/\delta u$ in the continuous case.

DEFINITION 3.2. [Constrained action] The action for semidiscrete EPDiff is defined in terms of three variables: the grid velocity $u \in \mathbb{R}^{d\times n_g}$; the particle positions $Q_\beta \in \Omega^{n_p}$; and the Lagrange multipliers $P_\beta \in T^*_Q\Omega^{n_p}$ which will become the particle momenta on the Hamiltonian side. The action is given by

$$\mathcal{A} = \int_0^T L(u) + \sum_\beta P_\beta \cdot \left(\dot{Q}_\beta - \sum_k u_k \psi_k(Q_\beta) \right) \mathrm{d}\, t\,. \tag{3.2}$$

This is the action for Lagrangian (Eq. (2.3)) when its particle velocities are required to satisfy the VPM tangent vector constraint given in Eq. (2.2).

PROPOSITION 3.1. *The variables* (u, P, Q) *which extremise the constrained action* $\mathcal{A}$ *in Eq. (3.2) satisfy*

$$\dot{Q}_\beta = \sum_k u_k \psi_k(Q_\beta)\,, \tag{3.3}$$

$$\dot{P}_\beta = -\, P_\beta \cdot \sum_k u_k \frac{\partial \psi_k}{\partial Q}(Q_\beta)\,, \tag{3.4}$$

$$\frac{\partial L}{\partial u_k} = \sum_\beta P_\beta \psi_k(Q_\beta)\,. \tag{3.5}$$

PROOF. After integration by parts, the first variation of $\mathcal{A}$ in (u, P, Q) is

$$\begin{aligned}\delta\mathcal{A} = \int_0^T & \sum_k \left(\frac{\partial L}{\partial u_k} - \sum_\beta P_\beta \psi_k(Q_\beta) \right) \cdot \delta u_k \\ & + \sum_\beta \left(\dot{Q}_\beta - \sum_k u_k \psi_k(Q_\beta) \right) \cdot \delta P_\beta \\ & - \sum_\beta \left(\dot{P}_\beta + P_\beta \cdot \sum_k u_k \frac{\partial \psi_k}{\partial Q}(Q_\beta) \right) \cdot \delta Q_\beta \,\mathrm{d}\, t\,,\end{aligned}$$

and the result follows by direct calculation. □

REMARK 3.1 (Left momentum map). Equation (3.5) in Proposition 3.1 bears a great resemblance to the momentum map for left action of the diffeomorphisms on embedded subspaces (HOLM and MARSDEN [2004]), which describes the singular solutions of continuum EPDiff equation. We will see later that Eq. (3.5) is the discrete version of that momentum map.

REMARK 3.2 (Grid momentum). The grid-momentum relation (Eq. (3.5)) allows one to obtain u from $\partial L/\partial \dot{Q}$ and Q by first calculating m, and then inverting the matrix H_{kl} in

$$\sum_l M_{kl} m_l = \frac{\partial L}{\partial u_k} = \sum_l H_{kl} u_l \,. \tag{3.6}$$

This is the discrete analog of the problem of solving for u from m in the elliptic relation,

$$m = \frac{\delta L}{\delta u} = (1 - \alpha^2 \Delta) u,$$

for the continuous case (HOLM and MARSDEN [2004]). Thus, the Lagrangian (Eq. (2.3)) is hyper-regular on the grid.

3.2. Hamiltonian structure

We obtain a Hamiltonian system by substituting u for P and Q *via* the momentum formula, a process summarized in the following proposition.

PROPOSITION 3.2. *The system of Eqs. (3.3–3.5) is canonically Hamiltonian with Hamiltonian function H given by*

$$H = \frac{1}{2} \sum_{k,l} M_{kl} m_l \cdot u_k \,, \tag{3.7}$$

where $u_k = \sum_{m,n} H^{-1}_{km} M_{mn} m_n$ *is defined in terms of* $\sum_l M_{kl} m_l$ *by inverting the matrix* H_{kl} *in Eq. (3.6), and where* m_k *is obtained* via *Eqs. (3.1) and (3.5).*

PROOF. Hamilton's canonical equations with H defined in Eq. (3.7) above may be expressed as Hamilton's canonical equations

$$\dot{P}_\beta = - \frac{\partial H}{\partial Q_\beta} \,, \qquad \dot{Q}_\beta = \frac{\partial H}{\partial P_\beta},$$

which becomes

$$\dot{P}_\beta = - P_\beta \cdot \sum_k u_k \frac{\partial \psi_k}{\partial Q_\beta}(Q_\beta) \,,$$

$$\dot{Q}_\beta = \sum_k u_k \psi_k(Q_\beta) \,,$$

recovering Eqs. (3.3) and (3.4). □

3.3. Constructing a fully discrete method

To construct a fully discrete method, we use the standard variational integrator approach as described in LEW, MARSDEN, ORTIZ and WEST [2003], applied to the constrained

action principle in Definition (3.2). We replace the integral over time by a Riemann sum over discrete time levels and define the map

$$\phi : \Omega^{n_p} \times \Omega^{n_p} \to T\Omega^{n_p},$$

which approximates $\dot{Q}_\beta$. We write the discrete action

$$\mathcal{A}_d = \Delta t \sum_{n=0}^{N} \left(L(u^n) - \sum_\beta P_\beta^n \cdot \phi(Q^n, Q^{n-1}) \right).$$

Minimization of the discrete action over u, P, and Q gives the numerical scheme.

For example, consider the choice

$$\phi(Q^n, Q^{n-1}) = \frac{Q^n - Q^{n-1}}{\Delta t}. \tag{3.8}$$

In this case, the discrete action becomes

$$\mathcal{A}_d = \Delta t \sum_{n=0}^{N} \left(\sum_{kl} H_{kl} u_k^n \cdot u_l^n + \sum_\beta P_\beta^n \cdot \left(\frac{Q_\beta^n - Q_\beta^{n-1}}{\Delta t} - \sum_k u_k^n \psi_k(Q_\beta^{n-1}) \right) \right),$$

which is minimized by the solutions

$$\sum_l H_{kl} u_l^n = \sum_\beta P_\beta^n \psi_k(Q_\beta^{n-1}), \tag{3.9}$$

$$Q_\beta^{n+1} = Q_\beta^n + \Delta t \sum_k u_k^{n+1} \psi_k(Q_\beta^n), \tag{3.10}$$

$$P_\beta^{n+1} = P_\beta^n - \Delta t P_\beta^{n+1} \cdot \sum_k u^{n+1} \frac{\partial \psi_k}{\partial Q}(Q_\beta^n). \tag{3.11}$$

This system is equivalent to the first-order symplectic Euler-A method (i.e., the first-order symplectic method which is implicit in P and explicit in Q) applied to the Hamiltonian system given in Proposition 3.2.

Higher-order variational schemes can be obtained by using a Runge-Kutta method to discretize Eq. (3.3). See Bou-Rabee and Marsden [2007a,b], Cotter and Holm [2008] for more details.

4. Discrete EP equation for VPM

In this section, we compute the discrete EPDiff equation directly on the Eulerian grid.

Proposition 4.1 (Discrete EP equation). *Let $(Q(t), P(t))$ be the solution to the canonical Hamiltonian system in Proposition 3.2. Then the grid momentum $m = (m_1, \ldots, m_{n_g}) \in$*

$\mathbb{R}^{d\times n_g}$ *defined by*

$$\sum_l M_{kl} m_l = \sum_\beta P_\beta \psi_k(Q_\beta) \tag{4.1}$$

satisfies the discrete EP equation

$$\dot{m}_k + (\mathrm{ad}^*_u\, m)_k = 0, \quad k = 1, \ldots, n_g,$$

where

- $(\mathrm{ad}^*_u\, m)_k$ *is defined by*

$$\langle \mathrm{ad}^*_u\, m, w\rangle_g = \langle P, [\mathrm{ad}_u\, w]^P\rangle_p. \tag{4.2}$$

- $[\mathrm{ad}_u\, w]^P_\beta$ *is defined by*

$$\begin{aligned}(\mathrm{ad}_u\, w)_\beta \equiv &\sum_k \psi_k(Q_\beta) u_k(t) \cdot \frac{\partial}{\partial Q_\beta}\left(\sum_l w_l(t)\psi_l(Q_\beta)\right)\\ &- \sum_k \psi_k(Q_\beta)(w_k(t)\cdot \frac{\partial}{\partial Q_\beta}\left(\sum_l u_l(t)\psi_l(Q_\beta)\right).\end{aligned} \tag{4.3}$$

- $\langle\cdot,\cdot\rangle_g$ *is the grid inner product defined by*

$$\langle f, g\rangle_g = \sum_{k,l} f_k \cdot M_{kl} g_l,$$

(i.e., *a discrete approximation of the* L^2 *inner product in the continuous case), and*

- $\langle\cdot,\cdot\rangle_p$ *is the particle inner product on* $T\Omega^{n_p}$ *defined by*

$$\langle F, G\rangle_p = \sum_\beta F_\beta \cdot G_\beta.$$

Remark 4.1. The operation $(\mathrm{ad}_u\, w)_\beta$ in Eq. (4.3) is the Lie bracket among PoU vector fields evaluated at the particle location Q_β. The operation $\mathrm{ad}^*_u\, m$ in Eq. (4.2) is its dual with respect to the pairing $\langle\cdot\,,\,\cdot\rangle_g$ on the grid.

Proof. For any vector field m defined on the grid, the time derivative of the momentum formula gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle m, w\rangle_k = \frac{\mathrm{d}}{\mathrm{d}t}\sum_{kl} m_k M_{kl} w_l,$$

$$
\begin{aligned}
\{\text{using momentum formula (Eq. (3.5))}\} &= \frac{\mathrm{d}}{\mathrm{d}t}\sum_l\sum_\beta P_\beta\psi_l(Q_\beta)\cdot w_l \\
&= \sum_l\left(P_\beta\frac{\partial\psi_l(Q_\beta)}{\partial Q_\beta}\cdot\dot{Q}_\beta+\sum_\beta\dot{P}_\beta\psi_l(Q_\beta)\right)\cdot w_l, \\
\{\text{using dynamical Eqs. (3.3) and (3.4)}\} &= \sum_l\sum_\beta\left(P_\beta\frac{\partial\psi_l(Q_\beta)}{\partial Q_\beta}\cdot\sum_k u_k\psi_k(Q_\beta)\right. \\
&\quad \left.- P_\beta\cdot\sum_k u_k\frac{\partial\psi_k(Q_\beta)}{\partial Q_\beta}\psi_l(Q_\beta)\right)\cdot w_l \\
&= \sum_\beta P_\beta\cdot\left(\sum_k u_k\frac{\partial\psi_k(Q_\beta)}{\partial Q_\beta}\cdot\sum_l w_l\psi_l(Q_\beta)\right. \\
&\quad \left.-\sum_k w_k\frac{\partial\psi_k(Q_\beta)}{\partial Q_\beta}\cdot\sum_l u_l\psi_l(Q_\beta)\right), \\
\{\text{using Eq. (4.3)}\} &= -\sum_\beta P_\beta(\mathrm{ad}_u\, w)_\beta \\
&= -\langle P,\mathrm{ad}_u\, w\rangle_p, \\
\{\text{using Eq. (4.2)}\} &= -\langle\mathrm{ad}^*_u\, m,\, w\rangle_k,
\end{aligned}
$$

hence the result. □

The discrete ad^* operator takes the form

$$
\begin{aligned}
(\mathrm{ad}^*_u\, m)_k &= \sum_m(M^{-1})_{km}\Big(\sum_l\sum_\beta\frac{\partial\psi_l}{\partial Q_\beta}(Q_\beta)(P_\beta\cdot u_l)\psi_m(Q_\beta) \\
&\quad -\sum_\beta P_\beta\frac{\partial\psi_m}{\partial Q_\beta}(Q_\beta)\cdot\sum_l u_l(t)\psi_l(Q_\beta)\Big) \\
&= \sum_m(M^{-1})_{km}\sum_\beta\Big(\psi_k(Q_\beta)P_\beta\cdot[\nabla u]^P_\beta-\sum_\beta P_\beta\frac{\partial\psi_m}{\partial Q_\beta}(Q_\beta)\cdot[u]_\beta\Big) \\
&= [m\cdot[\nabla u]^P]^G-[\nabla\cdot([u]^P m)]^G
\end{aligned}
$$

which is an approximation to the continuous ad^* operator using the particle-mesh calculus.

5. Left-action momentum map

First, recall that a canonical action of a Lie algebra A on a symplectic manifold $\mathcal{M}$ is a mapping from A to Hamiltonian vector fields on $\mathcal{M}$ which preserves the Lie brackets. Consider an element ξ of A and its action $\xi_{\mathcal{M}}$ on $\mathcal{M}$ which has Hamiltonian J. The momentum map J is related to the Hamiltonian J by

$$\langle J, \xi \rangle = J,$$

for all such elements ξ, where $\langle \cdot\,, \cdot \rangle : A^* \times A \to \mathbb{R}$ is the inner product between A and its dual A^*.

If A acts on a manifold $\mathcal{M}$, then we can define a canonical action of A on $T^*\mathcal{M}$ with Hamiltonian

$$J = \langle (P, Q), \xi_{\mathcal{M}} \rangle = P \cdot \xi_{\mathcal{M}}.$$

This is called the *cotangent lift* of the action to $T^*\mathcal{M}$. The definition of the momentum map for the cotangent lift of an action then becomes

$$\langle J, \xi \rangle = \langle (P, Q), \xi_{\mathcal{M}} \rangle. \tag{5.1}$$

We define the left-action of $\mathfrak{X}(\Omega)$ on Ω by

$$\xi \mapsto \xi_Q = \xi(Q) \cdot \frac{\partial}{\partial Q}.$$

The Hamiltonian for the cotangent-lifted left-action is then

$$J(\xi)(P, Q) = \langle (P, Q), \xi_Q \rangle.$$

We wish to obtain a momentum map which maps into the representation of D_{VPM} given by the map $\mathbb{R}^{d \times n_g} \to \mathfrak{X}(\Omega)$:

$$u \mapsto \sum_k u_k \psi_k(Q).$$

We do this by restricting $J(\xi)$ to elements of D_{VPM}:

$$\begin{aligned} J(u)(P, Q) &= \left\langle (P, Q), \sum_k u_k \psi_k(Q) \cdot \frac{\partial}{\partial Q} \right\rangle \\ &= \sum_{\beta k} P_\beta \cdot u_k \psi_k(Q), \end{aligned}$$

and this relation defines the left-action momentum map

$$J_k^L(P, Q) = \sum_\beta P_\beta \psi_k(Q_\beta). \tag{5.2}$$

As mentioned in Remark (3.1) this is again Eq. (3.5) derived earlier from constrained variations of the VPM action (Eq. (3.2)) with respect to the grid representatives of the velocity. This momentum map is the discrete version for VPM of a general result for Clebsch variational principles for ideal fluid dynamics (Holm and Kupershmidt [1983]; Holm, Marsden and Ratiu [1998]).

6. Right-action momentum map

The PoU vector field is defined over the whole space parameterized by the vectors $\{u_k\}_{k=1}^{n_g}$ associated with the mesh points. This means that we can follow how the flow map generated by the vector fields evolves over the whole space

$$\frac{\partial}{\partial t} g(x,t) = \sum_k u_k \psi_k(g(x,t)),$$

and we can follow the Jacobian of this map

$$\frac{\partial}{\partial t}\frac{\partial g(x,t)}{\partial x} = \sum_k u_k \nabla\psi_k(g(x,t)) \cdot \frac{\partial g(x,t)}{\partial x}.$$

In particular, we can evaluate this equation at each of our discrete points Q_β:

$$\dot{J}_\beta = \sum_k u_k \nabla\psi_k(Q_\beta) \cdot J_\beta, \qquad J_\beta = \frac{\partial g}{\partial x}(Q_\beta(0), t), \qquad J_\beta(0) = \mathrm{Id}. \tag{6.1}$$

A variation $\delta Q_\beta(0)$ in the initial conditions $Q_\beta(0)$ leads to a variation in the entire trajectory given by

$$\delta Q_\beta(t) = J_\beta(t)\delta Q_\beta(0). \tag{6.2}$$

This variation generates a symmetry of the equations as shown in the following lemma.

LEMMA 6.1. *The infinitesimal transformation given by Eqs. (6.1) and (6.2) together with*

$$\delta P_\beta = 0, \quad \beta = 1, \ldots, n_p, \qquad \delta u_k = 0, \quad k = 1, \ldots, n_k$$

is a symmetry of Eqs. (3.3)–(3.5).

PROOF. Since the equations have been derived from an action principle, we simply need to show that the infinitesimal transformation causes the action (Eq. (3.2)) to vanish. If we apply the transformation to the integrand (the Lagrangian), then we obtain

$$\begin{aligned}
\delta L &= \delta\left(\|u\|_g^2 + \sum_\beta P_\beta \cdot \left(\dot{Q}_\beta - \sum_{k=1}^{n_g} u_k \psi_k(Q_\beta)\right)\right) \\
&= \sum_\beta P_\beta \cdot \left(\delta\dot{Q}_\beta - \sum_{k=1}^{n_g} u_k \nabla\psi_k(Q_\beta) \cdot \delta Q_\beta\right) \\
&= \sum_\beta P_\beta \cdot \left(\dot{J}_\beta \cdot \delta Q_\beta(0) - \sum_{k=1}^{n_g} u_k \nabla\psi_k(Q_\beta) \cdot J_\beta \delta Q_\beta(0)\right) \\
&= \sum_\beta P_\beta \cdot \left(\sum_k u_k \nabla\psi_k(Q_\beta) \cdot J_\beta \cdot \delta Q_\beta(0) - \sum_{k=1}^{n_g} u_k \nabla\psi_k(Q_\beta) \cdot J_\beta \delta Q_\beta(0)\right) \\
&= 0,
\end{aligned}$$

and hence the result. □

This symmetry of the equations has an associated conserved momenta following Noether's theorem as described in the following proposition.

PROPOSITION 6.1. *For each $\beta = 1, \ldots, n_p$, the quantity $P_\beta J_\beta$ is conserved.*

PROOF. Applying the symmetry to the action principle and substituting the equations of motion (3.3)–(3.5) gives

$$\begin{aligned} 0 &= \delta \int_0^1 \|u\|_g^2 + \sum_\beta P_\beta \cdot \left(\dot{Q}_\beta - \sum_{k=1}^{n_g} u_k \psi_k(Q_\beta) \right) \mathrm{d}t \\ &= \int_0^1 \sum_\beta \left(\dot{Q}_\beta - \sum_{k=1}^{n_g} u_k \psi_k(Q_\beta) \right) \cdot \delta Q_\beta + \frac{\mathrm{d}}{\mathrm{d}t} \left(\sum_\beta P_\beta J_\beta \cdot \delta Q_\beta(0) \right) \mathrm{d}t \\ &= \sum_\beta \big(P_\beta(t_1) J_\beta(t_1) - P_\beta(t_0) J_\beta(t_0) \big) \cdot \delta Q_\beta(0). \end{aligned}$$

We obtain the result since $\delta Q_\beta(0)$ is an arbitrary vector. □

As described in LEW, MARSDEN, ORTIZ and WEST [2003], these conservation laws satisfied by the semidiscrete equations will be preserved by numerical methods provided that they are derived from a discrete variational principle, i.e., following the framework described in Section 3.3.

REMARK 6.1. The conserved momentum $P_\beta J_\beta$ is not a local quantity because J_β is obtained by integrating Eq. (6.1) along the entire solution trajectory for Q and P.

7. Kelvin's circulation theorem for discrete EPDiff

As discussed in the previous section, the discrete EPDiff Lagrangian is invariant under relabeling symmetries. This means that

$$\frac{\mathrm{d}}{\mathrm{d}t} \Big(P_\beta \cdot \frac{\partial g_t}{\partial x}(Q_\beta^0) \Big) = 0, \qquad \text{no sum on } \beta$$

for each $\beta = 1, \ldots, n_p$.

We can interpret this result to prove a discrete form of Kelvin's circulation theorem. Consider a loop $C(t)$ in Ω which is embedded in the flow, i.e.,

$$C(t) = g_t(C(0)).$$

We choose $C(t)$ so that some of the particles with trajectories $Q_\beta(t)$ are located at the initial time $t = 0$ on $Q_\beta(0) \in C(0)$. As $g_t(Q_\beta^0) = Q_\beta(t)$, those particles will stay on $C(t)$ for all time. Define the set Γ so that $\beta \in \Gamma$ if Q_β^0 is located on $C(0)$.

In order to discuss the circulation theorem, we need to introduce a discretization of density. As discussed in COTTER [2005], this is done by associating a constant $\tilde{D}_\beta$ with each particle so that the density on the grid may be written

$$D_k = \sum_\beta \tilde{D}_\beta \psi_k(Q_\beta).$$

This allows us to represent m/D evaluated at the location of particle β as $P_\beta/\tilde{D}_\beta$.

Next, we need to approximate line integration round $C(t)$. We do this by writing

$$\int_{C(t)} \frac{m}{D} \cdot \mathrm{d}x = \int_0^{2\pi} \frac{m \circ \gamma_t}{D \circ \gamma_t} \cdot \frac{\mathrm{d}\gamma_t}{\mathrm{d}s}\, \mathrm{d}s,$$

where $\gamma_t : [0, 2\pi] \to C(0)$ is a parameterisation of the loop $C(t)$. Substituting $\gamma_t = g_t \circ \gamma_0$ yields

$$\int_{C(t)} \frac{m}{D} \cdot \mathrm{d}x = \int_0^{2\pi} \frac{m \circ \gamma_t}{D \circ \gamma_t} \cdots \frac{\partial g_t}{\partial x}(\gamma_0(s)) \cdot \frac{\mathrm{d}\gamma_0}{\mathrm{d}s}\, \mathrm{d}s,$$

which we can approximate with a Riemann sum

$$\int_{C(t)} \frac{m}{D} \cdot \mathrm{d}x \approx \sum_\beta \frac{P_\beta}{\tilde{D}_\beta} \cdot \Delta x_\beta,$$

where

$$\Delta x_\beta = J_\beta \cdot \frac{\mathrm{d}\gamma_0}{\mathrm{d}s}(s_\beta) \Delta s_\beta,$$

with $\gamma_0(s_\beta) = Q_\beta^0$ and $\Delta s_\beta = s_{\beta+1} - s_\beta$.

Using this discretized line integration scheme, we can state our Kelvin circulation theorem as follows:

PROPOSITION 7.1. *Let $\{u_k(t)\}_{k=1}^{n_g}$ satisfies the discrete EP equations above, with $\{D_k(t)\}_{k=1}^{n_g}$ satisfying the discrete density equation. Let $C(t)$ be a closed loop advected in the flow generated by the velocity*

$$u(x, t) = \sum_k u_k(t) \psi_\beta(Q_\beta),$$

containing some subset of particles Q_β, with $\beta \in B \subset (1, \ldots, n_p)$. Define the discrete circulation sum

$$I(t) = \sum_{\beta \in B} \frac{P_\beta}{\tilde{D}_\beta} \cdot \Delta x_\beta,$$

then $I(t)$ satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t} I(t) = 0.$$

PROOF. The result proceeds directly from Corollary 6.2 for the right-action momentum map, which satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}(P_\beta \cdot J_\beta) = 0, \quad \forall \beta. \qquad \square$$

8. Numerical results

Convergence tests We begin by performing a convergence test for the 1D equations

$$m_t + u m_x + 2 m u_x = 0, \quad (1 - \alpha^2 \partial_x^2) u,$$

which, as discovered in CAMASSA and HOLM [1993], is completely integrable with the initial value problem dominated by peaked solitons (peakons) whose first derivatives are discontinuous. This property is illustrated in Fig. 8.1 which shows a numerical integration of the 1D equations starting from a smooth initial condition with singular peaked solitons emerging in finite time.

For our first convergence test, we use the result given in CAMASSA and HOLM [1993] that for an initial condition $u = (\pi/2)\mathrm{e}^x - 2\sinh x \arctan(\mathrm{e}^x) - 1$, with scaling constant $\alpha = 1$, the asymptotic speeds of the emitted peakons are $2/[(2n+1)(2n+3)]$, $n = 0, 1, 2, \ldots$. In particular, the asymptotic speed of the first peakon is $2/3$. Fig. 8.2 shows that the numerical calculation of the speed converges to the correct answer with a linear scaling for error against grid resolution.

For our second convergence test, we used the problem of an overtaking collision (illustrated in Fig. 8.3) between two right-propagating peakons. CAMASSA and HOLM [1993] gives a formula the phase shifts for such a collision (i.e., the asymptotic difference in positions for the larger and smaller soliton with and without the collision). A plot of the error in the phase shift against grid-size is given in Fig. 8.4.

We found that the performance of the method when solving for head-on peakon/antipeakon interactions was quite poor. During the collision, the two peakons approach each other and stick together once they are both within a grid width of each other. This appears to be an issue with representing the momentum using Lagrangian particles; so as to achieve a method with the correct results for the collision, the particle momenta would need to go to infinity during the collision. However, we can also view this as a benefit of the method. The peakon/antipeakon solution represents an instability in the equations, namely that a small perturbation of the solution can result in a peakon/antipeakon pair being created. As our method does not support this solution at the moment of collision, this type of instability does not pollute our numerical results.

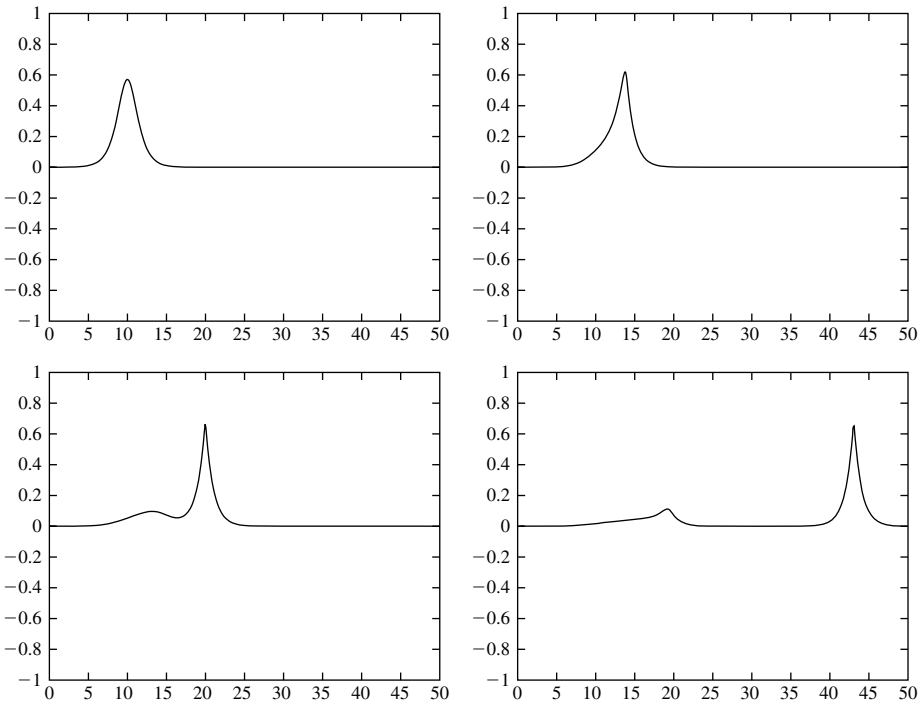


FIG. 8.1 Numerical solution to the one-dimensional EPDiff equation with initial data $u = (\pi/2)e^x - 2\sinh x \arctan(e^x) - 1$ and scaling constant $\alpha = 1$. The solutions are obtained using the VPM method with 500 grid points, 1000 particles, cubic B-splines for the basis functions, linear finite elements for the grid discretization of the Lagrangian, and a time-step of 0.1 with the first-order time discretization given by Eqs. ((3.9)–(3.11)). The figures show the velocity field at times 0, 5, 25, and 50. At $t = 5$, the smooth initial condition has "leaned to the right" and a discontinuous peak has formed. By $t = 25$, the peakon is well separated from the smooth part of the solution, and by $t = 50$, a second peak is starting to form.

8.1. 2D Flows

In this section, we show a few results obtained using the VPM method to discretize EPDiff in 2D with Lagrangian

$$L = \frac{1}{2}\int |\boldsymbol{u}|^2 + \alpha^2 |\nabla \boldsymbol{u}|^2 \, \mathrm{d}^2 x,$$

for a constant length scale α so that the velocity u is obtained from the momentum $\boldsymbol{m} = \frac{\delta l}{\delta \boldsymbol{u}}$ by inverting the modified Helmholtz operator

$$\boldsymbol{m} = (1 - \alpha^2 \Delta)\boldsymbol{u}.$$

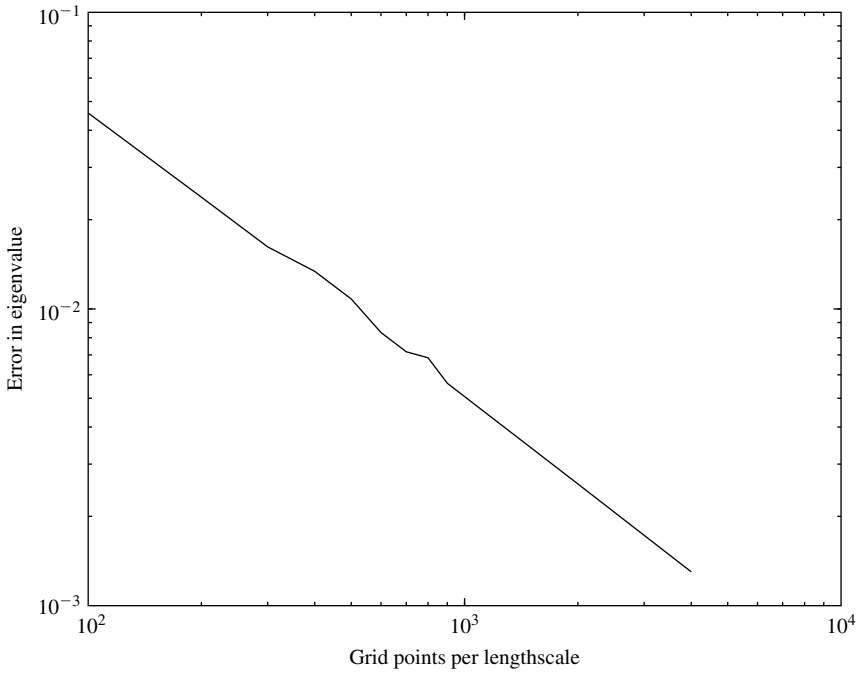


FIG. 8.2 Plot of error in calculating the asymptotic speed of the first emitted peakon from an initial condition $u = (\pi/2)e^x - 2\sinh x \arctan(e^x) - 1$ against grid resolution (measured as number of grid points in one characteristic length $\alpha = 1$), using the same method as Fig. 8.1. The number of particles used was twice the number of grid points, and the time-step used was scaled with the grid size to guarantee convergence of the fixed-point method for solving the linear system (i.e., not for accuracy). The logarithmic plot has slope -1 giving a linear scaling for the error with grid resolution.

In the first experiment, the initial condition for the momentum had a 2D "top-hat" profile

$$m = (m(x, y), 0), \quad m = \begin{cases} 1 & \text{if } a < x < b, \quad c < y < d, \\ 0 & \text{otherwise} \end{cases},$$

for constants a, b, c, *and* d so that the velocity has continuous gradients and has compact support. Fig. 8.5 shows the evolution of the velocity at subsequent times; it illustrates how EPDiff evolves to form singular filaments of momentum from smooth initial conditions.

8.2. Verification of conservation laws

The next set of numerical results demonstrates the conservation of the right-action momenta given in Section 6 and the connection with Kelvin's circulation theorem.

Fig. 8.6 shows the value of the momentum map for right-action $P_\beta \cdot J_\beta$ for a selection of particles from the flow in Fig. 8.5. All the particles have the same value because they all have the same initial momentum, and J is set to the identity initially. The figure shows that the numerical method preserves these conserved momenta up to roundoff

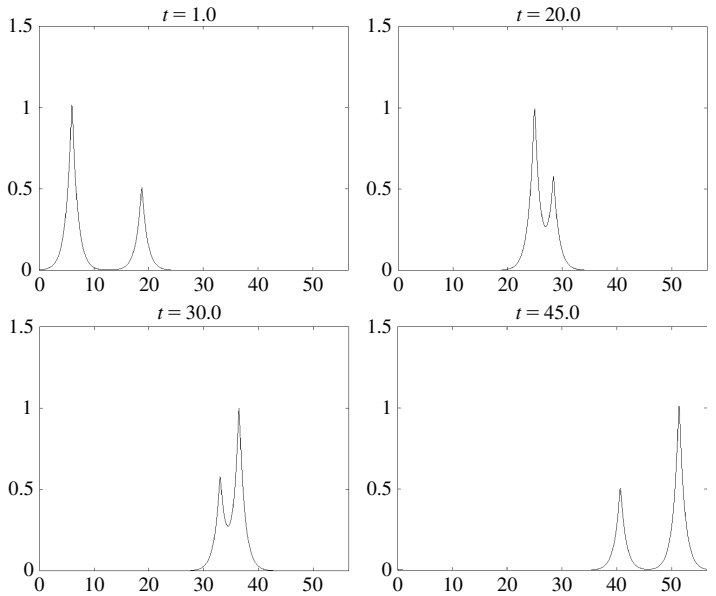


FIG. 8.3 Plots of the velocity during an overtaking collision between two peakons at times $t = 1$, $t = 25$, $t = 35$, and $t = 45$. The peakon to the left has greater velocity than the one to the right, so they collide. The momentum is transferred from the peakon behind to the one in front which accelerates away.

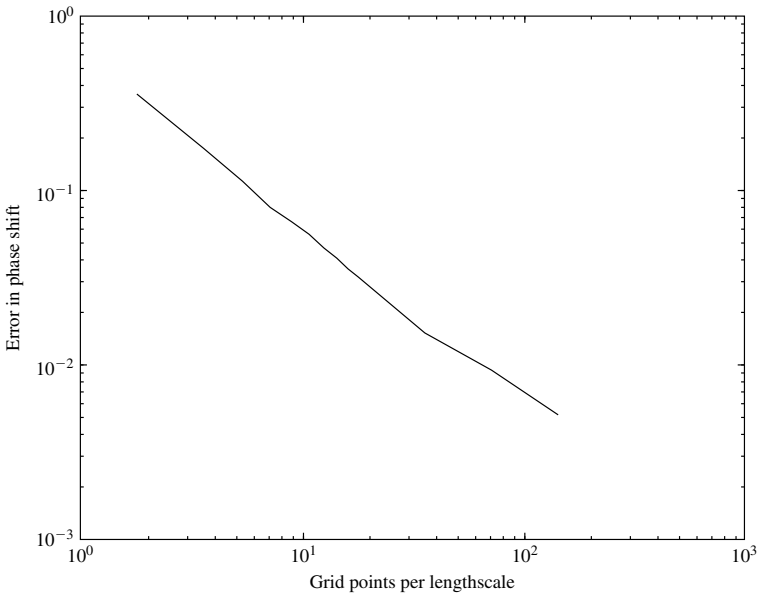


FIG. 8.4 Plot of error in calculating the fast shift in the faster peakon between a peakon of height 1.0 and a peakon of height 0.5 against grid resolution (measured as number of grid points in one characteristic length $\alpha = 1$). This plot shows linear convergence of the error in the numerical solution.

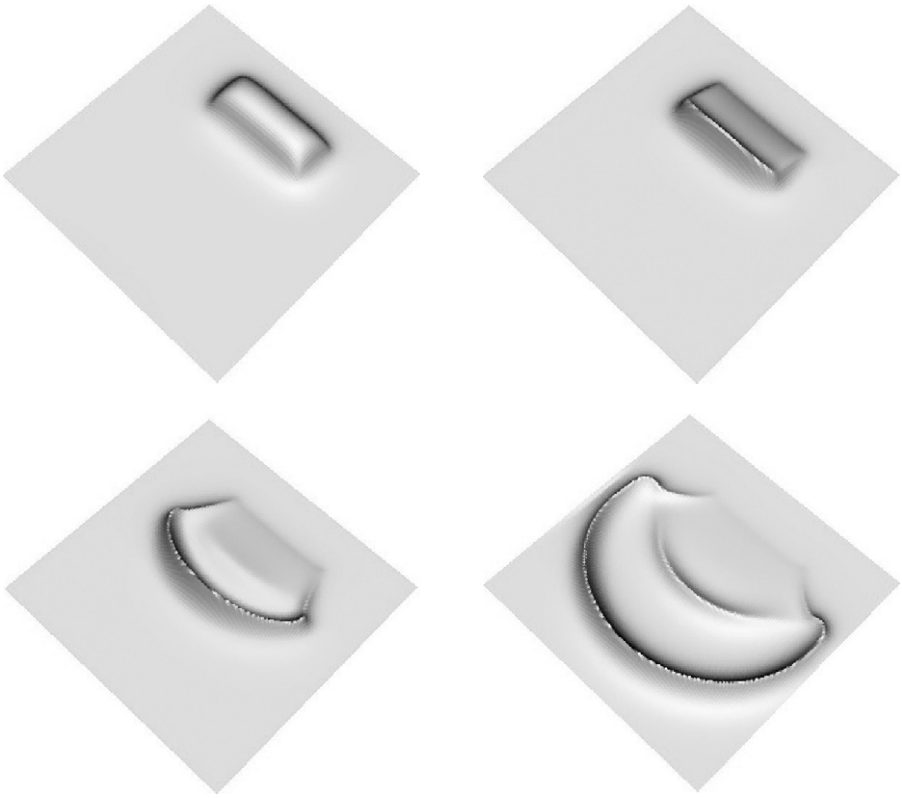


FIG. 8.5 Plots showing surfaces of velocity magnitude $|u|$ at times $t = 0, 0.45, 1.5$, and 3.4 with a C^2-smooth, compactly supported initial condition for $\boldsymbol{u}$. The momentum becomes supported on lines (so that the velocity has a "peaked" profile) which spread out. This is the 2D version of emerging peakons illustrated in Fig. 8.1. These results were obtained using 65536 particles, a 128×128 grid with periodic boundary conditions, tensor product B–spline basis functions, and piecewise-linear finite element discretization of the Lagrangian. The time-step is 1.0×10^{-3} and $\alpha = 0.2$.

error. This follows from Proposition 6.1 for the time continuous equations and the fact that conserved momentum maps are also conserved by variational integrators.

To verify the discrete circulation conservation discussed in Section 7, we took an arbitrary loop containing some of the particles and advected the loop with the flow shown in Fig. 8.5, using

$$Q^{n+1}(s) = Q^n(s) + \sum_k u_k \psi_k(Q(s)),$$

where s parameterizes the loop. During the course of the flow, this arbitrary circulation loop evolves, changing shape and length significantly. However, the circulation around the loop remains constant (up to numerical roundoff) as verified numerically in the following.

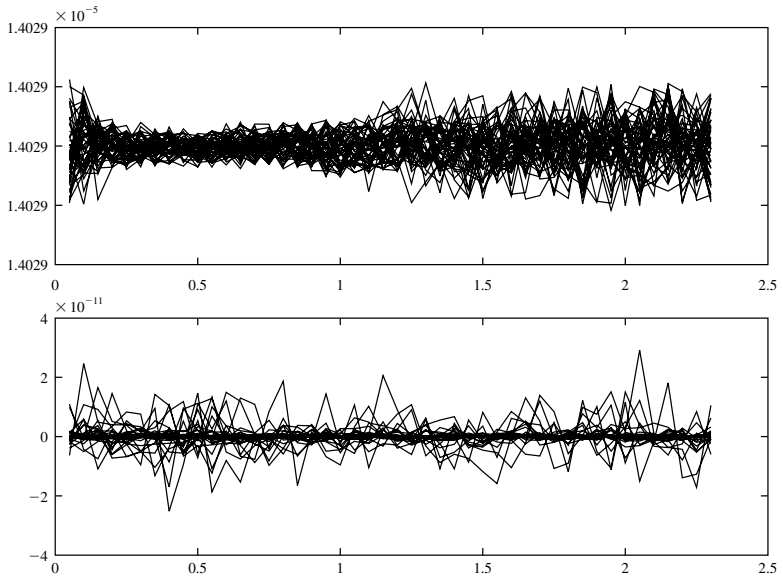


FIG. 8.6 Plots of $P_\beta \cdot J_\beta$ against time for a selection for particles with label β, illustrating that it is exactly conserved during the simulation. The upper plot is the x-component, and the lower plot is the y-component. Any variation seen is due to numerical roundoff error.

To write down the circulation integral, we choose an initial density

$$D_k = (M^{-1})_{kl} \sum_\beta \tilde{D}_\beta \psi_k(Q_\beta),$$

where the values of $\tilde{D}$ do not matter much as they are not coupled with the dynamics. Hence, we choose the values $\tilde{D}_\beta = 1$, $\beta = 1, \ldots, N_p$. To obtain the discretized loop integral

$$\sum_{\beta=1}^{N} \frac{P_\beta}{\tilde{D}_\beta} \cdot \Delta x_\beta,$$

we need to calculate Δx_β as discussed in Section 7. This is done by finding discrete line elements Δx_β^0 for the initial loop and then calculating Δx for subsequent time-steps using

$$\Delta x_\beta^n = J_\beta^n \cdot \Delta x_\beta^0.$$

8.3. Summary of circulation loop figures

- Fig. 8.7 shows the evolution in time of this circulation loop, under the flow induced by the expanding waves in Fig. 8.5.

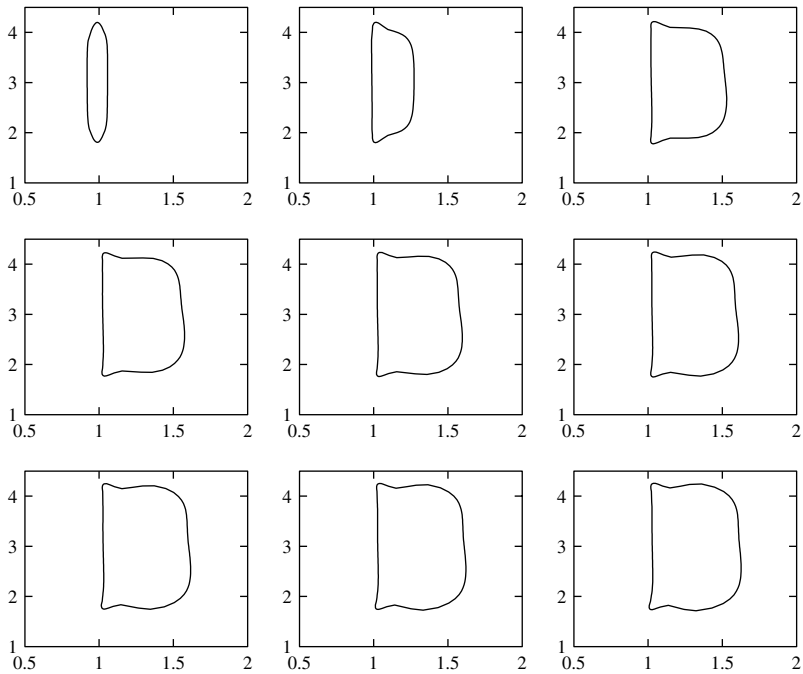


FIG. 8.7 Plots showing an embedded loop in the time-varying flow obtained from a solution of EPDiff, illustrated in Fig. 8.5, at times (left-to-right, top-to-bottom) $t = 0$, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, and 2.0.

- A plot showing initial and advected line elements is given in Fig. 8.8. This plot illustrates how the line elements evolve when a loop is stretched out by the flow. In the top and the bottom of the loop, where the stretching is greatest, one can see how the line elements extend to provide a numerical approximation to dx on the loop.
- Finally, a plot of the circulation integral is given in Fig. 8.9. This plot shows that the circulation round the loop is exactly preserved during the simulation (up to roundoff error in the calculation of the discrete integral).

9. Summary and outlook

In this chapter, we studied the VPM method applied to the EPDiff equation. We introduced a constrained variational principle for the method and gave discrete EP formulae on the Eulerian grid resulting from the variational principle, which show that the grid velocities and momenta satisfy the EPDiff equations in Eulerian form. Next, we looked at left- and right-actions of velocity vector fields on the Lagrangian particles and obtained corresponding momentum maps. The left-action, when restricted to the finite space of velocity fields used in the method, gives rise to a momentum map which is the same formula as used for calculating the grid momentum from the particle variables. The

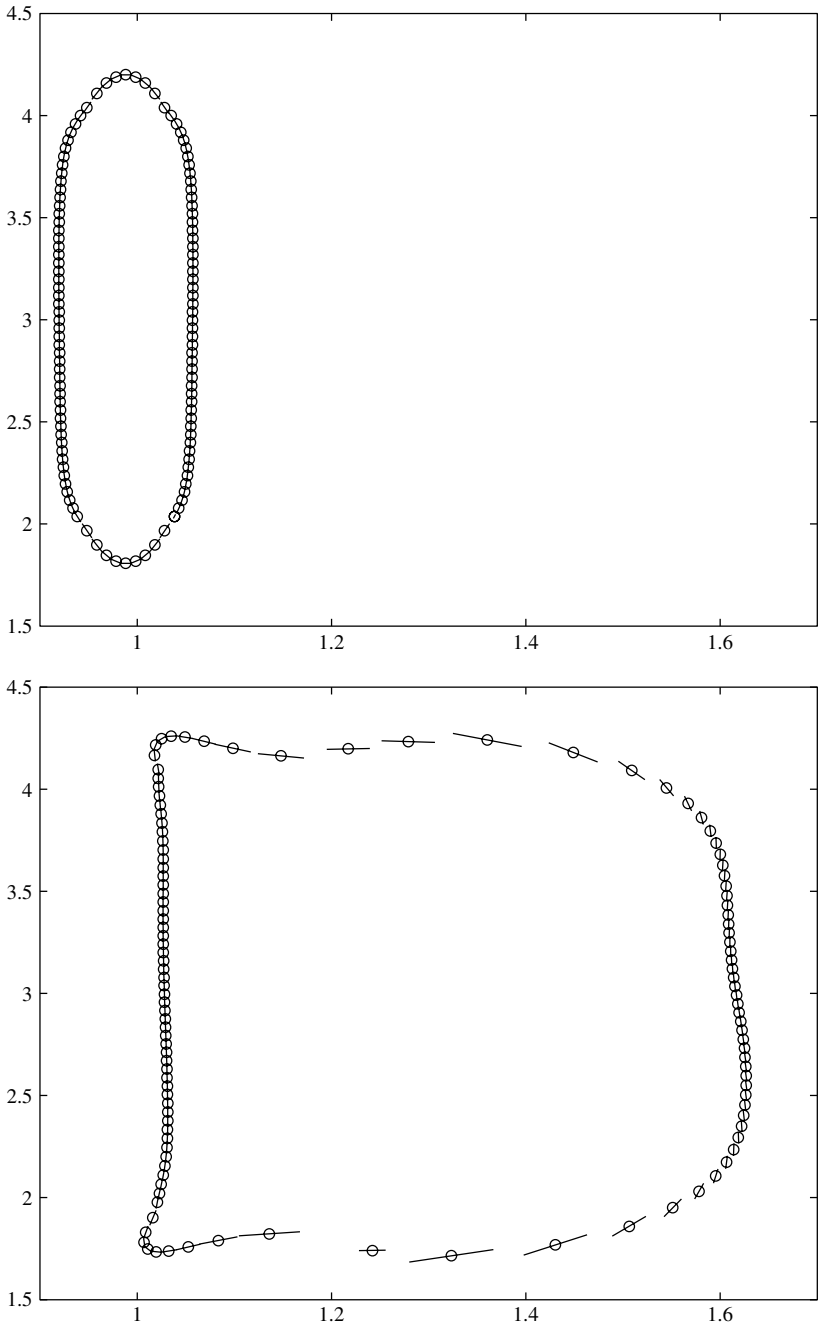


FIG. 8.8 Plot showing the curve with embedded line elements at time $t = 0$ and $t = 2.3$. The line elements at time $t = 2.3$ are the exact tangents to the curve which passes through the tangents given at time $t = 0$ and is then advected using the time independent flow given in Fig. 8.5.

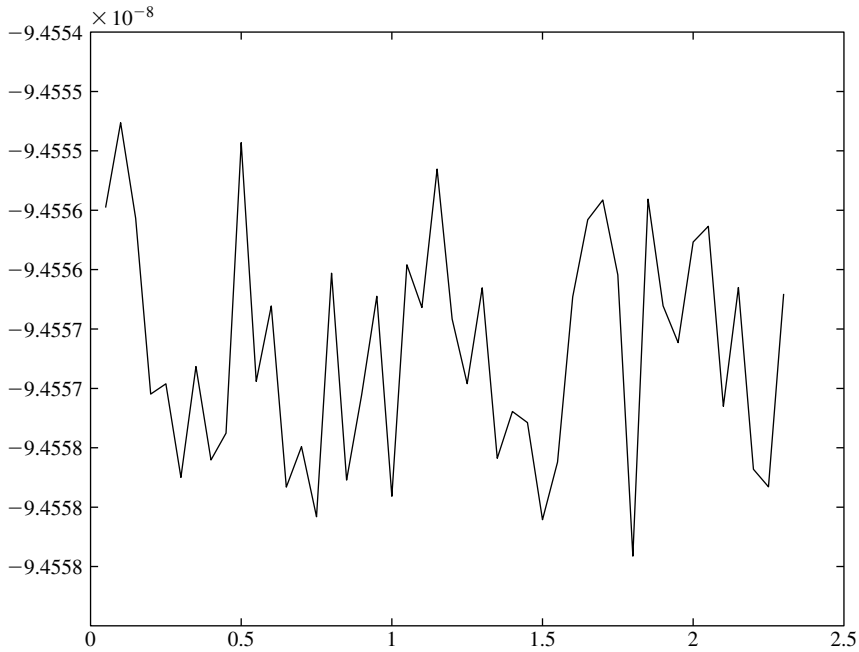


FIG. 8.9 Plot of the circulation integral around the loop illustrated in Fig. 8.8 against time. The total circulation is very small because the momentum is initially almost exactly tangent to the curve, and the conservation of PdQ ensures that it stays tangent during the whole simulation. The circulation is exactly preserved by the numerical method; any variations which can be seen here are due to roundoff error in the numerical scheme and in the calculation of the loop integral.

right-action can be interpreted as a discrete form of particle-relabeling since it corresponds to moving the particles in such a way so that the grid velocities remain constant. Finally, we gave some interpretation of these transformations in terms of matrices which determine the local deformation of infinitesimal line elements, thereby allowing us to write down discrete loop integrals on advected loops. This led to a discrete circulation theorem.

Our next aim is to find an extension of this work which gives a discrete circulation theorem for fluid PDEs which involve mass density and other advected quantities as well as velocity. The general approach, following the continuous theory, will be to

- specify the transformations corresponding to discrete relabeling,
- determine the transformation of density and other advected quantities under this discrete relabeling group,
- calculate the momentum densities obtained from these transformations, and
- show that the ratio of these momentum densities to the mass density is invariant.

Including advected quantities in this way will allow introduction of potential energy and hence linear dispersion effects into the numerical description of the internal wave interactions using the VPM method.

10. Acknowledgments

We are grateful to our colleagues Joel Fine and Matthew Dixon at Imperial College London for their advice and consultation regarding this problem. We are also grateful to the ONR-NLIWI program for partial funding of this endeavor, and to Tony Liu for use of the SAR images of the South China Sea taken from the Space Shuttle. DDH is also grateful for partial support from the Office of Science, US Department of Energy.

References

Arnold, V.I. (1966). Sur la géometrie differentielle des groupes de Lie de dimenson infinie et ses applications a l'hydrodynamique des fluids parfaits. *Ann. Inst. Fourier* (16), 319–361.

Bou-Rabee, N., Marsden, J.E. (2007). Hamilton-Pontryagin integrators on lie groups part I: introduction and structure-preserving properties. submitted.

Bou-Rabee, N., Marsden, J. E. (2007). Hamilton-Pontryagin integrators on Lie groups part II: numerical analysis and applications. submitted.

Camassa, R., Holm, D.D. (1993). An integrable shallow-water equation with peaked solitons. *Phys. Rev. Lett.* **71** (11), 1661–1664.

Camassa, R., Holm, D.D., Levermore, C.D. (1996). Long-time effects of bottom topography in shallow water. *Physica D*. **98** (2–4), 258–286.

Cendra, H., Marsden, J.E., Pekarsky, S., Ratiu, T.S. (2003). Variational principles for Lie-Poisson and Hamilton-Poincaré equations. *Moskow Math. Journ.* **3** (3), 833–867.

Cotter, C.J. (2005). A general approach for producing Hamiltonian numerical schemes for fluid equations. Available at 2005; http://www-arxiv.org/pdf/math.NA/0501468.

Cotter, C.J., Holm, D.D. (2008). Continuous and discrete Clebsch variational principles. to appear in Foundations of Computational Mechanics.

Foias, C., Holm, D.D., Titi, E.S. (2001). The Navier-Stokes-alpha model of fluid turbulence. *Physica D*. **152**, 505–519.

Frank, J., Gottwald, G., Reich, S. (2002). A Hamiltonian particle-mesh method for the rotating shallow-water equations. In: *Lecture Notes in Computational Science and Engineering,* vol. 26 (Springer-Verlag), pp. 131–142.

Holm, D.D., Kupfershmidt, B., Levermore, C.D. (1985). Hamiltonian differencing of fluid dynamics. *Adv. Appl. Math.* 52–84.

Holm, D.D., Kupershmidt, B. (1983). Poisson brackets and Clebsch representations for magnetohydrodynamics, multifluid plasmas, and elasticity. *Physica D*. **6**, 347–363.

Holm, D.D., Marsden, J.E. (2004). Momentum maps and measure valued solutions (peakons, filaments, and sheets) of the Euler-Poincaré equations for the diffeomorphism group. In: Marsden, J.E., Ratiu, T.S. (eds.), In: The Breadth of Symplectic and Poisson Geometry, A Festshrift for Alan Weinstein (Birkhäuser Boston, Boston, MA), pp. 203–235, Available at 2004; http://www-arxiv.org/abs/nlin.CD/0312048.

Holm, D.D., Marsden, J.E., Ratiu, T.S. (1998). The Euler-Poincaré equations and semidirect products with applications to continuum theories. *Adv. Math.* **137**, 1–81. Available at 1998; http://www-arxiv.org/abs/chao-dyn/9801015.

Holm, D.D., Rananather, J.T., Trouvé, A., Younes, L. (2004). Soliton dynamics in computational anatomy. *Neuroimage*. **23**, 170–178. Available at 2004; http://www-arxiv.org/abs/nlin.SI/0411014.

Lew, A., Marsden, J.E., Ortiz, M., West, M. (2003). An overview of variational integrators. In: Franca, L.P. (ed.), *Finite Element Methods: 1970s and Beyond* (CIMNE, Barcelona, Spain).

Marsden, J.E., Weinstein, A. (1983). Coadjoint orbits, vortices, and Clebsch variables for incompressible fluids. *Physica D*. **7**, 305–323.

Miller, M.I., Trouvé, A., Younes, L. (2002). On the metrics and Euler-Lagrange equations of computational anatomy. *Ann. Rev. Biomed. Eng.* **4**, 375–405.

Seliger, R.L., Whitham, G.B. (1968). Variational Principles in Continuum Mechanics. Proceedings of the Royal Society of London. Series A, *Mathematical and Physical Sciences*. **305**, Pages 1–25.

Yoder, J. A., Ankleson, S.G., Barber, R.T., Flament, P., Balch, W.M. (1994). A line in the sea. *Nature*. (371), 689–692.

Numerical Generation of Stochastic Differential Equations in Climate Models

Brian Ewald

Department of Mathematics, Florida State University, Tallahassee, Florida, USA

Cécile Penland

Physical Science Division/ESRL/NOAA 325 Broadway, Boulder, CO 80305-3328
November 2007

1. Introduction

It is, at present, impossible for a computer model to simulate weather and climate at all spatial and temporal scales. This would be no problem if the dynamical processes at widely different scales were independent of each other, but they are not. The modeler is therefore faced with the decisions of what physical processes to model explicitly, what processes to ignore, and what processes to parameterize. The word "parameterize" is used to indicate an approximation of a special kind; phenomena that are not explicitly treated in a numerical model are "parameterized" in a model when they are approximated by expressions involving phenomena that are. For example, the barotropic vorticity equation describes horizontal rotational motions only. The effects of divergent flow and vertical mass transport on these motions are parameterized by means of damping and forcing terms proportional to the horizontal rotational motions themselves. These parameterizations are not "ad hoc," as has sometimes been suggested, but are based on knowledge of data and physical principles.

Whether or not they are physically based, parameterizations are approximations and, as such, some are better than others. The more characteristics of a real process a parameterization can mimic, the better the approximation is expected to be. However, there are virtues in minimalism; resources expended in reproducing a marginally important effect are probably spent better elsewhere. Thus, returning to our example, we find that

Computational Methods for the Atmosphere and the Oceans
Special Volume (Roger M. Temam and Joseph J. Tribbia, Guest Editors) of
HANDBOOK OF NUMERICAL ANALYSIS, VOL. XIV
P.G. Ciarlet (Editor)
Copyright © 2009 Elsevier B.V.
All rights reserved
ISSN 1570-8659
DOI 10.1016/S1570-8659(08)00206-8

interactions between resolved and unresolved scales of vorticity are usually parameterized by hyperdiffusion in general circulation models with finite resolution in order to dissipate the numerical build-up of enstrophy at the smallest scales when these interactions are completely ignored. Accounting for the fact that these interactions can add enstrophy to the resolved scales as well as bleed it out from them, or that such interactions are more likely to be chaotically varying than temporally smooth, has been considered too expensive and difficult to be worth the trouble.

Accounting for the rapidly varying, chaotic nature of multiscale dynamical interactions in terms of random functions is the purview of stochastic theory. However, facing the importance of stochastic, subscale variability to weather and climate could happen only as computers became powerful enough to show that increasingly fine resolution in models never seemed to account for this variability completely. Further, powerful computers are needed to allow the rigorous application of stochastic theory in weather/climate models since numerical algorithms designed to optimize deterministic models are often not applicable to stochastic models. The art of numerical prediction and the theory of stochastic differential equations (SDEs) developed more or less contemporaneously, and certainly independently, over several decades so that the cultural shock of the two disciplines' discovery of each other has been severe. Happily, the last few years have seen the development of fruitful collaboration between geoscientists and applied mathematicians specializing in the development of numerical schemes for generating SDEs.

It is the purpose of this chapter to summarize numerical procedures for evaluating solutions of classical SDEs in climate prediction and research. Even this narrow focus is broad enough that we are likely to have overlooked relevant research, and we apologize to those researchers whose work falls into that category. The next section of this chapter discusses the central limit theorem, which directs how a system with scale separation may be approximated as an SDE. We do not provide a proof or even a rigorous statement of it, but rather set notation and examine some of its practical implications. The third section reviews the stochastic Taylor expansion and relates it to the development of stochastic numerical integration methods. The fourth section gives an overview of stochastic numerical methods as used in climate research, and the fifth presents examples. We conclude the chapter with a discussion of current directions in this field.

2. The central limit theorem

In the following, we discuss an extension of the traditional central limit theorem (e.g., Doob [1953], Wilks [1995]) usually employed by geoscientists to justify the use of Gaussian distributions. Informally, the classical central limit theorem states that the sum of independently sampled quantities is approximately Gaussian. In the version described below, we consider dynamical systems described by a slow timescale and faster timescales. The equations are averaged over a temporal interval large enough that the fast timescales collectively act as Gaussian stochastic forcing of the slow, coarse-grained system. In the mathematical literature (e.g., Feller [1966]), the fact that fine details of how the fast processes are distributed do not strongly affect the coarse-grained behavior of the slower dynamics is often called "the invariance principle." As the proof

of this theorem is far outside the scope of this review article, we state a commonly used form of it and refer the interested reader to the literature for details. GARDINER [1985] gives an heuristic description; we prefer PAPANICOLAOU and KOHLER [1974] for a technical statement of the theorem.

A dynamical system consisting of separated timescales may be written in the following manner:

$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \varepsilon \boldsymbol{G}(\boldsymbol{x}, t) + \varepsilon^2 \boldsymbol{F}(\boldsymbol{x}, t), \tag{2.1}$$

where $\boldsymbol{x}$ is an N-dimensional vector. In Eq. (2.1), the smallness parameter ε should not be taken as a measure of importance. It does measure the rapidity with which the terms on the right-hand side of Eq. (2.1) vary relative to each other; one can think of ε^2 as the ratio of the characteristic timescale of the first term to the characteristic timescale of the second. If we now cast Eq. (2.1) in terms of a scaled time coordinate,

$$\Delta s = \varepsilon^2 \Delta t, \tag{2.2}$$

Eq. (2.1) becomes

$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}s} = \frac{1}{\varepsilon}\boldsymbol{G}(\boldsymbol{x}, s/\varepsilon^2) + \boldsymbol{F}(\boldsymbol{x}, s/\varepsilon^2). \tag{2.3}$$

We further assume that the first term in Eq. (2.1) decays quickly "enough" in the time interval Δt. In the limit $\varepsilon \to 0$, $\Delta t \to \infty$ with $\varepsilon^2 \Delta t$ remaining finite, the central limit theorem states that Eq. (2.3) converges weakly to a "Stratonovich" (see below) SDE in the scaled coordinates:

$$\mathrm{d}\boldsymbol{x} = \boldsymbol{F}'(\boldsymbol{x}, s)\mathrm{d}s + \mathbf{G}'(\boldsymbol{x}, s) \cdot \mathrm{d}\boldsymbol{W}(s). \tag{2.4}$$

The symbol $\boldsymbol{W}$ in Eq. (2.4) is a K-dimensional vector, each component of which is an independent Wiener process, or Brownian motion, and the "$\cdot$" symbol indicates that it is to be interpreted in the sense of Stratonovich, discussed below. The symbol $\mathbf{G}'(\boldsymbol{x}, s)$ is a matrix, the first index of which corresponds to a component of x, and the second index of which corresponds to a component of $\mathrm{d}\boldsymbol{W}$. Using angle brackets to denote expectation values, we state the following properties of the vector Wiener process $\boldsymbol{W}$:

$$\boldsymbol{W}(s) \text{ is a vector of Gaussian random variables,} \tag{2.5a}$$

$$<\boldsymbol{W}(s)> = \mathbf{0}, \tag{2.5b}$$

$$<\boldsymbol{W}(s)\boldsymbol{W}^{\mathrm{T}}(t)> = \mathbf{I}\min(s, t), \tag{2.5c}$$

$$<\mathrm{d}\boldsymbol{W}(s)\mathrm{d}\boldsymbol{W}^{\mathrm{T}}(t)> = \mathbf{I}\,\delta(s - t). \tag{2.5d}$$

In Eq. (2.5), the symbol $\mathbf{I}$ denotes the identity matrix. The Wiener process is continuous, but is only differentiable in a generalized sense:

$$\mathrm{d}W_k = \xi_k \mathrm{d}t, \tag{2.6}$$

where "white noise" ξ_k is really a concept that can only be approximated in nature. As its name suggests, white noise has a flat spectrum and, therefore, infinite power. The Wiener process, on the other hand, has a power spectrum that decreases everywhere

with the square of the frequency. For more detailed descriptions of the properties of Wiener processes and white noise, we refer the interested reader to Arnold [1974], Chapter 3.

The white noise limit leading to Eq. (2.4) essentially represents the combined effect of the weakly correlated, rapid variations in the first term of Eq. (2.1) that occur within time Δt as the Gaussian variable in Eq. (2.4). This is why it is called the central limit theorem. For details and proof of the central limit theorem, we recommend, for example, the articles by Wong and Zakai [1965], Khasminskii [1966], and Papanicolaou and Kohler [1974]. Examples of geophysical applications may be found in Kohler and Papanicolaou [1977], Penland [1985], Majda et al. [1999], and Sardeshmukh, Penland and Newman [2001].

Let us examine the special case where the ith component of the rapidly varying term in Eq. (2.3) can be written as

$$G_i(\boldsymbol{x}, s/\varepsilon^2) = \sum_{k=1}^{K} G_{ik}(\boldsymbol{x}, s)\eta_k(s/\varepsilon^2), \tag{2.7}$$

where $\eta_k(s/\varepsilon^2)$ is a stationary, centered, and bounded random function. The integrated lagged covariance matrix of η is defined to be

$$C_{km} \equiv \int_0^\infty < \eta_k(t)\eta_m(t+t') > \mathrm{d}t', \quad k, m = 1, 2, \ldots, K, \tag{2.8}$$

where angle brackets denote expectation value. With these restrictions, the central limit theorem states that in the limit of long times ($t \to \infty$) and small $\varepsilon(\varepsilon \to 0)$, taken so that $s = \varepsilon^2 t$ remains fixed, the conditional probability density function (*pdf*) for $\boldsymbol{x}$ at time s given an initial condition $\boldsymbol{x}_o(s_o)$ satisfies the backward Kolmogorov equation (e.g., Horsthemke and Léfèver [1984], Bhattacharya and Waymire [1990])

$$\frac{\partial p(\boldsymbol{x}, s|\boldsymbol{x}_o, s_o)}{\partial s_o} = \mathcal{L} p(\boldsymbol{x}, s|\boldsymbol{x}_o, s_o), \tag{2.9}$$

where

$$\mathcal{L} = \sum_{i,j}^{N} a_{ij}(\boldsymbol{x}_o, s_o)\frac{\partial^2}{\partial x_{oi}\partial x_{oj}} + \sum_{i}^{N} b_i(\boldsymbol{x}_o, s_o)\frac{\partial}{\partial x_{oi}} \tag{2.10}$$

and

$$a_{ij}(\boldsymbol{x}, s) = \sum_{k,m=1}^{K} C_{km} G_{ik}(\boldsymbol{x}, s) G_{jm}(\boldsymbol{x}, s), \tag{2.11a}$$

$$b_i(\boldsymbol{x}, s) = \sum_{k,m=1}^{K} C_{km} \sum_{j=1}^{N} G_{jk}(\boldsymbol{x}, s)\frac{\partial G_{im}(\boldsymbol{x}, s)}{\partial x_j} + F_i(\boldsymbol{x}, s). \tag{2.11b}$$

In this limit, the conditional *pdf* also satisfies a forward Kolmogorov equation (called a "Fokker–Planck equation" in the scientific literature) in the scaled coordinates:

$$\frac{\partial p(\boldsymbol{x}, s|\boldsymbol{x}_o, s_o)}{\partial s} = \mathcal{L}^\dagger p(\boldsymbol{x}, s|\boldsymbol{x}_o, s_o), \tag{2.12}$$

where

$$\mathcal{L}^\dagger p = \sum_{i,j=1}^{N} \frac{\partial^2}{\partial x_i \partial x_j}\left[a_{ij}(\boldsymbol{x}, s)\, p\right] - \sum_{i}^{N} \frac{\partial}{\partial x_i}\left[b_i(\boldsymbol{x}, s)\, p\right]. \tag{2.13}$$

In short, the *moments* of $\boldsymbol{x}$ can be approximated by the moments of the solution to the "Stratonovich" SDE

$$\mathrm{d}x = \boldsymbol{F}(\boldsymbol{x}, s)\mathrm{d}s + \mathbf{G}(\boldsymbol{x}, s)\mathbf{S} \cdot \mathrm{d}\boldsymbol{W}. \tag{2.14}$$

In Eq. (2.14), $\mathbf{G}(\boldsymbol{x}, s)$ is the matrix whose (i, j)th element is $G^i_j(\boldsymbol{x}, s)$, and $\mathbf{S}$ is a matrix where the (k, m)th element of $\mathbf{S}\mathbf{S}^{\mathrm{T}}$ is C_{km}. Note that $\mathbf{S}$ is only unique up to its product with an arbitrary orthogonal matrix. Notice also that the usual factor of one-half found in most formulations of the Fokker–Planck equation has been absorbed into the definition of C_{km}.

Before proceeding, it may be well to discuss the physical systems described by Eq. (2.4). We have presented a sparse outline of how a dynamical system with two timescales, both finite and with one timescale much shorter than the other, may be described as a SDE we denoted with the name Stratonovich. This class of systems obeys the rules of classical Riemannian calculus and is appropriate for dynamical systems that we assume to be continuous from the outset, before we start thinking about treating the system as stochastic. That is, when constructing the calculus to describe the dynamics, the continuous limit is taken before the white noise limit is taken. Mathematically, these systems are described by integrals defined by limits of Riemann sums, where contributions from both the beginning and the end of each time interval are taken into account with equal weight. Details may be found in, for example, Kloeden and Platen [1992].

There is another class of systems, called "Ito systems," that arises when these limits are taken in opposite order. Physically, one has a discrete system at the smaller timescale, and the disrupting influence that is to be treated as white noise is uncorrelated between the discrete time steps. At a longer timescale, the discrete time steps are approximately continuous, and we construct a continuous calculus to handle the dynamical evolution. These Ito systems also involve Brownian motion, but they do not follow the classical Riemannian rules of calculus. In contrast to Stratonovich integrals, Ito integrals are defined by limits of Riemann sums involving contributions from only the beginning of each time interval.

The difference between Ito and Stratonovich calculus is important for scientists to understand because most of the physical phenomena they deal with are Stratonovich, while most mathematical references on stochastic numerical techniques are primarily interested in Ito schemes. As we shall see, there is a formal equivalence between Ito and Stratonovich descriptions of reality, so a theorem about an Ito process can generally be

carried over to the corresponding Stratonovich process. Unfortunately, the transformation from one description to the other can be prohibitively difficult.

Roughly speaking, the Wiener process (Eqs. (2.5) and (2.6)) wiggles around enough that infinitely many ways of integrating this function can be defined, each giving a different answer. The names "Ito" and "Stratonovich" label two of these calculi, each of which is a result of applying a limiting procedure to equations describing a physical system. Let us say we wish to evaluate the integral

$$I = \int_{t_0}^{T} W(t)\mathrm{d}W. \tag{2.15a}$$

Classically trained scientists might immediately write the solution to this equation as

$$I_S = [W^2(T) - W^2(t_0)]/2. \tag{2.15b}$$

However, this result comes from dividing the interval $(t_0, T]$ into n partitions

$$S_n = \sum_{i=1}^{n} W(\tau_i)[W(t_i) - W(t_{i-1})], \tag{2.16}$$

with $t_n = T$ and τ_i chosen arbitrarily in the interval $(t_{i-1}, t_i]$. As $n \to \infty$ and the partitions become progressively finer, S_n converges to I (Eq. (2.15a)). For regular deterministic processes, the answer does not depend on *where* τ_i is chosen in the interval $(t_{i-1}, t_i]$. However, for integrals over the Wiener process, it *does* matter (e.g., Arnold [1974]). The Stratonovich solution (Eq. (2.15b)) results as the mean square convergence of Eq. (2.16) when τ_i is chosen to be the midpoint between t_{i-1} and t_i. The Ito solution,

$$I_I = [W^2(T) - W^2(t_0)]/2 - (T - t_0)/2, \tag{2.17}$$

results as the mean square convergence of Eq. (2.16) when τ_i is chosen equal to the beginning of the interval t_{i-1}.

Which is the "correct" solution? Each can be, as can any of the other calculi corresponding to an infinite number of choices for where one chooses τ_i. Here, we concentrate on the two calculi that have been associated with naturally occurring phenomena: Ito and Stratonovich.

Looking at Eqs. (2.15b) and (2.17), one might note that there is an easy transformation between I_S and I_I. In fact, there is always a transformation between the solution to a Stratonovich SDE and an Ito SDE. The solution to the Stratonovich SDE (written in component form)

$$\mathrm{d}x_i = F_i(\boldsymbol{x}, t)\mathrm{d}t + \sum_{\alpha} G_{i\alpha}(\boldsymbol{x}, t) \cdot \mathrm{d}W_\alpha \tag{2.18a}$$

is equivalent to the Ito SDE

$$\mathrm{d}x_i = \left[F_i(\boldsymbol{x}, t) + \frac{1}{2} \sum_{\alpha j} G_{j\alpha}(\boldsymbol{x}, t) \frac{\partial G_{i\alpha}(\boldsymbol{x}, t)}{\partial x_j} \right] \mathrm{d}t + \sum_{\alpha} G_{i\alpha}(\boldsymbol{x}, t)\mathrm{d}W_\alpha. \tag{2.18b}$$

By "equivalent" it is meant that solving Eq. (2.18a) using Stratonovich calculus gives the same solution as solving Eq. (2.18b) using Ito calculus. That is, each equation evaluated for $\boldsymbol{x}$ using its appropriate calculus would describe an experimental outcome equally well as the other; the statistics of $\boldsymbol{x}$ in each case are the same. In Eq. (2.18), we have omitted the symbol "$\circ$" in keeping with standard mathematical notation of an Ito SDE.

As for Stratonovich systems, the transition *pdf* for an Ito system satisfies a Fokker–Planck equation (Eqs. (2.12) and (2.13)). However, the drift and diffusion terms, $\boldsymbol{b}$ and $\mathbf{a}$, are somewhat different. For the Ito SDE

$$\mathrm{d}\boldsymbol{x} = \boldsymbol{F}(\boldsymbol{x}, t)\mathrm{d}t + \mathbf{G}\mathrm{d}\boldsymbol{W}, \tag{2.19}$$

we have simply

$$\boldsymbol{b} = \boldsymbol{F}(\boldsymbol{x}, t) \tag{2.20a}$$

and

$$\mathbf{a} = \mathbf{G}\mathbf{G}^{\mathrm{T}}/2. \tag{2.20b}$$

For longer discussions on the difference between Ito and Stratonovich systems, we refer the reader to Arnold [1974], Horsthemke and Léfèver [1984], and Kloeden and Platen [1992]. In the following, we shall assume that the analytical work leading to an SDE has been done, the correct calculus has been identified, and that the scientist is ready to perform a numerical simulation of it.

3. The stochastic Taylor series and application to numerical schemes

3.1. Basic structure of the stochastic Taylor series

Stochastic Taylor series is an important theoretical and analytical tool in the study of SDEs and their numerical approximation. These are described in great detail in Kloeden and Platen [1992]; we repeat only their most salient properties here. To see how they are derived, we will first consider the derivation of a similar deterministic formula. For this, consider the following ordinary differential equation (ODE):

$$\frac{\mathrm{d}x}{\mathrm{d}t} = a(x). \tag{3.1}$$

We can then use the chain rule to write:

$$\frac{\mathrm{d}f(x)}{\mathrm{d}t} = f'(x)\frac{\mathrm{d}x}{\mathrm{d}t} = f'(x)a(x). \tag{3.2}$$

In Eq. (3.2) and in all equations of this section, prime denotes differentiation with respect to the argument. We can write all of this in a form analogous to that which we will use for the stochastic Taylor series, using the differential operator $L = a\dfrac{\mathrm{d}}{\mathrm{d}t}$, and we see that, if $\mathrm{d}x = a(x)\mathrm{d}t$, then

$$\mathrm{d}f(x) = Lf(x)\mathrm{d}t. \tag{3.3}$$

That is,

$$f(x(T)) = f(x(0)) + \int_0^T Lf(x(t))\mathrm{d}t. \tag{3.4}$$

If we apply this formula with Lf for f, we will have that

$$Lf(x(T)) = Lf(x(0)) + \int_0^T L^2 f(x(t))\mathrm{d}t. \tag{3.5}$$

We substitute this into the integral in the previous equation to arrive at

$$\begin{aligned} f(x(T)) &= f(x(0)) + \int_0^T Lf(x(0))\mathrm{d}t + \int_0^T \int_0^t L^2 f(x(s))\mathrm{d}s\mathrm{d}t, \\ &= f(x(0)) + Lf(x(0))T + \int_0^T \int_0^t L^2 f(x(s))\mathrm{d}s\mathrm{d}t. \end{aligned} \tag{3.6}$$

If we then use the change of variables formula (Eq. (3.4)) with $L^2 f$ for f, we get that

$$\begin{aligned} f(x(T)) &= f(x(0)) + Lf(x(0))T + \int_0^T \int_0^t L^2 f(x(0))\mathrm{d}s\mathrm{d}t \\ &\quad + \int_0^T \int_0^t \int_0^s L^3(x(u))\mathrm{d}u\mathrm{d}s\mathrm{d}t, \\ &= f(x(0)) + Lf(x(0))T + \frac{1}{2}L^2 f(x(0))T^2 + \int_0^T \int_0^t \int_0^s L^3(x(u))\mathrm{d}u\mathrm{d}s\mathrm{d}t. \end{aligned} \tag{3.7}$$

We can then continue this for as long as desired, yielding a Taylor series-like formula with an integral remainder term.

We will now derive stochastic Taylor series for the SDE

$$\mathrm{d}\boldsymbol{x} = \boldsymbol{F}(\boldsymbol{x}, t)\mathrm{d}t + \mathbf{G}(\boldsymbol{x}, t)(\cdot)\mathrm{d}\boldsymbol{W}, \tag{3.8}$$

where, as above, $\boldsymbol{x}$ and $\boldsymbol{F}$ are N-vectors, $\boldsymbol{W}$ is a K-dimensional Brownian motion, and $\mathbf{G}$ is an $N \times K$ matrix. The symbol "$(\cdot)$" is to be interpreted as "$\cdot$" if the Wiener process is to be integrated in the sense of Stratonovich, and is to be ignored if the noise is to be integrated in the sense of Ito. Instead of the chain rule, we have the Ito formula (also called the "stochastic chain rule:" GARDINER [1985], KLOEDEN and PLATEN [1992]) for any sufficiently smooth function $\boldsymbol{f}(\boldsymbol{x}, t)$:

$$\mathrm{d}\boldsymbol{f}(\boldsymbol{x}, t) = L^0\boldsymbol{f}(\boldsymbol{x}, t)\mathrm{d}t + \sum_{j=1}^{K} L^j\boldsymbol{f}(\boldsymbol{x}, t)\mathrm{d}\boldsymbol{W}_j(t). \tag{3.9}$$

Here, if Eq. (3.8) is an Ito equation,

$$L^0 = \frac{\partial}{\partial t} + \sum_{n=1}^{N} F_n \frac{\partial}{\partial x_n} + \frac{1}{2}\sum_{i,n=1}^{N}\sum_{j=1}^{K} G_{ij}G_{nj}\frac{\partial^2}{\partial x_i \partial x_n} \tag{3.10a}$$

and

$$L^j = \sum_{n=1}^{N} G_{nj} \frac{\partial}{\partial x_n}. \tag{3.10b}$$

Note that Eq. (3.10b) operated on the ith component f_i of $\boldsymbol{f}$ will be the ijth component of an $N \times K$ matrix. If the equation is Stratonovich, then Eq. (3.10b) is as in the Ito case, but

$$L^0 = \frac{\partial}{\partial t} + \sum_{n=1}^{N} F_n \frac{\partial}{\partial x_n}. \tag{3.10c}$$

We will now rewrite the SDE in integral form:

$$\boldsymbol{x}(T) = \boldsymbol{x}(0) + \int_0^T \boldsymbol{F}(\boldsymbol{x}, t)\mathrm{d}t + \int_0^T \mathbf{G}(\boldsymbol{x}, t)(\cdot)\mathrm{d}\boldsymbol{W}(t), \tag{3.11}$$

and the integral form of the Ito formula:

$$f(x(T), T) = \boldsymbol{f}(\boldsymbol{x}(0), 0) + \int_0^T L^0 \boldsymbol{f}(\boldsymbol{x}(t), t)\mathrm{d}t + \sum_j \int_0^T L^j \boldsymbol{f}(\boldsymbol{x}(t), t)(\cdot)\mathrm{d}W^j(t). \tag{3.12}$$

Then, as in the deterministic case above, we can use one of $L^0\boldsymbol{f}$ or $L^j\boldsymbol{f}$ for $\boldsymbol{f}$ in the Ito formula and substitute it back in. Note that now we have a choice as to what substitution to make, and that every time we make such a substitution, we replace one integral with two iterated integrals, and that we iterate different differential operators (L^0 and the L^js) rather than just one. As such, the formulas can quickly become unwieldy, and it is useful to have a compact notation.

For this purpose, we introduce the multi-indices $\alpha = (j_1, j_2, \ldots, j_l)$. Each j_k is either 0 or a number from 1 to K. If $j_k = 0$, j_k refers to L^0 if it is a superscript of L, or to $\mathrm{d}t$ if it is a superscript of $\mathrm{d}\boldsymbol{W}$. In other words, $\mathrm{d}t = \mathrm{d}W^0(t)$ by convention. If j_k is a number from 1 to K, it refers to L^{j_k} or $\mathrm{d}W^{j_k}(t)$. We also have the iterated differential operators

$$L^\alpha = L^{j_1} L^{j_2} \ldots L^{j_l} \tag{3.13}$$

and the iterated integrals

$$I_\alpha[\boldsymbol{f}(\boldsymbol{x}, t)]_{T_1}^{T_2} \equiv \int_{T_1}^{T_2} \int_{T_1}^{t} \cdots \int_{T_1}^{s} \boldsymbol{f}(\boldsymbol{x}(u), u)(\cdot)\mathrm{d}W^{j_1}(u)\mathrm{d}W^{j_2}(s) \cdots \mathrm{d}W^{j_l}(t). \tag{3.14}$$

If the integrand $\boldsymbol{f}(\boldsymbol{x}, t)$ is omitted from the iterated integral, it is assumed to be 1. Estimations of I_α for Ito integrals are generally different from those for Stratonovich integrals; we shall introduce those approximations when we get to examples of different schemes. Meanwhile, with this notation in hand, we can write any stochastic Taylor series in the form

$$\boldsymbol{f}(\boldsymbol{x}(T), T) = \boldsymbol{f}(\boldsymbol{x}(0), 0) + \sum_{\alpha \in A} L^\alpha \boldsymbol{f}(\boldsymbol{x}(0), 0) I_\alpha + \sum_{\beta \in B} I_\beta[\boldsymbol{f}(\boldsymbol{x}, t)], \tag{3.15}$$

where the integrals go from 0 to T. Note that here, B contains those multi-indices β for which $L^{\beta}\boldsymbol{f}$ could be the next substitution for $\boldsymbol{f}$ in Ito's formula, and A contains the αs which must have come earlier. When we use Eq. (3.15) to develop a numerical scheme, the approximation itself comprises multi-indices belonging to A, and the truncation error comprises the multi-indices belonging to B.

Before going on, we will clarify this notation with a one-dimensional example; that is, both $\boldsymbol{F}$ and $\mathbf{G}$ are scalars F and G, implying that the arbitrary function f and the Brownian motion W are scalars as well. If we apply Ito's formula (Eq. (3.12)) twice, with $L^0 f$ and then $L^1 f$ for f, we arrive at the expansion

$$\begin{aligned} f(x(T),T) &= f(x(0),0) + \int_0^T L^0 f(x(0),0)\mathrm{d}t + \int_0^T L^1 f(x(0),0)(\cdot)\mathrm{d}W(t) \\ &+ \int_0^T \int_0^t L^0 L^0 f(x(s),s)\mathrm{d}s\mathrm{d}t + \int_0^T \int_0^t L^1 L^0 f(x(s),s)(\cdot)\mathrm{d}W(s)\mathrm{d}t \\ &+ \int_0^T \int_0^t L^0 L^1 f(x(s),s)\mathrm{d}s(\cdot)\mathrm{d}W(t) + \int_0^T \int_0^t L^1 L^1 f(x(s),s)(\cdot)\mathrm{d}W(s)\mathrm{d}W(t) + \cdots \end{aligned} \tag{3.16}$$

This is more compactly written as

$$\begin{aligned} f(x(T),T) = f(x(0),0) &+ I_{(0)} L^0 f(x(0),0) + I_{(1)} L^1 f(x(0),0) \\ &+ I_{(0,0)}\Big[L^{(0,0)} f(x(t),t)\Big]_0^T + I_{(1,0)}\Big[L^{(1,0)} f(x(t),t)\Big]_0^T \\ &+ I_{(0,1)}\Big[L^{(0,1)} f(x(t),t)\Big]_0^T + I_{(1,1)}\Big[L^{(1,1)} f(x(t),t)\Big]_0^T . \end{aligned} \tag{3.17}$$

In Eq. (3.17), $I_0 = T$ and $I_1 = (W(T) - W(0))$. If we approximate our integral with terms involving only the initial condition, I_0 and I_1, then $A = \{(0), (1)\}$ and $B = \{(0,0), (1,0), (0,1), (1,1)\}$.

3.2. Application to numerical schemes

To create a numerical scheme for an SDE, we can expand differences using the stochastic Taylor series and drop the high-order terms. Of course, we break up the integration interval $(0, T]$ into discrete time steps so that the ith time step consists of an expansion analogous to Eq. (3.17) with $\boldsymbol{f}(\boldsymbol{x}(0), 0)$ replaced by $\boldsymbol{f}(\boldsymbol{x}(t_{i-1}), t_{i-1})$, $\boldsymbol{f}(\boldsymbol{x}(T), T)$ replaced by $\boldsymbol{f}(\boldsymbol{x}(t_i), t_i)$, T replaced by the time step Δ, and $(\boldsymbol{W}(T) - \boldsymbol{W}(0))$ replaced by a vector $\Delta\boldsymbol{W}$ of Gaussian random variables with each component having *variance* Δ. For example, consider Eq. (3.17), replacing $f(x(t), t)$ with $x(t)$ and dropping the remainder terms. Since $L^0 x = F$ and $L^1 x = G$, we are left with the stochastic Euler scheme (revisited in Section 4):

$$x(t_{i+1}) = x(t_i) + F(x(t_i), t_i)\Delta + G(x(t_i), t_i)\Delta W(t_i) + \text{Remainder}. \tag{3.18}$$

In the so-called strong schemes (see below), the estimation of ΔW is z_i, a Gaussian random deviate with mean zero and variance Δ. That is, roughly speaking, the deterministic part of the equation is updated with the time step, and the stochastic part is updated with the square root of the time step.

As another example, consider Eq. (3.15) with $A = \{(0), (1), (1, 1)\}$ and $B = \{(0, 0), (1, 0), (0, 1), (0, 1, 1), (1, 1, 1)\}$. That is, we expand the term

$$\begin{aligned}\int_0^T \int_0^t L^1 L^1 f(x(s), s)(\cdot)\mathrm{d}W(s)\mathrm{d}W(t) &= \int_0^T \int_0^t L^1 L^1 f(x(0), 0)(\cdot)\mathrm{d}W(s)\mathrm{d}W(t) \\ &+ \int_0^T \int_0^t \int_0^s L^0 L^1 L^1 f(x(u), u)\mathrm{d}u(\cdot)\mathrm{d}W(s)\mathrm{d}W(t) \\ &+ \int_0^T \int_0^t \int_0^s L^1 L^1 L^1 f(x(u), u)(\cdot)\mathrm{d}W(u)\mathrm{d}W(s)\mathrm{d}W(t) \qquad (3.19)\end{aligned}$$

and again write in compact form

$$\begin{aligned}I_{(1,1)}\Big[L^{(1,1)} f(x(t), t)\Big]_0^T &= I_{(1,1)} L^{(1,1)} f(x(0), 0) \\ &+ I_{(0,1,1)}\Big[L^{(0,1,1)} f(x(t), t)\Big]_0^T + I_{(1,1,1)}\Big[L^{(1,1,1)} f(x(t), t)\Big]_0^T . \qquad (3.20)\end{aligned}$$

Again taking $f(x(t), t) = x(t)$

$$L^{(1,1)} x(t) = \boldsymbol{G}(x(t), t)\frac{\partial \boldsymbol{G}(x(t), t)}{\partial x} \qquad (3.21a)$$

and either the Ito expression (KLOEDEN and PLATEN [1992])

$$I_{(1,1)} = (\Delta W^2 - \Delta^2)/2 \quad \text{(Ito)} \qquad (3.21b)$$

or the Stratonovich expression

$$I_{(1,1)} = \Delta W^2/2 \ \text{(Stratonovich)}, \qquad (3.21c)$$

we arrive at the so-called Milsteyn scheme

$$\begin{aligned}x(t_{i+1}) &= x(t_i) + F(x(t_i), t_i)\Delta + \boldsymbol{G}(x(t_i), t_i) z_i \\ &+ \boldsymbol{G}(x(t_i), t_i)\frac{\partial \boldsymbol{G}(x(t_i), t_i)}{\partial x} I_{(1,1)} + \text{Remainder}. \qquad (3.22)\end{aligned}$$

In Eq. (3.22), $I_{(1,1)}$ is appropriately estimated by either Eq. (3.21b) or (3.21c), again with ΔW represented by z_i.

3.3. Order of approximation

To obtain the order of the approximation, we need to know what order various terms are. It turns out that this depends on the sense, i.e., strong or weak, in which we want to approximately solve the SDE. The so-called strong schemes approximate the entire path of the solution process in an L^2-sense (see below). Weak schemes, in contrast, yield paths whose *moments* approximate all of the moments of the SDE; the paths themselves, however, may not be pointwise solutions to the SDE.

We shall clarify strong solutions first. If $\boldsymbol{x}(t)$ is the actual solution process and $\boldsymbol{y}(t)$ is our strong numerical approximation, then

$$\sqrt{< \sup_{0\leq t\leq T} |\boldsymbol{x}(t) - \boldsymbol{y}(t)|^2 >} = \boldsymbol{O}(\Delta t^{\gamma}), \tag{3.23}$$

where γ is the (strong) order of convergence. The vertical bars denote an appropriate norm, the Euclidean norm, for example. Note that this is indeed a strong convergence condition, usually much stronger than is needed or even necessarily desired. Strong schemes will, on average, approximate the path of $\boldsymbol{x}(t)$ that corresponds to the generated realization of the noise $\boldsymbol{W}(t)$. If we have no particular reason to believe that we are approximating the realization of the unresolved chaotic process that occurs in nature (and we generally call it noise because we do not have this belief), the added complexity and computational expense of strong schemes over weak schemes (see below) may not be warranted.

If one desires a strong scheme, then the correct order of an increment involving the iterated integral I_α depends on whether the integral is Ito or Stratonovich. For Ito integrals, the order is the number of deterministic terms (i.e., $j_k = 0$) plus half the number of stochastic terms (i.e., j_k between 1 and K), *except* that increments that are entirely deterministic (i.e., all the j_ks are 0) have effective order one-half smaller than this. For example, an increment with $\alpha = (0, 0, 1, 1, 0)$ has (strong) order 4, with $\alpha = (1, 0, 0)$ has order 2.5, and with $\alpha = (0, 0, 0, 0, 0)$ has order 4.5. It is understood that the iterated stochastic integrals are approximated using expressions appropriate to the Ito scheme, and that strong schemes can be developed for every integer and half-integer order.

For a particular example, consider Eq. (3.18). The first term on the right-hand side is, of course, order 0. The next two terms, with $\alpha = (0)$ and $\alpha = (1)$, are each of strong order 0.5. Thus, Eq. (3.18) represents an Ito scheme of strong order 0.5. The three next higher terms shown in Eq. (3.17) possess multi-indices $\alpha = (0, 0)$, $\alpha = (0, 1)$, and $\alpha = (1, 0)$. They are therefore of strong order 1.5. The last term, with $\alpha = (1, 1)$, is of strong order 1.

By the same reasoning, the Milsteyn scheme (Eq. (3.22)) is of strong order 1. Its next higher order terms in the remainder are of strong order 2 and 1.5. Hence, the Milsteyn scheme is an algorithm of strong order 1.

Similar counting of order is valid for strong Stratonovich integrals. However, only whole integer orders are valid, so all the half-order integers get bumped up to the next whole-order integer order. Note that the stochastic Euler scheme is not valid for Stratonovich integrals.

It is worth pointing out that it can be difficult to generate the necessary stochastic increments I_α for strong schemes (KLOEDEN and PLATEN [1992]) while still maintaining the formal accuracy to which the scheme was developed. For a one-dimensional noise, for example, the increment $I_{(0,1,1)}$, which appears in second-order schemes, is difficult to generate, since it is not Gaussian and has a complicated joint distribution with smaller order terms. For a multi-dimensional noise, the first-order increment $I_{(j,k)}$ is already difficult to generate.

We now turn to weak schemes. For so-called weak schemes, instead of approximating paths of solutions, we merely attempt to approximate the moments of solutions well. That is, we try to get the statistics correct. So, if $\boldsymbol{x}(t)$ is the actual solution process and $\boldsymbol{y}(t)$ is our numerical approximation, then

$$\left| < \boldsymbol{x}(t)^p > - < \boldsymbol{y}(t)^p > \right| = \boldsymbol{O}(\Delta t^\gamma) \tag{3.24}$$

for every $p = 1, 2, \ldots$, and where γ is the (weak) order of convergence. Note that weak schemes approximate every moment of the solution well.

It is important to note that what is said below will be valid *only* for Ito equations, and *not* for Stratonovich equations. This is not a major difficulty for sufficiently simple systems where it is possible to convert a Stratonovich equation to an Ito one (see Eq. 2.18).

If a weak scheme is desired, then the correct order of an increment involving the iterated integral I_α is simply the length of α. That is, both deterministic and stochastic terms in the increment have effective order 1. For high-order schemes this represents a considerable savings over strong schemes since many fewer increments are necessary. Weak schemes can be generated for any integral order. Note that all strong schemes of some order are also weak schemes, generally of a higher order. For example, the Euler scheme (Eq. (3.18)) is a strong scheme of order 0.5 and also a weak scheme of order 1.

A further savings can be realized with weak schemes since it is not necessary to simulate fully the stochastic increments. It is only necessary to simulate the first few moments (depending on the order of the scheme) of the increments.

For example, for a first-order weak scheme, such as the Euler scheme, it is only necessary that the simulated increments agree with the true stochastic increments up to the third moments, and these only need to agree up to first order in Δt (KLOEDEN and PLATEN [1992]). So, instead of generating a Gaussian random variable with mean 0 and variance Δt for the increment $\Delta W(t)$, it is sufficient to generate a coin-flip and generate a random variable $\Delta \widehat{W}$ such that

$$P(\Delta \widehat{W} = \sqrt{\Delta t}) = P(\Delta \widehat{W} = -\sqrt{\Delta t}) = \frac{1}{2}. \tag{3.25}$$

Note that even though we only approximate this increment correctly up through the third moment (and of these, only the second moment is nonzero), the scheme still approximates all moments of the solution well.

Second-order weak schemes need to get the first five moments of the stochastic increments correct, third-order schemes the first seven moments, etc.

To generate ΔW for a second-order scheme with a one-dimensional noise, the random variable $\Delta\widehat{W}$ with

$$P(\Delta\widehat{W} = \pm\sqrt{3\Delta t}) = \frac{1}{6}, \quad P(\Delta\widehat{W} = 0) = \frac{2}{3} \tag{3.26}$$

suffices, and for the iterated integral $I_{(1,0)}$, we can use $\Delta Z = \frac{1}{2}\Delta\widehat{W}\Delta t$ (KLOEDEN and PLATEN [1992]). If the scheme has a multi-dimensional noise, we will also need to generate the increments $I_{(j,k)}$ for $j \neq k$. For this, we can independently generate $V_{j,k}$ for $j < k$ such that

$$P(V_{j,k} = \pm\Delta t) = \frac{1}{2} \tag{3.27}$$

and $V_{k,j} = -V_{j,k}$, and approximate $I_{(j,k)}$ by $\Delta\widehat{W}^j\Delta\widehat{W}^k + V_{j,k}$. Generating all of this for just one time-step is faster than generating even one Gaussian random variable.

What can we do if we want a weak scheme for a Stratonovich SDE? We cannot use stochastic Taylor series to generate a weak scheme from scratch for a Stratonovich system since the weak order of a Stratonovich increment is not straightforward to estimate. However, we do have two possible options. First, as suggested above, we may be able to use Eq. (2.18) to convert from the Stratonovich system to the equivalent Ito system, thereby allowing use of weak Ito schemes. The other possibility is to make use of the fact that every strong scheme is also a weak scheme of the same order or higher. Thus, we can simply use a strong Stratonovich scheme and use it weakly, even to the extent of employing the simplifications of the stochastic increments as indicated for Ito weak schemes above.

4. Popular methods

The following summary of popular numerical algorithms is devoted to numerically integrating SDEs of the form

$$\mathrm{d}\boldsymbol{x} = \boldsymbol{F}(\boldsymbol{x}, t)\mathrm{d}t + \mathbf{G}(\boldsymbol{x}, s)(\cdot)\mathrm{d}\boldsymbol{W}(s). \tag{4.1}$$

Many definitions introduced in the preceding sections are reproduced for the convenience of the reader. This is not intended to be an exhaustive list, nor do we weigh the relative advantages and disadvantages of the various schemes. Our motivation in listing the schemes we have chosen is twofold. 1) Scientists often wish to "stochastify" an existing numerical model of a deterministic process. Several of the schemes discussed below (e.g., Ewald–Témam and the stochastic Runge–Kutta schemes) have been shown to converge to appropriate stochastic integrals by minimally adjusting well-known deterministic integration schemes. 2) If a stochastic model is to be written from scratch, one may wish to use a scheme that was developed specifically for SDEs. We have included schemes that we, the authors, have found satisfy one or the other need. As before, the first subscript of the matrix $\mathbf{G}$ corresponds to a component of $\boldsymbol{x}$; the second subscript corresponds to a component of $\mathrm{d}\boldsymbol{W}$. It will be made clear if an algorithm can only support one interpretation. In all cases, time step is denoted Δ. The symbol z_α is used to denote a centered (i.e., zero-mean) Gaussian random variable with *variance* Δ. If one chooses

to estimate a weak Ito integral, then z_α will be replaced by a random variable sampled as described in Section 3. However, in what follows, we shall assume that Gaussian estimations for the increment of the Wiener process are being used. The subscript on z_α is used to indicate that this random variable is used as the αth component of a vector Wiener process. The double stochastic integral over first the αth, and then the βth component of the vector Wiener process is denoted $I_{(\alpha,\beta)}$, and it will be clear whether that symbol represents Ito or Stratonovich integration. Finally, the scheme introduced by EWALD and TÉMAM [2003, 2005] has become known as the Ewald–Témam scheme in the meteorological literature (e.g., HANSEN and PENLAND [2006, 2007]) and that is the convention we follow here.

4.1. Euler method

This method converges only to Ito calculus. The order of its convergence was described in Section 3. Given Eq. (4.1), the algorithm is

$$x_i(t+\Delta) = x_i(t) + F_i(\boldsymbol{x}, t)\Delta + \sum_\alpha \boldsymbol{G}_{i\alpha}(\boldsymbol{x}, t) z_\alpha(t). \tag{4.2}$$

One generates the vector $\boldsymbol{z}(t)$ once at each time step and uses that same vector in updating every component of $\boldsymbol{x}$.

4.2. Heun method

This method (MCSHANE [1974], RÜMELIN [1982]) *does* converge, but only to Stratonovich calculus. The order of this convergence is of strong order 0.5 and weak order 1. As in the Euler scheme, we generate a vector $\boldsymbol{z}(t)$ once at each time step. Here, we use the Euler predictor as an intermediate variable

$$x_i'(t+\Delta) = x_i(t) + F_i(\boldsymbol{x}, t)\Delta + \sum_\alpha \boldsymbol{G}_{i\alpha}(\boldsymbol{x}, t) z_\alpha(t), \tag{4.3}$$

and then update $\boldsymbol{x}$ as follows:

$$\begin{aligned} x_i(t+\Delta) = x_i(t) &+ \frac{1}{2}\{F_i(\boldsymbol{x}, t)\Delta + F_i(\boldsymbol{x}', t+\Delta)\}\Delta + \cdots \\ &\cdots + \frac{1}{2}\sum_\alpha \{\boldsymbol{G}_{i\alpha}(\boldsymbol{x}, t) + \boldsymbol{G}_{i\alpha}(\boldsymbol{x}', t+\Delta)\} z_\alpha(t). \end{aligned} \tag{4.4}$$

The Heun method is a special case of the stochastic Runge–Kutta methods.

4.3. Runge–Kutta methods

The general $(m+1)$th order stochastic Runge–Kutta scheme has been published in RÜMELIN [1982], where its convergence properties are discussed, and is also examined in KLOEDEN and PLATEN [1992]. The reader should be aware that when KLOEDEN and PLATEN [1992] state that the Heun scheme, a second-order Runge–Kutta algorithm, "does not converge," they mean that the scheme does not converge to an Ito solution.

RÜMELIN [1982] provides explicit criteria by which the $(m+1)$th-order Runge–Kutta scheme may be judged to converge to Ito or Stratonovich calculus.

HANSEN and PENLAND [2006] considered a special case of the fourth-order scheme since that is the order most commonly used by scientists using Runge–Kutta algorithms in deterministic modeling. They also used RÜMELIN's [1982] criterion to show that this scheme converges to Stratonovich calculus. The stochastic version of this algorithm involves several intermediate steps:

$$\boldsymbol{K}_o = \boldsymbol{F}(\boldsymbol{x}(t), t), \tag{4.5a}$$

$$\mathbf{M}_o = \mathbf{G}(\boldsymbol{x}(t), t), \tag{4.5b}$$

$$\boldsymbol{x}' = \boldsymbol{x}(t) + \frac{1}{2}\boldsymbol{K}_o\Delta + \frac{1}{2}\mathbf{M}_o\boldsymbol{z}(t), \tag{4.5c}$$

$$\boldsymbol{K}_1 = \boldsymbol{F}(\boldsymbol{x}', t + \frac{1}{2}\Delta), \tag{4.5d}$$

$$\mathbf{M}_1 = \mathbf{G}(\boldsymbol{x}', t + \frac{1}{2}\Delta), \tag{4.5e}$$

$$\boldsymbol{x}'' = \boldsymbol{x}(t) + \frac{1}{2}\boldsymbol{K}_1\Delta + \frac{1}{2}\mathbf{M}_1\boldsymbol{z}(t), \tag{4.5f}$$

$$\boldsymbol{K}_2 = \boldsymbol{F}(\boldsymbol{x}'', t + \frac{1}{2}\Delta), \tag{4.5g}$$

$$\mathbf{M}_2 = \mathbf{G}(\boldsymbol{x}'', t + \frac{1}{2}\Delta), \tag{4.5h}$$

$$\boldsymbol{x}''' = \boldsymbol{x}(t) + \boldsymbol{K}_2\Delta + \mathbf{M}_2\boldsymbol{z}(t), \tag{4.5i}$$

$$\boldsymbol{K}_3 = \boldsymbol{F}(\boldsymbol{x}''', t + \Delta), \tag{4.5j}$$

$$\mathbf{M}_3 = \mathbf{G}(\boldsymbol{x}''', t + \Delta), \tag{4.5k}$$

$$\boldsymbol{x}(t+\Delta) = \boldsymbol{x}(t) + \frac{1}{6}(\boldsymbol{K}_o + 2\boldsymbol{K}_1 + 2\boldsymbol{K}_2 + \boldsymbol{K}_3)\Delta + \frac{1}{6}(\mathbf{M}_o + 2\mathbf{M}_1 + 2\mathbf{M}_2 + \mathbf{M}_3)\boldsymbol{z}(t). \tag{4.5l}$$

As before, the same vector of random numbers $\boldsymbol{z}(t)$ is used during the entire updating process.

4.4. Milsteyn method: explicit version

This method has versions corresponding to either Ito or Stratonovich calculus. Its strong order of convergence is of order 1, as is its weak order. The updating expression is

$$x_i(t+\Delta) = x_i(t) + F_i(\boldsymbol{x}, t)\Delta + \sum_{\alpha} G_{i\alpha}(\boldsymbol{x}, t)z_a + \sum_{j,\alpha,\beta} G_{j\alpha}\frac{\partial G_{i\beta}}{\partial x_j} I_{(\alpha,\beta)}. \tag{4.6}$$

In Eq. (2.12), $I_{(\alpha,\beta)}$ represents the double integral

$$I_{(\alpha,\beta)} = \int_t^{t+\Delta} \mathrm{d}W_\beta(t') \int_t^{t'} \mathrm{d}W_\alpha(t''). \tag{4.7}$$

When $\alpha = \beta$, we estimate $I_{(\alpha,\alpha)}$ as

$$I_{(\alpha,\alpha)} = (z_\alpha^2 - \Delta)/2 \tag{4.8a}$$

for Ito calculus and

$$I_{(\alpha,\alpha)} = z_\alpha^2/2 \tag{4.8b}$$

for Stratonovich calculus. When $\alpha \neq \beta$, $I_{(\alpha,\beta)}$ is the same for both Ito and Stratonovich calculus and must be approximated (KLOEDEN and PLATEN [1992]). In *no* case should it be ignored. In order to achieve a strong convergence order of unity, one chooses an integer $n \geq C/\Delta$, with C some positive constant, and estimates $I_{(\alpha,\beta)}$ as follows:

$$\begin{aligned} I_{(\alpha,\beta)} &= z_\alpha z_\beta/2 + \sqrt{\rho_n}(w_{\alpha,n} z_\beta - w_{\beta,n} z_\alpha) \ldots \\ &\quad \ldots + \frac{1}{2\pi} \sum_{m=1}^{n} \frac{1}{m} (v_{\alpha,m}(\sqrt{2} z_\beta + r_{\beta,m}) - v_{\beta,m}(\sqrt{2} z_\alpha + r_{\alpha,m})), \end{aligned} \tag{4.9a}$$

where

$$\rho_n = \frac{1}{12} - \frac{1}{2\pi^2} \sum_{m=1}^{n} \frac{1}{m^2}, \tag{4.9b}$$

and where, similar to the components of $\boldsymbol{z}$, the quantities $w_{\alpha,n}, v_{\alpha,m}, r_{\alpha,m}, w_{\beta,n}, v_{\beta,m}$, and $r_{\beta,m}$ are all independent Gaussian random variables having mean zero and variance Δ. Each of these random variables is also independent from the components of $\boldsymbol{z}$. The accuracy of the estimation increases with n, but this estimation can be very expensive computationally. In a large class of special cases, one loses no accuracy by approximating

$$I_{(\alpha,\beta)} = z_\alpha z_\beta/2. \tag{4.10}$$

This class of special cases is known as "commutative noise" and obtains when

$$\sum_k G_{k\alpha}(\boldsymbol{x}, t) \frac{\partial G_{i\beta}(\boldsymbol{x}, t)}{\partial x_k} = \sum_k G_{k\beta}(\boldsymbol{x}, t) \frac{\partial G_{i\alpha}(\boldsymbol{x}, t)}{\partial x_k}. \tag{4.11}$$

If the commutativity relation Eq. (3.5) does not hold, one must ensure that the approximation of $I_{(\alpha,\beta)}$ yields sufficient accuracy. The simplicity of Eq. (4.10) usually makes checking the commutativity relation worthwhile.

If accuracy of order 1 is necessary, the commutativity relation does not hold, and one is trying to estimate an Ito integral, then one of the weak schemes discussed in Section 3 may be appropriate.

4.5. Milsteyn method: implicit version

In this and in all implicit algorithms, remember this principle: *Never, ever, put the stochastic term in the part to be inverted.* Depending on your random number generator, you may get lucky for a particular simulation, but the scheme will eventually either blow up or give a spurious contribution so large that it will contaminate the rest of the solution. A cautionary tale is provided when we give examples. That being said, the implicit Milsteyn scheme is recognizable from the explicit version. That is, one chooses the level α_i of implicitness for the ith component of $\boldsymbol{x}$ and uses the updating equation

$$x_i(t+\Delta) = x_i(t) + \{\alpha_i F_i(\boldsymbol{x}, t+\Delta) + (1-\alpha_i)F_i(\boldsymbol{x}, t)\}\Delta \ldots$$
$$\cdots + \sum_{\beta} G_{i\beta}(\boldsymbol{x}, t) z_\beta + \sum_{k\beta\gamma} G_{k\beta}(\boldsymbol{x}, t) \frac{\partial G_{i\gamma}(\boldsymbol{x}, t)}{\partial x_k} I_{(\beta,\gamma)}. \tag{4.12a}$$

In Eq. (4.12a), one estimates the multiple stochastic integral as

$$I_{(\alpha,\beta)} = (z_\alpha z_\beta - \delta_{\alpha,\beta}\Delta)/2 \qquad \text{(Ito)} \tag{4.12b}$$

$$I_{(\alpha,\beta)} = z_\alpha z_\beta/2 \qquad \text{(Stratonovich)} \tag{4.12c}$$

if the noise is commutative. Otherwise, one either employs the more accurate approximations described in the explicit scheme or chooses a weak version. The order of convergence for the implicit Milsteyn scheme is the same as that for the explicit Milsteyn scheme.

4.6. Platen method: explicit version

The classic MILSTEYN [1974, 1978] method requires knowledge of how $\mathbf{G}(\boldsymbol{x}, t)$ changes with each component of $\boldsymbol{x}$. For very complicated multiplicative stochastic terms, analytical expressions may be difficult or impossible to evaluate. In this case, there is a two-step process introduced by Platen (KLOEDEN and PLATEN [1992]). For each component of the noise, one evaluates an intermediate value using no random numbers,

$$x'_{i,\alpha}(t+\Delta) = x_i(t) + F_i(\boldsymbol{x}, t)\Delta + G_{i\alpha}(\boldsymbol{x}, t)\sqrt{\Delta}. \tag{4.13a}$$

One then updates $\boldsymbol{x}$ as follows:

$$x_i(t+\Delta) = x_i(t) + F_i(\boldsymbol{x}, t)\Delta + \sum_{\alpha} G_{i\alpha}(\boldsymbol{x}, t) z_\alpha$$
$$+ \frac{1}{\sqrt{\Delta}} \sum_{\alpha,\beta} \{G_{i\beta}(\boldsymbol{x}'_\alpha, t) - G_{i\beta}(\boldsymbol{x}, t)\} I_{(\alpha,\beta)}. \tag{4.13b}$$

As in the Milsteyn method, the difference between Ito and Stratonovich calculus is made in the estimation of $I_{(\alpha,\beta)}$.

4.7. Implicit strong Runge–Kutta scheme

KLOEDEN and PLATEN [1992] call this algorithm the implicit order 1.0 strong Runge–Kutta scheme, although it is not a simple heuristic adaptation of the deterministic scheme, because it is derived using a strong stochastic Taylor expansion and involves intermediate variables. Again, one chooses the level α_i of implicitness for the ith component of $\boldsymbol{x}$. The supporting values are evaluated as in Eq. (3.7), and the updating equation is

$$x_i(t+\Delta) = x_i(t) + \{\alpha_i F_i(\boldsymbol{x}(t+\Delta), t+\Delta) + (1-\alpha_i) F_i(\boldsymbol{x}, t)\}\Delta \ldots$$

$$\cdots + \sum_{\beta} G_{i\beta}(\boldsymbol{x}, t) z_\beta + \frac{1}{\sqrt{\Delta}} \sum_{\alpha,\beta} \{G_{i\beta}(\boldsymbol{x}'_\alpha, t) - G_{i\beta}(\boldsymbol{x}, t)\} I_{(\alpha,\beta)}. \tag{4.14}$$

4.8. Ewald–Témam: explicit scheme

The explicit and implicit versions of the Ewald–Téman scheme have both strong and weak convergence order of one. The explicit Ewald–Témam is a modification of the Milsteyn scheme where the explicit tendency is replaced with a discretization (EWALD and TÉMAM [2003, 2005]):

$$x_i(t+\Delta) = x_i(t) + F_i(\boldsymbol{x}, t)\Delta + \sum_{\alpha} G_{i\alpha}(\boldsymbol{x}, t) z_a \ldots$$

$$\cdots + \sum_{j,\alpha,\beta} G_{j\alpha}(\boldsymbol{x}, t) \frac{\{G_{i\beta}(\boldsymbol{x} + \varepsilon_j \sqrt{\Delta}\hat{\boldsymbol{e}}_j, t) - G_{i\beta}(\boldsymbol{x}, t)}{\varepsilon_j \sqrt{\Delta}} I_{(\alpha,\beta)}. \tag{4.15}$$

In Eq. (4.15), $\hat{\boldsymbol{e}}_j$ is a unit vector corresponding to the component x_j. The vector ε has components less than or equal to unity, in units of $\boldsymbol{x}/\sqrt{t}$, and allows the modeler to adjust the discretized derivatives to the problem at hand. The comments already made concerning the double stochastic integral also apply here.

4.9. Ewald–Témam: implicit scheme

This scheme was devised to accommodate the architecture of extant climate models (EWALD and TÉMAM [2003, 2005]), including barotropic vorticity models (e.g., SARDESHMUKH and HOSKINS [1988]) and full general circulation models (e.g., SAHA, NADIGA, THIAW, WANG, WANG, ZHANG, VAN DEN DOOL, PAN, MOORTHI, BEHRINGER, STOKES, PENA, LORD, WHITE, EBISUZAKI, PENG and XIE [2006]). As such, this scheme is obviously different from any of the previous schemes described in this article. Deterministic climate models usually integrate the state vector first using a leapfrog step, followed by an implicit step. To implement the stochastic analog of this procedure, we rewrite Eq. (4.2) as

$$\mathrm{d}\boldsymbol{x} = \boldsymbol{F}_1(\boldsymbol{x}, t)\mathrm{d}t + \boldsymbol{F}_2(\boldsymbol{x}, t)\mathrm{d}t + \mathbf{G}(\boldsymbol{x}, t)(\cdot)\mathrm{d}\boldsymbol{W}. \tag{4.16}$$

In Eq. (3.10), $\boldsymbol{F}_1(\boldsymbol{x}, t)$ is the explicit part of the model and $\boldsymbol{F}_2(\boldsymbol{x}, t)$ is the implicit part. The implicit leapfrog scheme of EWALD and TÉMAM [2003, 2005] is as follows:

$$\boldsymbol{x}'(t+2\Delta) = \boldsymbol{x}(t) + 2\boldsymbol{F}_1(\boldsymbol{x}(t+\Delta), t+\Delta)\Delta + \boldsymbol{M}(\boldsymbol{x}(t), t) + \boldsymbol{M}(\boldsymbol{x}(t+\Delta), t+\Delta), \tag{4.17a}$$

$$\boldsymbol{x}(t+2\Delta) = \boldsymbol{x}'(t+2\Delta) + 2\boldsymbol{F}_2(\boldsymbol{x}(t+2\Delta), t+2\Delta)\Delta. \tag{4.17b}$$

In the updating expressions Eq. (4.17a), the ith component of the vector $\boldsymbol{M}(\boldsymbol{x}(t), t)$ is

$$M_i(\boldsymbol{x}, t) = \sum_{j\alpha\beta} G_{j\alpha}(\boldsymbol{x}, t)\frac{\partial G_{i\beta}(\boldsymbol{x}, t)}{\partial x_j} I_{(\alpha,\beta)} + \sum_{\alpha} G_{i\alpha}(\boldsymbol{x}, t) z_\alpha. \tag{4.17c}$$

Again, we approximate the derivative in Eq. (4.17c) as

$$\frac{\partial G_{i\beta}(\boldsymbol{x}, t)}{\partial x_j} = \frac{G_{i\beta}(\boldsymbol{x} + \varepsilon_j\sqrt{\Delta}\hat{\boldsymbol{e}}_j, t) - G_{i\beta}(\boldsymbol{x}, t)}{\varepsilon_j\sqrt{\Delta}}, \tag{4.17d}$$

with all symbols defined as in the explicit case.

As written, this scheme is of strong order 1 for both Ito and Stratonovich systems. If a weak order 1 Ito scheme is sufficient, the $I_{(\alpha,\beta)}$ terms in Eq. (4.17c) may be neglected.

5. Random number generators

All of the stochastic numerical schemes require random deviates. The Gaussian deviates required by strong schemes may be obtained by applying the Box–Müller (PRESS et al. [1992]) technique to output from pseudorandom number generators, which provide numbers evenly distributed on the interval (0,1), and the non-Gaussian deviates required by weak schemes also employ output from pseudorandom generators. Many computer languages have intrinsic random number generators that claim to fill this need, and some of them are dangerous for our purposes. According to GARCÍA [2000], C++ has a random number generator in the <stdlib.h> that is *not* intended for scientific programming; it repeats the same sequence of numbers after 33,000 calls. For comparison, a computer model with a triangular truncation of T382 in the horizontal and 64 levels in the vertical has more than 5,500,000 grid points. Perhaps the best random number generator is called the Mersenne Twister (MATSUMOTO and NISHIMURA [1998]), which is freely available from the website *http://www.math.sci.hiroshima-u.ac.jp/∼m-mat/MT/emt.html*. The period of this generator is an enormous 10^{6000}. We are unaware of any problems associated with this random number generator and would be grateful if anyone who finds any were to inform us.

6. Frequently asked questions

Before we begin exploring some common mistakes, let us note that when the noise is purely additive, many of the more complicated schemes are equivalent to the stochastic Euler scheme. Some of the examples below use additive noise for simplicity; multiplicative noise only exacerbates the problems discussed in this section.

6.1. Why do we need the square roots in the noise terms?

Without treating the system as a legitimately generated Wiener process, we do not have dynamical consistency. Consider the simple Ornstein–Uhlenbeck (OU) process below:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -\gamma x + \eta\xi. \tag{6.1}$$

In Eq. (6.1), white noise is denoted ξ, and η is a constant. It is easy to show that the stationary probability density is a centered Gaussian with variance $< x^2 > = \eta^2/2\gamma$. With values of $\gamma = 0.05$ and $\eta = 1$, $< x^2 > = 10$. Now we use the Euler method to numerically integrate Eq. (3.2), employing the relation $\mathrm{d}W = \xi \mathrm{d}t$. We shall also treat Eq. (6.1) naively, by treating $\eta\xi$ as a discrete white noise with unit variance at every time step and using the deterministic Euler scheme $x(t + \Delta) = x(t) + (-\gamma x(t) + \eta\xi)\Delta$. Figure 6.1 shows estimates of $< x^2 >$ for 10,000 samples from a time series sampled every 2.4 time units, using time steps of 0.2, 0.4, 0.6, and 0.8 time units. It is clear that the variance of the naively generated system increases monotonically with the time step, even for time steps an order of magnitude smaller than the standard deviation of the system to be generated. In fact, perusal of the schemes indicates that the variance for this system integrated naively should increase linearly. The stochastic Euler system, on the other hand, gives a reasonably accurate estimate of the variance. The moral of this story, of course, is that throwing random numbers into a deterministic integration scheme will provide noise with a variance dependent upon the time step.

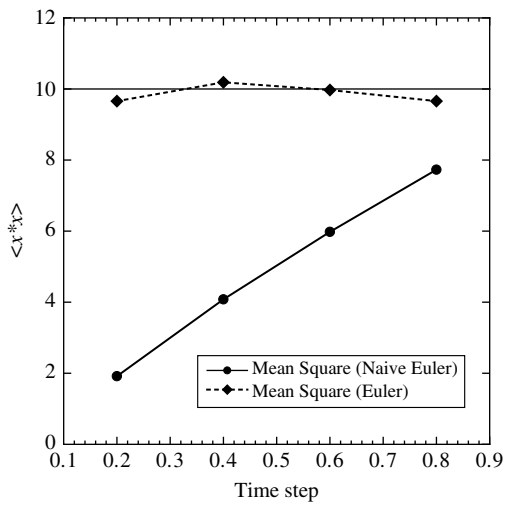


FIG. 6.1 Mean square of 10,000 samples of Eq. (6.1). Solid circles: Naive Euler scheme. Solid diamonds: stochastic Euler scheme, Eq. (4.1).

6.2. What happens if I use more than one random vector per time step?

There may be a temptation to hard-wire a random number generator into the code at every place the symbol z_α appears in the updating equation. If the system is multivariate, doing so is equivalent to using more than one random vector per time step. This is deadly if the correlations between variables are important. As an illustration, consider two damped oscillators coupled only through the driving noise:

$$\mathrm{d}\boldsymbol{x} = \mathbf{L}\boldsymbol{x}\mathrm{d}t + \mathbf{G}\mathrm{d}\boldsymbol{W}, \tag{6.2a}$$

where

$$\mathbf{L} = \begin{pmatrix} -0.5 & 1 & 0 & 0 \\ -1 & -0.5 & 0 & 0 \\ 0 & 0 & -0.5 & 1 \\ 0 & 0 & -1 & -0.5 \end{pmatrix} \tag{6.2b}$$

and

$$\mathbf{G} = \begin{pmatrix} \sqrt{2} & 0 & 0 & 0 \\ 0 & \sqrt{2} & 0 & 0 \\ 1/\sqrt{2} & 0 & \sqrt{1.5} & 0 \\ 0 & 1/\sqrt{2} & 0 & \sqrt{1.5} \end{pmatrix}. \tag{6.2c}$$

From the fluctuation–dissipation relationship (e.g., PENLAND and SARDESHMUKH [1995], NEWMAN, SARDESHMUKH and PENLAND [1997]), this system yields a covariance matrix

$$<\boldsymbol{x}\boldsymbol{x}^T> = \begin{pmatrix} 2 & 0 & 1 & 0 \\ 0 & 2 & 0 & 1 \\ 1 & 0 & 2 & 0 \\ 0 & 1 & 0 & 2 \end{pmatrix}. \tag{6.2d}$$

Generating the vector of random numbers before employing it in the time stepping procedure, as one ought to do, yielded the following sample covariance matrix:

$$\mathbf{C}_{\text{Sample}} = \begin{pmatrix} 1.979 & -0.006 & 0.993 & -0.029 \\ -0.006 & 1.990 & 0.016 & 1.001 \\ 0.993 & 0.016 & 2.024 & -0.013 \\ 0.002 & 1.001 & -0.013 & 2.010 \end{pmatrix}, \tag{6.2e}$$

whereas hard-wiring the random number generator into the updating equation yielded

$$\mathbf{C}_{\text{Hard-wire}} = \begin{pmatrix} 2.002 & 0.014 & -0.013 & 0.002 \\ 0.014 & 2.020 & 0.012 & -0.005 \\ -0.013 & 0.012 & 2.014 & 0.002 \\ 0.002 & -0.005 & 0.002 & 2.016 \end{pmatrix}. \tag{6.2f}$$

The vector of random numbers must be preserved during the entire course of the time step, or the correlations between state vector components are lost. This occurs with additive noise; the consequences of being cavalier with multiplicative noise can be dire indeed.

6.3. What happens if I use a naive implicit scheme?

This cautionary tale is reproduced from EWALD, PENLAND and TÉMAM [2004]. Consider the following forced diffusion equation:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = (k^2 - r)x + F, \tag{6.3a}$$

where $r = r_o + r_s\eta$, and where η is white noise. That is,

$$r\mathrm{d}t = r_o\mathrm{d}t + r_s\mathrm{d}W. \tag{6.3b}$$

An analytical solution for the stationary probability distribution function (*pdf*) is derived from the Fokker–Planck equation (HORSTHEMKE and LÉFÈVER [1984]) and is shown in Fig. 6.2 (heavy solid line) for the parameters $k = 0.1$, $r_o = 0.51$, $r_s = 0.5$, and $F = 0.5$. Also shown is the sample *pdf* from an integration of Eq. (6.3) using the implicit Ewald–Témam (filled circles). In this implementation, we have defined $a_1 = F$ and $a_2 = (k^2 - r_o)x$; further details may be found in EWALD, PENLAND and TÉMAM [2004]. For comparison is the sample *pdf* estimated from a time series generated using a naive application of an implicit scheme as follows:

$$x'(t + 2\Delta) = x(t) + 2a_1(x(t + \Delta), t + \Delta)\Delta, \tag{6.4a}$$

$$x(t + 2\Delta) = x(t) + 2a_2(x'(t + 2\Delta), t + 2\Delta)\Delta. \tag{6.4b}$$

Eq. (6.4) was used to integrate Eq. (6.3a) with η estimated as a Gaussian random variable of unit variance. This exercise shows the danger of throwing random numbers into a deterministic model and expecting to get useful results.

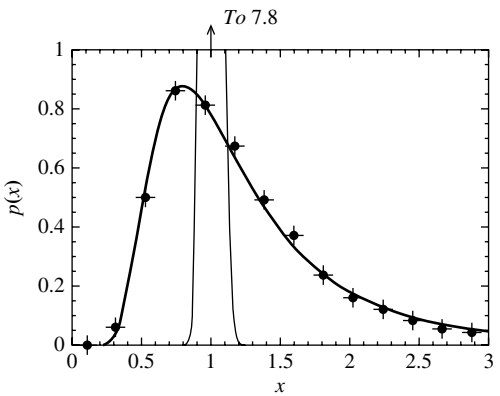


FIG. 6.2 Probability density function estimated from integrating Eq. (6.3). Heavy solid line: theoretically expected *pdf*. Filled circles with crosses: implicit Ewald–Témam. Light solid line: naive implicit scheme.

6.4. *How do I implement stochastic forcing in a dynamical model based on measured statistics?*

There is no single answer to this question. A stochastic parameterization cannot be implemented unless something is known about the timescales of the system to be parameterized. The common practice of simply augmenting a deterministic parameter in a model with an uncorrelated random component of specified, experimentally estimated, variance can cause most of the undesirable results discussed in this section (see, e.g., Fig. 6.2). Some of these effects can be ameliorated somewhat by using red noise, rather than white noise, but this is often extremely difficult in a high-dimensional, massively parallelized numerical model. Besides, using red noise rather than white noise does not guarantee the problems will go away. In the following, we consider two procedures that are appropriate for important special cases.

One solution to the quandary is possible when the timescale of the fast system is known at least to some approximation. In that case, the central limit theorem can be applied and the equivalent stochastic forcing identified. SARDESHMUKH, PENLAND and NEWMAN [2001], for example, applied this procedure to the linearized barotropic vorticity equation with stochastic fluctuations in the zonally symmetric mean velocity. After accounting for the effects of the annual cycle and El Niño, their four-times-daily measured velocities were found to have a spectrum varying roughly as the inverse square of the frequency. They, therefore, decided to approximate the fluctuations as an Ornstein–Uhlenbeck process with a timescale fast enough that the rest of the system would see those fluctuations as white. As is well known (e.g., HORSTHEMKE and LÉFÈVER [1984]) and as is readily verified by the formulas provided in Section 2, the white noise approximation of an Ornstein–Uhlenbeck process u with variance σ^2 and decay rate γ is

$$u\mathrm{d}t \approx (\sigma\sqrt{2/\gamma})\mathrm{d}W. \tag{6.5}$$

The larger σ and γ are, with σ^2/γ remaining constant, the better the approximation is.

It turned out that the decay time $\tau_s = 1/\gamma$ of the velocity fluctuations considered by SARDESHMUKH, PENLAND and NEWMAN [2001] was somewhat too large to present a legitimate candidate for a white noise parameterization (SARDESHMUKH, PENLAND and NEWMAN [2003]). However, when the decay time is small enough, Eq. (6.5) can be used to parameterize a physical quantity u stochastically.

A second procedure (HANSEN and PENLAND [2007]) can be used if the parameters of the stochasticity are diagnosed using data assimilation in, for example, a forecast model. That study was originally concerned with whether or not standard data assimilation procedures, such as various versions of the ensemble Kalman filter (ANDERSON [2001], BISHOP, ETHERTON and MAJUMDAT [2001], EVENSEN [1994], WHITAKER and HAMILL [2002]), could be applied to diagnosing the coefficients of random terms in a dynamical model. The answer was a reserved "yes," but the reservation had implications for stochastic integration.

The technique is best described using the example presented by HANSEN and PENLAND [2007], who considered a stochastic version of the chaotic Lorenz system:

$$\begin{aligned} \mathrm{d}x &= -a_0(x-y)\mathrm{d}t - a_s(x-y)\cdot \mathrm{d}W, \\ \mathrm{d}y &= (r_0 x - xz - y)\mathrm{d}t, \\ \mathrm{d}z &= (xy - b_0 z)\mathrm{d}t. \end{aligned} \tag{6.6}$$

In Eq. (6.6), the typical chaotic parameter values are $a_0 = 10$, $r_0 = 28$, and $b_0 = 8/3$. In addition, there is a stochastic component to be integrated in the sense of Stratonovich, with an addition parameter $a_s = 0.1$. Hansen and Penland [2007] integrated Eq. (6.6) using a stochastic fourth-order Runge–Kutta scheme (see Section 4c) using a time step (the details are important) of 0.00025 model units (mu). The resulting trajectory was taken as "truth." Employing the deterministic Lorenz model (i.e., Eq. (6.6) with $a_s = 0$) but augmented with the condition $\mathrm{d}a_0/\mathrm{d}t = 0$, as the assimilation model, the ensemble Kalman filter was used to assimilate the state vector (x, y, z, a_0). The deterministic Lorenz model was integrated using a standard fourth-order Runge–Kutta method with a time step of 0.01mu, and data from the "true" trajectory were assimilated every fifth time step, i.e., every $\tau_{obs} = 0.05$ mu. Data assimilation resulted in an estimate of $a_0 = 10.2$ with a standard deviation of that estimate being 0.42.

It is now that the central limit theorem (Section 2) becomes important. Our "truth" (Eq. 6.6) looks like Eq. (2.14). Our assimilation model looks like Eq. (2.1), with the x-component of $\varepsilon^2 \boldsymbol{F}$ identified as $a_0(x-y)$ and with $\varepsilon \boldsymbol{G}$ identified as $(x-y)$ times a centered (zero-mean) variable having a standard deviation of 0.42 in the same units as a_0. During the data assimilation, that variable is held fixed for $\tau_{obs} = 0.05$ mu after which another independent value is inserted into the assimilation model. Thus, the integrated lagged covariance (Eq. 2.8) of this variable is simply $(0.42)^2\tau_{obs}$, yielding a coefficient of the Wiener process equal to $(x-y)$ times the square root of that, or 0.098 $(x-y)$. This is not far from the true expression 0.1 $(x-y)$.

All this may be interesting, but the relation to stochastic integration is not obvious until one asks, "If the central limit theorem worked in the data assimilation scheme, why won't it work in forward integration?" In fact, it does. If we go to all the trouble to assimilate the stochastic parameter into a forecast model, the Stratonovich SDE is estimated to a remarkably accurate extent by a piecewise deterministic model, where the stochastic parameter is drawn from a Gaussian distribution having the *observed* mean and standard deviation (10.2 and 0.42, respectively, in the Hansen–Penland example). It is then injected into the forward integration, after which it is *held constant for exactly the assimilation period* ($\tau_{obs} = 0.05$ mu in this example) before it is replaced by another draw from the same distribution. Of course, for this to work, the assimilation period still has to be much smaller than the timescales of the dynamics the forecast model is trying to predict, or the central limit theorem is not satisfied.

7. Discussion

As the popularity of stochastic climate modeling increases, so does the temptation to do a quick and dirty job of it. The traditional identity of uncertainty as representing the extent to which a scientist "fails" to eliminate uncontrolled experimental variables causes

many scientists to harbor a certain discomfort with dynamically generated stochasticity, resulting in a very human denial of its importance. Once this denial can no longer be maintained, one may try to account for stochasticity in a numerical model while minimizing the effort needed to do so by simply adding a random number generator to the deterministic model. Unfortunately, it has been shown many times that this attitude can lead to wrong answers that look reasonable. We have tried to summarize approaches that combine necessary rigor and ease.

Ease is nearly as important as rigor. The immense practical difficulties of implementing stochastic numerical techniques in state-of-the art general circulation models are real and should be approached with sympathy and respect. It is for this reason that we have included not only numerical algorithms commonly employed for their theoretical advantages but also those algorithms that are similar to those already employed by geoscientists in numerical models. For example, the dynamical core of many deterministic models in climate research consists of a semi-implicit leapfrog scheme; the implicit Ewald–Témam scheme is therefore ideally suited for such models.

In this chapter, we have tried to clarify the reasons for the existence of two calculi found in nature. We have summarized some of their basic properties, including the differences and connections between them as well as a rule of thumb for the use of each: if the system one wishes to model derives its stochasticity from unresolved continuous chaotic processes, Stratonovich calculus is appropriate. If the process being modeled possesses discrete, uncorrelated components that are treated as approximately continuous, the appropriate calculus is Ito (Horsthemke and Léfèver [1984]). Practically, the difference between these calculi is in the definition of the integral over a Wiener process (Brownian motion) and the numerical approximation of that integral.

Although it is possible for the reader to use the algorithms presented here without understanding the basic theory behind them, we cannot recommend such a course. We have therefore presented a review of the stochastic Taylor series and its role in developing the numerical algorithms for integrating SDEs. In particular, we have noted that one need use strong schemes only if the actual trajectory of the system through phase space is required. If one is interested only in the statistics of the system, and this is usually the case, one may employ the more efficient weak schemes. As we have seen, a strong scheme of one order of convergence is equivalent to a weak scheme of at least that order. Sometimes, order of convergence is not the primary consideration in choosing a scheme; however, the coefficient in front of the order estimate can also make a difference. For this reason, if at all possible, we encourage modelers to experiment with different schemes on simple systems for which the analytic solution is known.

Finally, we urge the readers to refrain from taking "short cuts" since most of the theorems pertaining to deterministic numerical integrations do not apply to the stochastic case. The numerical integration of SDEs may be somewhat unfamiliar to traditional climate modelers, but stochastic schemes are well defined, straightforward, and increasingly common. For us, it is indeed welcome to see widespread advantage taken of these useful and elegant techniques.

References

ANDERSON, J.L. (2001). An ensemble adjustment Kalman filter for data assimilation. *Mon. Wea. Rev.* **129**, 2284–2903.

ARNOLD, L. (1974). *Stochastic Differential Equations: Theory and Applications* (John Wiley and Sons, New York), pp. 228.

BHATTACHARYA, R.N., WAYMIRE, E.C. (1990). *Stochastic Processes with Applications* (John Wiley and Sons, New York), pp. 672.

BISHOP, C.H., ETHERTON, B.J., MAJUMDAT S.J. (2001). Adaptive sampling with the ensemble transform Kalman filter. Part I: theoretical aspects. *Mon. Wea. Rev.* **129**, 420–436.

DOOB, J.L. (1953). *Stochastic Processes* (John Wiley and Sons, New York), pp. 654.

EVENSEN, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.* **99**, 10143–10162.

EWALD, B.D., TÉMAM, R. (2003). Analysis of stochsastic numerical schemes for the evolution equations of geophysics. *Appl. Math. Lett.* **16**, 1223–1229.

EWALD, B.D., TÉMAM, R. (2005). Numerical analysis of stochastic schemes in geophysics. *SIAM J. Numer. Anal.* **42**, 2257–2276.

EWALD, B., PENLAND, C., TÉMAM, R. (2004). Accurate integration of stochastic climate models. *Mon. Wea. Rev.* **132**, 154–164.

FELLER, W. (1966). *An Introduction to Probability Theory and its Applications, vol. 2* (John Wiley and Sons, New York), pp. 626.

GARCÍA, A.L. (2000). *Numerical Methods for Physics*, second ed. (Prentice Hall, Upper Saddle River, New York), pp. 423.

GARDINER, C.W. (1985) *Handbook of Stochastic Methods*, second ed. (Springer-Verlag, Berlin, Germany), pp. 442.

HANSEN, J., PENLAND, C. (2006). Efficient approximation techniques for integrating stochastic differential equations. *Mon. Wea. Rev.* in press.

HANSEN, J., PENLAND, C. (2007). On stochastic parameter estimation using data assimilation. *Physica D* in press.

HORSTHEMKE, W., LÉFÈVER, R. (1984). *Noise-Induced Transitions. Theory and Applications in Physics, Chemistry, and Biology* (Springer-Verlag, Berlin, Germany), pp. 318.

KHASMINSKII, R.Z. (1966). A limit theorem for solutions of differential equations with random right-hand side. *Theory Prob. Appl.* **11**, 390–406.

KLOEDEN, P.E., PLATEN, E. (1992). *Numerical Solution of Stochastic Differential Equations* (Springer-Verlag, Berlin, Germany), pp. 632.

KOHLER, W., PAPANICOLAOU, G.C. (1977). Wave propagation in a randomly inhomogeneous ocean. In: Keller, J.B., Papadakis, J.S. (eds.), *Chapter IV of Wave Propagation and Underwater Acoustics. Lecture Notes in Physics, vol. 70* (Springer Verlag, Berlin, Germany), pp. 153–223.

MAJDA, A.J., TIMOFEYEV, I., VANDEN EIJNDEN, E. (1999). Models for stochastic climate prediction. *Proc. Natl. Acad. Sci.* (USA), **96**, 14687–14691.

MATSUMOTO, M., NISHIMURA, T. (1998). Mersenne twister: a 623-dimensionally equidistributed uniform pseudorandom number generator. *ACM Trans. Model. Comput. Simul.* **8**, 3–30.

MCSHANE, E.J. (1974). *Stochastic Calculus and Stochastic Models* (Academic Press, New York).

Milsteyn, G.N. (1974). Approximate integration of stochastic differential equations. *Theor. Prob. Appl.* **19**, 557–562.

Milsteyn, G.N. (1978). A method of second-order accuracy integration of stochastic differential equations. *Theor. Prob. Appl.* **23**, 396–401.

Newman, M., Sardeshmukh, P.D., Penland, C. (1997). Stochastic forcing of the wintertime extratropical flow. *J. Atmos. Sci.* **54**, 435–455.

Papanicolaou, G.C., Kohler, W. (1974). Asymptotic theory of mixing stochastic differential equations. *Commun. Pure Appl. Math.* **27**, 641–668.

Penland, C. (1985). Acoustic normal mode propagation through a three-dimensional internal wave field. *J. Acoust. Soc. Am.* **78**, 1356–1365.

Penland, C., Sardeshmukh, P.D. (1995). The optimal growth of tropical sea surface temperature anomalies, *J. Climate*, **8**, 1999–2024.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P. (1992). Numerical Recipes in Fortran, *The Art of Scientific Computing*, second ed. (Cambridge University Press, Cambridge), pp. 963.

Rümelin, W. (1982). Numerical treatment of stochastic differential equations. *SIAM J. Numer. Anal.* **19**, 605–613.

Saha, S., Nadiga, S., Thiaw, C., Wang, J., Wang, W., Zhang, Q., van den Dool, H.M., Pan, H.-L., Moorthi, S., Behringer, D., Stokes, D., Pena, M., Lord, S., White, G., Ebisuzaki, W., Peng, P., Xie, P. (2006). The NCEP climate forecast system. *J. Clim.* **19**, 3483–3517.

Sardeshmukh, P.D., Hoskins, B.J. (1988). The generation of global rotational flow by steady idealized tropical divergence. *J. Atmos. Sci.* **45**, 1228–1251.

Sardeshmukh, P.D., Penland, C., Newman, M. (2001). Rossby waves in a fluctuating medium. In: Imkeller, P., von Storch, J.-S. (eds.), *Stochastic Climate Models.* In: Progress in Probability **49** (Birkhaueser Verlag, Basel, Switzerland).

Sardeshmukh, P.D., Penland, C., Newman, M. (2003). Drifts induced by multiplicative red noise with application to climate. *Europhys. Lett.* **63**, 498–504.

Whitaker, J.S., Hamill, T.M. (2002). Ensemble data assimilation withoug perturbed observations. *Mon. Wea. Rev.* **131**, 1485–1490.

Wilks, D.S. (1995). *Statistical Methods in the Atmospheric Sciences* (Academic Press, San Diego, CA), pp. 467.

Wong, E., Zakai, M. (1965). On the convergence of ordinary integrals to stochastic integrals. *Ann. Math. Stat.* **36**, 1560–1564.

Large-eddy Simulations for Geophysical Fluid Dynamics

Marcel Lesieur, Olivier Metais

*LEGI, BP 53, 38041 Grenoble-Cedex 9, France**

Abstract

We discuss the limits of turbulence direct-numerical simulations (DNS), and the need for large-eddy simulations (LES). We present spectral LES using subgrid models developed first in spectral space, then implemented in physical space. For a flow of constant density or weakly compressible, they are applied successively to inviscid isotropic turbulence, plane channels, and an obstacle with wall effect.

The second part of the chapter is more geophysical and reproduces with permission granted by Cambridge University Press the chapter on "Geophysical fluid dynamics" from the book *Large-eddy simulations of turbulence* by LESIEUR, MÉTAIS and COMTE [2005]. We first make a survey of flows encountered in Geophysics, with relevant climatic issues. Then the effect of spanwise rotation upon a plane channel is investigated with DNS and LES. For "moderate" (in terms of initial Rossby number based on the vorticity at the wall) rotation rates, one confirms the complete modification of the mean velocity profile into a linear profile over a large part of the channel, with a corresponding zero absolute mean vorticity. At "high" rotation rates, the channel flow becomes two-dimensional. Analogies with mixing layers and wakes submitted to spanwise rotation are also discussed, showing a universal character of rotating free- and wall-bounded shear flows. Finally, the development of a baroclinic jet in the atmosphere is studied. DNS and LES allow to show the formation of big quasi two-dimensional cyclonic vortices, with an intense vertical stretching of cyclonic vorticity in braids forming along the thermal fronts on the floor and the lid of the computational domain. Only LES did show secondary instabilities on the cold fronts of the braids. We discuss analogies with storm formation.

*Institut National Polytechnique de Grenoble, Université Joseph Fourier and CNRS. M. Lesieur is member of *Institut de France.*

Computational Methods for the Atmosphere and the Oceans
Special Volume (Roger M. Temam and Joseph J. Tribbia, Guest Editors) of
HANDBOOK OF NUMERICAL ANALYSIS, VOL. XIV
P.G. Ciarlet (Editor)

Copyright © 2009 Elsevier B.V.
All rights reserved

ISSN 1570-8659
DOI 10.1016/S1570-8659(08)00207-X

1. Large-eddy simulation an introductory survey

1.1. Limits of direct-numerical simulation in fluid turbulence

We work on the basis of fluid mechanics equations for a monophasic Newtonian flow (Navier-Stokes equations). The fluid may be compressible. Let us recall the notion of Kolmogorov scale $l_D = k_D^{-1}$, such that all motions whose spatial wavelength is smaller are damped by molecular viscosity. A direct-numerical simulation (DNS) is a deterministic solution of Navier-Stokes and the associated equations (total energy for instance): thanks to a proper projection on a spatiotemporal grid of partial-differential operators involved, one advances in time starting from an initial state, with prescribed conditions on the computational domain boundaries. The typical grid mesh Δx must be smaller than l_D. Notice also that numerical schemes have to be precise enough, and of the highest possible order. In developed turbulence of constant density, and if L is the large-scales typical size, the number of grid points necessary for a well-resolved DNS is $\approx (L/l_D)^3$. This is in fact proportional to a Reynolds number based on the so-called Taylor microscale raised to the $9/2$ power, which may be evaluated with the statistics of one velocity component. One finds $\approx 10^{15}$ points for a commercial-plane wing (see JIMENEZ [2000]), whose DNS will be possible within 30–50 years. In the atmospheric boundary layer of height 1000 m, the Kolmogorov scale is 10^{-3} m, which yields 10^{18} points. What about Jupiter, whose external gaseous circulation in the neighborhood of the great red spot is presented in Fig. 1.1? One is thus obliged to reduce considerably the number of system freedom degrees. As will be seen, Large-eddy simulations (LES) are a powerful tool for that.

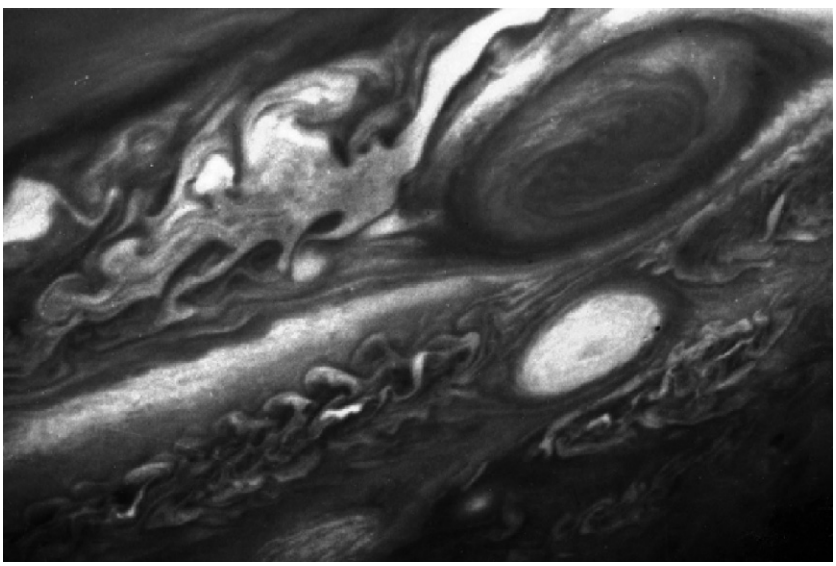


FIG. 1.1 Turbulence on Jupiter around the great red spot (picture courtesy of Jet Propulsion Laboratory, Pasadena). (See also color insert).

1.2. Incompressible LES formalism

For more details on these aspects, the reader is referred to LESIEUR, MÉTAIS and COMTE [2005], and LESIEUR [1997]. We assume that the fluid density is uniform and equal to ρ_0. Let $\Delta x > l_D$ be the spatial mesh size order of magnitude, and $G_{\Delta x}(\vec{x})$ a low-pass spatial filter of width Δx. This filter is chosen in order to eliminate subgrid scales of wavelength is $< \Delta x$. We introduce

$$\bar{f}(\vec{x}, t) = f * G_{\Delta x} = \int f(\vec{y}, t) G_{\Delta x}(\vec{x} - \vec{y}) \mathrm{d}\vec{y}. \tag{1.1}$$

An important remark is that the filter commutes with partial spatial and temporal derivatives (if Δx is uniform). We write now Navier-Stokes equations as

$$\frac{\partial u_i}{\partial t} + \frac{\partial}{\partial x_j}(u_i u_j) = -\frac{1}{\rho_0}\frac{\partial p}{\partial x_i} + \frac{\partial}{\partial x_j}(2\nu S_{ij}) \tag{1.2}$$

with

$$S_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right). \tag{1.3}$$

Filtered Navier-Stokes equations may be written as

$$\frac{\partial \bar{u}_i}{\partial t} + \frac{\partial}{\partial x_j}(\bar{u}_i \bar{u}_j) = -\frac{1}{\rho_0}\frac{\partial \bar{p}}{\partial x_i} + \frac{\partial}{\partial x_j}(2\nu \bar{S}_{ij} + \bar{u}_i \bar{u}_j - \overline{u_i u_j}), \tag{1.4}$$

where

$$T_{ij} = \bar{u}_i \bar{u}_j - \overline{u_i u_j} \tag{1.5}$$

is the subgrid-stresses tensor. In general, people make an eddy-viscosity assumption and write

$$T_{ij} = 2\nu_t(\vec{x}, t)\ \bar{S}_{ij} + (1/3) T_{ll}\ \delta_{ij} \tag{1.6}$$

so that Navier-Stokes LES equations write

$$\frac{\partial \bar{u}_i}{\partial t} + \frac{\partial}{\partial x_j}(\bar{u}_i \bar{u}_j) = -\frac{1}{\rho_0}\frac{\partial \bar{P}}{\partial x_i} + \frac{\partial}{\partial x_j}[2(\nu + \nu_t)\bar{S}_{ij}]. \tag{1.7}$$

There is a modified pressure called "macro-pressure":

$$\bar{P} = \bar{p} - (1/3)\rho_0 T_{ll}, \tag{1.8}$$

and the filtered continuity equation still holds (if the mesh is uniform)

$$\partial \bar{u}_j / \partial x_j = 0. \tag{1.9}$$

The mixing of a scalar (of molecular diffusivity κ_t) satisfying a Lagrangian Fourier heat equation may also be studied

$$\frac{\partial T}{\mathrm{d}t} + \frac{\partial}{\partial x_j}(T\ u_j) = \frac{\partial}{\partial x_j}\left\{\kappa \frac{\partial T}{\partial x_j}\right\}. \tag{1.10}$$

One makes the assumption of an eddy-diffusivity κ_t determined, thanks to a turbulent Prandtl number ν_t/κ_t. Thus, the scalar LES equation writes

$$\frac{\partial \bar{T}}{\mathrm{d}t} + \frac{\partial}{\partial x_j}(\bar{T}\,\bar{u}_j) = \frac{\partial}{\partial x_j}\left\{(\kappa + \kappa_t)\frac{\partial \bar{T}}{\partial x_j}\right\}. \tag{1.11}$$

In Smagorinsky's model [1963], the eddy viscosity is proportional to $(\Delta x)^2\sqrt{\bar{S}_{ij}\bar{S}_{ij}}$. There have been valuable improvements of this model by GERMANO, PIOMELLI, MOIN and CABOT [1991], with a local dynamic re-evaluation of the model "constant" through a double filtering. This Smagorinsky dynamic model has a proper behavior close to the wall in terms of expansions in powers of the distance to it.

1.3. Spectral LES

One of the problems associated with an eddy-viscosity model in physical space is the assumption of a spectral gap between resolved scales and subgrid scales, which is not realistic in general: even if a spectral gap is taken as initial conditions, nonlinear interactions will very quickly fill the gap. In this respect, the spectral eddy-viscosity concept is preferable. We write Navier-Stokes in Fourier space and introduce the spatial Fourier transform of the velocity $\hat{u}_i(\vec{k}, t)$ ($\vec{k}$ is the wavevector). Let k_C be a cutoff wavenumber. If Δx is characteristic of the grid mesh in physical space, subgrid scales $< \Delta x$ correspond to wavevectors of modulus $> k_C$. We consider a sharp filter which eliminates Fourier modes larger than k_C. In Navier-Stokes equations written in Fourier space, the nonlinear term corresponds to a convolution

$$\int_{\vec{p}+\vec{q}=\vec{k}} \hat{u}_j(\vec{p}, t)\hat{u}_m(\vec{q}, t)\mathrm{d}\vec{p}, \tag{1.12}$$

projected in a plane perpendicular to $\vec{k}$ in order to eliminate the pressure. If one writes now the equations for the subgrid modes $k \leq k_C$, the convolution may be split into explicit transfer

$$\int_{|\vec{p}|,|\vec{q}|<k_C}^{\vec{p}+\vec{q}=\vec{k}} \hat{u}_j(\vec{p}, t)\hat{u}_m(\vec{q}, t)\mathrm{d}\vec{p} \tag{1.13}$$

and subgrid-transfer

$$\int_{|\vec{p}|\mathrm{or}|\vec{q}|>k_C}^{\vec{p}+\vec{q}=\vec{k}} \hat{u}_j(\vec{p}, t)\hat{u}_m(\vec{q}, t)\mathrm{d}\vec{p}. \tag{1.14}$$

The latter, after division by $-k^2\hat{u}_i(\vec{k})$, permits to form a spectral eddy viscosity. A spectral eddy diffusivity for a scalar T transported by the flow may be defined also. However, spectral eddy coefficients thus defined involve unknown subgrid-scale quantities. Following KRAICHNAN [1976] who did it first for the test-field model, we determine these eddy coefficients at an energetic level, writing the evolution equations for the kinetic-energy and passive-scalar spectra given by the Eddy-Damped Quasi-Normal Markovian (EDQNM) theory (ORSZAG [1970]), ANDRÉ and LESIEUR [1977], LESIEUR [1997].

Splitting the transfers across k_C in the same way as done above, one can calculate the spectral eddy viscosity and eddy diffusivity. If k_C lies within a $k^{-5/3}$ range, we have

$$\nu_t(k|k_C) = 0.441\ {C_K}^{-3/2} X(k/k_C) \left[\frac{E(k_C)}{k_C}\right]^{1/2}, \tag{1.15}$$

where $E(k_C)$ is the kinetic-energy spectrum at k_C, and $X(k/k_C)$ a non-dimensional function equal to 1 up to $k/k_C \approx 1/3$ and sharply rising above ("plateau-peak" behavior, see CHOLLET and LESIEUR [1981]). The $[E(k_C)/k_C]^{1/2}$ "à-la-Heisenberg" scaling[1] of the eddy viscosity turns out to be essential for LES applications. We have also determined the EDQNM eddy diffusivity. It has the plateau-peak behavior, and the turbulent Prandtl number is approximately constant ($\approx$0.6) in Fourier space. In fact, such a value is the highest one permitted by adjustments of the constants arising in the passive-scalar spectrum EDQNM equation (see LESIEUR [1997, pp. 259–260]).

A major drawback of the plateau-peak model for LES applied to non-isotropic flows is the assumption of a $k^{-5/3}$ kinetic-energy spectrum at k_C. The spectral-dynamic model is an EDQNM generalization to spectra $\propto k^{-m}$ at the cutoff, with the exponent m not necessarily equal to 5/3. It is found

$$\nu_t(k|k_C) = 0.31{C_K}^{-3/2} \sqrt{3-m}\frac{5-m}{m+1} X(k/k_C) \left[\frac{E(k_C)}{k_C}\right]^{1/2}, \tag{1.16}$$

with an assumption of zero eddy viscosity for $m > 3$. The exponent m is determined through the LES with the aid of least-squares fits of the kinetic-energy spectrum close to the cutoff.

1.3.1. Decaying isotropic turbulence

Isotropic turbulence is a good approximation for small-scale atmosphere and ocean. As an example of the capabilities of the spectral-dynamic model, we present on Fig. 1.2 a decaying "Euler" isotropic turbulence simulation. The flow has a zero molecular viscosity and extends in a periodic box. Pseudospectral methods are used, and the initial kinetic-energy peak is at $k_i = 4$. The picture is taken at the time where enstrophy is maximum and is going to decay. One can see spaghetti-shaped vortices, visualized thanks to positive isosurfaces of

$$Q = \frac{1}{2}(\Omega_{ij}\Omega_{ij} - S_{ij}S_{ij}) = \frac{1}{4}(\vec{\omega}^2 - 2S_{ij}S_{ij}) = \frac{1}{2\rho}\nabla^2 p, \tag{1.17}$$

where S_{ij} and Ω_{ij} are, respectively, the deformation and rotation components of the velocity-gradient tensor. The left and bottom sides of the figure display Q on the sides of the box. This Q-criterion of HUNT, WRAY and MOIN [1988], where one looks at regions of space where rotation dominates deformation, is very efficient to identify coherent vortices. We have checked for isotropic turbulence that high vorticity-modulus distributions give the same vortical patterns. It would be very nice if the latter were the footprint of a singularity arising within Euler equations themselves, as proposed by many

[1] In fact, what Heisenberg proposed was an eddy viscosity proportional to $[E(k)/k]^{1/2}$ in a model equation for the kinetic-energy spectrum $E(k)$.

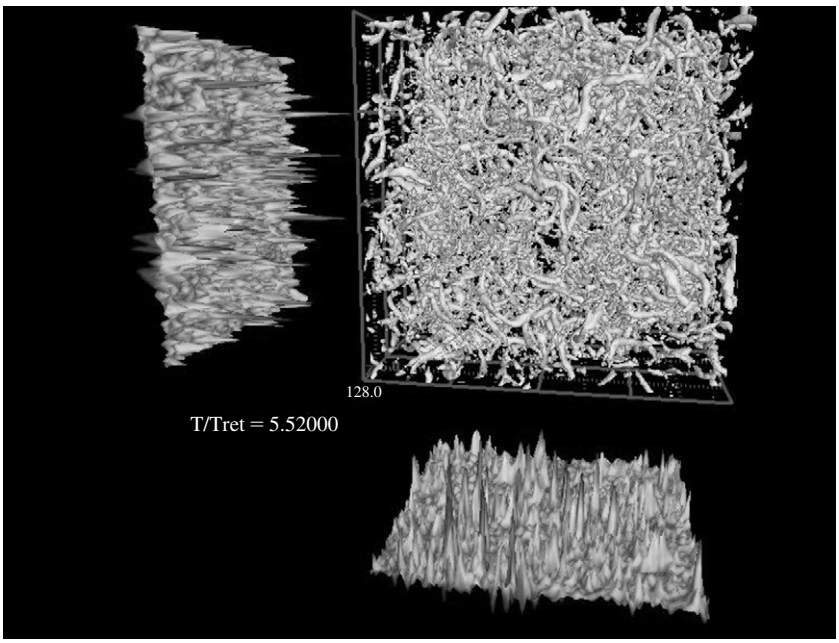


FIG. 1.2 Isotropic decaying turbulence. View of positive-Q isosurfaces (from LESIEUR, MÉTAIS and COMTE [2005], courtesy of Cambridge University Press). (See also color insert).

authors since the pioneering work of LERAY [1934]. The latter concerns Navier-Stokes, but LERAY ([1989], private communication) thought it should be understood in the sense of vanishing molecular viscosity.

1.3.2. Nonrotating plane channel

Another application of the spectral-dynamic model will be provided now in the plane Poiseuille flow case.

We consider a periodic (in the streamwise and spanwise directions) plane channel of constant density. It is well known that such a channel provides also a good example of a boundary layer above a flat plate. We use a mixed spectral-compact code, the compact scheme (of sixth order away from the walls) being employed in the transverse direction, while pseudo-spectral methods are used in the longitudinal and spanwise directions. Calculations carried out by LAMBALLAIS MÉTAIS and LESIEUR [1997] start with a parabolic Poiseuille velocity profile, to which a small three-dimensional (3D) white-noise perturbation is superposed, and are run up to complete statistical stationarity, assuming a constant bulk velocity U_m across the section. The width of the channel is $2h$, and the kinetic-energy spectrum allowing to determine the eddy-viscosity is calculated at each time-step with the aid of an average in planes parallel to the walls. One introduces wall units (v_* for velocity and ν/v_* for length).

We present a LES at $R_e = 2hU_m/\nu = 14000$ ($h^+ = 389$). There is a grid refinement close to the wall, in order to simulate accurately the viscous sublayer (first point at

$y^+ = 1$). The LES is compared statistically against a DNS at $h^+ = 395$ carried out by Kim (see ANTONIA, TEITEL, KIM and BROWNE [1992]). Figure 1.3 shows the mean velocity and root-mean-square (rms) velocity components, which agree well with the DNS. Here, the LES goes 70 times faster than the DNS, which is huge. In fact, our LES are in the continuation of DEARDORFF [1970] pioneering LES using Smagorinsky model with $C_S = 0.1$ and a wall law, and those of GERMANO, PIOMELLI, MOIN and CABOT [1991] using Smagorinsky's dynamic model.

1.4. Return to physical space

In practical applications, spectral methods may be too difficult to be used, and have to be replaced by finite-volume, finite-difference, or finite-elements methods. We have then defined in physical space eddy coefficients based on the same philosophy as spectral methods. In fact, the peak part of the spectral eddy viscosity can be formulated in physical space in the form of a hyperviscosity (LESIEUR and MÉTAIS [1996]), and this will be done later for the baroclinic jet simulation where we use a combination of the regular structure-function model and a hyperviscosity (see also GARNIER, MÉTAIS and LESIEUR [1998]). We recall subgrid models of the structure-function family (see MÉTAIS and LESIEUR [1992], LESIEUR [1997], for details). We first take a constant (without peak) spectral eddy viscosity, the value of the plateau being determined by subgrid energy-conservation considerations in an infinite $k^{-5/3}$ spectrum. This gives

$$\nu_{SF}(\vec{x}, t) = \frac{2}{3}\, C_K{}^{-3/2} \left[\frac{E(k_C, \vec{x}, t))}{k_C}\right]^{1/2}, \tag{1.18}$$

with $k_C = \pi/\Delta x$, Δx being the grid mesh for a regular orthogonal grid. $E(k_C, \vec{x}, t)$ is now a local kinetic-energy spectrum in physical space, evaluated thanks to the local second-order velocity structure function of the filtered field $F_2(\vec{x}, \Delta x)$. It yields

$$\nu_t^{SF}(\vec{x}, \Delta x) = 0.105\, C_K^{-3/2}\, \Delta x\, [F_2(\vec{x}, \Delta x)]^{1/2}, \tag{1.19}$$

where F_2 is not calculated by an ensemble average, but locally, using a statistical average of square-velocity differences between $\vec{x}$ and the six closest points surrounding $\vec{x}$ on the computational grid. The structure-function model (SF) predicts a good Kolmogorov spectrum in isotropic turbulence (MÉTAIS and LESIEUR [1992]), and there are various arguments permitting to conclude that it is about 20% less dissipative than Smagorinsky's model in isotropic or free-shear flows (see LESIEUR, MÉTAIS and COMTE [2005]). It has the same drawbacks as Smagorinsky for wall flows, preventing turbulence issuing from small perturbations to develop. Two improved versions of the SF model have been developed, the selective structure-function model and the filtered structure-function model (see, e.g., LESIEUR, MÉTAIS and COMTE [2005]).

1.4.1. Weakly compressible flows

Our LES have been extended to compressible Navier-Stokes equations, put under a conservative flux form. Equations involve mass, momentum, and total energy. They are still filtered by the "bar-filter" already introduced. Details may be found in LESIEUR,

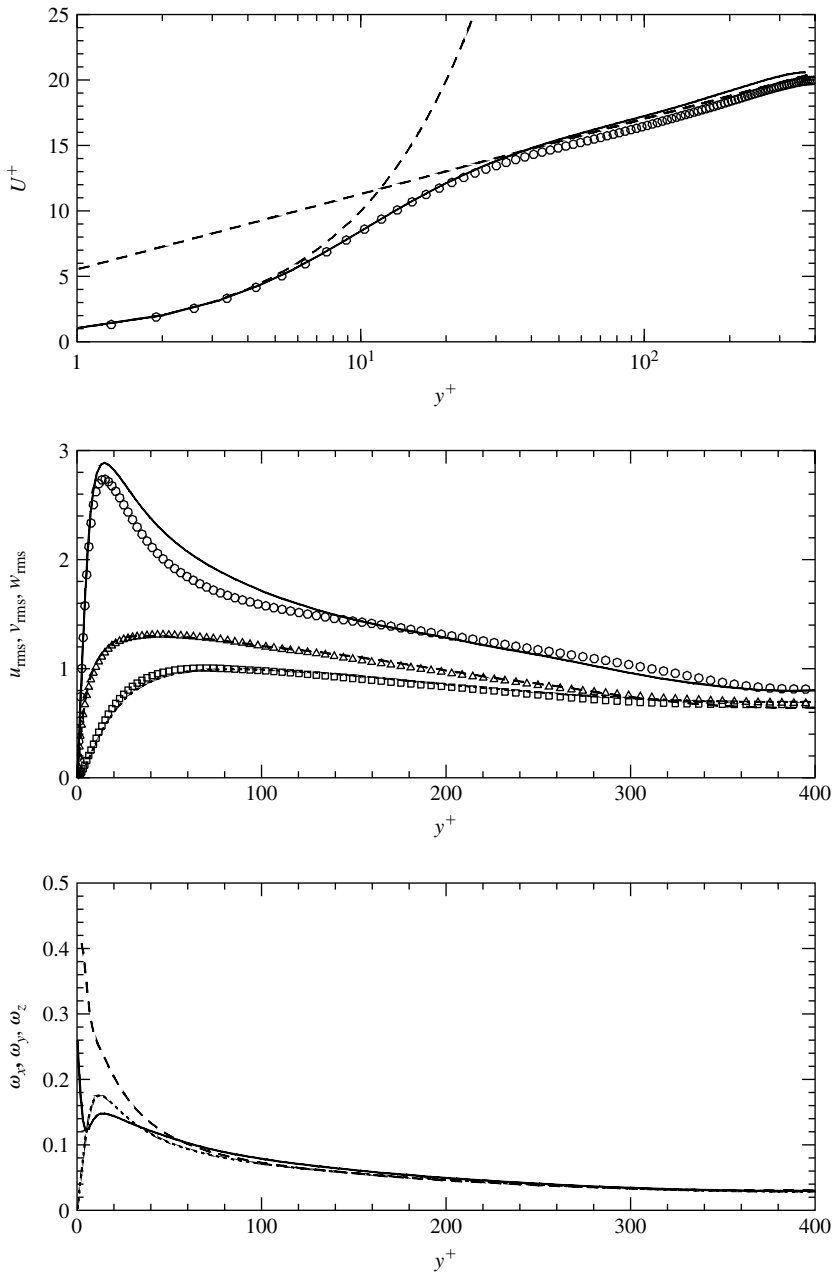


FIG. 1.3 Plane channel without rotation, spectral-dynamic model (straight lines, $h^+ = 389$) versus Kim's DNS (ANTONIA, TEITEL, KIM and BROWNE [1992]) (symbols, $h^+ = 395$); mean velocity (top), rms velocity components (bottom).

MÉTAIS and COMTE [2005]. The problem is greatly simplified by the introduction of a density-weighted Favre-filtering (FAVRE [1958]) called "tilde" such that

$$\overline{\rho f} = \bar{\rho}\tilde{f}, \tag{1.20}$$

as well as a "macro-temperature" related to the macro-pressure (already discussed) through the ideal-gases law.

We have developed a compressible Navier-Stokes solver called COMPRESS, using in space Mac Cormack finite-differences schemes of fourth-order for the nonlinear terms. We present here LES of low Mach number flows, which are in fact equivalent to incompressible LES.

The first one concerns a periodic channel with two spanwise tiny rectangular grooves on one of the walls (DUBIEF and DELCAYRE [2000]). The LES uses the filtered SF model. The Mach number is 0.3, and the wall Reynolds number $h^+ = 160$. The calculation's resolution is $200 \times 128 \times 64$. This low-Reynolds number simulation has been very well validated statistically by comparison with a DNS at same Reynolds. In this case, the DNS costs ten times more than the LES. One can see in Fig. 1.4 the structure of asymmetric hairpin-shaped coherent vortices propagating with the current. They consist in thin quasi-longitudinal vortices of length approximately 300 wall units and diameter 25. The vortices first creep along the ground, then rise at an angle of about 40°, due to self-induction. An animation provided in LESIEUR, MÉTAIS and COMTE [2005] shows how the vortices are transported by the flow. It is well recognized now that these vortices are

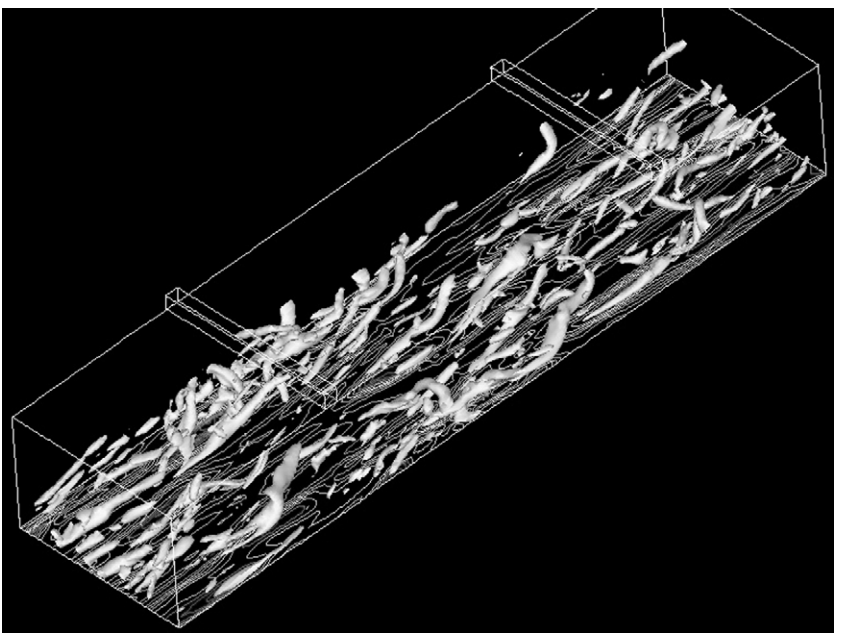


FIG. 1.4 $Q > 0$ isosurfaces in a weakly compressible channel (picture courtesy of Y. DUBIEF, from DUBIEF and DELCAYRE [2000]). (See also color insert).

associated with a system of longitudinal low- and high-speed velocity streaks extending from the wall to a distance of about 50 wall units, and whose spanwise wavelength is 100 wall units. An explanation for that is to say that if one considers the cross-section of a given longitudinal hairpin vortex, it will pump slow fluid from the wall through its foot, and send from its exterior fast fluid to the wall, inducing low-speed ascending (ejections) and high-speed descending (sweeps) in the flow. The same system of vortices and streaks is present close to each wall in the channel considered in Fig. 1.3. In fact, the peak at $y^+ \approx 15$ observed for the rms longitudinal velocity fluctuations shown in this figure corresponds to the maximum streak activity.

These boundary-layer vortices resemble very much analogous quasi-longitudinal "cigar-shaped" clouds oriented in the wind direction often observed in a neutral atmosphere. Indeed, vortices in adiabatic air are low-pressure and low-temperature regions where condensation is favored. The quasi-longitudinal vortices in turbulent boundary layers are asymmetric hairpins where one leg is favored, and this may be the reason why these clouds have generally one leg only.

The second low Mach number flow considered, discussed at length in LESIEUR, MÉTAIS and COMTE [2005], is an obstacle with wall effect. The Mach number is 0.2. We have a rectangular obstacle of thickness H, length 10 H, parallel to a flat plate and located at a distance of 0.2 H from it. The Reynolds number based on the incoming velocity U_0, H, and the molecular viscosity is 165000. Figure 1.5 shows a schematic

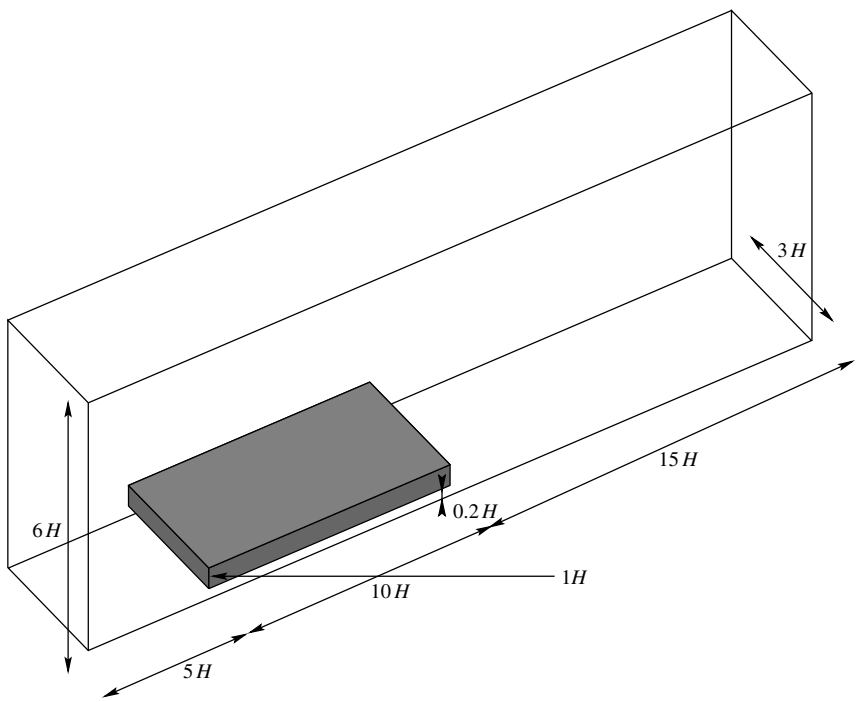


FIG. 1.5 Rectangular obstacle with wall effect (from LESIEUR, MÉTAIS and COMTE [2005], courtesy of Cambridge University Press).

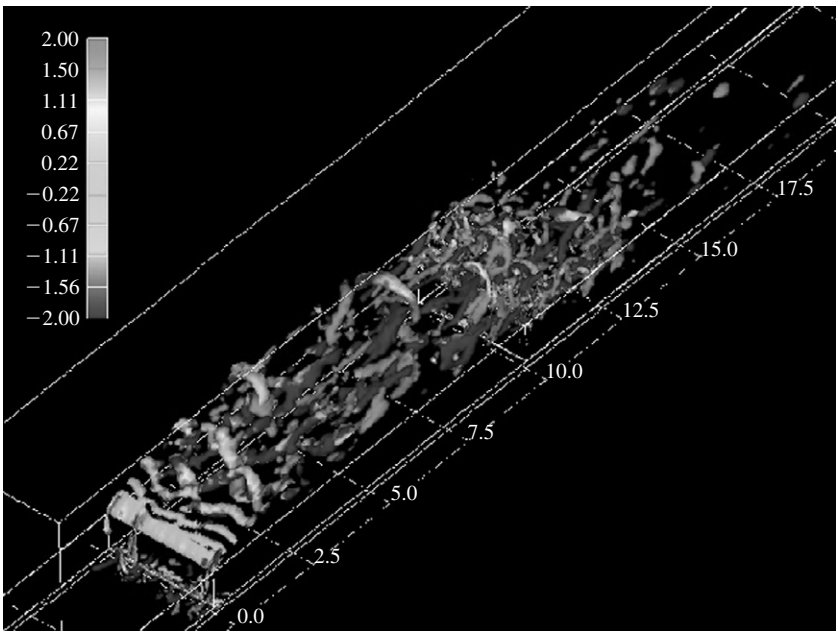


FIG. 1.6 Obstacle with wall effect. Positive Q isosurfaces with positive and negative longitudinal vorticity (from LESIEUR, MÉTAIS and COMTE [2005], courtesy of Cambridge University Press). (See also color insert).

view of the computational domain. The LES model is still the filtered SF model. The grid consists of 1,542,000 points, and the computational domain is split into four subdomains. Figure 1.6 displays the shedding behind the obstacle of large Λ-shaped vortices which impinge the lower wall downstream. In fact, the animation proposed in LESIEUR, MÉTAIS and COMTE [2005] indicates that this flow behaves upstream as an upward-facing step, with longitudinal hairpins swept first above the first ridge, then separating from the second ridge. A part of the current goes under the obstacle and mixes with the backward-facing step downstream. Applications of these calculations concern flows around vehicles and buildings.

2. Geophysical fluid dynamics

Foreword to the reader

The second part of the present article is based on the Chapter 8 (Geophysical fluid dynamics) taken from the book "Large-eddy simulations of turbulence," by MÉTAIS*, and* COMTE *(cambridge University Press, 2005). Our original manuscript of the chapter is reproduced with the kind authorization of Cambridge University Press. We had to incorporate some complementary explanations in order to make the section totally self-consistent. The reader is of course referred to the above book for a complete overview of large-eddy simulations in its fundamental, industrial, and geophysical aspects. The book contains also beautiful colored animations of a large class of LES computations.*

As stressed in LESIEUR, MÉTAIS and COMTE [2005], the LES concept was developed by meteorologists whose names where SMAGORINSKY [1963], LILLY [1962, 1966, 1967], and DEARDORFF [1970, 1972]. In fact, geophysical and astrophysical fluid dynamics contain an innumerable list of processes (generally 3D) which can be understood experimentally only, thanks to laboratory and in situ experiments, and numerically mostly by LES. We recall, for instance, that the Taylor microscale-based Reynolds number for small-scale atmospheric turbulence is larger than 10^4 so that DNS do not seem to be at hand in this case, even with the unprecedented development of computers.[2] As far as Earth is concerned, these processes are part of the extraordinary important issue constituted by climate modeling and prediction. It corresponds to a very complex system coupling dynamically and thermodynamically atmosphere (with water vapor, clouds, and hail), oceans (with salt and plankton), and ice, for periods of time going from seconds to hundreds of thousands of years. The question of global warming, which requires being able to predict the evolution, under the action of greenhouse-effect gases, of the Earth temperature (in the average or in certain particular zones) is vital for the survival of populations living close to the oceans and seas. Indeed, a global warming induces ice melt[3] which implies a sea-level elevation. It increases also evaporation, with as a result heavy rains and floods.

We first provide in this chapter a general introduction to Geophysical Fluid Dynamics (GFD). Then we will concentrate on two problems for which LES and DNS provide precious informations. The first is the effect of a fixed solid-body rotation upon a constant-density free-shear or wall-bounded flow. The second is the generation of storms through baroclinic instability in a dry atmosphere. The third problem which will not be discussed here is oceanic deep-water formation, essential link of the oceanic conveyor belt at the level of the Northern Atlantic, when the Gulf Stream water becomes saltier[4] and dives by gravity. LES related to the latter problem may be found in PADILLA-BARBOSA and MÉTAIS [2000].

2.1. Introduction to GFD

2.1.1. Rossby number

Definition We first consider a flow which rotates with a constant angular velocity $\Omega = f/2$. We work in the rotating frame. Let U be a characteristic relative velocity of the fluid, and L a characteristic scale of motion. We define the Rossby number as the ratio of characteristic inertial over Coriolis accelerations in Navier-Stokes equations, which yields

$$R_o = \frac{U}{fL}. \tag{2.1}$$

When $R_o \gg 1$, rotation is negligible. When $R_o \ll 1$, rotation dominates. In a rotating flow on a sphere at a latitude φ, the flow is approximately assumed to be equivalent to a

[2] Unless they become quantic, with binary informations carried out by atoms.
[3] Observed everywhere, from glaciers for many years and now to polar ice fields.
[4] Since a part of it is transformed into ice, which is fresh.

flow rotating around the local vertical axis with an angular velocity $\Omega \sin \varphi$ defined by the projection of the rotation vector $\vec{\Omega}$ on the local vertical. Then the Rossby number is still defined by Eq. (2.1), $f = 2\Omega \sin \varphi$ being the Coriolis parameter.

Here, one will take for U and L horizontal quantities.

2.1.2. Earth atmosphere

Large atmospheric scales They are called synoptic, with horizontal wavelengths of the order of or larger than several hundred kilometers. Since the effective thickness of the atmosphere is of the order of 15 km,[5] these motions are on a shallow layer and quasi two-dimensional (2D) on the Earth sphere. In medium latitudes, they correspond to quasi-horizontal vortices rotating around zones of high or low pressure, respectively, in the anticyclonic or cyclonic sense.[6] This circulation is driven by the geostrophic balance between pressure gradient and Coriolis force in the motion equations. Cyclonic vortices are more energetic than anticyclonic ones. The associated Rossby number ($U = 30\,\text{m/s}$, $L = 1000\,\text{km}$, $f = 10^{-4}$) is 0.3. The origin of these vortices comes from the baroclinic instability of easterly jet streams encountered by planes while flying at a 10 km elevation. The jet streams are due to the so-called thermal-wind balance, resulting both from geostrophic balance on the horizontal and hydrostatic balance between pressure gradient and gravity on the vertical (see LESIEUR [1997] for more details). The thermal-wind intensity is proportional to the horizontal north-south temperature gradient. Its regular intensity is of 30 m/s, but it may reach much higher values, such as the 110 m/s recorded in the days before the great storms on 26 and 28 December 1999 in Europe.

Tropical cyclones In lower latitudes, they correspond also to the production of cyclonic vortices, but the Rossby numbers are larger than in the former case. The main effect here is a huge evaporation of water at the sea surface if its temperature is above 26°. The water vapor thus produced rises due to thermal convection and condenses higher because of lower temperatures, the latent heat released being converted into horizontal kinetic energy which drives the system. With $U = 60\,\text{m/s}$, $L = 200\,\text{km}$, and $f \approx 10^{-4}\ \sin 23/\sin 45$, one gets $R_0 \approx 5$.

Hadley cells and trade winds The strong thermal convection above the ocean in the inter-tropical zone is responsible for the formation of two cells on both sides of the equator, called Hadley cells: warm air rises in the equatorial region, travels to higher latitudes where it cools, then goes down at the tropics level. While going down, it is deviated toward the west by Coriolis force and gives rise to trade winds. Weaker cells rotating in opposite sense are also observed, called Ferrel cells. This is in fact a year time-averaged view. Seasonal variations of Hadley cells lead to the monsoon phenomenon, where trade winds cross the equator and change sign, due to the opposite effect of Coriolis force. When these phenomena occur in a continent bordered to the west by an ocean, monsoon is accompanied by heavy rains.

[5] It corresponds to about 80% of the atmosphere mass. Above, air is more and more rarefied.

[6] Cyclonic here means the same sense of rotation as Earth, anticlockwise in the northern hemisphere and clockwise in the southern.

Ozone hole Let us mention the existence of Arctic and Antarctic circumpolar vortice. They are cyclonic and are, as the jet streams, driven by the thermal-wind balance. The north-south temperature gradient is very important at the level of poles because of the existence of ice. Therefore, the Antarctic circumpolar vortex is particularly intense. It is within this vortex that the seasonal phenomenon called "ozone hole" occurs, which is generally explained by the action of Chloro-Fluoro-Carbon (CFC) gases released by spray in medium latitudes. These CFC rise in the atmosphere up to the stratosphere (which is very stable from a thermal point of view, with a permanent inversion), then travel by quasi-horizontal turbulent diffusion everywhere, and particularly to the southern pole. For reasons still unclear, they can cross the border of the vortex (although it is very well marked), and penetrate inside. Ozone destruction occurs at the beginning of Austral Spring (end of September), through complex mechanisms which are far from being understood. This phenomenon is particularly marked since 1980. Due to the fact that ozone in the stratosphere protects us from ultraviolet radiation coming from the sun, drastic international measures have been taken to forbid the CFC production and use. However, the southern ozone hole does not show any tendency to disappear. There is also now a weaker Arctic ozone hole during spring.

Mesoscale and small-scale meteorology It involves wavelengths ranging from 10^{-3} m (the Kolmogorov scale) to several tenths of kilometer. These motions are strongly 3D. They are affected by thermal stratification and rotation sometimes. Thermal stratification may be stable in inversion zones or convectively unstable (in thunderstorms or tornadoes[7]). Durable inversion layers are often observed above cities located in troughs or surrounded by mountains, such as Grenoble, Los Angeles, or Mexico City, and are responsible for a strong industrial and car pollution. In this context, an interesting unstationary Reynolds-Averaged-Navier-Stokes (RANS) has been carried out by KENJERES and HANJALIC [2002]. Let us mention also the LES of FALLON, LESIEUR, DELCAYRE and GRAND [1997] using the selective structure-function model of a stably-stratified flow passing a straight backstep.

In an inversion situation, a wind crossing a mountain is going to give rise to internal gravity waves called lee waves, whose breaking up is an important source of turbulence. Since these waves propagate upwards, they are responsible for the so-called clear-air turbulence met by planes while passing above mountains at elevations of the order of 10 km. Finally, the atmosphere in contact with the ground gives rise to a turbulent boundary layer affected by rotation. It is a turbulent Ekman layer, whose typical height is 1 km, and may be again strongly felt while landing in a plane.

2.1.3. Oceanic circulation

Large-scale ocean Oceanic circulation in planetary scales is forced by winds. Trade winds, in particular, entrain surface water westwards. In the Northern Atlantic, for instance, the recirculation of the North-equatorial current within Mexico gulf is at the origin of the Gulf Stream. Since the dominant winds in higher latitudes and close to the

[7] Let us evaluate the Rossby number of a developed tornado in medium latitude: taking $U = 60$ m/s, $L = 1$ km, we have $R_o = 600$.

ocean are mainly westerly,[8] this yields the formation of a big anticyclonic recirculation cell characterized by a warm current rising on the western border of the basin, then crossing it up to the eastern border, and then going down to the equator. In the Northern Pacific, the equivalent of the Gulf Stream is the Kuroshivo, and the current going to the equator is the California current. The latter is characterized by upwellings of cold-deep water, due to the fact that Coriolis force deviates to the open sea; the warm surface water entrained by the wind which, by continuity, has to be replaced by deep water. Upwelling regions are also, because of the strong horizontal temperature gradients involved, submitted to baroclinic instability, with production of oceanic vortices. Let us mention in this respect the LES of TSENG and FERZIGER [2001], who look at the effect of coastal shape upon vortices generated by upwellings. The equivalent of California current in the Southern Pacific is the cold Humboldt current traveling to the equator along the Peru coast. Cold upwelling currents are extremely rich in fishes because they contain more oxygen. Upwelling exists also on western coasts of France, Portugal, and Africa, but the most famous of all is the Humboldt current, whose anomaly called El Nino is characterized by a reversal of the current, which becomes warm so that fishes disappear, with very bad consequences for the country economy. The quasi period of this unpredictable event is between 2 and 4 years. El Nino seems to be associated to a sort of nonlinear oscillation of the coupled system atmosphere/ocean, and has been observed for many centuries. The problem is that it seems to become more and more intense, may be as a consequence of global warming.

A few words again on the oceanic conveyor belt mentioned above. Right now, an important warming of the northern Atlantic Ocean is observed, with a worrying melting of the ice. This may reduce the deep-water formation, in such a way that the Gulf Stream would not dive anymore. On the other hand, we recall that the engine of the oceanic system is made of trade winds, which have no reason to be reduced by global warming, but rather amplified since they result from thermal convection. So Gulf Stream would certainly not disappear, but might, through continuity, have to find other routes than the present one.

Mesoscale and small-scale oceanography Oceanic currents are submitted to various instabilities responsible for the formation of eddies of scale $50 \sim 100$ km. Taking $U = 10$ cm/s, $L = 50$ km, oceanic vortices in medium latitudes have a Rossby number of 0.02, about ten times smaller than in the synoptic atmosphere. This is of the same order as Jupiter great red spot: here one takes $U = 100$ m/s, $L = 20000$ km, $\varphi = 45°$. Jupiter rotates faster than Earth, with an approximate period of 10 h. We have $f = 10^{-4} \times 24/10$, which gives $R_o \approx 0.02$, of same order as oceanic vortices (see SOMMÉRIA [2001]).

Since, as seen in LESIEUR [1997], the error in making the geostrophic-balance assumption is proportional to Rossby number, it shows that Earth oceanic or Jupiter vortices are closer to geostrophic balance than synotic atmosphere.

At smaller scales, the ocean may be the seat of intense 3D turbulence, due for instance to the breaking up of internal-gravity waves on the coast. Turbulent Ekman layers exist also close to the bottom and at the surface where the wind blows.

[8] Both because of Ferrel cells and jet streams.

2.1.4. Internal geophysics

Turbulence exists in the strongly heated liquid metal of Earth outer core. Here, the Rossby number may be calculated as follows in medium latitudes (CARDIN [2003], private communication). One takes $U = 10^{-3}$ m/s, $L = 1000$ km, which gives $R_o = 10^{-5}$. These flows are, because of their extreme slowness, the most rotation dominated of all those already considered up to now. They are electrically conductive and obey Magneto-Hydro-Dynamic (MHD) equations. Furthermore, they are submitted to a strong internal convection, whose result is certainly quasi 2D vortices of axis parallel to Earth axis of rotation.

2.2. *Effects of spanwise rotation on shear flows of constant density*

We study now with the aid of DNS or LES shear flows of uniform density rotating about a spanwise axis. The flow is assumed periodic in the streamwise x and spanwise z directions, $\bar{u}(y, t)$ being the longitudinal velocity at y averaged in the streamwise and spanwise directions. v and w are the velocity components along y and z. We recall that centrifugal forces are irrotational, and are therefore included into the pressure gradient (see LESIEUR [1997]). One defines now a local vorticity-based Rossby number

$$R_o(y, t) = -\frac{1}{2\Omega}\frac{\mathrm{d}\bar{u}}{\mathrm{d}y}, \tag{2.2}$$

as the ratio of the spanwise relative vorticity upon the entrainment vorticity. Regions with a positive or negative local Rossby will be called cyclonic or anticyclonic. We remind that the absolute vorticity vector $\vec{\omega}_a = \vec{\omega} + 2\Omega\vec{z}$ satisfies Helmholtz theorem in its conditions of applicability, which stresses that absolute-vortex elements follow the fluid parcels they contain. $R_o^{(i)}$ is the minimal value of $R_o(y, 0)$.

A pioneering linear-stability analysis of the problem in the inviscid case and for longitudinal modes (x-independent) was carried out by PEDLEY [1969]. Details are given in LESIEUR [1997], (p. 73), with analogies to centrifugal instabilities. PEDLEY [1969] shows that a necessary and sufficient condition for instability is that $R_o(y, 0) < -1$ somewhere in the flow.

2.2.1. Rotating channel

As in the nonrotating uniform-density case, we still consider constant flow-rate DNS or LES where we take as initial conditions a randomly perturbed parabolic profile. The axis $y = 0$ is the channel centerline. Thus, $R_o(y, 0)$ is linear and antisymmetric with respect to y. We choose the direction of $\vec{\Omega}$ such that the region of the channel $y > 0$ is initially cyclonic and $y < 0$ is anticyclonic. The minimum value $R_o^{(i)} = R_o(-h, 0)$ is always negative. LEZIUS and JOHNSTON [1976] have shown that such a flow is inviscidly unstable if $R_o^{(i)} < -1$. This is in fact equivalent in this case to PEDLEY [1969] result.

Figure 2.1, taken from LAMBALLAIS MÉTAIS and LESIEUR [1997], shows the distribution of the local Rossby number in DNS and LES as a function of $R_o^{(i)}$. The DNS is carried out at a Reynolds number based on the bulk velocity and $2h$ of 5000. The LES is done in the same conditions as the nonrotating LES using the spectral-dynamic model presented above at a Reynolds number of 14000. One sees that a $R_o(y) = -1$

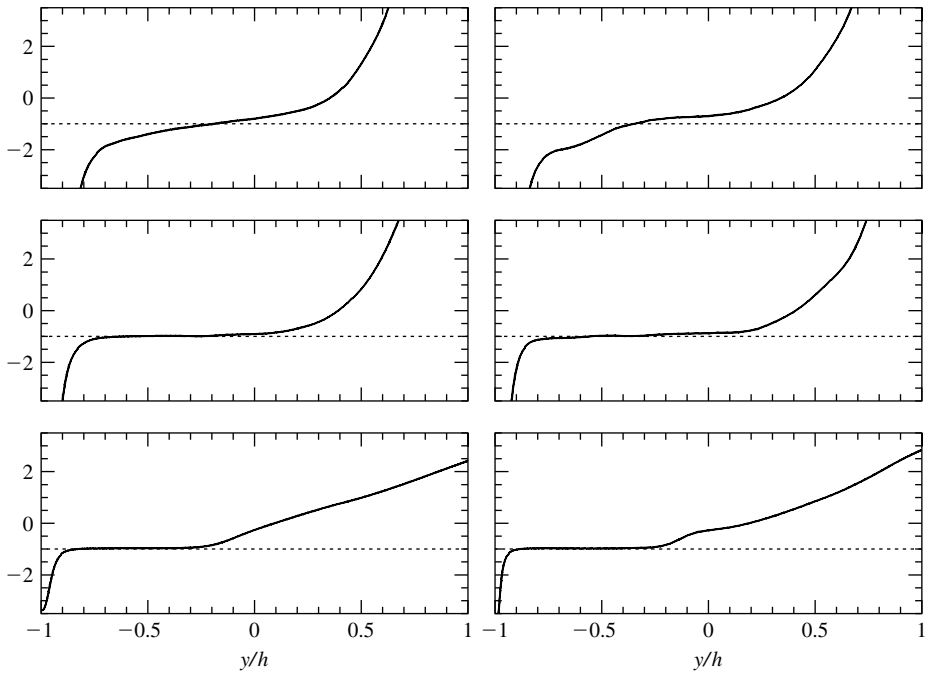


FIG. 2.1 Rotating channel of uniform density. Eventual local Rossby number distribution in the DNS (left) and LES (right). From top to bottom, minimal Rossby: −18, −6, and −2, from LAMBALLAIS MÉTAIS and LESIEUR [1997].

plateau forms for the three rotation rates in the LES case and the last two for the DNS. It yields

$$\frac{\mathrm{d}\bar{u}}{\mathrm{d}y} = 2\Omega, \tag{2.3}$$

which corresponds to a zero spanwise absolute vorticity. This result was shown experimentally by JOHNSTON, HALLEEN, and LEZIUS [1972]. Numerically, one should quote the DNS of KIM [1983], TAFTI and VANKA [1991], KRISTOFFERSEN and ANDERSSON [1993], NAKABAYASHI and KITOH [1996], and the LES of PIOMELLI and LIU [1995] using Smagorinsky's dynamic model. But the simulations of LAMBALLAIS LESIEUR and MÉTAIS [1997] investigate lower Rossby-number moduli (which means faster rotation rates) than above quoted authors. The DNSs of Lamballais are presented in LESIEUR [1997, pp. 432–433], with pictures of the vorticity modulus compared with the nonrotating case, and also the mean velocity profiles.

Lesieur notes "It is clear that the flow is quasi-laminar on the cyclonic side, while hairpins on the anticyclonic side are more and more inclined with respect to the wall as rotation is increased. It was also checked that longitudinal velocity fluctuations on this side are reduced when the Rossby number is increased, and that the corresponding streaks have disappeared at $R_o^{(i)} = -2$."

This order of magnitude of Rossby number (< -1) is what we call here "moderate" rotation. As stressed by LEZIUS and JOHNSTON [1976], there is a lower initial Rossby number (smaller than -1) under which the flow is stable again, and which decreases as Reynolds number is augmented.

If one increases the rotation rate,[9] there will be an interesting crossover at $R_o^{(i)} = -1$. Indeed, the case $R_o^{(i)} \geq -1$ ("fast" rotation) is totally different since it is stable and two-dimensionalizing from the point of view of the above instability. But LAMBALLAIS, LESIEUR and MÉTAIS [1996] have shown it may be subject to the growth of 2D Tollmien-Schlichting waves, as it is the case in particular for a DNS at $R_o^{(i)} = -0.1$: TS waves reach a nonlinear saturated state, and the flow is composed of purely 2D spanwise vortices of alternate-sign vorticity on the sides of the channel. The two rows of vortices on each wall are out-of-phase. This is analogous to a purely 2D solution studied by JIMENEZ [1990].

We remind finally that, for the channel, $R_o^{(i)}$ cannot exceed 0 by definition.

2.2.2. Rotating free-shear layers

Strong analogies exist with mixing layers and wakes rotating about a spanwise axis, where Pedley's analysis is also valid in the inviscid case. This was complemented by the viscous linear-instability studies of YANASE, FLORES, MÉTAIS and RILEY [1993]. They do show that Kelvin-Helmholtz instability is suppressed and replaced by the "Shear-Coriolis" instability, a purely longitudinal instability, if $R_o^{(i)}$ is, again, strictly lower than -1. As for the channel, there is a minimum lower bound for $R_o^{(i)}$ in this instability, which decreases as the Reynolds number increases. Iso-amplification rates for the mixing layer taken from this study are given in LESIEUR [1997, p. 424].

Rotating mixing layer For $R_o^{(i)} < -1$, and if the Reynolds number is high enough, the Shear-Coriolis instability is at hand. DNS and LES of FLORES [1993] and MÉTAIS, FLORES, YANASE, RILEY and LESIEUR [1995] show then in these conditions in the anticyclonic regions the stretching of intense purely longitudinal alternate vortices of absolute vorticity, which are such that the mean spanwise absolute vorticity becomes zero and the Rossby number equal to -1 in a large fraction of the region (see LESIEUR [1997, pp. 430–431]). This is clear in Fig. 2.2, showing the time evolution of the local Rossby number in a rotating mixing-layer DNS carried out by MÉTAIS, FLORES, YANASE, RILEY and LESIEUR [1995]. The Rossby number peaks initially, following the initial vorticity distribution. Then the amplitude of the peak decreases while the latter widens. At $t = 26.8$, a plateau close to -1 (slightly larger in fact) forms plateau, which is still there at $t = 35.7$.

Now we increase $R_o^{(i)}$. There is again a crossover $R_o^{(i)} = -1$, above which Kelvin-Helmholtz instability is at hand, with a strong two-dimensionalization: helical pairing is suppressed, as well as hairpin stretching between Kelvin-Helmholtz vortices. The structure of the mixing layer is a purely 2D mixing layer. Here, and contrary to the channel, $R_o^{(i)}$ may become positive. In this case, the mixing layer has become cyclonic,

[9] For the channel, $R_o^{(i)}$ is always negative so that increasing it corresponds to an increase in rotation rate.

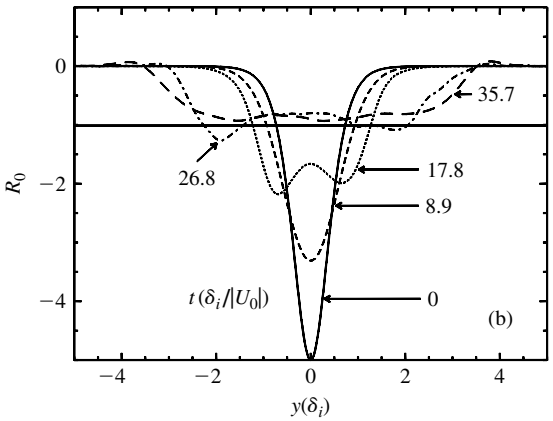


FIG. 2.2 Anticyclonic rotating mixing layer. Evolution with time of the local Rossby number in the DNS of (MÉTAIS, FLORES, YANASE, RILEY and LESIEUR [1995], picture courtesy of J. Fluid Mech.).

and increasing $R_o^{(i)}$ corresponds now to a reduction of the rotation rate. Our DNS and LES (see also LESIEUR, YANASE and MÉTAIS [1991]) do show that it remains very 2D up to $R_o^{(i)}$ of the order of 10. It is clear that if the Rossby number is increased and goes to infinity, 3D instabilities of the type found in the nonrotating case will develop again.

Rotating wake DNS and LES of FLORES [1993] and MÉTAIS, FLORES, YANASE, RILEY and LESIEUR [1995], as well as unpublished calculations done in Grenoble, permit to stress the following conclusions for the wake:

For $R_o^{(i)} < -1$, as in the mixing layer, and if the Reynolds number is high enough, the Shear-Coriolis instability is at hand. Since the wake has both cyclonic and anticyclonic sides, the Karman street is deeply modified. On the cyclonic side, one observes a 2D row of vortices without secondary hairpin-vortex stretching. On the anticyclonic side on the other hand, the anticyclonic Karman vortices do not exist anymore. They are replaced by the same longitudinal vortices as for the anticyclonic mixing layer at $R_o^{(i)} < -1$ discussed above, the local Rossby number becoming equal to -1 in this range. This is clear in Fig. 2.3, taken from a rotating-wake DNS at $R_o^{(i)} = -2.5$, and showing at times $t = 0$ and $t = 48.5$ the local Rossby number profiles. Here U_0 is the initial maximum deficit velocity and r_m a typical initial wake width. One sees how the initial antisymmetric Gaussian Rossby distribution is transformed, with the anticyclonic side a plateau slightly greater than -1, as in the anticyclonic mixing layer.

The crossover at $R_o^{(i)} = -1$ exists again. Above, the wake becomes a purely 2D Karman street, and there is no great difference in the wake structure between $R_o^{(i)} = -1$ and $R_o^{(i)} = -0.1$. As for the channel, $R_o^{(i)}$ cannot exceed zero.

Universality of free-shear layers These results demonstrate a very interesting universality of free or wall-bounded shear layers rotating about a spanwise axis as far as the three following points are concerned:

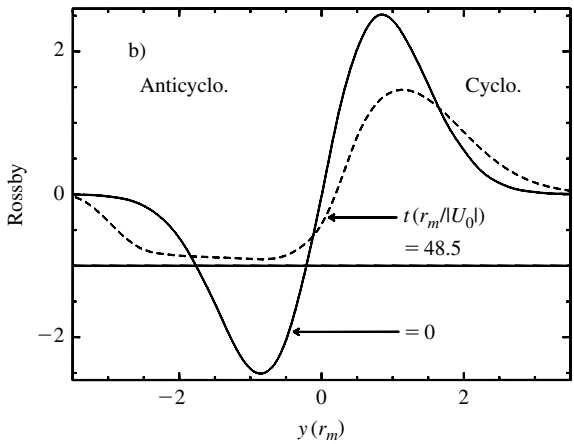


FIG. 2.3 Rotating-wake DNS at $R_o^{(i)} = -2.5$, local Rossby number profiles at $t = 0$ and $t = 48.5$.

There is a crossover Rossby number $R_o^{(i)} = -1$ separating a structure dominated by PEDLEY's [1969] longitudinal mode if the Reynolds number is high enough, from a 2D structure.

Under the crossover, there is a region of space where one observes the establishment of a $R_o(y) = -1$ plateau. This point, well predicted by LES and DNS, still poses severe problems to one-point closure models (see, e.g., NAGANO and HATTORI [2002]).

In this region, the flow evolves into a set of purely longitudinal absolute vortices.

This last point has been demonstrated numerically. The explanation provided by many authors for the channel case[10] is based on the fact that turbulence evolves toward a marginally stable state from the point of view of linear-stability theory. Another explanation is given by LESIEUR, DELCAYRE and LAMBALLAIS [1999] in terms of nonlinear reorientation of absolute vortices. They propose an exact analysis based on Euler equations, where x-independance[11] is assumed. The evolution equations (following the motion) of the absolute vorticity $\vec{\omega}_a$ of components $\omega_1 = \partial w/\partial y - \partial v/\partial z$, $\omega_2 = \partial u/\partial z$, $\omega_3 + f = -\partial u/\partial y + f$ can be written for this x-independent solution as

$$\frac{D}{Dt}\vec{\omega}_a = \overline{\overline{F}} \otimes \vec{\omega}_a, \tag{2.4}$$

[10] Results obtained in Grenoble for the rotating free-shear flows are new, and have not been commented by other people.

[11] Indeed, we have seen that the linear-stability analysis shows that this longitudinal mode dominates shear instabilities in this case.

with

$$\overline{\overline{F}} = \begin{pmatrix} 0 & f & 0 \\ 0 & \dfrac{\partial v}{\partial y} & \dfrac{\partial v}{\partial z} \\ 0 & \dfrac{\partial w}{\partial y} & \dfrac{\partial w}{\partial z} \end{pmatrix} = \overline{\overline{F}}_1 + \overline{\overline{F}}_2, \tag{2.5}$$

$$\overline{\overline{F}}_1 = \begin{pmatrix} 0 & f & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{2.6}$$

and

$$\overline{\overline{F}}_2 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \dfrac{\partial v}{\partial y} & \dfrac{\partial v}{\partial z} \\ 0 & \dfrac{\partial w}{\partial y} & \dfrac{\partial w}{\partial z} \end{pmatrix}. \tag{2.7}$$

The action of $\overline{\overline{F}}_1$ upon $\vec{\omega}_a$ is to leave its projection $\vec{\omega}_n$ upon the y, z plane unchanged, and to stretch ω_1 as

$$\frac{D\omega_1}{Dt} = f\omega_2. \tag{2.8}$$

We have also

$$\frac{D}{Dt}\vec{\omega}_n = \overline{\overline{F}}_2 \otimes \vec{\omega}_n. \tag{2.9}$$

The tensor $\overline{\overline{F}}_2$ is in fact the velocity-gradient tensor in the y, z plane, and we can apply the same analysis as in Chapter 2 of Lesieur, Métais and Comte [2005] for the stretching of various vectors. Indeed, during the linear stage of evolution, DNS of Métais, Flores, Yanase, Riley and Lesieur [1995] concerning an anticyclonic mixing layer of initial Rossby -5 show the growth of the longitudinal mode, with absolute vortex filaments in phase and inclined approximately 45° above the horizontal plane. This produces concentrations of longitudinal vorticity in the y, z plane. Let us assume that a nonlinear regime is reached where longitudinal vorticity concentrations are strong enough to form vortices, whose core is "elliptic" in the sense that the eigenvalues of $\overline{\overline{F}}_2$ (or $-\overline{\overline{F}}_2|^t$) are pure imaginaries. Rotation of $\vec{\omega}_n$ about $\vec{x}$ (in the sense of the sign of the longitudinal vorticity) will therefore dominate deformation in Eq. (2.9), implying an increase of the

spanwise absolute-vorticity component (which is negative). The Rossby number (which was lower than -1) will increase also. We have here for the absolute-vorticity vector an interesting mechanism of longitudinal self-reorientation, possible only in a nonlinear regime.

GFD applications These results have important applications in geophysical and astrophysical fluid dynamics when density differences may be neglected.

In the mesoscale Earth's atmosphere, typical Rossby numbers moduli are larger than one. Then anticyclonic vortices should be destroyed and cyclonic ones two-dimensionalized. This is observed in the wake (visualized by clouds) of some islands, which display very asymmetric Karman streets.

In the ocean, cyclonic and anticyclonic mesoscale eddies which do not result from baroclinic instability should be two-dimensionalized. This might be the case in particular for detached vortices behind capes. An example is the Strait of Gibraltar, separating the Atlantic Ocean from the Mediterranean Sea. The Atlantic is much fresher than the Mediterranean.[12] Therefore, Mediterranean water will dive into the Atlantic through Gibraltar while Atlantic water will enter the Mediterranean remaining at the surface. The coast of Morocco will act as a backward-facing step, and vortices will be shed on the north coast of Algeria. These vortices have been observed. They are anticyclonic, but since the associated Rossby numbers are very low, they should be two-dimensionalized.

On Jupiter, where there is no evidence of baroclinic instability, vortices should be two-dimensionalized whatever their sign. It is the case in particular of the great red spot, which is anticyclonic.

2.3. *Storm formation*

As already stressed, storm formation results from baroclinic instability of a jet stream in thermal-wind balance resulting from a horizontal north-south temperature gradient. The simplest model for such a study is Eady model.

Notice that we are going to change the notations with respect to the preceding section: x, y, and z will be, respectively, the zonal, meridional, and vertical directions, u, v, and w being the velocity components in these directions.

2.3.1. *Eady model*

Details are given in Drazin and Reid [1981]. One considers a channel rotating about the vertical axis z with an angular velocity $f/2$, studied within Boussinesq approximation. The flow is stably stratified along the vertical and the horizontal, with constant temperature gradients. The channel has a width L and a depth H. Periodicity is assumed in the x direction. Free-slip boundary conditions on the lateral walls and upper lid are taken, with a no-slip condition at the ground. One assumes initially a thermal-wind balance corresponding to a (x, y) independent zonal westerly velocity rising linearly from 0 at the ground to U on the lid. Let $R_o = U/(fL)$ be the Rossby number, and $F = U/(NH)$ be

[12] Indeed, the latter is very warm and undergoes a strong evaporation.

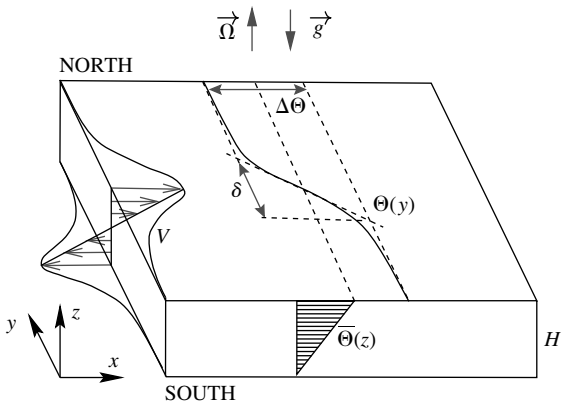


FIG. 2.4 Baroclinic jet configuration (picture courtesy of E. Garnier).

the Froude number, N being the Brunt-Vaisala frequency (see details in LESIEUR [1997, p. 49]). A linear-stability analysis shows that instability occurs when $R_o/F < 0.76$.

2.3.2. Baroclinic jet

The configuration of this simulation (DNS or LES) is displayed in Fig. 2.4. It is a baroclinic jet in a channel. One still work within Boussinesq approximation.

Like in Eady model, there is initially a positive constant vertical temperature gradient. The difference with Eady model is a hyperbolic-tangent velocity profile of width δ in the meridional direction.[13] The initial velocity field is in thermal-wind balance. It corresponds to a zonal jet varying linearly with the vertical, from $-V$ at the ground ($z = 0$) to $+V$ on the lid ($z = H$). We stress that a Galilean transformation of velocity V would yield a zero velocity at the ground, as in reality, and a velocity $2V$ at the lid. The Rossby and Froude numbers are, respectively, $V/f\delta$ and V/NH. The DNS carried out by GARNIER, MÉTAIS and LESIEUR [1998] allowed first to carry out a linear-stability analysis (by cancellation of the nonlinear terms). It shows that instability occurs for $R_o/F < 1.5$. Afterwards, simulations are carried out for $R_o/F = 0.5$, corresponding to a physically realistic situation in the Earth atmosphere. DNS of Fig. 2.5 show the formation of quasi 2D cyclonic vortices and weaker anticyclonic vortices. They display also the production of intense cyclonic vertical vorticity within the fronts separating the cold fluid to the north from the warm fluid to the south, at the ground and under the lid. Such a production may be explained as follows: one can show in the framework of Boussinesq approximation (and neglecting viscosity) that relative vertical vorticity ω satisfies

$$\frac{D\omega}{Dt} = (\omega + f)\frac{\partial w}{\partial z} + \omega_1 \frac{\partial w}{\partial x} + \omega_2 \frac{\partial w}{\partial y}, \tag{2.10}$$

[13] This width is defined exactly in the same way as for the vorticity thickness in a mixing layer.

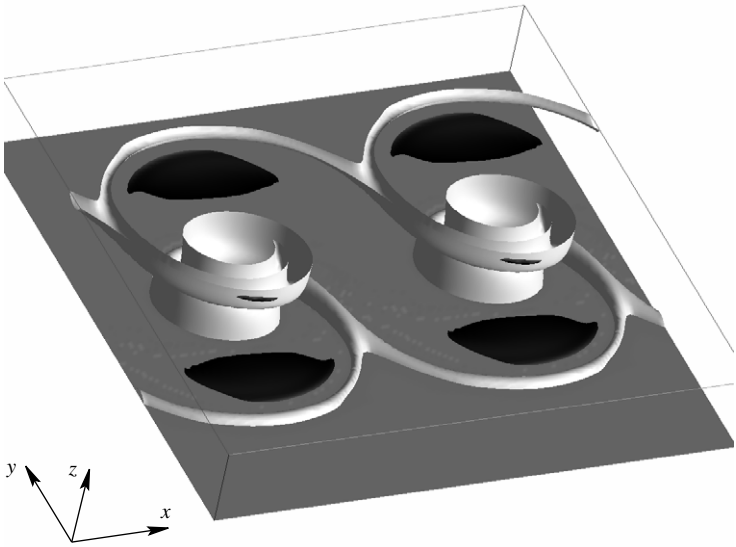


FIG. 2.5 DNS of the baroclinic jet: light grey, cyclonic relative vertical vorticity; black, anticyclonic vorticity (Picture courtesy of E. Garnier).

where ω_1 and ω_2 are here the components of $\vec{\omega}$ along x and y. It was shown by GARNIER, MÉTAIS and LESIEUR [1998] in their simulations that the last two terms in the r.h.s. of Eq. (2.10) can be neglected, and the equation reduces to

$$\frac{D}{Dt}\omega \approx (f+\omega)\frac{\partial w}{\partial z}. \tag{2.11}$$

One understands thus how the warm fluid at the ground arriving in contact of the cold fluid will be obliged to rise, yielding $\partial w/\partial z > 0$ in the r.h.s. of Eq. (2.11). Here f is positive, and the Rossby number modulus of the calculation ($\approx |\omega|/f$) is low enough so that $f + \omega > 0$ and ω will grow. Starting with a weak $|\omega|$, there will be growth of cyclonic vorticity and damping of anticyclonic-vorticity modulus. The same occurs under the lid, where cold fluid will sink under warm fluid, with again $\partial w/\partial z > 0$. This is illustrated by an animation 8-1, which is another DNS done by Garnier at Rossby and Froude numbers both equal to 0.1. It presents the vertical relative vorticity (pink positive – cyclonic, blue negative – anticyclonic). First, one can see the double Bickley jet at the ground and on the lid. Afterwards the progressive formation of quasi 2D cyclonic vortices is observed, with formation of intense cyclonic vorticity braids along the fronts.

GARNIER, MÉTAIS and LESIEUR [1998] carried out LES using a subgrid model combining the SF and a hyperviscosity. A primary instability occurs as in the DNS. Then there is a secondary instability of the cold front, with production of two secondary vortices in two days (see Fig. 2.6). Their vorticity of the order of 2 to 3 times that of primary vortices. They show also that the secondary instability is associated with regions where

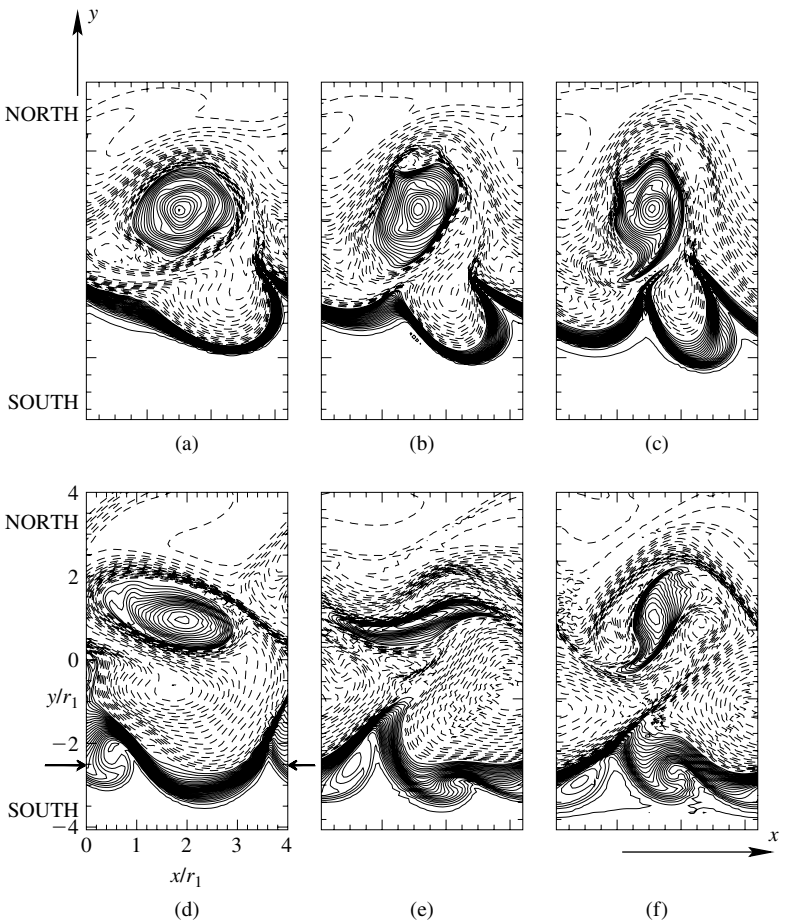


FIG. 2.6 Time evolution in two days of the ground temperature in the baroclinic jet LES. Time goes as indicated by letters (Picture courtesy of E. Garnier).

the ratio of local Rossby over Froude numbers is smaller than 1.5, the critical value from the point of view of their linear-stability study.

As stressed in LESIEUR, MÉTAIS and GARNIER [2000], there are strong analogies with storms which struck Europe on December 26th (see Figs. 2.7 and 2.8) and 28th 1999.

The situation on December 25th (Fig. 2.7) indicates a big cyclonic perturbation arriving on Scandinavia after having crossed Great Britain. The corresponding winds, of the order of 30 m/s, are typical of a "regular" storm issuing from a primary baroclinic instability, and which cross the channel sea very often. It is clear that France is split by the cold front at the level of Brittany. One should remind that this Christmas 99 evening, a warm wind coming from the South was blowing above Grenoble, with temperatures of the order of 15°C very unusual for the season. This wind was induced by the vortex sitting on Scandinavia. The December 26th storm (Fig. 2.8) is a small vortex of ≈ 400 km

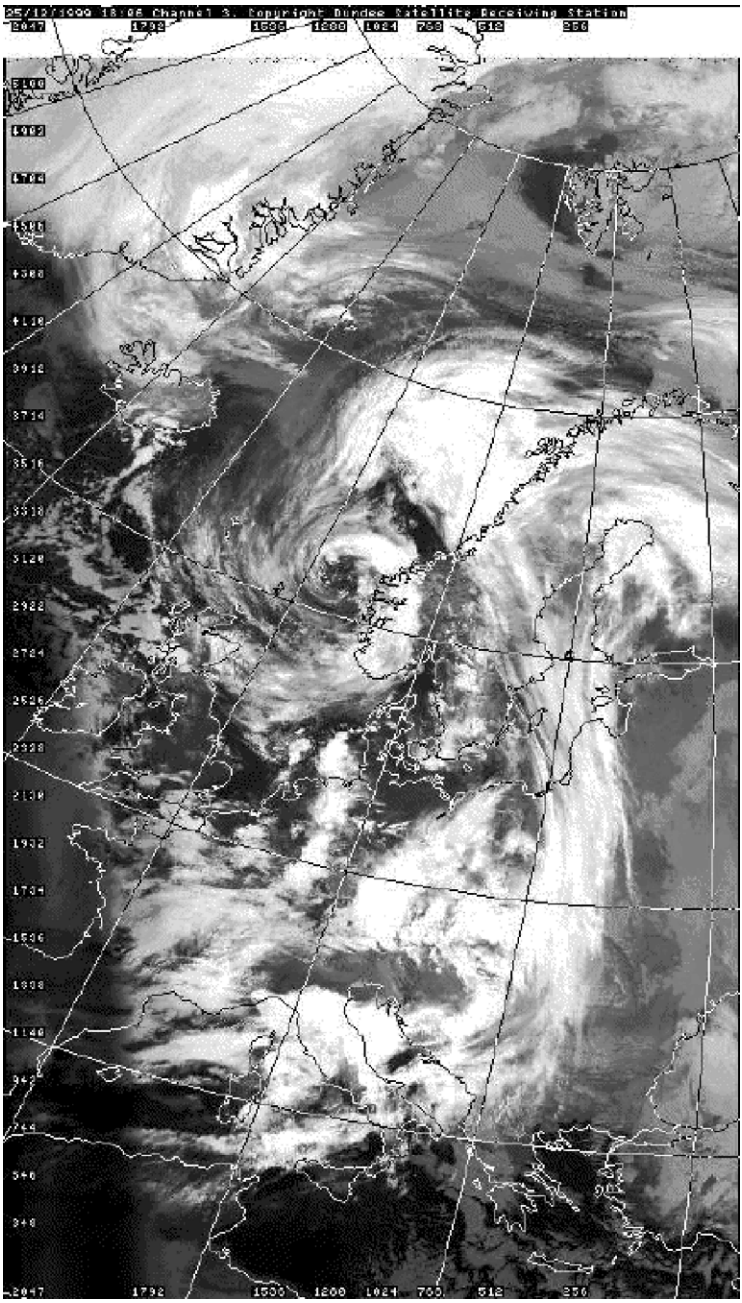


FIG. 2.7 Satellite view of Europe on December 25th, 18 h (Picture courtesy of Dundee University).

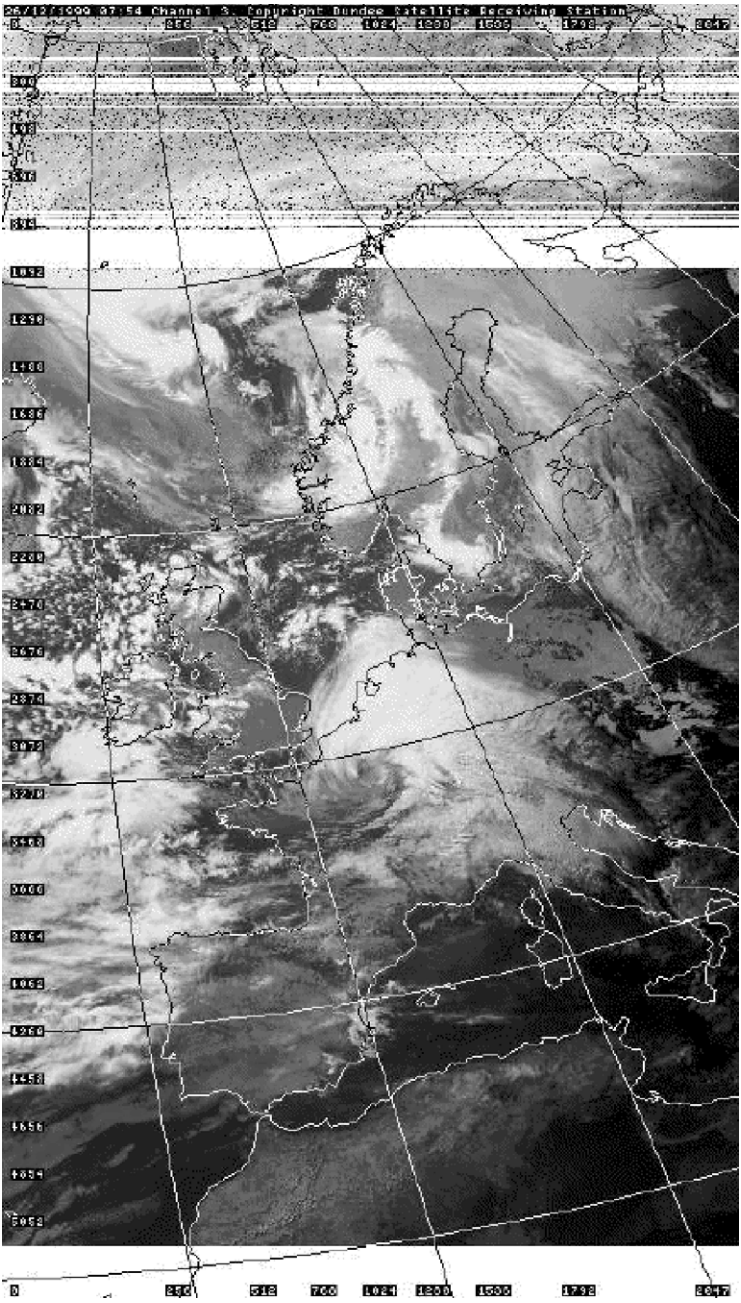


FIG. 2.8 Satellite view of Europe on December 26th, 8 h (Picture courtesy of Dundee University).

diameter, which formed on Brittany and, according to Météo-France, crossed France from west to east at a speed of 28 m/s. One may think that the relative velocity at the border of this vortex is $\approx$ 28 m/s, which induces winds of velocity 56 m/s to the south of the vortex and weak to the north. Since there was a second vortex of the same kind two days after, these storms present analogies with the secondary vortices observed in the above baroclinic jet LES. Such velocities were totally unexpected from the point of view of roofs and electric and telephone lines. We recall that forces exerted by a wind upon a body are proportional to the squared velocity when the wake of the body is turbulent. Our constructions could resist winds $\approx$ 30 m/s but not 60 m/s, which exert forces four times larger.

Among the differences between the baroclinic-jet simulations and the December 1999 severe storms, LESIEUR, MÉTAIS and GARNIER [2000] note pressure troughs measured in the storms, and not observed in the simulations.

One of the questions posed by LESIEUR, MÉTAIS and GARNIER [2000] concerns the requirements needed by a numerical weather-forecast code in order to be able to capture analogous phenomena correctly and predict the severe vertical vortex stretching in the thermal fronts. One may wonder whether the hydrostatic approximation is sufficient, and if Boussinesq- or anelastic-type approximations of the Navier-Stokes equation should not be preferred. Gravity waves should be filtered out in some way. Meshes smaller than 50 km should be used horizontally, and numerical schemes in the vertical direction should not be too diffusive. The use of LES seems compulsory in order to obtain secondary instabilities, since the latter were dissipated by molecular viscosity in the baroclinic-jet DNS. The question of data assimilation to define the proper initial field is another controversial topic. The buildup of forecasting tools able to predict accurately the wind velocities associated with such severe storms is of essential importance for populations living in midlatitude regions.

It is difficult to know whether these events are associated with a warming climate. Climatologists stress that polar regions warm faster than equatorial ones, leading to meridional temperature gradients which should become in the mean weaker rather than stronger, as they certainly were during the December 1999 storms.

References

André, J.C., Lesieur, M. (1977). Influence of helicity on high Reynolds number isotropic turbulence. *J. Fluid Mech.* **81**, 187–207.

Antonia, R.A., Teitel, M., Kim, J., Browne, L.W.B. (1992). Low Reynolds number effects in a fully-developed turbulent channel flow. *J. Fluid Mech.* **236**, 579–605.

Chollet, J.P., Lesieur, M. (1981). Parameterization of small scales of three-dimensional isotropic turbulence utilizing spectral closures. *J. Atmos. Sci.* **38**, 2747–2757.

Deardorff, J.W. (1970). A numerical study of three-dimensional turbulent channel flow at large Reynolds number. *J. Fluid Mech.* **41**, 453–480.

Deardorff, J.W. (1972). Numerical investigation of neutral and unstable planetary boundary layers. *J. Atmos. Sci.* **29**, 91–115.

Drazin, P.G., Reid, W.H. (1981). *Hydrodynamic Stability* (Cambridge University Press).

Dubief, Y., Delcayre, F. (2000). On coherent-vortex identification in turbulence. *J. Turb.* **1**, 11.

Fallon, B., Lesieur, M., Delcayre, F., Grand, D. (1997). Large-eddy simulations of stable-stratification effects upon a backstep flow. *Eur. J. Mech. B/Fluids*. **16**, 525–644.

Favre, A. (1958). Equations statistiques des gaz turbulents, C.R. Acad. Sci., Paris, **246**: masse, quantité de mouvement, pp 2576–2579; énergie cinétique, énergie cinétique du mouvement macroscopique, énergie cinétique de la turbulence, pp. 2839–2842; enthalpies, entropie, températures, pp. 3216–3219.

Flores, C. (1993). Etude numérique de l'influence d'une rotation sur lesécoulements cisaillés libres, PhD thesis, Grenoble, France.

Garnier, E., Métais, O., Lesieur, M. (1998). Synoptic and frontalcyclone scale instabilities in baroclinic jet flows. *J. Atmos. Sci.* **55**, 1316–1335.

Germano, M., Piomelli, U., Moin, P., Cabot, W. (1991). A dynamic subgrid-scale eddy-viscosity model. *Phys. Fluids A.* **3**, 1760–1765.

Hunt, J., Wray, A., Moin, P. (1988). Eddies, stream, and convergence zones in turbulent flows, Center for Turbulence Research Rep, CTR S88, pp. 193.

Jimenez, J. (2000). Turbulence. In: Batchelor, G.K., Moffatt, H.K., Woster, M.G. (eds.), *Perspectives in Fluid Mechanics* (Cambridge University Press), pp. 231–283.

Jimenez, J. (1990). Transition to turbulence in two-dimensional Poiseuille flow. *J. Fluid Mech.* **218**, 265–297.

Johnston, J.P., Halleen, R.M., Lezius, D.K. (1972). Effects of spanwise rotation on the structure of two-dimensional fully developed turbulent channel flow. *J. Fluid Mech.* **56**, 553–557.

Kenjeres, S., Hanjalic, K. (2002). Combined effects of terrain orography and thermal stratification on pollutant dispersion in a town valley: a T-RANS simulation. *J. Turb.* **3**, 26.

Kim, J. (1983). The effect of rotation on turbulence structure. In: *Proc. 4th Symp. on Turbulent Shear Flows, Karlsruhe*, pp. 6.14–6.19.

Kraichnan, R.H. (1976). Eddy viscosity in two and three dimensions. *J. Atmos. Sci.* **33**, 1521–1536.

Kristoffersen, R., Andersson, H.I. (1993). Direct simulations of low-Reynolds-number turbulent flow in a rotating channel. *J. Fluid Mech.* **256**, 163–197.

Lamballais, E., Métais, O., Lesieur, M. (1997). Spectral-dynamic model for large-eddy simulations of turbulent rotating channel flow. *Theor. Comp. Fluid Dyn.* **12**, 149–177.

Lamballais, E., Lesieur, M., Métais, O. (1996). Effects of spanwise rotation on the vorticity stretching in transitional and turbulent channel flow. *Int. J. Heat. Fluid Flow*. **17**, 324–332.

Leray, J. (1934). Sur le mouvement d'un fluide visqueux emplissant l'espace. *J. Acta Math.* **63**, 193–248.

LESIEUR, M., MÉTAIS, O., COMTE, P. (2005). *Large-Eddy Simulations of Turbulence* (Cambridge University Press).

LESIEUR, M., MÉTAIS, O., GARNIER, E. (2000). Baroclinic instability and severe storms. *J. Turb.* **1**, 2.

LESIEUR, M., DELCAYRE, F., LAMBALLAIS, E. (1999). Spectral eddy-viscosity based LES of shear and rotating flows. In: Kerr, R., Kimura, Y. (eds.), *Developments in geophysical turbulence* (Kluwer), pp. 235–252.

LESIEUR, M. (1997). *Turbulence in Fluids*, third ed. (Kluwer Springer), pp. 515.

LESIEUR, M., MÉTAIS, O. (1996). New trends in large-eddy simulations of turbulence. *Ann. Rev. Fluid Mech.* **28**, 45–82.

LESIEUR, M., YANASE, S., MÉTAIS, O. (1991). Stabilizing and destabilizing effects of a solid-body rotation on quasi-two-dimensional shear layers. *Phys. Fluids A*. **3**, 403–407.

LEZIUS, D.K., JOHNSTON, J.P. (1976). Roll-cell instabilities in rotating laminar and turbulent channel flows. *J. Fluid Mech.* **77**, 153–175.

LILLY, D.K. (1962). On the numerical simulation of buoyant convection. *Tellus*. **14**, 148–172.

LILLY, D.K. (1966). On the application of the eddy-viscosity concept in the inertial subrange of turbulence. NCAR ms # 123.

LILLY, D.K. (1967). The representation of small-scale turbulence in numerical simulation experiments. In: *Proc. IBM Sci. Comput. Symp. on Envir. Sci.* IBM Form 320–1951, pp. 195–210.

MÉTAIS, O., FLORES, C., YANASE, S., RILEY, J.J., LESIEUR, M. (1995). Rotating free shear flows part 2: numerical simulations. *J. Fluid Mech.* **293**, 41–80.

MÉTAIS, O., LESIEUR, M. (1992). Spectral large-eddy simulations of isotropic and stably-stratified turbulence. *J. Fluid Mech.* **239**, 157–194.

NAGANO, Y., HATTORI, H. (2002). An improved turbulence model for rotating shear flows. *J. Turb.* **3**, 6.

NAKABAYASHI, K., KITOH, O. (1996). Low Reynolds number fully developed two-dimensional turbulent channel flow with system rotation. *J. Fluid Mech.* **315**, 1–29.

ORSZAG, S.A. (1970). Analytical theories of turbulence. *J. Fluid Mech.* **41**, 363–386.

PADILLA-BARBOSA, J., MÉTAIS, O. (2000). Large-eddy simulations of deep-ocean convection: analysis of the vorticity dynamics. *J. Turb.* **1**, 9.

PEDLEY, T.J. (1969). On the stability of viscous flow in a rapidly rotating pipe. *J. Fluid Mech.* **35**, 97–115.

PIOMELLI, U., LIU, J. (1995). Large-eddy simulation of rotating channel flows using a localized dynamic model. *Phys. Fluids A*. **7** (4), 839–848.

SMAGORINSKY, J. (1963). General circulation experiments with the primitive equations. *Mon. Wea. Rev.* **91** (3), 99–164.

SOMMÉRIA, J. (2001). Two-dimensional turbulence. In: LESIEUR, M., YAGLOM, A., DAVID, F. (eds.), *New Trends in Turbulence* (EDP-Springer), pp. 554.

TAFTI, D.K., VANKA, S.P. (1991). A numerical study of the effects of spanwise rotation on turbulent channel flow. *Phys. Fluids A*. **3** (4), 642–656.

TSENG, Y., FERZIGER, J. (2001). Effects of coastal geometry and the formation of cyclonic/anti-cyclonic eddies on turbulent mixing in upwelling simulation. *J. Turb.* **2**, 14.

YANASE, S., FLORES, C., MÉTAIS, O., RILEY, J. (1993). Rotating free-shear flows. I. Linear stability analysis. *Phys. Fluids*. **5**, 2725–2737.

Two Examples from Geophysical and Astrophysical Turbulence on Modeling Disparate Scale Interactions

Pablo Mininni

NCAR, P.O. Box 3000, Boulder, CO 80307-3000, U.S.A.

Annick Pouquet

NCAR, P.O. Box 3000, Boulder, CO 80307-3000, U.S.A.

Peter Sullivan

NCAR, P.O. Box 3000, Boulder, CO 80307-3000, U.S.A.

Abstract

Turbulent flows are ubiquitous, and as manifestations of one of the last outstanding unsolved problems of classical physics, they form today the focus of numerous investigations. In view of the very large number of modes that are excited, a variety of modeling techniques can be used in conjunction with state of the art numerical methods. A few of the issues that need to be addressed by models of turbulence, such as the presence of strong localized structures, the degree of nonlocality of nonlinear interactions, the slow return to isotropy and homogeneity, and the interactions between eddies and waves, are reviewed here; all implicate a large number of scales in interactions. Two specific modeling examples are given, one for waves and eddies in oceanic flows, and one for the generation of magnetic fields in planetary and stellar bodies, both using variants of Lagrangian-averaged methods. Finally, it is also argued that in order to understand geophysical turbulence, there is a strong need for combining modeling methods and sophisticated numerical techniques, such as high-accuracy adaptive mesh refinement.

Computational Methods for the Atmosphere and the Oceans
Special Volume (Roger M. Temam and Joseph J. Tribbia, Guest Editors) of
HANDBOOK OF NUMERICAL ANALYSIS, VOL. XIV
P.G. Ciarlet (Editor)

Copyright © 2009 Elsevier B.V.
All rights reserved

ISSN 1570-8659
DOI 10.1016/S1570-8659(08)00208-1

1. Geophysical turbulence

1.1. Introduction

Flows in nature are often in a turbulent state driven by large-scale forcing (e.g., novae explosions in the interstellar medium) or by instabilities (e.g., convection in the sun). Such flows involve a huge number of coupled modes of different scales leading to great complexity both in their temporal dynamics and in the physical structures that emerge. Many scales are excited in turbulent flows, for example, from the planetary scale to a few meters for convective clouds in the atmosphere, and much smaller scales when considering microprocesses such as droplet formation. Turbulence is linked to many issues in the geosciences, e.g., in meteorology, oceanography, climatology, ecology, solar terrestrial interactions and fusion; its study is not limited to inquiries in physics; turbulence plays an equally prominent multifaceted role in industrial flows, e.g., through the presence of seed particles or bubbles, in studies of combustive and chemically reactive flows, in the area of aeronautical engineering (e.g., aircraft safety and performance), or in epidemiology.

Nonlinearities prevail in turbulent flows when the Reynolds number Re – which measures, as a control parameter, the amount of active temporal or spatial scales – is large (Frisch [1995], Lesieur [1997], Leslie [1973], Monin and Yaglom [1975], Pope [2000], Tennekes and Lumley [1972]). The number of degrees of freedom (dof) increases as $\mathsf{Re}^{9/4}$ for $\mathsf{Re} \gg 1$ in the Kolmogorov framework (Kolmogorov [1941], hereafter K41) and for geophysical flows, often $\mathsf{Re} > 10^8$. The ability to probe large Re numerically and to examine in detail the large-scale behavior of turbulent flows depends critically on the ability to resolve such a large number of modes or else to model them adequately. Theory demands that computations of turbulent flows reflect a clear scale separation between the energy-containing, inertial, and dissipative ranges; convergence studies on compressible flows show that to achieve this, it is necessary to compute on grids with at least 2048^3 dof (Sytine, Porter, Woodward, Hodson and Winkler [2000]). Today such computations are barely feasible (Kurien and Taylor [2005], Yeung, Donzis and Sreenivasan [2005]), and modeling is a necessity; although in the highest Re to date achieved with a run on a 4096^3 grid (Kaneda, Ishihara, Yokokawa, Itakura and Uno [2003]), one does see the emergence of scaling laws in turbulence, as for example a constant energy dissipation rate.

There are many features of turbulent flows that need modeling, for example, the presence of strong localized structures, the high degree of nonlocality of nonlinear interactions in Fourier space (i.e., the (relative) importance of interactions between widely separated scales versus between comparable scales), the slow return to isotropy in the small scales or a measurable influence of waves on eddies. The purpose of this paper is to show, using two specific examples, how one can take into account some of these detailed features of turbulent flows in modeling techniques. We present in the following a few ideas and do not attempt at an exhaustive review of turbulence or of modeling for turbulent flows (see Frisch [1995], Lesieur [1997], Pope [2000], and references therein). The next section describes some of the issues that one should address when modeling turbulent flows at high Re; Section 3 deals with waves and eddies in the oceanic boundary

layer (OBL), Section 4 with the coupling of the velocity to the magnetic induction in the magnetohydrodynamic (MHD) framework, and Section 5 contains concluding remarks including a brief discussion on adaptive mesh refinement (AMR) tools.

1.2. A few issues

The Navier-Stokes equation governing an incompressible velocity field $\mathbf{v}$ with pressure to density ratio P, forcing function $\mathbf{F}$ and viscosity ν is

$$\frac{\partial \mathbf{v}}{\partial t} + \nabla P = -\mathbf{v} \cdot \nabla \mathbf{v} + \nu \nabla^2 \mathbf{v} + \mathbf{F}, \tag{1.1a}$$

$$\nabla \cdot \mathbf{v} = 0. \tag{1.1b}$$

For a characteristic velocity U_0 and characteristic length scale L_0, one defines the Reynolds number as $\mathsf{Re} = \mathsf{U_0 L_0}/\nu$. This is the only governing parameter of the problem, in absence of any other external forcing such as buoyancy or rotation. Analyses and reviews of turbulent flows in diverse contexts can be found in many books (op. cit.). Here, we want to stress some of the aspects of such flows that may pose problems when studying them numerically with the help of models.

Modeling is an ancient art at the intersection of mathematics and physics; more recently, in attempts on understanding the long-term evolution of weather and climate, modeling has also evolved toward including chemistry and bio-geo processes. Numerical modeling is more recent and proteiform; having nevertheless acquired its credentials, it can be viewed as being complementary to theoretical, observational, and experimental approaches to our understanding of turbulent flows as they appear in the industrial context, in geophysics and in astrophysics.

Numerical modeling has at least two faces. One is that of direct numerical simulations (or DNS) where all scales are treated accurately and explicitly, with or without AMR or multigrid methods. A current requirement for DNS of turbulent flows to be well resolved is that the ratio of the Kolmogorov wavenumber k_D beyond which dissipation prevails to the maximum wavenumber $k_{\max}$ of the computation be of the same order of magnitude, where $k_D \sim [\epsilon/\nu^3]^{1/4}$ with ϵ and ν, respectively, the energy injection rate and the viscosity and having assumed a Kolmogorov energy spectrum $E(k) \sim k^{-5/3}$. With $k_{\max} \sim 1000$ as can be achieved today, this means that less than 10% of the Fourier modes are in the inertial range since the number of modes in a Fourier shell centered around a given wavenumber K scales as K^2 in three-space dimensions. Several ways have been devised around this difficulty. One of them is a decimation algorithm whereby the number of modes after a cutoff wavenumber, say in the inertial range, is successively cut by a factor 2 in adjacent Fourier shells (Meneguzzi, Politano, Pouquet and Zolver [1996]), leading to a number of modes varying as $\ln \mathsf{Re}$ as opposed to $\mathsf{Re}^{9/4}$. Little employed, it might be worth reconsidering, perhaps in conjunction with other methods such as AMR or large eddy simulations (LES).

Since the disadvantage of DNS until now is that only modest Re in three dimensions (3D) can be analyzed with sufficient accuracy, one way around it is thus to resort to LES (Germano, Piomelli, Moin and Cabot [1991], Lesieur and Métais [1996], Meneveau and Katz [2000], Sagaut [2006]). Such techniques are widely used in

engineering, as well as in atmospheric sciences and to a lesser extent in astrophysics, and few such models have been developed in MHD (see AGULLO, MÜLLER, KNAEPEN and CARATI [2001], POUQUET, FRISCH and LÉORAT [1976], YOSHIZAWA [1985]). In LES, the effects of the small scales on the large scales are incorporated through various (sometimes, many) parameters that can be tuned in search of an agreement with observations. In an LES computation, the Reynolds number is not generally known except that it is assumed very large. Two well-known transport coefficients included in an LES are the eddy viscosity, whereby the small scales are dissipating the large scales because of the energy cascade, and the eddy noise, whereby the small scales, for example through beating, create a small amount of energy at large scale with a prescribed spectrum[1].

The degree of resolution of an LES (see POPE [2004] for a review) can be defined as

$$M = \frac{E_C}{E_R + E_C},$$

where E_C and E_R are, respectively, the kinetic energy in the residual (unresolved) modes and the resolved kinetic energy at a given Reynolds number. For a DNS, $M = 0$ and for a Reynolds-averaged simulation (RANS), $M = 1$. If we now assume a Kolmogorov energy spectrum and if we further assume that the cutoff wavenumber $k_{\rm cut}$ of the LES is in the inertial range, it is easy to show that in the limit of infinite Reynolds number, $M_\infty \sim k_{\rm cut}^{2/3}$; with $k_{\rm cut} = 125$, we have $M_\infty \sim 0.04$ and with an optimistic but realizable in the future $k_{\rm cut} = 1000$, one finds that M_∞ drops to 1%. In other words, LES will be well resolved if the cutoff wavenumber can be taken in the inertial range.

However, in the traditional LES approach, the Reynolds number is not known and one rather attempts at modeling the behavior of the flow in the limit of very large Re. A different approach sometimes coined quasi-DNS is to use modeling tools that allow for a larger Reynolds number than what a DNS would give on a given grid by using a variety of techniques that can be viewed as filtering the small scales. One such model, the Lagrangian-averaged α model (CHEN, HOLM, FOIAS, OLSON, TITI and WYNNE [1998], HOLM [2002a], HOLM, MARSDEN and RATIU [1998a]), has been used with success for fluids (CHEN, HOLM, FOIAS, OLSON, TITI and WYNNE [1998]) and MHD (MININNI, MONTGOMERY and POUQUET [2005b,c]), whereby it was shown that one may reasonably expect a gain of a factor 4 in linear resolution, hence a gain of 64 in memory and 128 in CPU time when reproducing the large-scale and small-scale features of a 3D flow at a given Re (GRAHAM, MININNI and POUQUET [2005]). Quasi-DNS is in between the two aforementioned methods. It attempts at conserving, within the confines of the model, some of the mathematical structures of the primitive equations such as the existence of invariance properties (e.g., total energy), topological constraints (invariance of helicity, Kelvin theorem), or the existence of exact solutions such as Alfvén waves in MHD. Such invariant properties may occur in an altered way (e.g., not using the same norm); for

[1] In climate simulations, one needs to model physical phenomena which are not incorporated directly, such as convective clouds, by opposition to LES where one models the unresolved scales of a given well-posed problem; one example is the attempt, not quite successful yet, at reproducing both the amplitude and the frequency of the El Niño Southern Oscillation in the Pacific Ocean.

example, the Lagrangian-averaged model introduces a scale, α, beyond which fluctuations in the Lagrangian path of fluid particles are ignored. To lowest order, this model preserves energy, but in the $\mathcal{H}_1$ norm instead of the $\mathcal{L}_2$ norm. This methodology turns out to be a successful way to model the primitive equations, based on the generalized lagrangian mean (GLM) method (ANDREWS and MCINTYRE [1978a]).

In the following, we give two examples of such an averaging procedure, for an OBL problem in which the Stokes drift arises as an extra term appearing in the Navier-Stokes equation and in the case of magnetic field generation (or dynamo problem) in the MHD approximation. The OBL problem is described in some detail, in particular, concerning the interactions of eddies and waves. In the case of MHD, a brief review of existing models is first given and then the same Lagrangian-averaged model as described in Section 2 is put to good use in the context of coupling to magnetic fields; a few results are also given in the context of small magnetic Prandtl numbers. Finally, Section 4 is the conclusion in which some thought is also given on the potential of combining modeling and AMR. It is to be noted that even though these two problems stem from vastly different physical conditions, similarities in their modeling (as, e.g., the concept of Stokes drift) and thus in the implementation of LES can be found, similarities linked to the basic multiscale coupling underlying their dynamical nonlinear evolution. It is perhaps the simple central purpose of this paper to bring once more to light such similarities, giving a form of universality of concepts underlying the understanding of geophysical turbulence under drastically varied physical circumstances.

2. Waves, currents, and turbulence in the OBL

The air-sea interface and more broadly the ocean mixed layer (or OBL) plays an important role in geophysical flows, e.g., it controls tropical cyclone intensity (EMANUEL [2004]) and modulates flux exchange in the global ocean circulation (LARGE, MCWILLIAMS and DONEY [1995]). The top of the ocean has air-sea fluxes, surface gravity waves, boundary-layer turbulence, Ekman currents, and significant amounts of small-scale bubbles when the surface is disrupted by intermittent breaking waves[2]. The physics of these phenomena is closely linked despite the difference in length scales $\mathcal{O}(1\,\text{mm} \sim 100\,\text{m})$. Progress by direct computational methods has been impeded by the structural complexity of upper ocean flows, disparity in magnitude and timescale among the flow components, and by high Reynolds number Re. The current generation of bulk (ensemble average) computational models of the OBL used in large-scale prediction codes do not account for the dynamic nature of the surface wave field. These 1D (vertical column) parameterizations of the OBL (e.g., LARGE, MCWILLIAMS and DONEY [1995]) use vertical mixing rules partially based on past experience with the atmospheric boundary layer. This perception of the OBL is in part a necessity as the fierce measuring environment does not permit easy probing of the important interactions that determine the turbulent fluxes of momentum, scalars, and energy across the temporally and spatially evolving air-sea interface. While many aspects of surface waves are well known, a clear understanding

[2] See BAUMERT, SIMPSON and SÜNDERMANN [2005] for a review of marine turbulence.

of their relations with fluxes, turbulence, and currents has long been elusive. However, the impact of waves on OBL dynamics is acknowledged in more recent bulk models (e.g., CRAIG and BANNER [1994], MCWILLIAMS and SULLIVAN [2000]), improvements motivated by better measurements and findings from turbulence resolving simulations with wave effects as described here.

Despite the passive appearance of the sea surface under light winds, $U_a \approx [4 - 6]\mathrm{ms}^{-1}$, the OBL exhibits spectacular and nonintuitive disparate scale interactions between fast oscillating surface waves and slower moving underlying currents and turbulence. Motivation for a deeper understanding of the interactions between waves and currents is the appearance of Langmuir circulations near the ocean surface (LANGMUIR [1938], LEIBOVICH [1983], THORPE [2003]). These counter-rotating vortices, roughly aligned parallel with the wind direction and perpendicular to the wave crests, are dynamics unique to the OBL (see Fig. 2.1). The prevailing theoretical interpretation of Langmuir circulations is that the Stokes drift of the wave field $\mathbf{u}^s$ generates "vortex forces" that stretch, tilt, and amplify background vorticity $\boldsymbol{\omega}$ in the OBL (LEIBOVICH [1983]); preferential horizontal tilting of vertical vortex filaments by Stokes drift leads to streamwise vortices. This remarkable coupling was first predicted by a novel asymptotic analysis (CRAIK and LEIBOVICH [1976]) and later explored by turbulence resolving computations, primarily by the technique of LES (e.g., LI, GARRETT and SKYLLINGSTAD [2005], MCWILLIAMS, SULLIVAN and MOENG [1997], NOH, MIN and RAASCH [2004], SKYLLINGSTAD and DENBO [1995]).

For winds $U_a > 6\,\mathrm{ms}^{-1}$, visible white capping of the sea surface also takes place and intermittent wave breaking (e.g., MELVILLE and MATUSOV [2002]) coexists with Langmuir circulations. Breaking waves transfer momentum from the wind-generated wave field to the oceanic turbulence and currents; this is the dominant mechanism for the net momentum exchange between winds and currents usually associated with the wave-averaged surface wind stress (DONELAN [1998]). There are recent attempts to model this additional wave effect in turbulence simulations as a random field of stochastic impulses (NOH, MIN and RAASCH [2004], SULLIVAN, MCWILLIAMS and MELVILLE [2004, 2005]). The computational model of the OBL described here ignores these complications and instead simply focuses on average wave-current interactions embodied in the Craik-Leibovich equations at an Re sufficiently high to support 3D turbulence. This framework allows the coupling of waves, currents, and turbulence to be examined in numerical simulations and exposes surprising interactions between irrotational waves and vortical turbulence. Although the present interest is on 3D flows, the Craik-Leibovich ideas have also been evaluated in 2D models using constant eddy viscosity as a surrogate for turbulence (e.g., LI, ZAHARIEV and GARRETT [1995]).

2.1. *Wave-current interactions*

The Craik-Leibovich perturbation equations (CRAIK and LEIBOVICH [1976], LEIBOVICH [1983]) are a multiscale expansion of the full Navier-Stokes (Eqs. (1.1a) and (1.1b)) assuming that the total velocity field

$$\mathbf{v} = \epsilon \mathbf{u}^w + \epsilon^2 \mathbf{u} \tag{2.1}$$

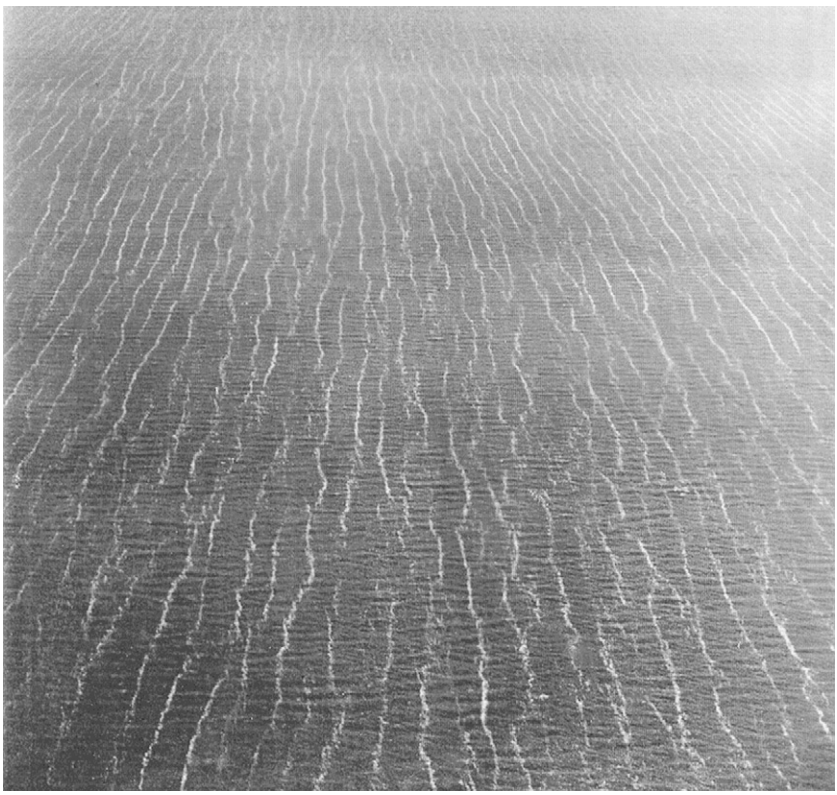


FIG. 2.1 Foam lines tracking the formation of Langmuir circulations in the Great Salt Lake. The forward facing Y junctions result from the merging of plus-and minus-signed streamwise vortices beneath the water surface (photograph courtesy of S. Monismith).

is decomposed into a large amplitude irrotational component associated with the wave field $\mathbf{u}^w$ and a smaller amplitude rotational component $\mathbf{u}$ identified with the currents; the small parameter ϵ is the wave slope ak of a typical component of the wave spectrum. Insertion of Eq. (2.1) into the Navier-Stokes equations and temporally averaging over the rapid oscillations of the wave field leads to a set of perturbation equations in the small parameter ϵ. The first correction due to the presence of a wave field is a vortex force $\mathbf{u}^s \times \boldsymbol{\omega}$ (where vorticity $\boldsymbol{\omega} = \nabla \times \mathbf{u}$), a dynamic Stokes pressure (see Eq. (2.7)), and an additional advection of any scalar field $\mathbf{u}^s \cdot \nabla q$. The Stokes drift (PHILLIPS [1977])

$$\mathbf{u}^s(z) = \int^T \mathbf{u}^w \, \mathrm{d}t \cdot \nabla \mathbf{u}^w \tag{2.2}$$

which appears in these new terms is the horizontal Lagrangian motion due to the wave field; the averaging time T is long compared with the timescale of the oscillations. Building on the Craik-Leibovich framework, MCWILLIAMS and RESTREPO [1999], MCWILLIAMS, RESTREPO and LANE [2004] carry the perturbation to higher order and

show how wave-averaged effects enter into rotating systems and contribute to scalar transport. HOLM [1996] draws an analogy between the vortex force and a fluid plasma driven by a rapidly varying external electromagnetic field. At the lowest order of the perturbation, the Craik-Leibovich phase-averaged equations are equivalent to the GLM theory developed by ANDREWS and MCINTYRE [1978a]. PHILLIPS [2002] has developed a theoretical model for time-varying wave fields midway between the original Craik-Leibovich equations and the GLM.

2.2. *LES and the OBL*

In addition to surface waves, the OBL forced by wind stress and thermal stratification at large Re is unstable to small perturbations and fully developed 3D turbulence, spread over a wide spectrum of scales, fills the mixed layer. Turbulent fluxes of momentum and density modify the mean currents, the overlying stratification, and set the rate at which cooler (denser) fluid is entrained at the base of the thermocline. Given the wide range of scales in the OBL, it is not computationally feasible to explicitly resolve all scales of motion from the largest energy containing eddies down to the smallest viscous dissipation scales. The technique of LES is then a practical computational approach to modeling OBL turbulence at high Re. The basis for LES was pioneered by DEARDORFF [1980], LEONARD [1974], and LILLY [1967] for atmospheric flows, and is now widely practiced in many flavors in the geophysical and engineering communities (e.g., see MOENG and STEVENS [2000], MOENG and SULLIVAN [2002], SAGAUT [2006]).

The LES equation set is formally obtained by applying a low-pass spatial filter $\mathcal{G}(\mathbf{x}, \mathbf{x}')$ to the equations of motion in order to eliminate scales smaller than can be supported by the available grid resolution. Total and filtered (or resolved) variables $(a, \overline{a})$ are related by the linear convolution operation

$$\overline{a}(\mathbf{x}) = \int \mathcal{G}(\mathbf{x}, \mathbf{x}')\, a(\mathbf{x}')\, \mathbf{dx}', \tag{2.3}$$

and thus the small-scale [or subgrid-scale (SGS)] fluctuation $a' = a - \overline{a}$. Low-pass filtering the time tendency and advective terms in the transport equations for momentum ρu_i and scalar concentration ρq for an incompressible fluid of density ρ exposes key ideas in LES:

$$\frac{\partial u_i}{\partial t} + \frac{\partial u_j u_i}{\partial x_j} \to \frac{\partial \overline{u}_i}{\partial t} + \frac{\partial \overline{u_j u_i}}{\partial x_j} \equiv \frac{\partial \overline{u}_i}{\partial t} + \frac{\partial \overline{u}_j \overline{u}_i}{\partial x_j} + \frac{\partial \tau_{ij}}{\partial x_j} \tag{2.4a}$$

$$\frac{\partial q}{\partial t} + \frac{\partial u_j q}{\partial x_j} \to \frac{\partial \overline{q}}{\partial t} + \frac{\partial \overline{u_j q}}{\partial x_j} \equiv \frac{\partial \overline{u}_i}{\partial t} + \frac{\partial \overline{u}_j \overline{q}}{\partial x_j} + \frac{\partial Q_j}{\partial x_j}. \tag{2.4b}$$

In Eqs. (2.4) and (2.5), filtering products of total flow variables generate additional SGS momentum and scalar fluxes

$$\mathbf{T} \equiv \tau_{ij} = \overline{u_i u_j} - \overline{u}_i \overline{u}_j, \tag{2.5a}$$

$$\mathbf{Q} \equiv Q_i = \overline{q u_i} - \overline{q}\, \overline{u}_i. \tag{2.5b}$$

($\mathbf{T}$, $\mathbf{Q}$) are spatially averaged correlations and represent the effect of small-scale velocity $\mathbf{u}' = \mathbf{u} - \overline{\mathbf{u}}$ and scalar $q' = q - \overline{q}$ fluctuations on resolved scales. The trace of $\mathbf{T}$ is the SGS energy $e = \tau_{ii}/2$, see Eq. (2.16). These SGS fluxes are unknown and must be modeled in terms of filtered variables to close the equation set. A detailed discussion of filtering as it applies to LES and its implication for SGS modeling is given by POPE [2000, Chapter 13].

The Craik-Leibovich equations are a subset of the full OBL dynamics including waves, currents, and turbulence under the decomposition and assumptions outlined in Section 2.1. At the lowest order of the perturbation expansion, the water surface remains flat and there is no feedback from the currents (or turbulence) to the wave field, i.e., for the purposes of modeling, the wave field is externally imposed on the OBL. To include the vortex force in the LES framework, the filtering is now interpreted as a combination of low-pass spatial and temporal filters (MCWILLIAMS, SULLIVAN and MOENG [1997]).

2.2.1. Model equations with wave effects

An example of an LES model used to examine marine turbulence in the OBL is described by MCWILLIAMS, SULLIVAN and MOENG [1997]. It is based on a conventional atmospheric LES code (MOENG [1984], SULLIVAN, MCWILLIAMS and MOENG [1996]) but modified to account for surface wave effects: the flow model is incompressible Boussinesq equations with a single-point, second-moment turbulent kinetic energy (TKE) closure subgrid-scale parameterization and a flat upper surface. The added wave effects are the vortex force and Lagrangian mean advection associated with Stokes drift and a wave-averaged increment to the pressure that arise through conservative wave-current interaction. Additional acceleration and energy generation due to nonconservative wave breaking can also be included in a fuller model (SULLIVAN, MCWILLIAMS and MELVILLE [2004, 2005]). In the limit of vanishing molecular viscosity and diffusivity, the LES equations for momentum, density (and likewise for any material scalar concentration q), and conservation of mass are the following (MCWILLIAMS, SULLIVAN and MOENG [1997]):

$$\frac{D\overline{\mathbf{u}}}{Dt} + \mathbf{f} \times \overline{\mathbf{u}} + \nabla\pi + \hat{\mathbf{z}} g \frac{\overline{\rho}}{\rho_*} + \nabla \cdot \mathbf{T} = \mathbf{u}^s \times (\mathbf{f} + \overline{\boldsymbol{\omega}}), \tag{2.6a}$$

$$\frac{D\overline{q}}{Dt} + \nabla \cdot \mathbf{Q} = -\mathbf{u}^s \cdot \nabla \overline{q}, \tag{2.6b}$$

$$\nabla \cdot \overline{\mathbf{u}} = 0, \tag{2.6c}$$

where the wave-averaged effects are isolated on the right-hand side. In Eq. (2.6), ($\overline{\mathbf{u}}$, $\overline{\boldsymbol{\omega}}$) are the resolved-scale velocity and vorticity; $\overline{\rho}$ is the density, and the generalized pressure field is

$$\pi = \frac{\overline{p}}{\rho_*} + \frac{2}{3}e + \frac{1}{2}\left[(\overline{u}_i + u_i^s)(\overline{u}_i + u_i^s) - \overline{u}_i \overline{u}_i\right]. \tag{2.7}$$

The vortex force is nondivergent and hence contributes to the right-hand side of the pressure Poisson equation (see Section 2.2.2 and MCWILLIAMS, SULLIVAN and MOENG [1997]). Note with this flux form of the momentum equations, π reduces to $\overline{p}/\rho_* + 2e/3$ in the absence of a wave field.

Other variables appearing in this equation set are the Coriolis vector $\mathbf{f} = f\hat{\mathbf{z}}$[3]; reference density, ρ_*; gravitational acceleration, g; and subgrid-scale momentum and density fluxes, $(\mathbf{T}, \mathbf{Q})$. These subgrid-scale fluxes are modeled using the eddy viscosity prescription described in MOENG [1984] and SULLIVAN, MCWILLIAMS and MOENG [1994], which implies that the principal feedback of e on $\overline{\mathbf{u}}$ and $\overline{\rho}$ occurs through subgrid-scale mixing with eddy viscosity and diffusivities $(\nu_t, \nu_q) \propto e^{1/2}$. An important pathway for energy exchange between the resolved and subgrid scales is the SGS production with wave effects

$$\mathbf{T} : (\mathbf{S} + \nabla \mathbf{u}^s) = \tau_{ij} S_{ij} + \tau_{ij} \frac{\partial u_i^s}{\partial x_j}, \tag{2.8}$$

where the resolved scale rate of strain tensor

$$\mathbf{S} \equiv S_{ij} = \frac{1}{2}\left(\frac{\partial \overline{u}_i}{\partial x_j} + \frac{\partial \overline{u}_j}{\partial x_i}\right). \tag{2.9}$$

LES has a closure problem analogous to that in Reynolds-averaged Navier-Stokes (RANS) modeling (POPE [2000]), but in LES, the SGS fluxes are stochastic variables in space and time (e.g., SULLIVAN, HORST, LENSCHOW, MOENG and WEIL [2003]). The literature on SGS models for LES is abundant and remains an active research area (e.g., see GEURTS [2001], MENEVEAU and KATZ [2000], WYNGAARD [2004]). Here a simple prescription is employed, where subgrid-scale fluxes are modeled as down gradient processes

$$\tau_{ij} - \delta_{ij}\tau_{kk} = -2\nu_t S_{ij}, \quad Q_i = -\nu_q \frac{\partial \overline{q}}{\partial x_i} \tag{2.10}$$

with eddy viscosity and diffusivity

$$\nu_t = C_k \, l \, e^{1/2} \quad \text{and} \quad \nu_q = (1 + 2l/\triangle)\nu_t. \tag{2.11}$$

The transport equation for SGS energy, needed in the eddy viscosity, is constructed from equations for the total and resolved energy, which include stratification and wave-average effects:

$$\frac{\partial e}{\partial t} = \overline{u_i \frac{\partial u_i}{\partial t}} - \overline{u}_i \frac{\partial \overline{u}_i}{\partial t}. \tag{2.12}$$

The subgrid-scale TKE equation

$$\frac{De}{Dt} + \mathbf{T} : \mathbf{S} + \mathbf{Q} \cdot \hat{\mathbf{z}} + \mathcal{E} - \nabla \cdot (2\nu_t \nabla e) = -\mathbf{u}^s \cdot \nabla e - \mathbf{T} : (\nabla \mathbf{u}^s) \tag{2.13}$$

[3] Usually the Coriolis vector is approximated as $\mathbf{f} = 2\Omega \sin\phi \hat{\mathbf{z}}$ with Ω the Earth's rotation and ϕ a local latitude. The consequences of including a horizontal component of rotation is discussed by MCWILLIAMS and HUCKLE [2006] and references therein.

contains advection, production, buoyancy, dissipation, and diffusion, terms that appear in a conventional closure formulation (DEARDORFF [1980]). Viscous dissipation is parameterized as

$$\mathcal{E} = C_\epsilon \frac{\mathrm{e}^{3/2}}{l}, \tag{2.14}$$

under the assumption that the filter cutoff scale $\triangle$ lies in the Kolmogorov inertial range where the energy spectrum function $E(\kappa) \sim \mathcal{E}^{2/3}\kappa^{-5/3}$ and κ is the wavenumber. The modeling constants based on inertial range matching (MOENG and WYNGAARD [1988]) are $(C_k, C_\epsilon) = (0.1, 0.93)$. $\triangle$ is related to the mesh spacing $\triangle\mathbf{x}$ by $\triangle = (\frac{3}{2}\triangle x\, \frac{3}{2}\triangle y\, \triangle z)^{1/3}$, where the 3/2 factor accounts for dealiasing of the solutions in a mixed finite-difference pseudospectral code. SCOTTI, MENEVEAU and LILLY [1993] shows that this estimate of $\triangle$ is acceptable even for modestly nonisotropic grid meshes. The mixing length scale $l = \triangle$ but is reduced in regions of strong stratification (DEARDORFF [1980]).

Finally, boundary conditions need to be specified to close the set of LES equations. Periodic conditions in the horizontal directions (x, y) are assumed for all variables as is customary for turbulent simulations. At the sea surface, $z = 0$, a constant surface stress $\boldsymbol{\tau}$ is typically imposed based on a measured bulk aerodynamic formula using winds U_a at a nominal height $z = 10\,\mathrm{m}$ above the water (LIU, KATSAROS and BUSINGER [1979]). Consistent with a flat upper surface, the vertical velocity $\overline{w} = 0$ at $z = 0$. The mixed layer is bounded below by a stably stratified thermocline, and then the appropriate boundary conditions as $z \to -\infty$ are $\overline{\mathbf{u}} = 0$ along with negligible momentum fluxes and a zero vertical gradient for SGS energy.

The LES model described above includes wave effects without actually resolving the oscillatory wave motions themselves. The implicit assumption is that the wave quantities are unaffected by the currents; this is explicit in the particular rules specified for $\mathbf{u}^s$ (see Section 2.3). Further simplifications are the neglect of the positive buoyancy caused by air entrained into bubbles in breaking waves and the assumption of a linearized equation of state where density is proportional to temperature θ.

2.2.2. *Numerical techniques*

The numerical method used to integrate the coupled set of Eqs. (2.6a)–(2.6c) is based on the scheme described by MOENG [1984], SULLIVAN, MCWILLIAMS and MOENG [1996]. The equation set is advanced in time in physical space using a pseudospectral approximation for all horizontal derivatives and a second-order finite differencing scheme in the vertical. Dealiasing of the solutions (e.g., CANUTO, HUSSAINI, QUARTERONI and ZANG [1988]) is performed using the "2/3-rule" with a sharp spectral cutoff filter. The time-stepping of all variables $(\overline{\mathbf{u}}, \overline{q}, e)$ is a low-storage, third-order Runge-Kutta scheme (SULLIVAN, MCWILLIAMS and MOENG [1996]) using a fractional step method. The code adjusts the time-step dynamically based on a fixed Courant-Fredrichs-Lewy number.

Since the wave influences on deforming the oceanic free surface are averaged out in this formulation, the wave effects on pressure enter indirectly into the model since π is calculated by solving the pressure-Poisson equation derived from taking the divergence of the momentum equation and applying the incompressibility condition. This is

the standard method of determining the pressure with numerical techniques based on fractional step methods (e.g., see ARMFIELD and STREET [1999]). The elliptic Poisson pressure equation is solved by first applying a 2D fast fourier transform (FFT) in horizontal planes followed by tridiagonal matrix inversion in the vertical direction. This step is accomplished with a custom-built parallel algorithm using forward and backward matrix transposes.

Parallelization is accomplished using a combination of the message passing interface (MPI) and OpenMP programming paradigms (AOYAMA and NAKANO [1999], CHANDRA, DAGUM, KOHR, MAYDAN, MCDONALD and MENON [2001]). The vertical domain is first naturally decomposed into finite-size vertical blocks using MPI on each computational node. To achieve a high degree of parallelization, the OpenMP implementation is then applied to each node. An advantage of this approach is that a standard FFT algorithm can be applied since memory on a node (vertical block) is shared, i.e., FFTs are not split across processors.

2.3. *Simulations with wave influences*

External inputs to an ocean LES are the wind stress (usually parameterized in terms of the wind speed) and the wave spectrum, the latter is needed to determine the profile of Stokes drift. Given the atmospheric wind, past LES typically estimate $u^s(z)$ from a single monochromatic wave (MCWILLIAMS, SULLIVAN and MOENG [1997], NOH, MIN and RAASCH [2004], SKYLLINGSTAD and DENBO [1995]) chosen to match a dominant component in the wave spectrum. In general, winds and waves are not in local equilibrium owing to numerous influences, e.g., shallow water in coastal regions, swell propagation from distant storms and fronts, and variable light winds; hence, the wave field is not easily categorized in terms of a few bulk atmospheric parameters. However, a cleanly posed problem is the special situation of wind-wave equilibrium where the wave field is well developed and its energy density, peak frequency, and spectral shape are changing slowly with time. In this case, the empirically determined wave height spectrum (ALVES, BANNER and YOUNG [2003])

$$F\left(\frac{\sigma}{2\pi} = n\right) = \frac{\alpha_w\, g^2}{(2\pi)^4\, n^4\, n_p} \exp\left[-\left(\frac{n}{n_p}\right)^{-4}\right] \tag{2.15}$$

depends only on wind speed. Here n is the frequency, σ is the radial frequency, α_w is the "Phillips constant", g is the gravitational acceleration, and n_p is the peak in the spectrum related to a reference atmospheric wind $U_a(z = 10\,\mathrm{m})$ by

$$\nu = \frac{n_p\, U_a}{g}. \tag{2.16}$$

Empirical constants that appear in the above expressions are $\nu \approx 0.123$ and $\alpha_w \approx 6.15 \times 10^{-3}$. MCWILLIAMS and RESTREPO [1999, Eq. (62)] show how to incorporate a full wave spectrum in the estimate of u^s. The Stokes drift is

$$u^s(z) = \frac{2}{g} \int_0^\infty F(\sigma)\, \sigma^3 \exp\left[\frac{2\sigma^2 z}{g}\right] \mathrm{d}\sigma, \tag{2.17}$$

which can be evaluated by simple numerical quadrature. Here, we show typical results from an ocean LES with the Stokes profile given by Eq. (2.21) with grid resolution (300, 300, 128) in a computational domain of (300, 300, 110) m at a wind speed $U_a = 15\,\mathrm{ms}^{-1}$ and water friction velocity $u_* = 0.0187\,\mathrm{ms}^{-1}$. A comparison with a baseline calculation with no Stokes drift highlights the important role the surface wave field plays in the OBL.

The snapshot of the flow field, given in Fig. 2.2, provides striking evidence of the impact of the vortex forces as vigorous coherent structures develop near the ocean surface. Animations show the persistence of the bulk flow pattern as well as temporal and spatial mergers of the structures. Closer inspection of the $w-$contours identifies alternating regions of strong downwelling and upwelling elongated in the prevailing wind direction (in this LES, the surface wind stress $\boldsymbol{\tau}$ is aligned with $\hat{\mathbf{x}}$). The patterns in Fig. 2.2 are signatures of Langmuir circulations which are readily observed in Fig. 2.1. McWilliams, Sullivan and Moeng [1997], using lower resolution LES, identify and track the lifecyles of these Langmuir cells. They find the Langmuir cells merge and broaden laterally with increasing depth, i.e., the cells are not confined to the near surface region.

The coherent structures generated by the surface wave field play important roles in the overall mixing of the OBL. They enhance the turbulence variances (especially near the surface), alter the important vertical fluxes of momentum and scalars, and modify the mean current and thermal structure of the OBL. This is borne out in the

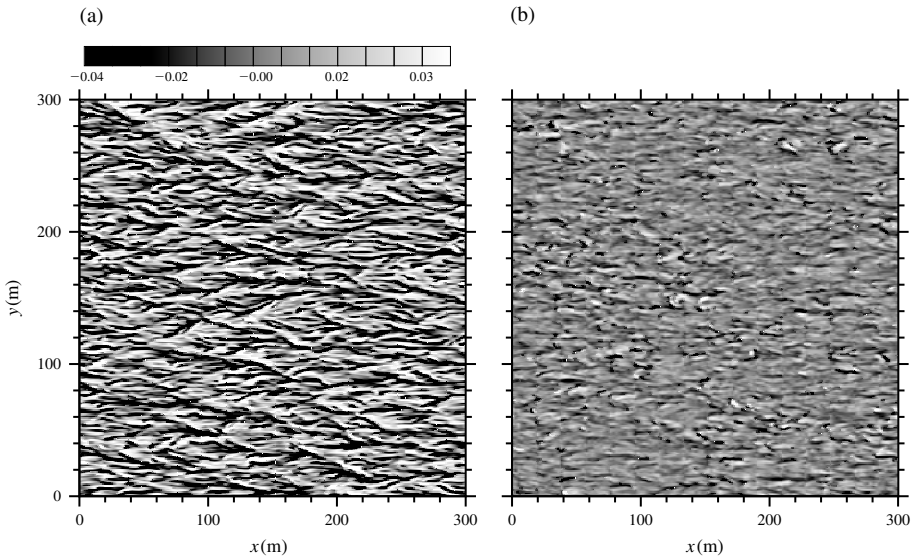


FIG. 2.2 Visualization of vertical velocity field $\overline{w}$ at $z = -1.14\,\mathrm{m}$ in $x - y$ planes for LES with (a) and without (b) vortex forces for a wind speed $U_a = 15\,\mathrm{ms}^{-1}$. Note the formation of streaky structures, signatures of Langmuir circulations, roughly aligned with the $x-$direction in the presence of wave-averaged effects. The ranges of the gray-scale color bar (in ms^{-1}) are identical between the two panels.

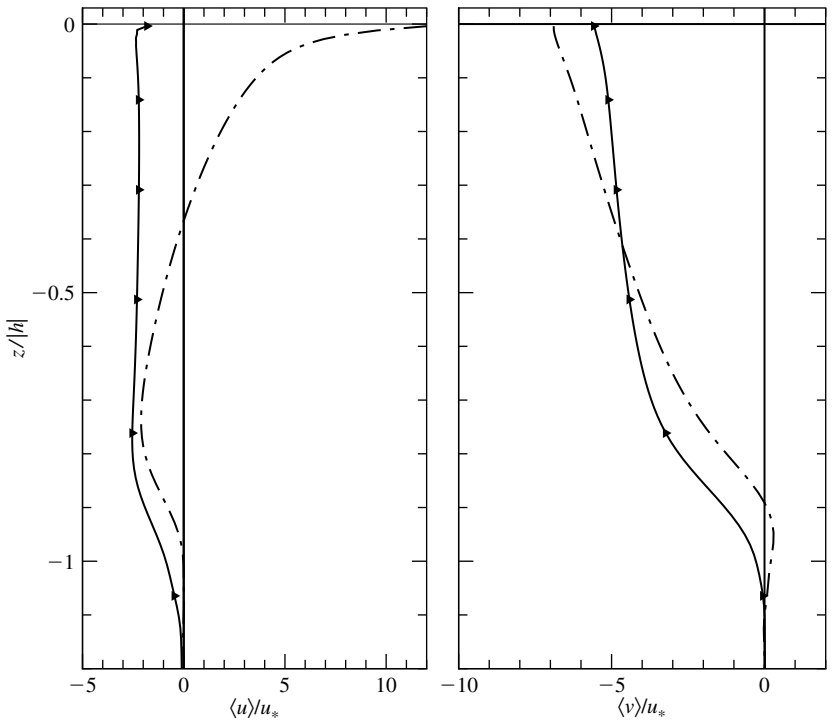


FIG. 2.3 Vertical profiles of the mean streamwise $\langle u \rangle$ and spanwise $\langle v \rangle$ currents normalized by friction velocity u_*. Results with wave-average effects are shown by solid lines, and results with no waves are indicated by dash-dot lines. The mixed layer depth $h \approx -51$ m.

vertical profiles of mean currents, TKE, and dissipation shown in Figs. 2.3–2.5. The Langmuir cells act as nonlocal mixing agents, qualitatively similar to plumes in the atmospheric boundary layer, e.g., the mean current profile $\langle u \rangle$ is uniform over almost the entire depth of the OBL. Modification of the current profiles at $z = h$, compared to LES with no wave effects, alters the entrainment of cool water from below the thermocline. This is one of the wave processes not included in the majority of simple 1D parameterizations of the OBL. In a broad sense, the wave field elevates the injection and dissipation of turbulence energy as shown in Figs. 2.4 and 2.5. Near the surface, the spanwise rotation of the Langmuir cells increases the spanwise variance $\langle v'^2 \rangle$ and the important vertical variance $\langle w'^2 \rangle$ by factors of at least 5 and 4, respectively. This leads to the overall enhancement of TKE and a mandatory increase in dissipation. Field observations (TERRAY, DONELAN, AGRAWAL, DRENNAN, KAHMA, WILLIAMS, HWANG and KITAIGORODSKII [1996]) of dissipation confirm at least a factor of 10 increase in the near surface dissipation compared to classical wall-layer scaling. Much of this enhanced dissipation results from breaking waves not considered here (SULLIVAN, MCWILLIAMS and MELVILLE [2005]).

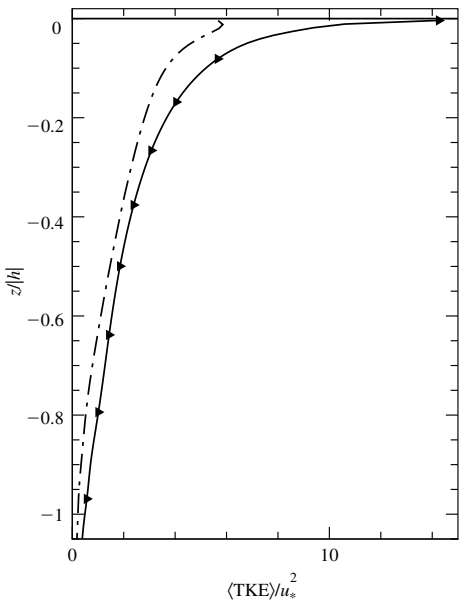


FIG. 2.4 Vertical profiles of total turbulent kinetic energy resolved plus subgrid-scale contributions $(\overline{u}_i\overline{u}_i/2 + e)$. Simulations with and without wave effects are indicated by solid and dash-dot lines, respectively.

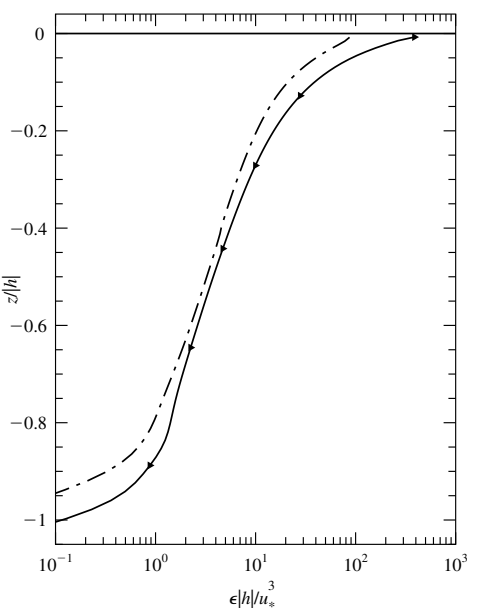


FIG. 2.5 Vertical profiles of dissipation profiles below the ocean surface for large eddy simulations with and without wave effects (solid and dash-lines, respectively). Dissipation $\mathcal{E}$ is estimated from the subgrid-scale model given by Eq. (2.14).

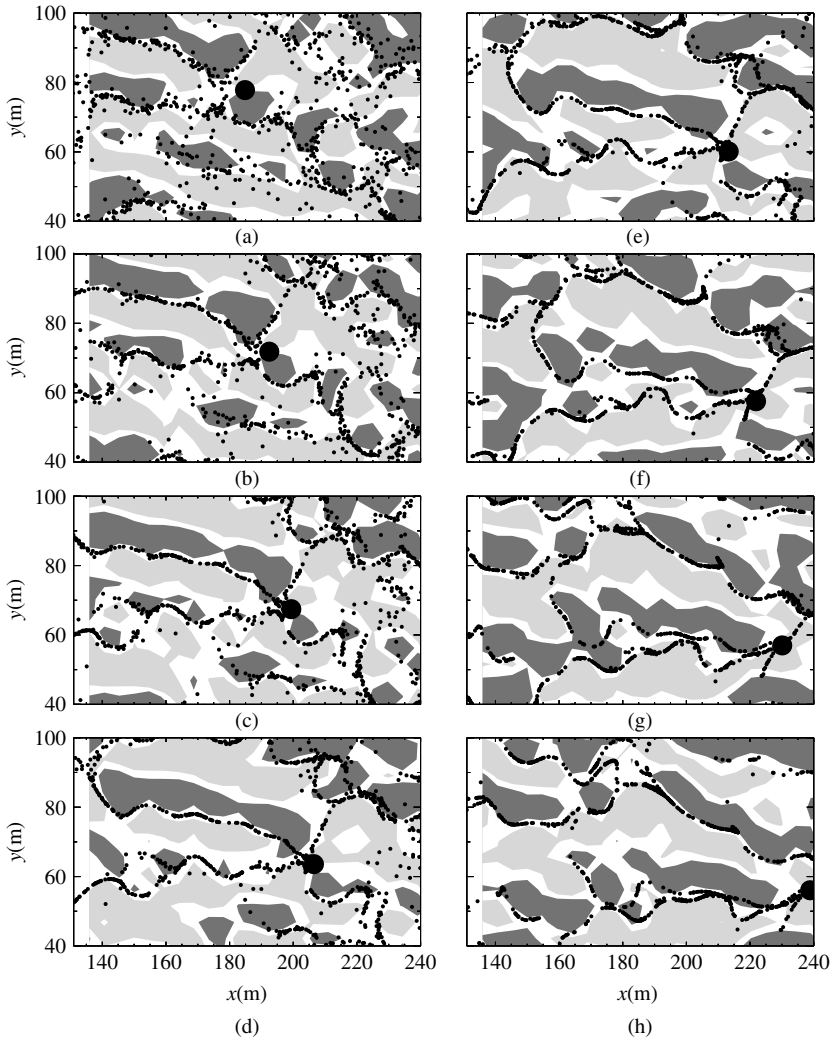


FIG. 2.6 Temporal and spatial variation of surface particles and near-surface streamwise vorticity contours in LES with vortex forces. Positive and negative signed contours at $z = -1.8\,\mathrm{m}$ are indicated by (dark, light) shading. Surface particles are denoted by small solid dots. The position of a single particle that is near the early-time Y junction of two particle lines is indicated by large solid dot; note the merger that occurs to the left of this particle. Time increases from top to bottom then left to right (MCWILLIAMS, SULLIVAN and MOENG [1997]).

The time and spatial evolution of surface particles, in Fig. 2.6, provides evidence for the formation of 3D Langmuir circulations near the sea surface. The patterns show convergence at downwelling lines that meander laterally at the top of the ocean surface. The lines join at forward oriented Y junctions in a manner striking similar to the

observations in Fig. 2.1. This results from lateral coalescence of streamwise positive- and negative-signed vorticity.

Turbulence resolving simulations and a novel asymptotic analysis have helped to reveal the coupling between physical processes of disparate timescales in the OBL. Their combination illustrates the important role of coherent structures in turbulent flows and provides new insights into connections between waves and vortical motions. The LES model outlined above relies heavily on the framework of the Craik-Leibovich equations which naturally begs the question as to how well the assumptions in Section 2.1 hold. The DNS of ZHOU [1999] suggest that, at least for a single monochromatic wave the Craik-Leibovich assumptions hold up quite well. More complete simulations of the turbulent OBL with a resolved spectrum of surface waves and dissipative effects due to breaking are a future challenge for simulations owing to the intense computational demands of widely separated time and space scales.

3. The dynamo problem

3.1. From fluids to MHD

The dynamics of magnetic fields (see for an introduction MOFFATT [1978], PARKER [1979], SOWARD [1983]) has many similarities with Navier-Stokes turbulence: recall the Batchelor analogy between vorticity and induction, both undergoing stretching through velocity gradients. Hence, one expects growth of a magnetic seed as one observes growth of the vorticity in a fluid. The question then arises as to whether the dynamics of fluids with magnetic fields is the same as in the absence of magnetic field. For example, does the coupling to a magnetic field actually break the Kolmogorov scaling, or do we observe the same statistical properties in the fluid case and in the MHD case? What kind of structures prevail at small scale, e.g., are there vortex filaments? Some of the answers to these questions may stem from the numerous observations of our environment and from numerical simulations and experiments as well. Magnetic fields are embedded in a medium which is turbulent because Reynolds numbers are large for geophysical flows. Electric fields and ionospheric currents play a dynamic role in the evolution of the atmosphere above 100 km, and the input of energy from the magnetosphere during magnetic storms can affect the thermosphere and ionosphere on global scales. The solar wind and the solar convection zone can both be viewed as a laboratory for turbulence, with numerous satellites that observe its properties (in the heliospheric case, see BRUNO and CARBONE [2005] for a review; in the solar case, see PARKER [1979] for the fundamentals, and BRANDENBURG and SUBRAMANIAN [2005] for an exhaustive review). In the solar wind, the magnetic energy spectrum is close to Kolmogorovian $E_{\mathrm{m}}(k) \sim k^{-1.70}$ (MATTHAEUS and GOLDSTEIN [1982]), whereas in the magnetosphere of Jupiter, the spectra (determined using Galileo spacecraft data) are anisotropic (SAUR, POLITANO, POUQUET and MATTHAEUS [2002]), viz. $\sim k_{\perp}^{-2}$ (where $\mathbf{k}_{\perp}$ refers to wavevectors perpendicular to $\mathbf{B_0}$, the uniform field).

Plasmas in stars, the interplanetary medium, the Earth's magnetopause, and conducting fluids in planetary cores are often studied using the MHD approximation. The

MHD equations derive from Maxwell's equations for subrelativistic velocities, hence the displacement current can be neglected. They consist of Eq. (1.1), with the force $\mathbf{F}$ in this section given by $\mathbf{F} = \mathbf{j} \times \mathbf{b} + \mathbf{F}'$, where $\mathbf{j} \times \mathbf{b}$ is the Lorentz force and $\mathbf{F}'$ the external mechanical forcing. Under these approximations, from the Maxwell equations, the evolution of the magnetic induction is given by

$$\frac{\partial \mathbf{b}}{\partial t} = \eta \nabla^2 \mathbf{b} + \nabla \times (\mathbf{v} \times \mathbf{b}), \tag{3.1a}$$

$$\nabla \cdot \mathbf{b} = 0. \tag{3.1b}$$

Here, $\mathbf{b} = \mathcal{B}/\sqrt{\mu_0 \rho_0}$ is the Alfvén velocity, $\mathcal{B}$ is the magnetic induction, μ_0 the permeability, and η the magnetic diffusivity. The condition $\nabla \,.\, \mathbf{b} = \mathbf{0}$ indicates the lack of magnetic monopoles. Since displacement currents in the Maxwell equations are neglected, the current density in dimensionless units is $\mathbf{j} = \nabla \times \mathbf{b}$. Equation (1.1) with the Lorentz force, together with Eq. (3.1) support Alfvén waves with $\mathbf{u} = \pm \mathbf{b}$. Note that these waves are also exact nonlinear solutions of the MHD equations.

Equation (3.1) is called the induction equation and expresses simple physics: magnetic field lines in a conducting fluid behave as material lines, which are advected by the velocity field. From Eq. (3.1), a theorem equivalent to Kelvin's circulation theorem for neutral fluids can be proven,

$$\frac{\mathrm{d}\Phi}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t} \iint_S \mathbf{b} \cdot \mathrm{d}\mathbf{S} = \iint \eta \nabla^2 \mathbf{b} \cdot \mathrm{d}\mathbf{S}. \tag{3.2}$$

This result, known as Alfvén's theorem, implies that for the ideal case ($\eta = 0$) the magnetic flux Φ through a surface S advected by the fluid velocity $\mathbf{v}$ is conserved. As a result, if the fluid motions stretch the surface S and change its area, the magnetic energy has to increase to maintain the constant flux. This process is a fundamental component of dynamo action as will be discussed later.

In addition, in 3Ds and under proper boundary conditions (e.g., periodic boundary conditions), the MHD equations have three ideal ($\nu = 0 = \eta$) quadratic invariants: the total energy

$$E = \frac{1}{2} \int \left(v^2 + b^2 \right) \mathrm{d}^3 x, \tag{3.3}$$

the cross helicity

$$H_C = \frac{1}{2} \int \mathbf{v} \cdot \mathbf{b} \, \mathrm{d}^3 x, \tag{3.4}$$

and the magnetic helicity

$$H_M = \frac{1}{2} \int \mathbf{A} \cdot \mathbf{b} \, \mathrm{d}^3 x, \tag{3.5}$$

where $\mathbf{A}$ is the vector potential, $\mathbf{b} = \nabla \times \mathbf{A}$. While in MHD turbulence the total energy has a direct cascade to small scales (as in hydrodynamic turbulence), the magnetic helicity can display an inverse cascade. In an inverse cascade process (Alexakis, Mininni and Pouquet [2006], Brandenburg [2001], Frisch, Pouquet, Léorat and Mazure

[1975], HOSSAIN, MATTHAEUS and MONTGOMERY [1983], LILLY [1969], MENEGUZZI, FRISCH and POUQUET [1981], POUQUET and PATTERSON [1978], POUQUET, FRISCH and LÉORAT [1976]), excitations externally injected at the small scales are preferentially transferred to the larger scales and pile up there, creating coherent macroscopic magnetic structures at large scales where none were initially present. The difficulties associated with inverse cascades in LES models of MHD turbulence will be discussed in Section 3.2.

As in Section 1.2, we can define a magnetic Reynolds number as $\mathsf{Rm} = \mathsf{U}_0\mathsf{L}_0/\eta$. The ratio of the mechanic Reynolds number Re to the magnetic Reynolds number Rm is the magnetic Prandtl number Pm. Plasmas in stellar interiors, the interplanetary medium, or the magnetopause and conducting fluids in planetary cores are characterized by mechanic and magnetic Reynolds numbers much larger than one (see e.g., LEAMON, MATTHAEUS, SMITH, ZANK, MULLAN and OUGHTON [2000], MATTHAEUS, DASSO, WEYGAND, MILANO, SMITH and KIVELSON [2005] for an estimation of Re in the solar wind). In addition, the magnetic Prandtl number Pm is in general different than one. As a few examples, the magnetic Prandtl number in the solar convective region is estimated to be $\mathsf{Pm} \approx 10^{-5} - 10^{-6}$ (PARKER [1979]), and in the Earth's core, $\mathsf{Pm} \approx 10^{-5}$. Liquid sodium experiments are also characterized by small values of Pm (NOGUCHI, PARIEV, COLGATE, BECKLEY and NORDHAUS [2002], PÉTRÉLIS, BOURGOIN, MARIÉ, BURGUETE, CHIFFAUDEL, DAVIAUD, FAUVE, ODIER and PINTON [2003], SISAN, SHEW and LATHROP [2003], SPENCE, NORNBERG, JACOBSON, KENDRICK and FOREST [2006]). On the other hand, the magnetic Prandtl number in the interplanetary medium or the magnetopause is estimated to be much larger than one (HAUGEN, BRANDENBURG and DOBLER [2003, 2004], SCHEKOCHIHIN, BOLDYREV and KULSRUD [2002], SCHEKOCHIHIN, COWLEY, TAYLOR, MARON and MCWILLIAMS [2004], SCHEKOCHIHIN, HAUGEN, BRANDENBURG, COWLEY, MARON and MCWILLIAMS [2005]).

The large values of the mechanic and magnetic Reynolds numbers forbid simulations using realistic values for Re, Rm, and Pm. As an example, simulations of the geodynamo (GLATZMAIER and ROBERTS [1995, 1996], KONO, and ROBERTS [2002], ROBERTS and GLATZMAIER [2001]) or the solar convective region (CATTANEO [1999]) are often done for $\mathsf{Pm} \sim 1$, and values of Re and Rm are small compared with the values observed in nature. The development of subgrid models and LES for MHD turbulence could alleviate this situation, although difficulties inherent to the MHD equations make this development more complex than in the hydrodynamic case.

In this section, we review recent attempts to obtain LES for MHD flows. We discuss some of the models with particular emphasis in the Lagrangian-average MHD (LAMHD) equations (HOLM [2002a,b], MININNI, MONTGOMERY and POUQUET [2005b,c], MONTGOMERY and POUQUET [2002]), and a LES proposed for the particular regime $\mathsf{Pm} \ll 1$ (PONTY, POLITANO and PINTON [2004]).

3.2. LES for MHD

While SGS models of hydrodynamic turbulence have a rich history, models for MHD flows are still in their infancy (see e.g., MÜLLER and CARATI [2002a]). One of the main difficulties in MHD is that hypotheses often made in hydrodynamics (e.g., locality

of interactions in Fourier space) are not necessarily valid for magnetofluids. The general expression of the MHD energy spectrum is not known. And several regimes can be expected according to whether the system is mechanically or magnetically forced, whether the fields are statistically aligned or not, etc.

In Section 2.2, we discussed the expression of the SGS tensors used in some models for neutral flows. Here we will focus on the MHD case. The general problem of obtaining an LES for conducting fluids can be stated as follows. We split the fields into resolved and unresolved components,

$$\mathbf{v} = \overline{\mathbf{u}} + \mathbf{u}, \tag{3.6}$$

$$\mathbf{b} = \overline{\mathbf{B}} + \mathbf{B}. \tag{3.7}$$

Here, $\mathbf{u}$ and $\mathbf{B}$ are the unresolved components of the total fields $\mathbf{v}$ and $\mathbf{b}$. The resolved components $\overline{\mathbf{u}}$ and $\overline{\mathbf{B}}$ are defined through a filter; then $\overline{\mathbf{u}} = \mathcal{G}_v * \mathbf{v}$ and $\overline{\mathbf{B}} = \mathcal{G}_b * \mathbf{b}$, where $*$ denotes a convolution. The inverse of these expressions are given in terms of differential operators. In traditional LES, a spatial filter is used. For convenience, we will consider differential filters, i.e., $\mathcal{G}_v^{-1}$ and $\mathcal{G}_b^{-1}$ are linear differential operators. Based on the properties of the Navier-Stokes equations, GERMANO [1986a,b] and SAGAUT [2006] recommend the use of elliptic differential operators to filter in the spatial domain. The same considerations apply to the MHD case.

The MHD equations for the resolved fields can be written as

$$\frac{\partial \overline{\mathbf{u}}}{\partial t} = -\overline{\mathbf{u}} \cdot \nabla \overline{\mathbf{u}} + \overline{\mathbf{J}} \times \overline{\mathbf{B}} - \nabla \overline{P} + \nu \nabla^2 \overline{\mathbf{u}} - \nabla \cdot \tau_u, \tag{3.8}$$

$$\frac{\partial \overline{\mathbf{B}}}{\partial t} = \nabla \times \left(\overline{\mathbf{u}} \times \overline{\mathbf{B}} \right) + \eta \nabla^2 \overline{\mathbf{B}} - \nabla \cdot \tau_b, \tag{3.9}$$

where $\overline{\mathbf{J}} = \nabla \times \overline{\mathbf{B}}$, and the tensors τ_v and τ_b are the Leonard subgrid-stress (SGS) tensors (LEONARD [1974], SAGAUT [2006]) given by

$$\tau_v = \overline{\mathbf{v}\mathbf{v}} - \overline{\mathbf{u}}\,\overline{\mathbf{u}} - \left(\overline{\mathbf{b}\mathbf{b}} - \overline{\mathbf{B}}\,\overline{\mathbf{B}} \right), \tag{3.10}$$

$$\tau_b = \overline{\mathbf{v}\mathbf{b}} - \overline{\mathbf{u}}\overline{\mathbf{B}} - \left(\overline{\mathbf{b}\mathbf{v}} - \overline{\mathbf{B}}\overline{\mathbf{u}} \right). \tag{3.11}$$

The problem of obtaining an LES for MHD reduces to find expressions for these two subgrid-stress tensors based on the dynamics of the resolved scales and the knowledge of the statistical properties of the unresolved turbulent scales.

Perhaps the simplest example of the problems found when trying to derive expressions for these tensors in MHD can be explained considering mean field theory (MFT) (KRAUSE and RAEDLER [1980]). In MFT, only the induction equation is studied (with the velocity field considered given) to obtain an equation for the evolution of the mean magnetic field $\overline{\mathbf{B}}$. The magnetic and velocity fields are decomposed as in Eqs. (3.6) and (3.7), where the overbars are defined as averages that satisfy Taylor's hypotheses and interpreted as large-scale fields, while the fields $\mathbf{u}$ and $\mathbf{B}$ are considered to be turbulent small-scale

fluctuations with zero mean. After using first-order approximations to close the equations for the mean magnetic field, the following equation is obtained:

$$\frac{\partial \overline{\mathbf{B}}}{\partial t} = \nabla \times \left(\overline{\mathbf{U}} \times \overline{\mathbf{B}} + \alpha \overline{\mathbf{B}} \right) + \beta \nabla^2 \overline{\mathbf{B}}. \tag{3.12}$$

Here, $\alpha = \tau \overline{\mathbf{u} \cdot \nabla \times \mathbf{u}}/3$ and $\beta = \tau \overline{u^2}/3$, where τ is a characteristic timescale of the small-scale fluctuations (the magnetic diffusivity η, assumed very small, was dropped against the transport coefficient β). These simple expressions are only valid when the magnetic energy is small compared with the kinetic energy. General nonlinear expressions were obtained in Pouquet, Frisch and Léorat [1976] using the eddy-damped quasi-normal Markovian (EDQNM) closure model (Lesieur [1997], Orszag [1977]). While the coefficient β describes an enhanced diffusivity due to the turbulent fluctuations, the term proportional to α is a source of magnetic energy in the large scales due to helical fluctuations in the small-scale velocity field. In more detailed models (see, e.g., Pouquet, Frisch and Léorat [1976]), this can be interpreted as an inverse cascade of magnetic helicity. The strong impact of the small scales on the large scales in MHD turbulence makes derivations of LES more complex.

In this context, some LES have been derived using closure theories. Also using the EDQNM closure and assuming both the kinetic and magnetic energy spectra follow a $k^{-5/3}$ power law in the inertial range, Zhou, Schilling and Ghosh [2002] derived expressions for eddy damping and backscatter. Nonhelical and statistically stationary turbulence was assumed. Verma [2004] and Verma and Kumar [2004] used renormalization theory to obtain turbulent diffusivities and viscosities, and then applied the expressions as effective coefficients to construct an LES; backscatter of energy was neglected. Yoshizawa [1987, 1990] used multiscale modifications to the direct interaction approximation (Kraichnan [1959], McComb [1990]) to compute effective transport coefficients and derive an LES. While the LES in Yoshizawa [1987] is of the Smagorinsky type (Leonard and cross terms were neglected, and as a result the LES has no backscatter of energy from the subgrid scales to the resolved scales; see Clark, Ferziger and Reynolds [1979], Deardorff [1970], Smagorinsky [1963]), in Yoshizawa [1990], extensions were considered to take into account cross terms and electromotive forces due to small-scale fluctuations.

Instead of using spectral closures, another approach used to derive expressions for the SGS tensors in MHD turbulence is to extend models previously used in LES of hydrodynamic turbulence, in many cases based on dimensional arguments. The drawback with this approach is that our knowledge of the statistical properties of small scales in MHD is more limited, and we lack the experiments to validate some of the hypotheses made. Theobald, Fox and Sofia [1994] derived a purely resistive LES model for MHD using analogous expressions to SGS models of hydrodynamic turbulence. Note that a resistive SGS model cannot capture dynamo action by small-scale (unresolved) turbulent fluctuations (e.g., the α coefficient in MFT). Also based on the idea of an effective eddy viscosity, Agullo, Müller, Knaepen and Carati [2001] and Müller and Carati [2002a] derived an LES for MHD and compared the results against simulations of free decaying 3D turbulence without helicity. The expressions in these models are based on

the Smagorinsky model. Later, MÜLLER and CARATI [2002b] extended these dissipative models to also consider the case in which the SGS tensors can be a source of energy in the resolved scales, using the change in the local cross helicity to define the tensors. The idea is that the MHD equations favor configurations in which the velocity and the magnetic field are parallel or antiparallel, a process known as Alfvenization. Taking into account this process, results representing an improvement over the approach of AGULLO, MÜLLER, KNAEPEN and CARATI [2001] and MÜLLER and CARATI [2002a] were obtained for the decay of nonhelical MHD turbulence. Note, as also discussed in Sections 2.1 and 2.2.1, the importance of capturing the effect of the waves in the flow when doing SGS modeling (in this case, the alignment of the velocity and magnetic fields is naturally associated to Alfvén waves in MHD flows). The expressions for the subgrid tensors in this model are

$$\tau_v \approx -2\nu_{\text{eff}}\mathbf{S}_v, \quad \tau_b \approx -2\eta_{\text{eff}}\mathbf{J}, \tag{3.13}$$

with $\mathbf{S}_v = (\nabla\overline{\mathbf{u}} + \nabla\overline{\mathbf{u}}^T)/2$ and $\mathbf{J} = (\nabla\overline{\mathbf{B}} - \nabla\overline{\mathbf{B}}^T)/2$. The turbulent viscosity ν_{eff} and diffusivity η_{eff} are given by

$$\nu_{\text{eff}} = C\Delta^2 \left|\mathbf{S}_v : \mathbf{S}_b\right|^{1/2}, \tag{3.14}$$

$$\eta_{\text{eff}} = D\Delta^2 \text{sign}\left(\overline{\mathbf{J}}\cdot\overline{\boldsymbol{\omega}}\right)\left|\overline{\mathbf{J}}\cdot\overline{\boldsymbol{\omega}}\right|^{1/2}. \tag{3.15}$$

Here, $\mathbf{S}_b = (\nabla\overline{\mathbf{B}} - \nabla\overline{\mathbf{B}}^T)/2$, $\overline{\mathbf{J}} = \nabla\times\overline{\mathbf{B}}$, and $\overline{\boldsymbol{\omega}} = \nabla\times\overline{\mathbf{u}}$. In MÜLLER and CARATI [2002b], the parameters C and D are computed using a dynamic procedure, and Δ is the filtering length on the LES. The allowed change in sign of η_{eff} following changes in the alignment of $\overline{\mathbf{u}}$ and $\overline{\mathbf{B}}$ allows for a backscatter of magnetic energy from the small (unresolved) scales to the large (resolved) scales. This can be interpreted as a simple model of the inverse cascades in MHD turbulence (see, e.g., POUQUET [1996] for review).

In the context of MHD flows at low Pm, other SGS models have been proposed. PONTY, MININNI, MONTGOMERY, PINTON, POLITANO and POUQUET [2005] used a modified LES (see PONTY, POLITANO and PINTON [2004] for details) where only the velocity field at scales smaller than the magnetic diffusion scale was modeled using a turbulent effective viscosity dependent on the wavenumber. The expression for the effective viscosity in Fourier space is based on CHOLLET and LESIEUR [1981] and is given by

$$\nu_{\text{eff}}(k,t) = 0.27[1 + 3.58(k/k_c)^8]\sqrt{E_V(k_c,t)/k_c}, \tag{3.16}$$

where $k_c = 2\pi/\Delta$ is the largest resolved wavenumber, and $E_V(k,t)$ is the kinetic energy spectrum at time t. Since the value of Rm for $\text{Pm} \ll 1$ is often moderate, the induction equation is solved directly (i.e., all scales down to the magnetic dissipation scale are explicitly solved in the model). As a result, the equations solved in the LES are

$$\frac{\partial\overline{\mathbf{u}}}{\partial t} = -\overline{\mathbf{u}}\cdot\nabla\overline{\mathbf{u}} + \overline{\mathbf{J}}\times\overline{\mathbf{B}} - \nabla p + \nu_{\text{eff}}\nabla^2\overline{\mathbf{u}}\,, \tag{3.17}$$

$$\frac{\partial\overline{\mathbf{B}}}{\partial t} = \nabla\times\left(\overline{\mathbf{u}}\times\overline{\mathbf{B}}\right) + \eta\nabla^2\overline{\mathbf{B}}\,. \tag{3.18}$$

We will discuss some applications of this SGS model in Section 3.4.4. Modifications to the model based on second-order closures of turbulence (POUQUET, FRISCH and LÉORAT [1976]) that take into account variable energy spectra are being implemented presently (BAERENZUNG, POLITANO, PONTY and POUQUET [2006]).

A similar SGS model has been used in SCHEKOCHIHIN, HAUGEN, BRANDENBURG, COWLEY, MARON and MCWILLIAMS [2005] to study dynamo action with delta correlated in time random forcing but with the expression of the effective viscosity ν_{eff} based on the Smagorinsky model. SCHEKOCHIHIN, HAUGEN, BRANDENBURG, COWLEY, MARON and MCWILLIAMS [2005] also used hyperviscosity to study the regime of $\mathsf{Pm} \ll 1$, although it should be remarked that these methods are known to give wrong growth rates for the magnetic energy in the anisotropic case (GROTTE, BUSSE and TILGNER [2000]). Also, hyperdiffusivity is known to change the saturation amplitude of the large-scale magnetic field in helical dynamos (BRANDENBURG and SARSON [2002]).

3.3. A more formal approach

The LES approaches to MHD listed in the previous section are based on traditional LES for hydrodynamic turbulent flows or on spectral closures. As previously mentioned, in a LES only the large scales are resolved, while the effect of the unresolved small scales is considered through the subgrid stress tensors τ. This is equivalent to a reduction in the number of degrees of freedom of the flow. But to have a proper description of the large scales, the large-scale equations have to be closed. The closure consists on approximating the tensors τ based on the information contained in the resolved scales. Since the unresolved scales are unknown, in LES the equations are closed based on the knowledge of hydrodynamic turbulence (e.g., the slope of the energy spectrum in the inertial range).

Our knowledge of MHD turbulence is more limited. Systematic experiments to explore MHD turbulence in the laboratory are rather recent (NOGUCHI, PARIEV, COLGATE, BECKLEY and NORDHAUS [2002], PÉTRÉLIS, BOURGOIN, MARIÉ, BURGUETE, CHIFFAUDEL, DAVIAUD, FAUVE, ODIER and PINTON [2003], SISAN, SHEW and LATHROP [2003], SPENCE, NORNBERG, JACOBSON, KENDRICK and FOREST [2006]); from simulations (ALEXAKIS, MININNI and POUQUET [2005, 2006], BRANDENBURG [2001], GHOSH, MATTHAEUS and MONTGOMERY [1998], GRAPPIN, POUQUET and LÉORAT [1983], HOSSAIN, MATTHAEUS and MONTGOMERY [1983], KINNEY, MCWILLIAMS and TAJIMA [1995], MENEGUZZI, FRISCH and POUQUET [1981], MININNI, ALEXAKIS and POUQUET [2005a], POUQUET, MENEGUZZI and FRISCH [1986], TING, MATTHAEUS and MONTGOMERY [1986]) and observations (e.g., in the solar wind, see BRUNO and CARBONE [2005]), it is clear that MHD turbulence has several regimes depending on the ratio of the several ideal invariants given above. As a result, the traditional LES approach can be expected to apply only to particular cases.

The central problem in DNS of MHD turbulence is that the flow may have a tendency to develop small scales and thin current sheets in a finite time. In hydrodynamic turbulence, several regularizations were proposed, to derive from the Navier-Stokes equations a set of equations that can be treated mathematically and prove existence and uniqueness of solutions. These regularized models have smooth solutions, and thus its gradients can

be controlled by an external parameter (e.g., a filtering length). Although they are not traditional LES, in practice, they can be used as LES to decrease the number of degrees of freedom of the system in a numerical simulation.

Several of these regularizations can be mentioned (see, e.g., CLARK, FERZIGER and REYNOLDS [1979], LERAY [1934]). An important advance in the classical method of regularization and averages for dynamical systems was the development of the GLM equations (ANDREWS and MCINTYRE [1978a,b]), where a slow and fast decomposition of the Lagrangian particle trajectory in general form was introduced. The resulting GLM equations are nonlinear and exact, but they are not closed. For the Navier-Stokes equation, a Lagrangian-averaged model, or LANS model (also called α model), was introduced as a closure of the GLM equations that is based on applying (1) Taylor's hypothesis of "frozen-in" turbulence and (2) Lagrangian averaging to the GLM decomposition in Hamilton's principle. The model equations then result directly in the Eulerian representation as a closure for GLM from the variational principle given by the Lagrangian-averaged Euler-Poincaré theorem (HOLM, MARSDEN and RATIU [1998a,b]). Note the GLM method, with a decomposition of mean and fast fluctuations around the mean, has several points in common with the LES described in Sections 2.1 and 2.2.1. The Stokes drift, given by Eq. (2.2), appears as the result of averaging horizontal Lagrangian motions over the fast oscillations. As previously mentioned, at the lowest order of the perturbation, the Craik-Leibovich phase-averaged equations and the GLM equations are equivalent.

The hydrodynamic α model (FOIAS, HOLM and TITI [2001, 2002], HOLM [2002a,b], HOLM, MARSDEN and RATIU [1998b], ILYIN and TITI [2003]) was validated against simulations of hydrodynamic turbulence in pipes and periodic boundaries in CHEN, HOLM, FOIAS, OLSON, TITI and WYNNE [1998], CHEN, HOLM, MARGOLIN and ZHANG [1999], MOHSENI, KOSOVIĆ, SHKOLLER and MARSDEN [2003]. It differs from traditional LES in that the filter is applied to the Lagrangian of the ideal fluid. As a result, the modification to the equations is conservative and Hamiltonian properties of the system (e.g., the quadratic invariants or Kelvin's circulation theorem) are preserved.

The LANS model was recently extended to the ideal MHD case (HOLM [2002a,b]), and its viscous extension was tested against DNS of MHD turbulence (MININNI, MONTGOMERY and POUQUET [2005b,c]). In short, as its hydrodynamic counterpart, the LAMHD model is derived by Lagrangian averaging ordinary MHD along particle trajectories (HOLM [2002b]). In the context of magnetoconvection, the LAMHD equations were also studied in JONES and ROBERTS [2005]. Moreover, it was shown in GRAHAM, HOLM, MININNI and POUQUET [2006] that the LAMHD equations preserve the original scaling properties of the MHD equations, and as a result, the model does not depend on an explicit assumption on the functional form of the energy spectrum in MHD turbulence. On the other hand, new problems appear as new criteria are needed to set the filtering scales to properly model the unresolved scales at a given Reynolds number.

The incompressible LAMHD equations are

$$\frac{\partial \mathbf{v}}{\partial t} = -\overline{\mathbf{u}} \cdot \nabla \mathbf{v} - \mathbf{v} \cdot \nabla \overline{\mathbf{u}}^T + \mathbf{j} \times \overline{\mathbf{B}} - \nabla \mathcal{P} + \nu \nabla^2 \mathbf{v} + \mathbf{F}', \tag{3.19}$$

$$\frac{\partial \overline{\mathbf{B}}}{\partial t} = \nabla \times \left(\overline{\mathbf{u}} \times \overline{\mathbf{B}} \right) + \eta \nabla^2 \mathbf{b}. \tag{3.20}$$

The first two terms on the right-hand side of Eq. (3.19) can be rewritten as $\overline{\mathbf{u}} \times \boldsymbol{\omega}$ (where $\boldsymbol{\omega} = \nabla \times \mathbf{v}$) as the vortex force in Section 2.1. The pressure $\mathcal{P}$ in the incompressible case ($\nabla \cdot \overline{\mathbf{u}} = \nabla \cdot \overline{\mathbf{B}} = \nabla \cdot \mathbf{v} = \nabla \cdot \mathbf{b} = 0$) is to be determined from the relevant Poisson equation. The overline denotes smoothed fields related to the unsmoothed fields by

$$\mathbf{v} = \left(1 - \alpha_v^2 \nabla^2\right) \overline{\mathbf{u}} = \mathcal{H}_v^{-1} \overline{\mathbf{u}}, \tag{3.21}$$

$$\mathbf{b} = \left(1 - \alpha_b^2 \nabla^2\right) \overline{\mathbf{B}} = \mathcal{H}_b^{-1} \overline{\mathbf{B}}. \tag{3.22}$$

The Helmholtz operator $\mathcal{H}^{-1} = 1 - \alpha^2 \nabla^2$ is elliptic. Inverting this operator, the filter can be written as a convolution

$$\overline{\mathbf{u}} = \int \frac{\mathrm{e}^{-|\mathbf{x}' - \mathbf{x}|/\alpha_v}}{4\pi\alpha_v^2 |\mathbf{x}' - \mathbf{x}|} \mathbf{v}(\mathbf{x}') \mathrm{d}^3 x' = \mathcal{H}_v * \mathbf{v}, \tag{3.23}$$

$$\overline{\mathbf{B}} = \int \frac{\mathrm{e}^{-|\mathbf{x}' - \mathbf{x}|/\alpha_b}}{4\pi\alpha_b^2 |\mathbf{x}' - \mathbf{x}|} \mathbf{b}(\mathbf{x}') \mathrm{d}^3 x' = \mathcal{H}_b * \mathbf{b}. \tag{3.24}$$

Equations (3.19) and (3.20) can also be derived using a simple filtering approach (Montgomery and Pouquet [2002]) whereby the velocity and magnetic field are filtered but not their sources (vorticity, through the Biot-Savart law, and current density, through Ampère's law). The coefficients α_v and α_b are to be interpreted as the smallest active length scale participating in the nonlinear interactions; smaller scales are swept along by the larger ones (see, e.g., Foias, Holm and Titi [2001] and references therein). In other words, α_v characterizes the correlation length between the instantaneous Lagrangian fluid trajectory and its mean (time average), while α_b is its magnetic counterpart. These two parameters need not be equal (in Section 3.4.4, we will discuss a criterion to set these to scales). The traditional MHD system corresponding to the primitive equations is obtained by setting $\alpha_v = 0$ and $\alpha_b = 0$.

As in the model discussed in Section 2.1, the LAMHD equations are conservative. However, unlike the LES discussed in Section 2.2.1, LAMHD simulations so far have been done in general without adding eddy viscosities and diffusivities. In 3Ds, the LAMHD equations have three quadratic ideal invariants ($\nu = \eta = 0$). These are the energy

$$E = \frac{1}{2} \int \left(\overline{\mathbf{u}} \cdot \mathbf{v} + \mathbf{B} \cdot \overline{\mathbf{B}}\right) \mathrm{d}^3 x, \tag{3.25}$$

the cross helicity

$$H_C = \frac{1}{2} \int \mathbf{v} \cdot \overline{\mathbf{B}} \, \mathrm{d}^3 x, \tag{3.26}$$

and the magnetic helicity,

$$H_M = \frac{1}{2} \int \overline{\mathbf{A}} \cdot \overline{\mathbf{B}} \, \mathrm{d}^3 x. \tag{3.27}$$

Also, the LAMHD equations satisfy the theorems of Kelvin and Alfvén. The Kelvin circulation theorem for the incompressible LAMHD motion involves both the smooth and unsmooth velocities,

$$\frac{\mathrm{d}}{\mathrm{d}t}\oint_{c(\overline{\mathbf{u}})}\mathbf{v}\cdot\mathrm{d}\mathbf{x}=\oint_{c(\overline{\mathbf{u}})}(\mathbf{j}\times\overline{\mathbf{B}}+\nu\nabla^2\mathbf{v})\cdot\mathrm{d}\mathbf{x}\,. \tag{3.28}$$

Hence, the Lorentz force and viscous force can each generate circulation of $\mathbf{v}$ around material loops moving with the smoothed velocity $\overline{\mathbf{u}}$. The LAMHD equations also satisfy an Alfvén's theorem. If S is a surface bounded by a closed line C, where C is advected by the smoothed field $\overline{\mathbf{u}}$, then

$$\frac{\mathrm{d}\overline{\Phi}}{\mathrm{d}t}=\frac{\mathrm{d}}{\mathrm{d}t}\iint_S\overline{\mathbf{B}}\cdot\mathrm{d}\mathbf{S}=-\iint_S\eta\nabla^2\mathbf{b}\cdot\mathrm{d}\mathbf{S}. \tag{3.29}$$

As a result, the smoothed flux $\overline{\Phi}$ is conserved in the ideal case $\eta=0$. Then, stretching of material loops by the smoothed velocity $\overline{\mathbf{u}}$ increases the amplitude of the smoothed magnetic field $\overline{\mathbf{B}}$ and the magnetic energy.

The LAMHD model modifies the nonlinear terms in the equations for ordinary MHD. By a short sequence of manipulations, we may recast the LAMHD equations into a form which is reminiscent of an LES turbulence model. In "LES form," the equations are:

$$\frac{\partial\overline{\mathbf{u}}}{\partial t}=-\overline{\mathbf{u}}\cdot\nabla\overline{\mathbf{u}}+\overline{\mathbf{J}}\times\overline{\mathbf{B}}-\nabla\overline{\mathcal{P}}'+\nu\nabla^2\overline{\mathbf{u}}-\alpha_v^2\nabla\cdot\mathcal{H}_v*\tau_v, \tag{3.30}$$

$$\frac{\partial\overline{\mathbf{B}}}{\partial t}=\nabla\times\left(\overline{\mathbf{u}}\times\overline{\mathbf{B}}\right)+\eta\nabla^2\left(1-\alpha_b^2\nabla^2\right)\overline{\mathbf{B}}, \tag{3.31}$$

where the subgrid stress tensor τ_v is given by

$$\begin{aligned}\tau_v=&\left(\nabla\overline{\mathbf{u}}\cdot\nabla\overline{\mathbf{u}}^T+\nabla\overline{\mathbf{u}}\cdot\nabla\overline{\mathbf{u}}-\nabla\overline{\mathbf{u}}^T\cdot\nabla\overline{\mathbf{u}}\right)\\&-\left(\nabla\overline{\mathbf{B}}\cdot\nabla\overline{\mathbf{B}}^T+\nabla\overline{\mathbf{B}}\cdot\nabla\overline{\mathbf{B}}-\nabla\overline{\mathbf{B}}^T\cdot\nabla\overline{\mathbf{B}}\right).\end{aligned} \tag{3.32}$$

For $\alpha_v=\alpha_b$, these equations have solutions $\overline{\mathbf{u}}=\pm\overline{\mathbf{B}}$, $\mathbf{v}=\pm\mathbf{b}$, i.e., Alfvenic solutions with the magnetic field parallel or antiparallel to the velocity field. As a result, in the LAMHD equations, the properties of Alfvenization in the small scales are obtained directly from the original MHD equations without the need of extra hypotheses. Moreover, τ_v vanishes if the system is in an Alfvenic state. This is to be expected in a SGS tensor for MHD since for aligned fields the nonlinear terms in the MHD equations are zero, and as a result, there is no transfer of energy to the small scales.

A few more properties of the tensor τ_v in the LAMHD equations are worth mentioning. The first parenthesis on the right-hand side of Eq. (3.32) is the hydrodynamic contribution to the SGS tensor, associated with the nonlinear term $\mathbf{v}\cdot\nabla\mathbf{v}$ in the momentum equation. The second parenthesis on the right-hand side of Eq. (3.32) is the magnetic contribution due to the Lorentz force $\mathbf{j}\times\mathbf{b}$ in the unresolved scales. Moreover, the first term in each parenthesis corresponds to the tensor-diffusivity model of Leonard [1974]. This term is generic (Winckelmans, Wray, Vasilyev and Jeanmart [2001]): for all regular symmetric filters with a nonzero second-order moment, this term is always the first

term of the reconstruction expansion for the SGS tensor. However, although this term gives a significant contribution to τ_v, it is incomplete and extra terms are needed in the reconstruction of the SGS tensor if one is to recover sturctural properties of the equations. As an example, in the hydrodynamic version of Eq. (3.32) ($\mathbf{b} = \overline{\mathbf{B}} = 0$), the three terms in the first parenthesis are needed for the Kelvin circulation theorem to hold in the subgrid model.

3.4. Some examples and applications

Here we review a few problems of interest in MHD turbulence for astrophysical and geophysical flows in which LES can help to study numerically a region of parameter space more realistic than what traditional DNS can do. We also discuss briefly some of the details of the implementation of the LAMHD equations in numerical methods. The first practical problem consists on the modeling of strong gradients in the small scales, a feature of turbulent flow often referred to as intermittency. The generation of strong gradients in MHD flows is believed to be associated with the occurrence of strong reconnection events in which magnetic field topology changes and magnetic energy is dissipated, a phenomenon of relevance in the solar corona, the solar wind, and the magnetosphere. Only the LAMHD equations are discussed in this context. The second problem we review is the generation of magnetic fields by dynamo action in media where the kinematic viscosity is much smaller than the magnetic diffusivity, a problem of relevance for the Earth's core. For this problem, two different subgrid models are considered.

3.4.1. Numerical methods

The LAMHD equations (as well as LES based on spectral filters) are easy to implement in spectral or pseudospectral methods. As an example, in Fourier based pseudospectral methods, the Helmoltz differential operator in Eqs. (3.21) and (3.22) can be inverted to obtain

$$\hat{\overline{\mathbf{u}}}(k) = \frac{\hat{\mathbf{v}}(k)}{1 + \alpha_v^2 k^2}, \tag{3.33}$$

$$\hat{\overline{\mathbf{B}}}(k) = \frac{\hat{\mathbf{b}}(k)}{1 + \alpha_b^2 k^2}, \tag{3.34}$$

where the hat denotes Fourier transformed. In this way, the filter reduces to an algebraic operation. As a result, Eqs. (3.19) and (3.20) can be solved numerically at almost no extra cost. If other numerical methods are used and the LES form of the LAMHD equations is preferred, the convolution $\mathcal{H}_v * \tau_v$ or an inversion of the Helmoltz operator is needed. This inversion requires a solution of a Poisson equation but can be circumvented, for example, by expanding the inverse of the Helmoltz operator into higher orders of the Laplacian operator (Zhao and Mohseni [2005]):

$$\left(1 - \alpha^2 \nabla^2\right)^{-1} = 1 + \alpha^2 \nabla^2 + \alpha^4 \nabla^4 + \ldots \tag{3.35}$$

3.4.2. Validation of the LAMHD equations

The LAMHD equations were validated against DNS in three paradigmatic problems of MHD turbulence: selective decay, dynamic alignment, and inverse cascades. Validations were done in 2 (Mininni, Montgomery and Pouquet [2005b]) and 3Ds (Mininni, Montgomery and Pouquet [2005c]). In 3Ds, dynamo action was also studied (Mininni [2006], Mininni, Ponty, Montgomery, Pinton, Politano and Pouquet [2005d], Ponty, Mininni, Montgomery, Pinton, Politano and Pouquet [2005]).

Selective decay and dynamic alignment correspond to free decaying simulations (no external forcing). It is well known that for decaying MHD turbulent situations, the presence of enough initial H_M or H_C can lead to a late-time state in which the ratios $|H_M/E|$ or $|H_C/E|$ can be close to maximal. The first situation, which corresponds to selective decay (Kinney, McWilliams and Tajima [1995], Ting, Matthaeus and Montgomery [1986]), leads to a late-time quasi-steady state in which the remaining energy is nearly all magnetic and is nearly all condensed into the longest wavelength modes allowed by the boundary conditions. The second situation, which corresponds to dynamic alignment (Ghosh, Matthaeus and Montgomery [1998], Grappin, Pouquet and Léorat [1983], Pouquet, Meneguzzi and Frisch [1986]), leads to a late-time quasi-steady state in which **v** and **B** are nearly parallel or antiparallel. In inverse cascades, the system is forced at intermediate scales, and the propagation of perturbations to the largest available scale in the system is observed. Dynamo action will be discussed in more detail in Section 3.4.4.

All these examples show the richness of solutions and behaviors possible in an MHD system. As previously indicated, this makes the development of LES in MHD difficult since the spectral properties and dynamics of the system depend on the forcing used or the initial conditions. The LAMHD equations were shown to capture the long-wavelength spectra in all these problems, allowing for a significant reduction of computer time and memory at the same kinetic and magnetic Reynolds numbers. We refer the reader to Mininni, Montgomery and Pouquet [2005b,c] for more details.

3.4.3. Intermittency

It was shown in Graham, Mininni and Pouquet [2005] for two-dimensional (2D) MHD flows (which has intermittency properties similar to the 3D case since the energy cascades to small scales in 2D as well as 3D because of the breaking of the conservation of vorticity through the Lorentz force) that the LAMHD equations reproduce well the intermittency of the flow as measured either by the cancellation exponent quantifying the folding of current sheets (Ott, Du, Sreenivasan, Juneja and Suri [1992], Sorriso-Valvo, Carbone, Noullez, Politano, Pouquet and Veltri [2002]) or by the anomalous exponents of structure functions (Graham, Holm, Mininni and Pouquet [2006]). Also, a form of the Kármán-Howarth theorem was proved for LAMHD, indicating that LAMHD has in scales larger than α_v and α_b the same scaling properties than the original MHD equations.

The comparison of DNS with simulations of the LAMHD equations at smaller spatial resolution showed that, if high-order statistics are to be preserved (e.g., eight-order structure functions) then the linear resolution in the LAMHD simulations (compared

with the DNS) cannot be decreased by more than a factor of 2. In order to preserve fifth- and up to sixth-order moments, the linear resolution in the LAMHD simulations can be a factor 4 smaller than in the DNS. In 2D simulations, this decreases the memory requirements by a factor of 16 and the computing time by a factor of 64. Extra gains are obtained in 3D problems. And further gains in the linear resolution can be obtained if only second-order correlators (e.g., the energy spectrum) are of interest.

3.4.4. Dynamos at low magnetic Prandtl number

Dynamo processes result from the motions of an electrically conducting fluid amplifying and maintaining a magnetic field starting from an arbitrarily small value. They have long been of interest for geophysics and astrophysics (Dikpati and Charbonneau [1999], Glatzmaier and Roberts [1995], Kono, and Roberts [2002], Krause and Raedler [1980], Moffatt [1978], Nandy and Choudhuri [2002], Parker [1979], Soward [1983]) and more recently for laboratory experiments in liquid metals (Gailitis, Lielausis, Platacis Dement'ev, Cifersons, Gerbeth, Gundrum, Stefani, Christen and Will [2001], Noguchi, Pariev, Colgate, Beckley and Nordhaus [2002], Pétrélis, Bourgoin, Marié, Burguete, Chiffaudel, Daviaud, Fauve, Odier and Pinton [2003], Sisan, Shew and Lathrop [2003], Spence, Nornberg, Jacobson, Kendrick and Forest [2006], Steiglitz and Müller [2001]).

As previously mentioned, the generation of magnetic fields in celestial bodies occurs in media for which the viscosity ν and the magnetic diffusivity η are vastly different. For example, in the interstellar medium, Pm has been estimated to be as large as 10^{14}, whereas in stars such as the Sun and for planets such as the Earth, it can be very low ($\mathsf{Pm} < 10^{-5}$, the value for the Earth's liquid iron core). At the same time, the Reynolds numbers Re and Rm can be very large, and the flow is highly complex and turbulent, with prevailing nonlinear effects rendering the problem difficult to address. To study this problem, several approaches have been used. One way to circumvent the difficulty of resolving a vast number of scales is to use SGS models. Here we review a recent combination of DNS, LAMHD, and traditional LES (Ponty, Politano and Pinton [2004]). The approach allowed recently the study of turbulent MHD flows down to $\mathsf{Pm} = 5 \times 10^{-3}$ (Mininni [2006]).

In a dynamo simulation, only mechanical energy is injected into the system to sustain a turbulent steady state against dissipation. The magnetic energy is amplified from a small initial magnetic field and sustained against Ohmic dissipation solely by induction due to the turbulent motions of the conducting fluid. Note that this poses a hard problem for SGS models: as the system evolves in time, it suffers a transition from hydrodynamic turbulence to fully developed MHD turbulence.

Several forcing functions can be used and were explored in the literature. We focus on a particular forcing, given by the Taylor-Green flow; it combines a well-defined structure at large scales and turbulent fluctuations at small scales. The expression of the forcing $\mathbf{F}'$ is

$$\mathbf{F}'(k_0) = 2F \begin{bmatrix} \sin(k_0\, x)\cos(k_0\, y)\cos(k_0\, z) \\ -\cos(k_0\, x)\sin(k_0\, y)\cos(k_0\, z) \\ 0 \end{bmatrix}, \tag{3.36}$$

with $k_0 = 2$ so that dynamo action is free to develop at scales larger or smaller than the forcing scale $k_f = k_0\sqrt{3}$. This force generates flow cells that have locally differential rotation and helicity, two key ingredients for dynamo action.

Granted that the stretching and folding of magnetic field lines by velocity gradients overcome dissipation, dynamo action takes place above a critical magnetic Reynolds number Rm^c. For values of Rm larger than this critical value, a small magnetic field is amplified exponentially. For values of Rm smaller than Rm^c, the magnetic field is dissipated. A problem of interest in this context is to predict the behavior of Rm^c as Pm is lowered to values of interest for geophysical applications, as well as for dynamo experiments in the laboratory using liquid sodium. Using DNS at the highest resolution allowed by present computers, only values of $\mathsf{Pm} \sim 0.1$ had been reached for most flows.

Any numerical study of this problem begins with DNS. For simplicity, 3D periodic domains were considered. Then a pseudospectral algorithm can be used to solve the equations. In particular, the simulations that we discuss used an explicit second-order Runge-Kutta method to advance in time, and a classical dealiasing rule: the last resolved wavenumber is $k_{\max} = N/3$, where N is the linear resolution. Resolutions from 64^3 to 512^3 grid points were used to cover Pm from $1-0.2$. It is clear DNS are limited in the Reynolds numbers and the (lowest) value of Pm they can reach. To extend the range in Pm studied, the LAMHD equations and LES were used.

To use the LAMHD equations, a criterion is needed for the filtering lengths α_v and α_b. In the context of the dynamo, the length α_v is adjusted to have a well-resolved kinetic energy spectrum before the small magnetic field is introduced. In general, this is achieved by setting $1/\alpha_v \sim k_{\max}/2$ (Geurts and Holm [2002]). Then, the filtering scale for the magnetic field is set as the ratio of the Kolmogorov dissipation scale for both fields: $\alpha_v/\alpha_b = \mathsf{Pm}^{3/4}$ (Ponty, Mininni, Montgomery, Pinton, Politano and Pouquet [2005]). In this way, at a given magnetic Prandtl number, it is enforced that given a resolved (filtered) velocity field, the (filtered) magnetic field will also be well resolved when the nonlinear turbulent MHD regime is reached. Finally, in order to study still lower Pm, the LES model given by Eqs. (3.16)–(3.18) was implemented.

Fig. 3.1 shows the threshold Rm^c for dynamo action as a function of Pm^{-1}, obtained from numerical simulations. Values of Rm^c for Pm in the range from 1 to 0.2 were obtained from DNS. For Pm in the range 0.4–0.09, the LAMHD equations were used. Finally, the LES was used to study the range of Pm that lies between $\sim 1/10$ and $\sim 1/100$. Overlapping DNS and LAMHD simulations in the range 0.2–0.4 were used to verify the matching of the first two methods. The error bars also indicate that the three methods are consistent with the LAMHD equations systematically overestimating the value of Rm^c, and the LES underestimating it. However, the combination of DNS and two different subgrid models proved useful to study a problem that is numerically out of reach if only DNS is used.

At $P_M = 1$, the magnetic field is self-sustained for $\mathsf{Rm} > \mathsf{Rm}^c \approx 30$. As P_M is lowered, it is observed that the threshold for dynamo action reaches $\mathsf{Rm}^c \approx 70$ at $\mathsf{Pm} \approx 0.5$ and then increases sharply to $\mathsf{Rm}^c \approx 220$ at $\mathsf{Pm} \approx 0.3$; at lower Pm, it does not increase anymore but drops slightly to a value of 200 at $\mathsf{Pm} \approx 0.2$. This sharp increase and stabilization is followed by a plateau in the value of Rm^c for values of Pm down to ~ 0.01.

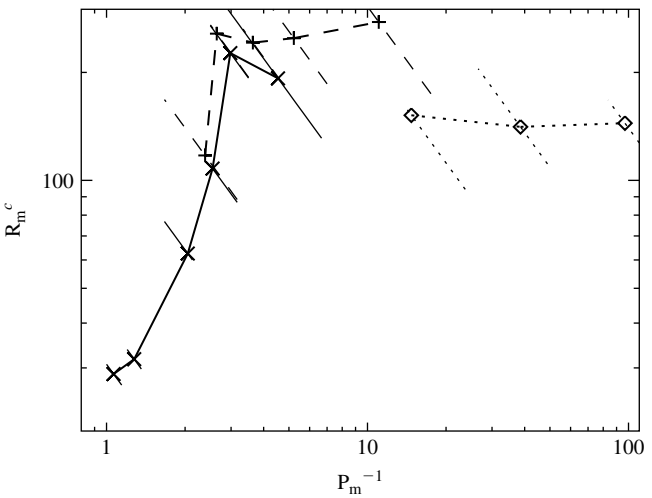


FIG. 3.1 Behavior of Rm^c for dynamo action with the inverse of Pm. Symbols are: $\times$ (DNS), $+$ (LAMHD), and $\diamond$ (LES). Transverse thin lines indicate error bars in the determination of Rm^c (modified from PONTY, MININNI, MONTGOMERY, PINTON, POLITANO and POUQUET [2005]).

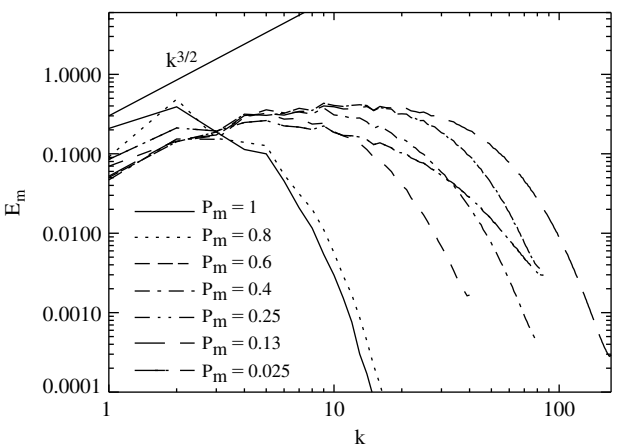


FIG. 3.2 Magnetic spectra for $\mathsf{Pm} = 1$ to $\mathsf{Pm} = 0.025$, at early times in the simulations. The curves are shifted vertically for better comparison. Spectra for values of Pm in the range 0.6–1 are from DNS, 0.13–0.4 are from LAMHD simulations, and $\mathsf{Pm} = 0.025$ from LES (modified from PONTY, MININNI, MONTGOMERY, PINTON, POLITANO and POUQUET [2005]).

As a result, the combination of the different methods allowed the finding in simulations of an asymptotic regime of MHD turbulence at small magnetic Prandtl number.

Fig. 3.2 shows the magnetic energy spectra $E_M(k)$ at early times in the simulations when the magnetic energy is small compared with the kinetic energy. All these runs have $\mathsf{Rm} > \mathsf{Rm}^c$, i.e., magnetic energy is growing as the result of the dynamo process.

Depending on the value of Pm, results from DNS, LAMHD simulations, and LES are shown.

Note that the magnetic field grows in a broad range of modes, and the peak in the magnetic energy spectrum moves progressively to smaller scales as Pm is decreased and Re is increased. Even with the magnetic spectrum peaking in the small scales, both subgrid models are able to capture the behavior of $E_M(k)$, as well as the trend followed by its peak as the value of Pm is changed. The LAMHD equations are also able to follow the evolution of the system until the growth of magnetic energy saturates nonlinearly, and the system reaches a MHD turbulent steady state.

4. Discussion

4.1. AMR

Turbulent flows with a wide range of excited scales and a wide range of parameters unattainable with DNS have to be modeled, and such models must be tested and improved using experiments, observations, and DNS at the highest Re achievable accurately. This is where recent progress in dynamically AMR comes into play. However, in view of the complexity of turbulent flows, AMR methods have but barely been developed and used for complex flows at high Re and it is not clear yet whether they will succeed, i.e., whether they will provide accurate representations of turbulent flows at a numerical cost (both memory wise and CPU wise) that is lower than performing the same computation on a regular grid. Note that the potential savings in memory may be as important as the savings in computer run time[4].

Do DNS of turbulence using AMR require high order discretization? In low-order adaptive computations (Berger and Colella [1989], Berger and Oliger [1984]), refinement criteria are often governed by physical indicators which do not always anticipate sudden events; moreover, structures can be smoothed out before they ever develop in the numerical calculation (see Kritsuk, Norman and Padoan [2006] in the context of supersonic turbulence). On the other hand, global spectral methods have the ability to capture multiple scales with little dissipation or dispersion. This property of high-order methods is likely to make them indispensable for accurate and controlled modeling of turbulence governed by Re because of extreme and localized events (Rosenberg, Fournier, Fisher and Pouquet [2006]).

Since the mid-1980s, there have been attempts to combine the geometrical flexibility of low-order methods with the high accuracy of high order ones through the development of the spectral element method (Patera [1984]) and the discontinuous Galerkin method (Cockburn, Karniadakis and Shu [2003]). The transition to high order is slowly spreading through a number of disciplines: engineering (e.g., Deville, Fischer and Mund [2002], Henderson [1995], Hesthaven and Warburton [2004],

[4]The direct visualization of AMR grids is a reasonably well-solved problem, and a few rudimentary end-user tools exists: ChomboVis (Ligocki [2000]), VAPOR (Clyne and Rast [2005]), and the FLASH Code viewer, FlashView (Chicago [2006]) to name a few.

KARNIADAKIS and SHERWIN [2005], LOTH, FISCHER, ARSLAN, BERTRAM, LEE, ROYSTON, SONG, SHAALAN and BASSIOUNY [2003], LU, BERZINS, GOODYER and JIMAC [2005], REMACLE, FLAHERTY and SHEPHAD [2003]), atmospheric sciences (DENNIS, FOURNIER, SPOTZ, ST.-CYR, TAYLOR, THOMAS and TUFO [2005], FOURNIER, TAYLOR and TRIBBIA [2004]), ocean sciences (e.g., HAIDVOGEL, CURCHITSER, ISKANDARANI, HUGHES and TAYLOR [1997], ISKANDARANI, HAIDVOGEL and LEVIN [2003], LEVIN, ISKANDARANI and HAIDVOGEL [2000]), and geophysics (FOURNIER, BUNGE, HOLLERBACH and VILOTTE [2005]).

AMR methods with spectral elements use a posteriori error estimators based on the numerical quality of the solution to serve as indicators of changes in the solution requiring better resolution. This may be paramount when accuracy is essential such as for the search of singularities. Vorticity and current density dynamics may be linked to the development of singularities for fluids and MHD (see DOMBRE and PUMIR [1995] for an introduction). There are many studies of possible singular behavior in turbulence, e.g., through analysis, numerical simulations and modeling (e.g., BEALE, KATO and MAJDA [1984], BHATTACHARJEE, NG and WANG [1995], CANTWELL [1992], CONSTANTIN and TITI [1994], GRAUER, MARLIANI and GERMASCHEWSKI [1998], KERR [1993, 2005], PELZ [2001], VIEILLEFOSSE [1984]). VIEILLEFOSSE [1984] proposed that the pressure Hessian $P_{\mu,\nu} := \partial_\mu \partial_\nu P$ be isotropic, i.e., $\sim \delta_{\mu,\nu}$, leading to singularities (see KLAPPER, RADO and TABOR [1996] for its extension to MHD and LI and MENEVEAU [2005] for generalization of the model viewed as an LES). The structures that develop in a turbulent flow (vortex sheets and filaments, VINCENT and MENEGUZZI [1991]; current sheets in MHD, MININNI, POUQUET and MONTGOMERY [2006], POLITANO, POUQUET and SULEM [1995]) can be identified and followed numerically (and experimentally, see DOUADY, COUDER and BRACHET [1991]), and the maxima of the vorticity and current monitored in order to see whether criteria developed in BEALE, KATO and MAJDA [1984] are met, but these computations, especially when using AMR (GRAUER, MARLIANI and GERMASCHEWSKI [1998]), are difficult and require high accuracy.

However, it should be noted that the fact that the field which is integrated numerically (say, the velocity) is C_1 in a spectral element method means that it is continuous but its derivatives are not: one takes an averaged edge values between contiguous elements, leading possibly to energy dissipation. Refining further until the discontinuity becomes comparable to roundoff errors may be way too costly, so this aspect must be studied in detail and is an open field of investigation presently. This may otherwise introduce potentially spurious indication of singularity in the flow so that particular care should be taken when dealing with the search for singularities with AMR codes.

Multiresolution-based methods, like the spectral element method, for adaptivity should also be useful for structure identification. It was shown in PORTER, POUQUET and WOODWARD [1998] that, by looking at vorticity using a band-pass filter centered on the inertial-range, the curvature of vorticity participates in the nonlinear transfer of energy. One can envision the dynamics of vorticity "worms" as advected local orthonormal frames following vorticity concentrations that in turn are carried along Lagrangian fluid paths. A vortex line can be tagged using the Frenet-Serret local frame which

gives the relationship between the tangent, normal and bi-normal to the curve and its curvature κ and torsion τ (the latter, a pseudoscalar); the temporal dynamics of κ and τ can be analyzed (Constantin, Procaccia and Segel [1995]). The same can be done in MHD (Brandenburg, Procaccia and Segel [1995], Pouquet [1996]). This view of vorticity dynamics would benefit from a hybrid computational approach. New theories (Gibbon, Holm, Kerr and Roulstone [2006], Ohkitani [1993]) emphasize that vorticity alignment with $P_{\mu,\nu}$ and the strain rate tensor $S_{\mu,\nu}$ are the sources of change of vorticity direction and magnitude following a Lagrangian coordinate system to which it is attached (Gibbon, Holm, Kerr and Roulstone [2006], Ohkitani [1993]). $P_{\mu,\nu}$ and $S_{\mu,\nu}$ are essentially Eulerian quantities but drive the Lagrangian evolution of the coordinate frame attached to the vorticity. The ability to compute numerically $P_{\mu,\nu}$ and $S_{\mu,\nu}$ will help diagnose the governing properties of vorticity dynamics that lie at the heart of turbulence.

These localized structures can be viewed as obstacles on the trajectory of fluid particles and as such as the source of their intermittent behavior as measured in a Lagrangian framework. It is conceivable that the trajectory of fluid particles, viewed as a random walk, may keep a memory of their path because of such obstacles and may thus lead to the large-scale large-time behavior observed in experiments (Mordant, Delour, Leveque, Arneodo and Pinton [2002]), in models (Aringazin and Mazhitov [2004]), in numerical simulations (Sorriso-Valvo, Carbone, Noullez, Politano, Pouquet and Veltri [2002]), or in nature (Matthaeus and Goldstein [1986]). These experiments and DNS are essential for future modeling.

The error estimators in spectral element methods not only give criteria for refinement and/or coarsening but also determine whether high- or low-order resolution is needed; in other words, they indicate whether h-refinement (element splitting or merging as in low-order methods) or p-refinement (elemental discretization function order increase as in high-order methods) is prescribed (see Mavriplis [1990, 1994] for an implementation in combustion). For a moderate level of accuracy, the low-order methods may be more efficient, but when high accuracy is needed as in the case of turbulence, the exponentially converging high-order methods are essential (Rosenberg, Pouquet and Mininni [2007]). Hence, high-order methods are definitely needed, but the combination of high and low order through dynamic AMR has the potential to provide the best efficiency in simulating turbulence that spans multiple scales, includes regions of smooth as well as sharp features, and engenders the creation and decay of complex structures. Furthermore, by using the spectral-element method combined with an analysis of localized interactions between scales, we can quantitatively model and investigate many important phenomena that involve scale interactions localized in parts of the domain, and that heretofore were mainly only described qualitatively or heuristically, atmospheric blocking events providing one such example (Fournier [2005]).

4.2. Concluding remarks

In the vocabulary of quantum field theory, the turbulent transport coefficients (eddy viscosity and eddy noise) were labeled absorption and emission terms; in the context of stochastic modeling, these terms, in a generalized Langevin-like equation for turbulence,

may be associated respectively today with multiplicative noise (introducing a term $\nu_{turb}\mathbf{u}$ in the dynamical equation) and additive noise (simply adding a noise term in the Navier-Stokes equation). It is known that both are important, and that in fact, they should have some amount of correlation (see DUBRULLE, LAVAL, NAZARENKO and ZABORONSKI [2004] for an illustration of that fact, and BERNER [2005] for an application to planetary waves). This correlation can be easily understood in the context of weak turbulence theory and two-point closures for turbulent flows since both eddy terms due to the small scales will be functionals of the (same) energy spectrum and its derivatives (the primitive equations being closed at the level of two-point correlation functions, resulting in an explicit dynamical evolution for the energy spectra)[5].

One purpose for the turbulence community might be to assemble tools in different aspects of this difficult problem and to exploit the synergy of the assembled tools in order to significantly advance the state of the art. Such tools could be based on recent advances in high-order dynamic adaptive refinement methods, high-performance scalable computing including implementation on terascale (and soon to be petascale) platforms and analysis and visualization of large data sets on irregular grids, turbulence modeling, and multiresolution analysis. The combination of AMR and modeling (or more generally, filtering) has seldom been tested (DEVILLE, FISCHER and MUND [2002]) but may prove to be a very efficient way to pursue investigations of turbulent flows.

Progress in 2D turbulence has been accomplished (see, e.g., BERNARD, BOFFETTA, CELANI and FALKOVICH [2006]) but the 3D case remains a formidable challenge. Applications to multiscale properties of turbulence such as intermittency and its relation to coherent structures, stretching and curvature of vorticity, interactions between widely separated scales, and finally low magnetic Prandtl number dynamos have been reviewed briefly here. This approach will lead to enhanced understanding and to more accurate modeling of turbulent flows in a variety of contexts once the multiresolution, multistructure, multiphysics interactions are better apprehended.

5. Acknowledgments

We would like to thank Alex Alexakis, Aimé Fournier, Jonathan Graham, Jim McWilliams, Catherine Mavriplis, David Montgomery, Hélène Politano, Yannick Ponty, Duane Rosenberg, and Mark Taylor for useful discussions. NCAR is sponsored by the National Science Foundation. PM is supported by NSF CMG grant 0327888, and PPS is partially supported by the Physical Oceanography Section of the Office of Naval Research. Computations reported in this paper were performed at NCAR and at Pittsburgh.

[5]For an introduction to the literature on stochastic modeling, see GARDINER [1983], MAJDA, TIMOFEYEV and VANDEN-EIJNDEN [1999].

References

AGULLO, O., MÜLLER, W.-C., KNAEPEN, B., CARATI, D. (2001). Large eddy simulation of decaying magnetohydrodynamic turbulence with dynamic subgrid-modeling. *Phys. Plasmas* **8**, 3502–3505.

ALEXAKIS, A., MININNI, P.D., POUQUET, A. (2005). Shell-to-shell energy transfer in magnetohydrodynamics. I. Steady state turbulence. *Phys. Rev. E* **72**, 046301.

ALEXAKIS, A., MININNI, P.D., POUQUET, A. (2006). On the inverse cascade of magnetic helicity. *Astrophys. J.* **640**, 335–343.

ALVES, J.G.M., BANNER, M.L., YOUNG, I.R. (2003). Revisiting the Pierson-Moskowitz asymptotic limits for fully developed wind waves. *J. Phys. Oceanogr.* **33**, 1301–1323.

ANDREWS, D.G., MCINTYRE, M.E. (1978a). An exact theory of nonlinear waves on a Lagrangian-mean flow. *J. Fluid Mech.* **89**, 609–646.

ANDREWS, D.G., MCINTYRE, M.E. (1978b). On wave action and its relatives. *J. Fluid Mech.* **89**, 647–664.

AOYAMA, Y., NAKANO, J. (1999). RS/6000 SP: Practical MPI programming, Technical Report IBM Redbook SG24-5380-00, International Business Machines.

ARINGAZIN, A., MAZHITOV, M. (2004). One-dimensional Langevin models of fluid particle acceleration in developed turbulence. *Phys. Rev. E* **69**, 026305–1–17.

ARMFIELD, S., STREET, R. (1999). The fractional step method for the Navier-Stokes equations on staggered grids: the accuracy of three variations. *J. Comp. Phys.* **153**, 660–665.

BAERENZUNG, J., POLITANO, H., PONTY, Y., POUQUET, A. (2008). Spectral modeling of magnetohydrodynamic turbulent flows. *Phys. Rev. E* **78**, 026310.

BAUMERT, H.Z., SIMPSON, J., SÜNDERMANN, J. (2005). *Marine Turbulence: Theories, Observations, and Models* (Cambridge University Press).

BEALE, J., KATO, T., MAJDA, A. (1984). Remarks on the breakdown of smooth solutions for the three-dimensional Euler equations. *Comm. Math. Phys.* **94**, 61–66.

BERGER, M., COLELLA, P. (1989). Local adaptive mesh refinement for shock hydrodynamics. *J. Comp. Phys.* **82**, 64–84.

BERGER, M., OLIGER, J. (1984). Adaptive mesh refinement for hyperbolic partial differential equations. *J. Comp. Phys.* **53**, 484–512.

BERNARD, G., BOFFETTA, G., CELANI, A., FALKOVICH, G. (2006). Conformal invariance in two-dimensional turbulence. *Nat. Phys.* **2**, 124–128.

BERNER, J. (2005). Linking nonlinearity and nongoussianity of planetary wave behavior by the Fokker-Planck equation. *J. Atmos. Sci.* **62**, 2098–2117.

BHATTACHARJEE, A., NG, D., WANG, X. (1995). Finite-time vortex singularity and Kolmogorov spectrum in a symmetric three-dimensional spiral model. *Phys. Rev. E* **52**, 5110–5123.

BRANDENBURG, A. (2001). The inverse cascade and nonlinear alpha-effect in simulations of isotropic helical hydromagnetic turbulence. *Astrophys. J.* **550**, 824–840.

BRANDENBURG, A., SARSON, G.R. (2002). Effect of hyperdiffusivity on turbulent dynamos with helicity. *Phys. Rev. Lett.* **88**, 055003.

BRANDENBURG, A., SUBRAMANIAN, K. (2005). Astrophysical magnetic fields and nonlinear dynamo theory. *Phys. Rep.* **417**, 1–210.

BRANDENBURG, A., PROCACCIA, I., SEGEL, D. (1995). The size dynamics of magnetic flux structures in magnetohydrodynamic turbulence. *Phys. Plasmas* **2**, 1148–1156.

BRUNO, R., CARBONE, V. (2005). The solar wind as a turbulence laboratory. Living Rev. Solar Phys. 2. Available at 2005; http://www.livingreviews.org/lrsp-2005-4.

Cantwell, B. (1992). Exact solutions of a restricted euler equation for the velocity gradient tensor. *Phys. Fluids A* **4**, 782–793.

Canuto, C., Hussaini, M.Y., Quarteroni, A., Zang, T.A. (1988). *Spectral Methods in Fluid Dynamics* (Springer Verlag).

Cattaneo, F. (1999). On the origin of magnetic fields in the quiet photosphere. *Astrophys. J.* **515**, L39–L42.

Chandra, R., Dagum, L., Kohr, D., Maydan, D., McDonald, J., Menon, R. (2001). *Parallel Programming in OpenMP* (Academic Press).

Chen, S.Y., Holm, D.D., Foias, C., Olson, E.J., Titi, E.S., Wynne, S. (1998). The Camassa-Holm equations as a closure model for turbulent channel and pipe flows. *Phys. Rev. Lett.* **81**, 5338–5341.

Chen, S.Y., Holm, D.D., Margolin, L.G., Zhang, R. (1999). Direct numerical simulations of the Navier-Stokes alpha model. *Physica D* **133**, 66–83.

Chicago, U. (2006). Available at 2006; http://www-flash.uchicago.edu/webside/codesupport/flash-view.

Chollet, J.-P., Lesieur, M. (1981). Parametrization of small scales of three-dimensional isotropic turbulence using spectral closures. *J. Atmos. Sci.* **38**, 2747–2757.

Clark, R.A., Ferziger, J.H., Reynolds, W.C. (1979). Evaluation of subgrid-scale models using an accurately simulated turbulent flow. *J. Fluid Mech.* **91**, 1–16.

Clyne, J., Rast, M. (2005). A prototype discovery environment for analyzing and visualizing terascale turbulent fluid flow simulations. In: Erbacher, R.F., Roberts, J.C., Grohn, M.T., Borner, K. (eds.), *Visualization and Data Analysis 2005* (SPIE, Bellingham, WA), pp. 284–294, Available at 2005; http://www-vapor.ucar.edu.

Cockburn, B., Karniadakis, G.E., Shu, C. (2003). *Discontinuous Galerkin Methods* (Springer Verlag).

Constantin, P., Titi, E. (1994). Onsager's conjecture on the energy conservation for solutions of Euler's equation. *Comm. Math. Phys.* **165**, 207–209.

Constantin, P., Procaccia, I., Segel, D. (1995). Creation and dynamics of vortex tubes in three-dimensional turbulence. *Phys. Rev. E* **51**, 3207–3222.

Craig, P.D., Banner, M.L. (1994). Modeling wave-enhanced turbulence in the ocean surface layer. *J. Phys. Oceanogr.* **24**, 2546–2559.

Craik, A., Leibovich, S. (1976). A rational model for Langmuir circulations. *J. Fluid Mech.* **73**, 401–426.

Deardorff, J.W. (1970). A numerical study of three-dimensional turbulent channel flow at large Reynolds numbers. *J. Fluid Mech.* **41**, 453–465.

Deardorff, J.W. (1980). Stratocumulus-capped mixed layers derived from a three-dimensional model. *Bound. Layer Meteorol.* **18**, 495–527.

Dennis, J., Fournier, A., Spotz, W., St.-Cyr, A., Taylor, M., Thomas, S., Tufo, H. (2005). High resolution mesh convergence properties and parallel efficiency of a spectral element atmospheric dynamical core. *Int. J. High Perform. Comp. Appl.* **19**, 225–235.

Deville, M., Fischer, P., Mund, E. (2002). *High Order Methods for Incompressible Fluid Flow* (Cambridge University Press).

Dikpati, M., Charbonneau, P. (1999). A Babcock-Leighton flux transport dynamo with solar-like differential rotation. *Astrophys. J.* **518**, 508–520.

Dombre, T., Pumir, A. (1995). Singularities and turbulence: a tentative dictionary. In: Tabeling, P., Cardoso, O. (eds.), *Turbulence* (NATO ASI Series, Ser. B Physics), pp. 341–351.

Donelan, M.A. (1998). Air-water exchange processes. In: *Physical Processes in Lakes and Oceans, Volume 54 of Coastal and Estuarine Studies* (American Geophysical Union), pp. 19–36.

Douady, S., Couder, Y., Brachet, M. (1991). Direct observation of the intermittency of intense vorticity filaments in turbulence. *Phys. Rev. Lett.* **67**, 983–986.

Dubrulle, B., Laval, J.-P., Nazarenko, S., Zaboronski, O. (2004). A model for rapid stochastic distortions of small scale turbulence. *J. Fluid Mech.* **520**, 1–21.

Emanuel, K. (2004). Tropical cyclone energetics and structure. In: *Atmospheric Turbulence and Mesoscale Meteorology* (Cambridge University Press), pp. 165–191.

Foias, C., Holm, D.D., Titi, E.S. (2001). The Navier-Stokes-alpha model of fluid turbulence. *Phys. D* **152**, 505–519.

Foias, C., Holm, D.D., Titi, E.S. (2002). The three-dimensional viscous Camassa-Holm equations and their relation to the Navier-Stokes equations and turbulence theory. *J. Dyn. Diff. Equ.* **14**, 1–35.

Fournier, A. (2005). Instantaneous wavelets energetics transfers between atmospehric blocking and local eddies. *J. Clim.* **13**, 2151–2171.

Fournier, A., Taylor, M., Tribbia, J. (2004). The spectral element atmosphere model (SEAM): high-resolution parallel computation and localized resolution of regional dynamics. *Mon. Wea. Rev.* **132**, 726–748.

Fournier, A., Bunge, H., Hollerbach, R., Vilotte, J.-P. (2005). A Fourier-spectral element algorithm for thermal convecction in rotating axisymmetric containers. *J. Comp. Phys.* **204**, 462–489.

Frisch, U. (1995). *Turbulence. The Legacy of A.N. Kolmogorov* (Cambridge Univ. Press).

Frisch, U., Pouquet, A., Léorat, J., Mazure, A. (1975). On the possibility of an inverse cascade in MHD helical turbulence. *J. Fluid Mech.* **68**, 769–778.

Gailitis, A., Lielausis, O., Platacis, E., Dement'ev, S., Cifersons, A., Gerbeth, G., Gundrum, T., Stefani, F., Christen, M., Will, G. (2001). Magnetic field saturation in the riga dynamo experiment. *Phys. Rev. Lett.* **86**, 3024–3027.

Gardiner, C. (1983). *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences* (Springer-Verlag).

Germano, M. (1986a). Differential filters for the large eddy numerical simulation of turbulent flows. *Phys. Fluids* **29**, 1755–1757.

Germano, M. (1986b). Differential filters of elliptic type. *Phys. Fluids* **29**, 1757–1758.

Germano, M., Piomelli, U., Moin, P., Cabot, W. (1991). A dynamic subgrid scale eddy viscosity model. *Phys. Fluids A* **3**, 1760–1765.

Geurts, B.J. (2001). *Modern Simulation Strategies for Turbulent Flow* (R. T. Edwards).

Geurts, B.J., Holm, D.D. (2002). Alpha-modeling strategy for les of turbulent mixing. In: Drikakis, D., Geurts, B.J. (eds.), *Turbulent Flow Computation* (Kluwer, Dordrecht, The Netherlands), pp. 237–278.

Ghosh, S., Matthaeus, W.H., Montgomery, D.C. (1998). The evolution of cross helicity in driven/dissipative two-dimensional magnetohydrodynamics. *Phys. Fluids* **31**, 2171–2184.

Gibbon, J., Holm, D., Kerr, R., Roulstone, I. (2006). Quarternions and particle dynamics in the Euler fluid equations, preprint Imperial College.

Glatzmaier, G.A., Roberts, P.H. (1995). A three-dimensional self-consistent computer simulation of a geomagnetic field reversal. *Nature* **377**, 203–209.

Glatzmaier, G.A., Roberts, P.H. (1996). Rotation and magnetism of Earth's inner core. *Science* **274**, 1887–1891.

Graham, J.P., Mininni, P.D., Pouquet, A. (2005). Cancellation exponent and multifractal structure in two-dimensional magnetohydrodynamics: direct numerical simulations and Lagrangian averaged modeling. *Phys. Rev. E* **72**, 045301(R).

Graham, J.P., Holm, D.D., Mininni, P.D., Pouquet, A. (2006). Inertial range scaling, Kármán-Howarth theorem, and intermittency for forced and decaying Lagrangian averaged magnetohydrodynamic equations in two dimensions. *Phys. Fluids* **18**, 045106.

Grappin, R., Pouquet, A., Léorat, J. (1983). Dependence on correlation of MHD turbulence spectra. *Astron. Astrophys.* **126**, 51–56.

Grauer, R., Marliani, C., Germaschewski, K. (1998). Adaptive mesh refinement for singular solutions of the incompressible Euler equation. *Phys. Rev. Lett.* **80**, 4177–4180.

Grotte, F.H., Busse, F.H., Tilgner, A. (2000). Effects of hyperdiffusivities on dynamo simulations. *Geophys. Res. Lett.* **27**, 2001–2004.

Haidvogel, D., Curchitser, E., Iskandarani, M., Hughes, R., Taylor, M. (1997). Global modeling of the ocean and atmosphee using the spectral element method. *Atmos. Ocean* **35**, 505–531.

Haugen, N.E.L., Brandenburg, A., Dobler, W. (2003). Is nonhelical hydromagnetic turbulence peaked at small scales? *Astrophys. J.* **597**, L141–L144.

Haugen, N.E.L., Brandenburg, A., Dobler, W. (2004). Simulations of nonhelical hydromagnetic turbulence. *Phys. Rev. E* **70**, 016308.

Henderson, R. (1995). Unstructured spectral element methods for simulation of turbulent flows. *J. Comp. Phys.* **122**, 191–217.

Hesthaven, J., Warburton, T. (2004). Discontinuous Galerkin methods for the time-domain Maxwell's equations: an introduction, preprint Available at 2004; http://www-dam.brown.edu/scicomp/publications/reprts/Y2004/BrownSC2004-04.pdf (2004), 18.

Holm, D.D. (2002a). Averaged Lagrangians and the mean effects of fluctuations in ideal fluid dynamics. *Physica D* **170**, 253–286.

Holm, D.D. (2002b). Lagrangian averages, averaged Lagrangians, and the mean effects of fluctuations in fluid dynamics. *Chaos* **12**, 518–530.

Holm, D.D. (1996). The ideal Craik-Leibovich equations. *Physica D* **98**, 415–441.

Holm, D.D., Marsden, J.E., Ratiu, T.S. (1998a). The Euler-Poincaré equations and semidirect products with applications to continuum theories. *Adv. Math.* **137**, 1–81.

Holm, D.D., Marsden, J.E., Ratiu, T.S. (1998b). Euler-Poincaré models of ideal fluids with nonlinear dispersion. *Phys. Rev. Lett.* **80**, 4173–4176.

Hossain, M., Matthaeus, W.H., Montgomery, D. (1983). Long-time states of inverse cascades in the presence of a maximum length scale. *J. Plasma Phys.* **30**, 479–493.

Ilyin, A.A., Titi, E.S. (2003). Attractors to the two-dimensional Navier-Stokes-α model: an α-dependence study. *J. Dyn. Diff. Equ.* **15**, 751–777.

Iskandarani, M., Haidvogel, D., Levin, J. (2003). A three-dimensional spectral elemant basin m model for the solutions of the hydrostatic primitive equations. *J. Comp. Phys.* **186**, 397–425.

Jones, C.A., Roberts, P.H. (2005). Turbulence models and plane layer dynamos. In: Soward, A.M., Jones, C.A., Hughes, D.W., Weiss, N.O. (eds.), *Fluid Dynamics and Dynamos in Astrophysics and Geophysics* (CRC Press, Boca Raton, FL), pp. 295–330.

Kaneda, T., Ishihara, M., Yokokawa, K., Itakura, K., Uno, A. (2003). Energy dissipation rate and energy spectrum in high resolution direct numerical simulations of turbulence in a periodic box. *Phys. Fluids* **15**, L21–L24.

Karniadakis, G., Sherwin, S. (2005). *Spectral/hp Element Methods for CFD* (Oxford Univ. Press).

Kerr, R. (1993). Evidence for a singularity of the three-dimensioanl, incompressible Euler equations. *Phys. Fluids A* **5**, 1725–1746.

Kerr, R. (2005). Velocity scaling of collapsing Euler vortices. *Phys. Fluids* **17**, 075103.

Kinney, R., McWilliams, J.C., Tajima, T. (1995). Coherent structures and turbulent cascades in two-dimensional incompressible magnetohydrodynamic turbulence. *Phys. Fluids* **2**, 3623–3639.

Klapper, I., Rado, A., Tabor, M. (1996). Lagrangian study of dynamics and singularity formation at magnetic null points in ideal three-dimensional magnetohydodynamics. *Phys. Plasmas* **3**, 4281–4283.

Kolmogorov, A. (1941). Energy dissipation in locally isotropic turbulence. *Dokl. Akad. Nauk* **32**, 19–20.

Kono, M., Roberts, P.H. (2002). Recent geodynamo simulations and observations of the geomagnetic field. *Rev. Geophys.* **40**, 1–53.

Kraichnan, R.H. (1959). The structure of isotropic turbulence at very high Reynolds number. *J. Fluid Mech.* **5**, 497–543.

Krause, F., Raedler, K.-H. (1980). *Mean-Field Magnetohydrodynamics and Dynamo Theory* (Pergamon Press, New York).

Kritsuk, A., Norman, M., Padoan, P. (2006). Adaptive mesh refinement for supersonic molecular cloud turbulence. *Astrophys. J. Lett.* **638**, L25–L28.

Kurien, S., Taylor, M. (2005). Direct numerical simulations of turbulence: data generation and statistical analysis. *Los Alamos Sci.* **29**, 142–151.

Langmuir, I. (1938). Surface motion of water induced by wind. *Science* **87**, 119–123.

Large, W.G., McWilliams, J.C., Doney, S.C. (1995). Oceanic vertical mixing: a review and a model with a nonlocal boundary layer parameterization. *Rev. Geophys.* **32**, 363–403.

Leamon, R.J., Matthaeus, W.H., Smith, C.W., Zank, G.P., Mullan, D.J., Oughton, S. (2000). MHD-driven kinetic dissipation in the solar wind and corona. *Astrophys. J.* **537**, 1054–1062.

Leibovich, S. (1983). The form and dynamics of Langmuir circulations. *Ann. Rev. Fluid Mech.* **15**, 391–427.

Leonard, A. (1974). Energy cascade in large-eddy simulations of turbulent fluid flows. *Adv. Geophys. A* **18**, 237–248.

Leray, J. (1934). Sur le mouvement d'un liquide visqueux emplissant l'espace. *Acta Math.* **63**, 193–248.

Lesieur, M. (1997). *Turbulence in Fluids* (Kluwer Academic Press).

Lesieur, M., Métais, O. (1996). New trends in large eddy simulations of turbulence. *Ann. Rev. Fluid Mech.* **28**, 45–82.

Leslie, D. (1973). *Developments in the Theory of Turbulence* (Clarendon Press).

Levin, J., Iskandarani, M., Haidvogel, D. (2000). A nonconforming spectral element ocean model. *Int. J. Numer. Meth. Fluids* **34**, 495–525.

Li, M., Zahariev, K., Garrett, C. (1995). Role of Langmuir circulation in the deepening of the ocean surface mixed layer. *Science* **270**, 1955–1957.

Li, M., Garrett, C., Skyllingstad, E. (2005). A regime diagram for classifying turbulent large eddies in the upper ocean. *Deep Sea Res. I* **52**, 259–278.

Li, Y., Meneveau, C. (2005). Origin of non-Gaussian statistics in hydrodynamic turbulence. *Phys. Rev. Lett.* **95**, 164502.

Ligocki, T. (2000). Implementing a visualization tool for adaptive mesh refinement data using VTK. In: *Proceedings of Visualization Development Environments* pp. 27–28.

Lilly, D.K. (1969). Numerical simulation of two-dimensional turbulence. *Phys. Fluids* **12** (Suppl. II), 240–249.

Lilly, D.K. (1967). The representation of small-scale turbulence in numerical simulation experiments. In: Goldstine, H.H. (ed.), *Proc. IBM Scientific Computing Sympo. on Environmental Sciences* (IBM, Yorktown Heights, New York), pp. 195–210.

Liu, W.T., Katsaros, K.B., Businger, J.A. (1979). Bulk parameterization of air-sea exchanges in heat and water vapor including the molecular constraints at the interface. *J. Atmos. Sci.* **36**, 1722–1735.

Loth, F., Fischer, P., Arslan, N., Bertram, C., Lee, S., Royston, T., Song, R., Shaalan, W., Bassiouny, H. (2003). Transitional flow at the venous anastomosis of an arteriovenous graft: potential relationship with activation of the ERK1/2 Mechanotransduction pathway. *J. Biomed. Eng.* **125**, 49–61.

Lu, H., Berzins, M., Goodyer, C., Jimac, P. (2005). High order discontinuous Galerkin method for elastohydrodynamic lubrication line contact problems. *Comm. Num. Meth. Eng.* **21**, 643–650.

Majda, A., Timofeyev, I., Vanden-Eijnden, E. (1999). Models for stochastic climate prediction. *Proc. Nat. Acad. Sci.* **96**, 14687–14691.

Matthaeus, W.H., Dasso, S., Weygand, J.M., Milano, L.J., Smith, C.W., Kivelson, M.G. (2005). Spatial correlation of solar-wind turbulence from two-point measurements. *Phys. Rev. Lett.* **95**, 231101.

Matthaeus, W., Goldstein, M. (1982). Measurement of the rugged invariants of magnetohydrodynamic turbulence in the solar wind. *J. Geophys. Res.* **87**, 6011–6028.

Matthaeus, W., Goldstein, M. (1986). Low-frequency 1/f noise in the interplanetary magnetic field. *Phys. Rev. Lett.* **57**, 495–499.

Mavriplis, C. (1990). *A Posteriori Error Estimators for Adaptivve Specctral Element Techniques, vol. 29* (Vieweg).

Mavriplis, C. (1994). Adaptive mesh strategies for the spectral element method. *Comp. Meth. Appl. Mech. Eng.* **116**, 77–86.

McComb, W.D. (1990). *The Physics of Fluid Turbulence* (Oxford Univ. Press, Clarendon, Jamaica).

McWilliams, J.C., Huckle, E. (2006). Ekman layer rectification. *J. Phys. Oceanogr.* **36**, 1646–1659.

McWilliams, J.C., Restrepo, J.M. (1999). The wave-driven ocean circulation. *J. Phys. Oceanogr.* **29**, 2523–2540.

McWilliams, J.C., Sullivan, P.P. (2000). Vertical mixing by Langmuir circulations. *Spill Sci. Technol. Bull.* **6**, 225–237.

McWilliams, J.C., Sullivan, P.P., Moeng, C.-H. (1997). Langmuir turbulence in the ocean. *J. Fluid Mech.* **334**, 1–30.

McWilliams, J.C., Restrepo, J.R., Lane, E.M. (2004). An asymptotic theory for the interaction of waves and currents in shallow coastal water. *J. Fluid Mech.* **511**, 135–178.

Melville, W.K., Matusov, P. (2002). Distribution of breaking waves at the ocean surface. *Nature* **417**, 58–63.

Meneguzzi, M., Frisch, U., Pouquet, A. (1981). Helical and nonhelical turbulent dynamos. *Phys. Rev. Lett.* **47**, 1060–1064.

Meneguzzi, M., Politano, H., Pouquet, A., Zolver, M. (1996). A sparse mode spectral method for the simulation of turbulent flows. *J. Comp. Phys.* **123**, 32–44.

Meneveau, C., Katz, J. (2000). Scale-invariance and turbulence models for large-eddy simulations. *Ann. Rev. Fluid Mech.* **32**, 1–32.

Mininni, P.D. (2006). Turbulent magnetic dynamo excitation at low magnetic Prandtl number. *Phys. Plasmas* **13**, 056502.

Mininni, P.D., Alexakis, A., Pouquet, A. (2005a). Shell-to-shell energy transfer in magnetohydrodynamics. II. Kinematic dynamo. *Phys. Rev. E* **72**, 046302.

Mininni, P.D., Montgomery, D.C., Pouquet, A. (2005b). A numerical study of the alpha model for two-dimensional magnetohydrodynamic turbulent flows. *Phys. Fluids* **17**, 035112.

Mininni, P.D., Montgomery, D.C., Pouquet, A. (2005c). Numerical solutions of the three-dimensional magnetohydrodynamic α model. *Phys. Rev. E* **71**, 046304.

Mininni, P.D., Ponty, Y., Montgomery, D.C., Pinton, J.-F., Politano, H., Pouquet, A. (2005d). Dynamo regimes with a nonhelical forcing. *Astrophys. J.* **626**, 853–863.

Mininni, P., Pouquet, A., Montgomery, D. (2006). Small-scale structures in three-dimensional magneto-hydrodynamic turbulence. *Phys. Rev. Lett.* **97**, 244503.

Moeng, C.-H. (1984). A large-eddy simulation model for the study of planetary boundary-layer turbulence. *J. Atmos. Sci.* **41**, 2052–2062.

Moeng, C.-H., Stevens, B. (2000). Representing the stratocumuls-topped boundary layer in GCMs. In: *General Circulation Model Development* (Academic Press), pp. 577–604.

Moeng, C.-H., Sullivan, P.P. (2002). Large-eddy simulation. In: *Encyclopedia of Atmospheric Sciences* (Academic Press), pp. 1140–1150.

Moeng, C.H., Wyngaard, J.C. (1988). Spectral analysis of large-eddy simulations of the convective boundary layer. *J. Atmos. Sci.* **45**, 3573–3587.

Moffatt, H.K. (1978). *Magnetic Field Generation in Electrically Conducting Fluids* (Cambridge Univ. Press, Cambridge).

Mohseni, K., Kosović, B., Shkoller, S., Marsden, J.E. (2003). Numerical simulations of the Lagrangian averaged Navier-Stokes equations for homogeneous isotropic turbulence. *Phys. Fluids* **15**, 524–544.

Monin, A., Yaglom, A. (1975). *Statistical Fluid Mechanics: Mechanics of turbulence, vol. 1 1971. vol. 2 1975* (MIT Press).

Montgomery, D., Pouquet, A. (2002). An alternative interpretation for the holm 'alpha model.' *Phys. Fluids* **14**, 3365–3366.

Mordant, N., Delour, J., Leveque, E., Arneodo, A., Pinton, J. (2002). Long time correlations in Lagrangian dynamics: a key to intermittency in turbulence. *Phys. Rev. Lett.* **89**, 254502–1–4.

Müller, W.-C., Carati, D. (2002a). Dynamic gradient-diffusion subgrid models for incompressible magnetohydrodynamic turbulence. *Phys. Plasmas* **9**, 824–834.

Müller, W.-C., Carati, D. (2002b). Large-eddy simulation of magnetohydrodynamic turbulence. *Comp. Phys. Comm.* **147**, 544–547.

Nandy, D., Choudhuri, A.R. (2002). Explaining the latitudinal distribution of sunspots with deep meridional flow. *Science* **296**, 1671–1673.

Noguchi, K., Pariev, V.I., Colgate, S.A., Beckley, H.F., Nordhaus, J. (2002). Magnetorotational instability in liquid metal Couette flow. *Astrophys. J.* **575**, 1151–1162.

Noh, Y., Min, H.S., Raasch, S. (2004). Large eddy simulation of the ocean mixed layer: the effects of wave breaking and Langmuir circulation. *J. Phys. Oceanogr.* **34**, 720–735.

Ohkitani, K. (1993). Eigenvalue problems in three-dimensional Euler flows. *Phys. Fluids A* **5**, 2570–2572.

Orszag, S.A. (1977). Statistical theory of turbulence. In: Balian, R., Peube, J.L. (eds.), *Fluid Dynamics 1973, Les Houches Summer School of Theoretical Physics* (Gordon and Breach, New York), pp. 237–374.

Ott, E., Du, Y., Sreenivasan, K.R., Juneja, A., Suri, A.K. (1992). Sign-singular measures: fast magnetic dynamos, and high-Reynolds-number fluid turbulence. *Phys. Rev. Lett.* **69**, 2654–2657.

Parker, E.N. (1979). *Cosmical Magnetic Fields* (Clarendon Press, New York).

Patera, A. (1984). A spectral element method for fluid dynamics: laminar flow in a channel expansion. *J. Comp. Phys.* **54**, 468–488.

Pelz, R. (2001). Symmetry and the hydrodynamic blow-up problem. *J. Fluid Mech.* **444**, 299–320.

Pétrélis, F., Bourgoin, M., Marié, L., Burguete, J., Chiffaudel, A., Daviaud, F., Fauve, S., Odier, P., Pinton, J.-F. (2003). Nonlinear magnetic induction by helical motion in a liquid sodium turbulent flow. *Phys. Rev. Lett.* **90**, 174501.

Phillips, O.M. (1977). *Dynamics of the Upper Ocean* (Cambridge University Press).

Phillips, W.R.C. (2002). Langmuir circulations beneath growing or decaying surface waves. *J. Fluid Mech.* **469**, 317–342.

Politano, H., Pouquet, A., Sulem, P. (1995). Current and vorticity dynamics in three-dimensional turbulence. *Phys. Plasmas* **2**, 2931–2939.

Ponty, Y., Politano, H., Pinton, J.-F. (2004). Simulation of induction at low magnetic Prandtl number. *Phys. Rev. Lett.* **92**, 144503.

Ponty, Y., Mininni, P.D., Montgomery, D.C., Pinton, J.-F., Politano, H., Pouquet, A. (2005). Numerical study of dynamo action at low magnetic Prandtl numbers. *Phys. Rev. Lett.* **94**, 164502.

Pope, S.B. (2000). *Turbulent Flows* (Cambridge University Press).

Pope, S. (2004). Ten questions concerning the large-eddy simulations of turbulent flows. *New J. Phys.* **6**, 35–58.

Porter, D., Pouquet, A., Woodward, P. (1998). Intermittency in compressible flows. In: Frisch, U. (ed.), *ETC7. Advances in Turbulence VII. Fluid Mechanics and its Applications* (Kluwer), pp. 255–258.

Pouquet, A. (1996). Turbulence, Statistics and Structures: an introduction. In: Chiuderi, C., Einaudi, G. (eds.), *Vth European School in Astrophysics, San Miniato, vol. 468 of Lecture Notes in Physics: Plasma Astrophysics* (Springer-Verlag), pp. 163–217.

Pouquet, A., Patterson, G.S. (1978). Numerical simulation of helical magnetohydrodynamic turbulence. *J. Fluid Mech.* **85**, 305–323.

Pouquet, A., Frisch, U., Léorat, J. (1976). Strong MHD helical turbulence and the non-linear dynamo effect. *J. Fluid Mech.* **77**, 321–354.

Pouquet, A., Meneguzzi, M., Frisch, U. (1986). The growth of correlations in MHD turbulence. *Phys. Rev. A* **33**, 4266–4276.

Remacle, J.-F., Flaherty, J., Shephad, M. (2003). *An Adaptive Discontinuous Galerkin Technique with Qn Orthogonal Basis Applied to Rayleigh-Taylor Flow Instabilities, vol. 45* (SIAM).

Roberts, P.H., Glatzmaier, G.A. (2001). The geodynamo, past, present and future. *Geophys. Astrophys. Fluid Dyn.* **94**, 47–84.

Rosenberg, D., Fournier, A., Fisher, P., Pouquet, A. (2006). Geophysical-astrophysical spectral element adaptive refinement (GASpAR): object-oriented h-adaptive code for geophysical fluid dynamics simulations *J. Comp. Phys.* **215**, 59–80.

Rosenberg, D., Pouquet, A., Mininni, P. (2007). Adaptive mesh refinement with spectral accuracy for magnetohydrodynamics in two space dimensions. *New J. Phys.* **9**, 304.

Sagaut, P. (2006). *Large Eddy Simulation for Incompressible Flows*, Springer Verlag, Berlin, Germany.

Saur, J., Politano, H., Pouquet, A., Matthaeus, W. (2002). Evidence for weak MHD turbulence in the middle magnetoosphere of Jupiter. *Astron. Astrophys.* **386**, 699–708.

Schekochihin, A.A., Boldyrev, S., Kulsrud, R.M. (2002). Spectra and growth rates of fluctuating magnetic fields in the kinematic dynamo theory with large magnetic Prandtl numbers. *Astrophys. J.* **567**, 828–852.

Schekochihin, A.A., Cowley, S.C., Taylor, S.F., Maron, J.L., McWilliams, J.C. (2004). Simulations of the small-scale turbulent dynamo. *Astrophys. J.* **612**, 276–307.

Schekochihin, A., Haugen, N., Brandenburg, A., Cowley, S., Maron, J., McWilliams, J. (2005). The onset of a small-scale turbulent dynamo at low magnetic Prandtl numbers. *Astrophys. J.* **625**, L115–L118.

Scotti, A., Meneveau, C., Lilly, D.K. (1993). Generalized Smagorinsky model for anisotropic grids. *Phys. Fluids A* **5**, 2306–2308.

Sisan, D.R., Shew, W.L., Lathrop, D.P. (2003). Lorentz force effects in magnetoturbulence. *Phys. Earth Plan. Int.* **135**, 137–159.

Skyllingstad, E.D., Denbo, D.W. (1995). An ocean large-eddy simulation of Langmuir circulations and convection in the surface mixed layer. *J. Geophys. Res.* **100**, 8501–8522.

Smagorinsky, J. (1963). General circulation experiments with the primitive equations. I: The basic experiment. *Mon. Wea. Rev.* **91**, 99–165.

Sorriso-Valvo, L., Carbone, V., Noullez, A., Politano, H., Pouquet, A., Veltri, P. (2002). Analysis of cancellation in two-dimensional magnetohydrodynamic turbulence. *Phys. Plasmas* **9**, 89–95.

Soward, A. (ed.) (1983). *Solar and Planetary Magnetism* (Gordon and Breach).

Spence, E.J., Nornberg, M.D., Jacobson, C.M., Kendrick, R.D., Forest, C.B. (2006). Observation of a turbulence-induced large scale magnetic field. *Phys. Rev. Lett.* **96**, 055002.

Steiglitz, R., Müller, U. (2001). Experimental demonstration of a homogeneous two-scale dynamo. *Phys. Fluids* **13**, 561–564.

Sullivan, P.P., McWilliams, J.C., Moeng, C.-H. (1994). A subgrid-scale model for largeeddy simulation of planetary boundary-layer flows. *Bound. Layer Meteorol.* **71**, 247–276.

SULLIVAN, P.P., MCWILLIAMS, J.C., MOENG, C.-H. (1996). A grid nesting method for largeeddy simulation of planetary boundary layer flows. *Bound. Layer Meteorol.* **80**, 167–202.

SULLIVAN, P.P., HORST, T.W., LENSCHOW, D.H., MOENG, C.-H., WEIL, J.C. (2003). Structure of subfilter-scale fluxes in the atmospheric surface layer with application to largeeddy simulation modeling. *J. Fluid Mech.* **482**, 101–139.

SULLIVAN, P.P., MCWILLIAMS, J.C., MELVILLE, W.K. (2004). The oceanic boundary layer driven by wave breaking with stochastic variability. I: Direct numerical simulations. *J. Fluid Mech.* **507**, 143–174.

SULLIVAN, P.P., MCWILLIAMS, J.C., MELVILLE, W.K. (2005). Surface waves and ocean mixing: insights from numerical simulations with stochastic surface forcing. In: *14th 'Aha Huliko'a Hawaiian Winter Workshop on Rogue Waves* pp. 147–154.

SYTINE, I., PORTER, D., WOODWARD, P., HODSON, S., WINKLER, K. (2000). Convergence tests for the Piecewise Parabolic Method and Navier-Stokes solutions for homogeneous compressible turbulence. *J. Comp. Phys.* **158**, 225–238.

TENNEKES, H., LUMLEY, J. (1972). *A First Course in Turbulence* (MIT Press).

TERRAY, E.A., DONELAN, M.A., AGRAWAL, Y.C., DRENNAN, W.M., KAHMA, K.K., WILLIAMS, A.J., HWANG, P.A., KITAIGORODSKII, S.A. (1996). Estimates of kinetic energy dissipation under breaking waves. *J. Phys. Oceanogr.* **26**, 792–807.

THEOBALD, M.L., FOX, P.A., SOFIA, S. (1994). A subgrid-scale resistivity for magnetohydrodynamics. *Phys. Plasmas* **1**, 3016–3032.

THORPE, S.A. (2003). Langmuir circulation. *Ann. Rev. Fluid Mech.* **36**, 55–79.

TING, A.C., MATTHAEUS, W.H., MONTGOMERY, D. (1986). Turbulent relaxation processes in magnetohydrodynamics. *Phys. Fluids* **29**, 3261–3274.

VERMA, M.K. (2004). Statistical theory of magnetohydrodynamic turbulence: recent results. *Phys. Rep.* **401**, 229–380.

VERMA, M.K., KUMAR, S. (2004). Large-eddy simulations of fluid and magnetohydrodynamic turbulence using renormalized parameters. *Pramana J. Phys.* **63**, 553–561.

VIEILLEFOSSE, P. (1984). Internal motion of a small element of fluid in an inviscid flow. *Phys. A* **125**, 150–162.

VINCENT, A., MENEGUZZI, M. (1991). The spatial structure and statistical properties of homogeneous turbulence. *J. Fluid Mech.* **225**, 1–20.

WINCKELMANS, G.S., WRAY, A.A., VASILYEV, O.V., JEANMART, H. (2001). Explicit-filtering large-eddy simulation using the tensor-diffusivity model supplemented by a dynamic Smagorinsky term. *Phys. Fluids* **13**, 1385–1403.

WYNGAARD, J.C. (2004). Toward numerical modeling in the Terra Incognita. *J. Atmos. Sci.* **61**, 1816–1826.

YEUNG, P., DONZIS, D., SREENIVASAN, K. (2005). High Reynolds number simulation of turbulent mixing. *Phys. Fluids* **17**, 081703.

YOSHIZAWA, A. (1985). Statistical theory for magnetohydrodynamic turbulent shear flows. *Phys. Fluids* **28**, 3313–3320.

YOSHIZAWA, A. (1987). Subgrid modeling for magnetohydrodynamic turbulent shear flows. *Phys. Fluids* **30**, 1089–1095.

YOSHIZAWA, A. (1990). Self-consistent turbulent dynamo modeling of reversed field pinches and planetary magnetic fields. *Phys. Fluids B* **2**, 1589–1600.

ZHAO, H., MOHSENI, K. (2005). A dynamic model for the Lagrangian-averaged Navier-Stokes-α equations. *Phys. Fluids* **17**, 075106.

ZHOU, H. (1999). Numerical simulation of Langmuir circulations in a wavy domain and its comparison with the Craik-Leibovich theory, PhD thesis, Stanford University.

ZHOU, Y., SCHILLING, O., GHOSH, S. (2002). Subgrid scale and backscatter model for magnetohydrodynamic turbulence based on closure theory: theoretical formulation. *Phys. Rev. E* **66**, 026309.

Data Assimilation for Geophysical Fluids

Jacques Blum

Laboratoire Jean-Alexandre Dieudonné, Université de Nice Sophia-Antipolis, Parc Valrose, 06108 Nice cedex 2, France

François-Xavier Le Dimet

INRIA and Laboratoire Jean-Kuntzmann, Université de Grenoble, 38041, Grenoble Cedex 9, France

I. Michael Navon

Department of Scientific Computing, The Florida State University, Tallahassee, Florida 32306-4120, USA

Abstract

The ultimate purpose of environmental studies is the forecast of its natural evolution. A prerequisite before a prediction is to retrieve at best the state of the environment. Data assimilation is the ensemble of techniques which, starting from heterogeneous information, permit to retrieve the initial state of a flow. In the first part, the mathematical models governing geophysical flows are presented together with the networks of observations of the atmosphere and of the ocean. In variational methods, we seek for the minimum of a functional estimating the discrepancy between the solution of the model and the observation. The derivation of the optimality system, using the adjoint state, permits to compute a gradient which is used in the optimization. The definition of the cost function permits to take into account the available statistical information through the choice of metrics in the space of observation and in the space of the initial condition. Some examples are presented on simplified models, especially an application in oceanography. Among the tools of optimal control, the adjoint model permits to carry out sensitivity studies, but if we look for the sensitivity of the prediction with respect to the observations, then a second-order analysis should be considered. One of the first methods used for assimilating data in oceanography is the nudging method, adding a forcing term in the equations. A variational variant of nudging method is described and also a so-called

Computational Methods for the Atmosphere and the Oceans
Special Volume (Roger M. Temam and Joseph J. Tribbia, Guest Editors) of
HANDBOOK OF NUMERICAL ANALYSIS, VOL. XIV
P.G. Ciarlet (Editor)

Copyright © 2009 Elsevier B.V.
All rights reserved

ISSN 1570-8659
DOI 10.1016/S1570-8659(08)00209-3

"back and forth" nudging method. The proper orthogonal decomposition method is introduced in order to reduce the cost of the variational method. For assimilating data, stochastic methods can be considered, being based on the Kalman filter extended to nonlinear problems, but the inconvenience of this method consists in the difficulty of handling huge covariance matrices. The dimension of the systems used for operational purposes (several hundred of millions of variables) requires to work with reduced variable techniques. The ensemble Kalman filter method, which is a Monte-Carlo implementation of the Bayesian update problem, is described. A considerable amount of information on geophysical flows is provided by satellites displaying images of their evolution, the assimilation of images into numerical models is a challenge for the future: variational methods are successfully considered in this perspective.

1. Introduction: specificity of geophysical flows

The mathematical modeling of geophysical flows has experienced a tremendous development during the last decades, mainly due to the growth of the available computing resources and to the development of networks of remote or in situ observations. The domain of modeling has been extended to complex flows such as the atmosphere with some chemical species or coupled media such as the atmosphere and the ocean. A tentative extension of the domain of prediction is also under way with seasonal prediction on one hand and short term and very accurate prediction on the other hand: nowcasting mainly devoted to extreme events. Geophysical fluids such as air, atmosphere, ocean, surface, or underground water are governed by the general equations of fluid dynamics: mass and energy conservation, behavior laws. Nevertheless some specific factors must be taken into account such as follows:

- Uniqueness of a situation. Each geophysical episode is unique. A given situation has never existed before and will not exist in the future. A field experiment cannot be duplicated. It means that environmental sciences are not strictu sensu experimental sciences: an hypothesis cannot be validated by repetitions of an experiment. Geophysical models should be tested and validated with data associated to distinct episodes.
- Nonlinearity. Geophysical processes are nonlinear due to their fluid component, and furthermore, they include some other nonlinear processes such as radiative transfer. Nonlinearity implies interactions and energy cascades between spatial and temporal scales. Seeking a numerical solution to the equations requires discretizing these equations and therefore cutting off some of the scales. A major problem comes from the fact that subscale processes could be associated with large fluxes of energy. For example, a cumulonimbus cloud has a characteristic size of 10 km in the horizontal and vertical directions. The typical grid size of a general circulation model (GCM) is of the order of 40 km, therefore, larger than the characteristic dimension of a cumulonimbus, a thunderstorm cloud. The total energy (thermal and mechanical) of such a cloud is considerable. By the same token typical vertical velocities of a GCM are of the order of some centimeters or decimeters per second,

in a cumulonimbus cloud there exist observations of vertical velocities of the order of 100 meters per second. Therefore, it will be crucial to represent the fluxes of energy associated to subgrid processes by some additional terms in the equations. Parametrization of subgrid effects will include some empirical coefficients that should be tuned in such a way that the model produces a good forecast.

- Initial and boundary conditions. The general equations are not sufficient to carry out a prediction. Some additional information, such as initial and boundary conditions, should be provided. Most of the geophysical fluids do not have any natural boundaries. In the same way, there are no natural initial conditions. Nevertheless, these conditions are essential for carrying out a prediction, more especially as the system is turbulent and hence very much dependent on the initial condition.

Therefore, it is crystal clear that modeling will have to take into account observations. If, for instance, we consider a measurement of the wind at a given site, the same data can be used either in a GCM or in a very local model. According to the context, i.e., the model, a different confidence will be attributed to the observation. It does not make to have a model without data or data without model (otherwise known as "Lions' theorem").

Data assimilation is the ensemble of techniques combining in an optimal way (in a sense to be defined) the mathematical information provided by the equations and the physical information given by the observation in order to retrieve the state of a flow. The concept of data assimilation can be extended to other sources of information, e.g., statistics of error on the observations and/or on the error of prediction. Another source of information is provided by images originating from space observations, which as of the present time are not optimally used. The goal of data assimilation is to link together these heterogeneous (in nature, quality, and density) sources of information in order to retrieve a coherent state of the environment at a given date.

The equations of the model (shallow water, quasi-geostrophic (QG), or general primitive equations) are of the first order with respect to time. In a GCM, there is no lateral boundary condition. Assuming that all the regularity conditions are fulfil (if they were known), an initial condition will be sufficient to integrate the model and to get the forecast.

Originally, the problem of data assimilation was to determine the initial condition from observations. Since the same mathematical tools are used, data assimilation also includes the estimation of some model parameters or of some boundary conditions.

As a first approximation, three types of methods are considered:

- Interpolation methods. These methods interpolate the measurement from the points of observation toward the grid points. The interpolation can be weighted by the statistics of the observations (covariance matrices). The method is simple to implement, but it is not justified by physical arguments: the retrieved fields will not be necessarily coherent from the physical viewpoint, e.g., the initial condition may be located outside of the attractor, therefore the associated solution will contain gravity waves until it reaches the attractor. Until a recent date, these methods were the most commonly used in operational meteorology. For a review on optimal interpolation, one can refer to KALNAY [2003].

- Variational methods. Data assimilation is set as being a problem of constrained optimization, then the tools of optimal control are carried out to solve it. At present, these methods are operationally used at the European Center for Medium Range Weather Forecasting (ECMWF) (Reading, UK), Météo-France, the National Center for Environmental Prediction (USA), Japan and Canada (see RABIER [2005], KALNAY [2003]).
- Stochastic methods. The basic idea is to consider the fields as the realization of a stochastic process and carry out Kalman filtering (KF) methods. The main difficulty stems from the fact that the covariance matrices of the state variables have huge dimensions for operational models. The ensemble Kalman filtering (EnKF) methods were devised to address this issue and are presently seeing a major development at different research centers.

Section 2 will be devoted to the presentation of a certain number of simplified models for the geophysical flows and to a description, with appropriate figures, of data available for the atmosphere or the ocean. In Section 3, the variational method, mentioned above and often called Four-Dimensional Variational Data Assimilation (4D-VAR) is explained in detail and two examples of solutions for shallow water or QG models will be given. Section 4 is devoted to a second-order adjoint analysis, which enables in particular to perform a sensitivity analysis on the results of the variational method. In Section 5, the so-called nudging method is explained, with a special emphasis on the optimal nudging method, which uses the variational technique. In order to reduce the cost of these 4D-VAR methods, the Proper Orthogonal Decomposition (POD) is introduced in Section 6, and the application of this reduced-space basis to variational methods is presented. In Section 7, KF is introduced and a special emphasis is given to the EnKF, which is widely used in operational data assimilation. Finally, the recent problem of assimilation of images in meteorology or oceanography is tackled in Section 8 by a variational technique.

2. Models and data for geophysical flows

2.1. Models

The equations governing the geophysical flows are derived from the general equations of fluid dynamics. The main variables used to describe the fluids are as follows:

- The components of the velocity
- Pressure
- Temperature
- Humidity in the atmosphere, salinity in the ocean
- Concentrations for chemical species

The constraints applied to these variables are as follows:

- Equations of mass conservation
- Momentum equation containing the Coriolis acceleration term, which is essential in the dynamic of flows at extra tropical latitudes

- Equation of energy conservation including law of thermodynamics
- Law of behavior (e.g., Mariotte's Law)
- Equations of chemical kinetics if a pollution type problem is considered

These equations are complex; therefore, we cannot expect to obtain an analytical solution. Before performing a numerical analysis of the system, it will be necessary to

- Simplify the equations. This task will be carried out on physical basis. For example, three-dimensional (3D) fields could be vertically integrated using hydrostatic assumptions in order to obtain a two-dimensional (2D) horizontal field which is more tractable for numerical purposes.
- Discretize the equations. The usual discretization methods are considered: finite differences, finite elements, or spectral methods.

Several of these techniques may be simultaneously used. For instance, in the ARPEGE model designed by Meteo-France, the horizontal discretization is spectral in longitude, a truncated development in Legendre's polynomial along latitude, while the vertical one is based on a finite difference scheme. The horizontal nonlinear terms (advection) are computed using finite differences and then transformed onto a spectral base. As mentioned above, it will be necessary to estimate the subgrid fluxes of energy and matter. A parametrization of these phenomena will be included in the model, which will contain some empirical parameters, difficult to estimate and to adjust because they are associated to some complex physical processes which cannot be directly measured.

In mathematics, it is usual to study the convergence of discrete models toward a continuous one when the discretization parameter goes to zero. Does this approach make sense for these problems? According to the value of this parameter, the physics of the problem will change and another parametrization will be necessary. With a grid size of 100 km, the cumulonimbus clouds will not be explicit in the model. With a grid size of 100 m, the physics of convective clouds, including the water cycle under gaseous, liquid, and solid phases, should be explicitly taken into account. According to the value of the discretization parameter, the domain of validity of the approximation will change. In this framework, the convergence of discretization schemes is beyond the scope of actual problems.

In meteorology as well as in oceanography or hydrology, it is usual to use a basic "toy" model for numerical experiments: the Saint-Venant's equations (or shallow water equations). These 2D horizontal equations are obtained after a vertical integration of the 3D fields, assuming the hydrostatic approximation, this is equivalent to neglecting the vertical acceleration. The density is supposed to be constant; therefore, there is no thermodynamic effect. They assume the form:

$$\begin{aligned}
&\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} - fv + \frac{\partial \phi}{\partial x} = 0,\\
&\frac{\partial v}{\partial t} + u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y} + fu + \frac{\partial \phi}{\partial y} = 0,\\
&\frac{\partial \phi}{\partial t} + \frac{\partial u\phi}{\partial x} + \frac{\partial v\phi}{\partial y} = 0,
\end{aligned} \tag{2.1}$$

where u and v are the horizontal components of the velocity, ϕ is the geopotential (proportional to the pressure), and f is the Coriolis parameter for earth's rotation.

An important property of these equations is the existence of an attractor, on which the solution will orbit. For very simplified models such as Lorenz' equations of 3Ds in the space phase, the structure of this attractor is already topologically complex. For more realistic models, it is known that the attractor is, at midlatitudes, in the vicinity of the geostrophic equilibrium, which is defined by the equality of the gradient of pressure and the Coriolis force. From a practical point of view, the attractor is characterized by a weak development of gravity waves. Therefore, if the initial condition does not belong to a close neighborhood of the attractor, then the integration of the model will give rise to undue gravity waves until the solution reaches the attractor.

Another "toy" model, which filters the gravity waves, is the QG model where the dominant terms are the pressure gradient and the Coriolis force, which cancel each other in the geostrophic balance. It consists in a first-order expansion of the Navier-Stokes equation with respect to the Rossby number. It is an approximate model with respect to the full primitive equation model, in particular because thermodynamics are discarded. However, it has been shown to be able to realistically reproduce the statistical properties of midlatitude ocean circulation including the very energetic jet and mesoscale features, typical of regions like the Gulf Stream. The equations of this model will be given in the next section.

Operational models in meteorology and oceanography are of very large dimension with 10^7 to 10^9 variables, hence the implementation of efficient numerical methods is a challenge for high performance computing. Future developments of coupled models ocean-atmosphere will dramatically increase the need for efficient numerical methods for coupled models.

2.2. Data

At the present time, many sources of data are used. Around 300 millions of data are screened every day by the ECMWF located in Reading (UK). An exhaustive information can be found on the ECMWF site http://www.ecmwf.int/products/forecasts/d/charts/monitoring/coverage/.

In meteorology, the data collected for operational use are as follows:

- Ground observations: wind, temperature, pressure, humidity. These data are collected on a dedicated network, on ships, and also in airports (see Fig. 2.1). The number of observations varies from day to day. To give an order of magnitude, about 200,000 data are measured daily, but after a quality control process, only 30,000 are used in the assimilation.
- Pilot balloons provide information on the wind.
- Radiosondes provide data on the vertical structure of the atmosphere: wind, temperature, pressure, and humidity. This network is displayed in Fig. 2.2. The figures of synoptic measurements and radiosonde clearly show the heterogeneity of the data density: North America and Europe have a good coverage, while information is very sparse on the oceans. This lack of information is compensated by drifting buoys.

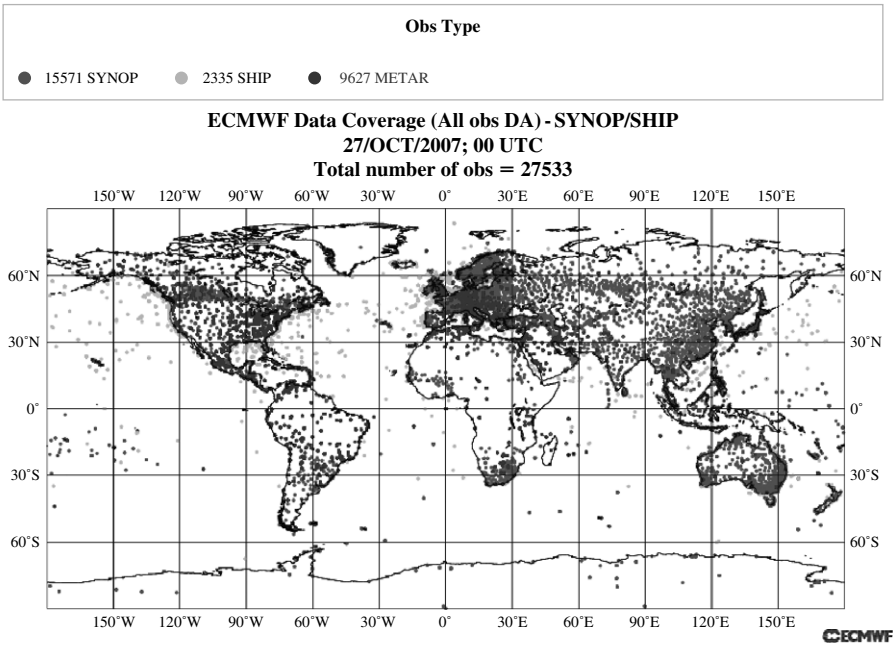


FIG. 2.1 SYNOP/SHIP data: synoptic networks in red, airport data in blue, and ship data in green. (See also color insert).

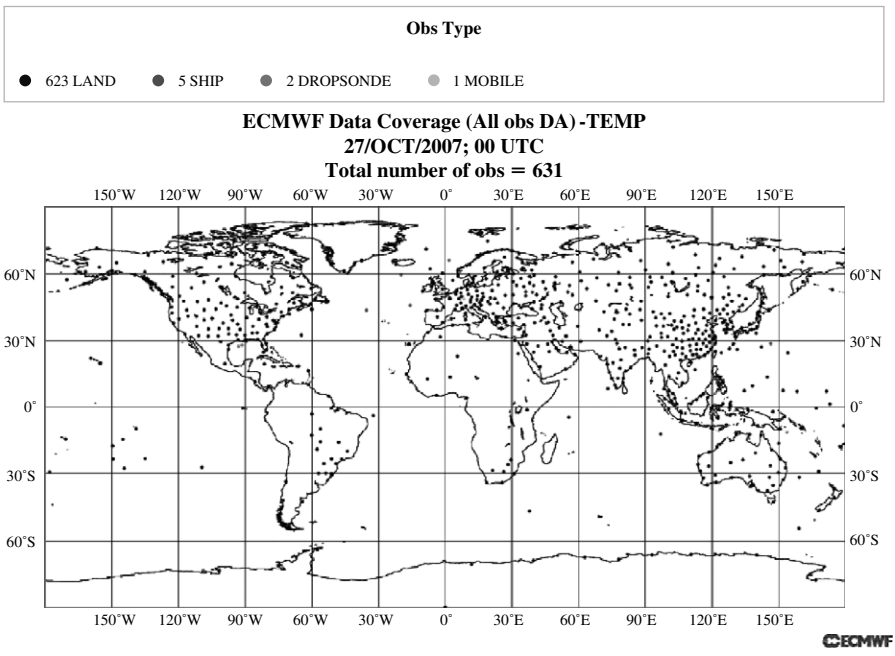


FIG. 2.2 Radiosonde measurements. (See also color insert).

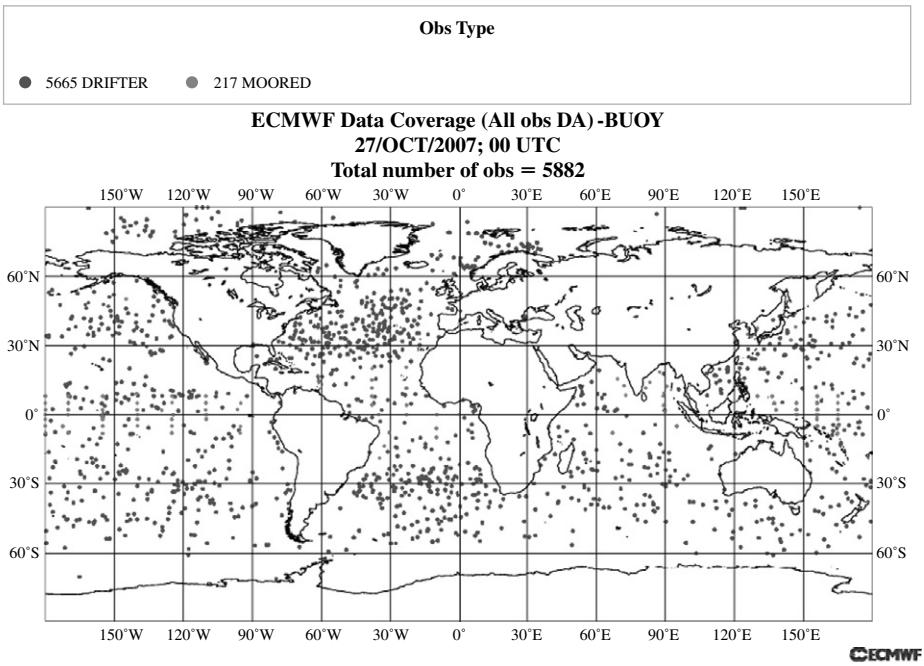


FIG. 2.3 Drifting and moored buoys. (See also color insert).

- Drifting buoys measure the temperature of the air and of the ocean, salinity, and wind. The network is completed by moored buoys, mainly located in the most energetic part of the ocean (see Fig. 2.3). These data are also used for oceanic models. Around 6000 data provided by buoys are daily assimilated at ECMWF.

At the present time (2007), around 99 percent of screened data originates from 45 satellites, but only 94 percent of assimilated data comes from satellites. Two main categories of satellites are used:

- Geostationary satellites provide information on the wind by estimating the shifting of clouds considered as Lagrangian tracers. To make this measurement useful, the altitude of the clouds must be known (it is derived, by solving an inverse problem, from the vertical temperature profiles). Figure 2.4 displays the areas of observation covered by ten geostationary satellites. Around 300,000 observations are assimilated.
- Polar-orbiting satellites (NOAA, EUMETSAT) are used for the estimation of the vertical temperature profiles, basically radiances are measured, then temperatures are estimated as the solution of an inverse problem. Figure 2.5 displays the trajectories of six satellites on October 27th 2007. Around 400,000 observations are assimilated.

In oceanography, data are much scarcer than in meteorology. Electromagnetic waves do not penetrate deeply in the ocean; therefore, remote sensing is much more difficult. The

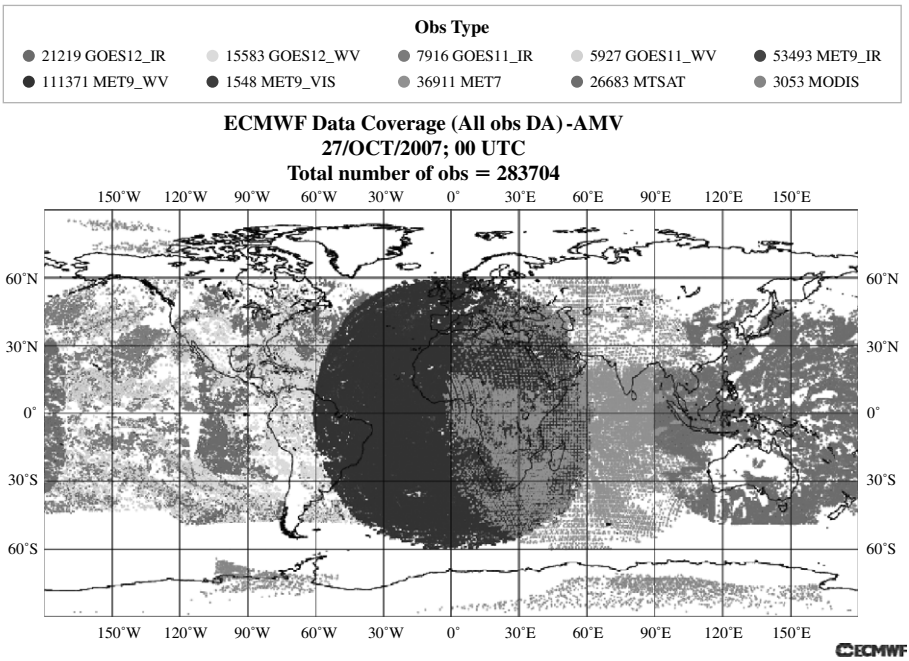


FIG. 2.4 Observations from ten geostationary satellites. (See also color insert).

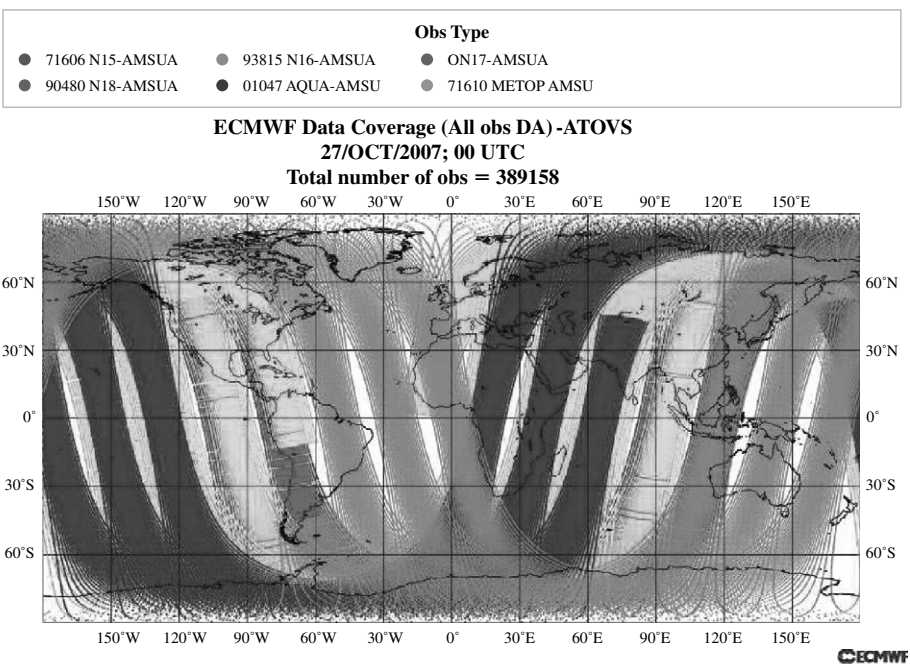


FIG. 2.5 Trajectories of six polar-orbiting satellites. (See also color insert).

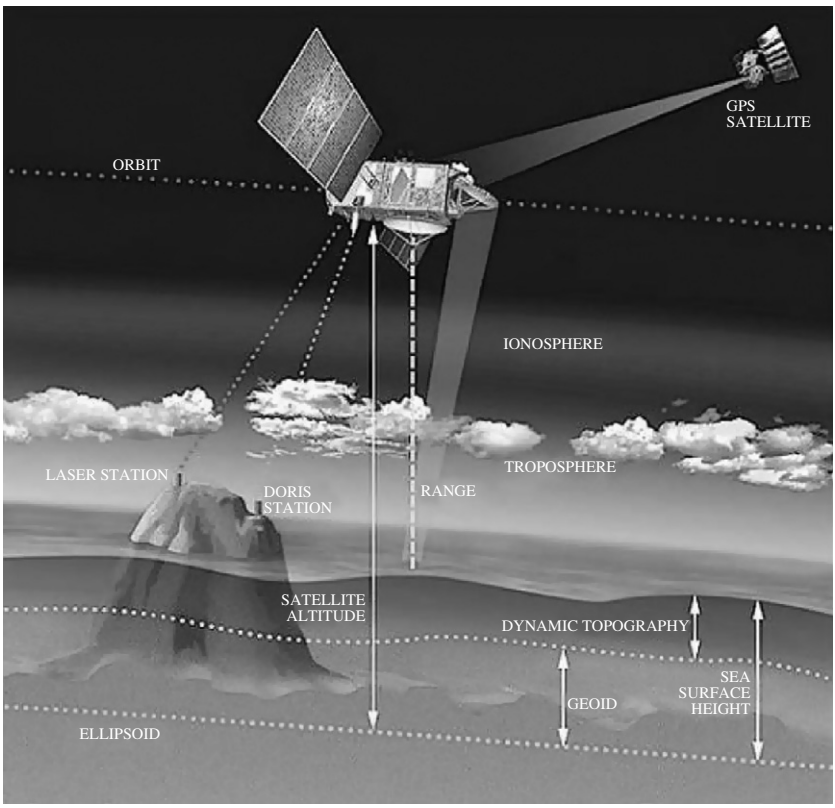


FIG. 2.6 Satellite altimetry (CNES figure from Aviso web site http://www.aviso.oceanobs.com). (See also color insert).

development of operational oceanography is based on in situ measurements (temperature, salinity) and on altimetric satellites (ERS, Topex-Poseidon, Jason) measuring the surface elevation of the ocean with precision on the order of some centimeter (see Fig. 2.6). Lagrangian floats measure also the position of the drifting buoys with a given time periodicity (NODET [2006]).

Both in meteorology and oceanography, observations are heterogeneous in nature, density, and quality. A rough estimation of the number of screened data is 300 millions and of assimilated data is around 18 millions. This number has to be compared with the 800 millions of variables of the ECMWF operational model in 2007. Therefore, retrieving the state of the atmosphere from observations is clearly an ill-posed problem. Some a priori information has to be provided to estimate an initial state.

3. Variational methods

Variational methods were introduced by SASAKI [1958]. These methods consider the equations governing the flow as constraints, and the problem is closed by using a

variational principle, e.g., the minimization of the discrepancy between the model and the observations. In the following, we will consider that the model is discrete with respect to space variables. Optimal Control Techniques (LIONS [1971]) were proposed by LE DIMET [1980], LE DIMET and TALAGRAND [1986], and COURTIER and TALAGRAND [1987].

3.1. Ingredients

The various ingredients of a variational method are as follows:

- A state variable $X \in \mathcal{X}$ which describes the evolution of the medium at the grid points. X depends on time and is for operational models of large dimension (3.10^7 for the ECMWF model).
- A model describing the evolution of the fluid. Basically, it is a system of nonlinear differential equations which is written as

$$\begin{cases} \dfrac{\mathrm{d}X}{\mathrm{d}t} = F(X, U) \\ X(0) = V \end{cases} . \tag{3.1}$$

- A control variable $(U, V) \in \mathcal{P}$ space of control. Most of the time the control is the initial condition or/and some internal variables of the model: parameters or boundary conditions. We will assume that when a value has been fixed for the parameter, then the model has a unique solution. For sake of simplicity, we will not consider constraints on the state variable. Nevertheless, humidity and salinity cannot be negative; therefore, the set of controls does not necessarily have the structure of a vector space.
- Observations $X_{obs} \in \mathcal{O}_{obs}$. They are discrete and depend on time and space and are not, from either geographical or physical point of view, in the same space as the state variable. Therefore, we will introduce some operator C mapping the space of state into the state of observations. In practice, this operator can be complex.
- A cost function J measuring the discrepancy of the solution of the model associated to (U, V) and the observations.

$$J(U, V) = \frac{1}{2} \int_0^T \|C.X(U, V) - X_{obs}\|^2 \mathrm{d}t \tag{3.2}$$

The choice of the norm is important because it allows introduction of some a priori information like the statistics of the fields through the covariance matrix which is positive definite. In practice, some additional term is added to the cost function, e.g., the so-called background term which is the quadratic difference between the initial optimal variable and the last prediction. This term acts like a regularization term in the sense of Tikhonov (TIKHONOV and ARSENIN [1977]).

Then the problem of variational data assimilation (VDA) can be set as

$$\begin{cases} \text{Find} \quad U^*, V^* \in \mathcal{P} \quad \text{such that} \\ J(U^*, V^*) = \inf\limits_{(U,V)\in\mathcal{P}} J(U, V). \end{cases} \tag{3.3}$$

3.2. *Optimality system*

With respect to (U, V), we have a problem of unconstrained optimization. Problem in Eq. (3.3) will have a unique solution if J is strictly convex, lower semicontinuous and if

$$\lim_{||(U,V)||\to+\infty} J(U, V) \to +\infty.$$

When J is differentiable, a necessary condition for (U^*, V^*) to be a solution is given by the Euler-Lagrange equation:

$$\nabla J(U^*, V^*) = 0,$$

where ∇J is the gradient of J with respect to (U, V).

Furthermore, the determination of ∇J permits one to implement optimization methods of gradient type.

Let $(u, v) \in \mathcal{P}$, $\widehat{X}$ be the Gâteaux-derivative (directional derivative) of X in the direction (u, v) that is the solution of

$$\begin{cases} \dfrac{\mathrm{d}\widehat{X}}{\mathrm{d}t} = \left[\dfrac{\partial F}{\partial X}\right].\widehat{X} + \left[\dfrac{\partial F}{\partial U}\right].u, \\ \widehat{X}(0) = v \end{cases} \tag{3.4}$$

where $\left[\dfrac{\partial F}{\partial X}\right]$ is the Jacobian of the model with respect to the state variable. This Eq. (3.4) is known as the *linear tangent model*.

By the same token, we get the directional derivative of J:

$$\widehat{J}\,(U, V, u, v) = \int_0^T \big(C.X - X_{obs}, C\widehat{X}\big)\mathrm{d}t.$$

We will get the gradient by exhibiting the linear dependence of $\widehat{J}$ with respect to (u, v). For this purpose, we introduce $P \in \mathcal{X}$ the so-called *adjoint variable*, to be defined later. Let us take the inner product of Eq. (3.4) with P, then integrate between 0 and T. An integration by part shows that if P is defined as the solution of

$$\begin{cases} \dfrac{\mathrm{d}P}{\mathrm{d}t} + \left[\dfrac{\partial F}{\partial X}\right]^T.P = C^t\big(C.X - X_{obs}\big) \\ P\,(T) = 0, \end{cases} \tag{3.5}$$

then the gradient is given by

$$\nabla J = \begin{pmatrix} \nabla_U J \\ \nabla_V J \end{pmatrix} = \begin{pmatrix} -\left[\dfrac{\partial F}{\partial U}\right]^t . P \\ -P\,(0) \end{pmatrix}.$$

Therefore, the gradient is obtained by a backward in time integration of the adjoint model.

3.3. Optimization

The determination of (U^*, V^*) is carried out by performing a descent-type unconstrained optimization method. Given a first guess (U_0, V_0), we define a sequence by

$$\begin{pmatrix} U_n \\ V_n \end{pmatrix} = \begin{pmatrix} U_{n-1} \\ V_{n-1} \end{pmatrix} + \rho_n D_n.$$

D_n is the direction of descent. Usually conjugate gradient or Newton type methods are used. ρ_n is the step size defined by

$$\rho_n = ArgMin J \left(\begin{pmatrix} U_{n-1} \\ V_{n-1} \end{pmatrix} + \rho D_n \right),$$

This problem looks simple: it is the minimization of a function of one variable. For a nonlinear problem, it entails a high computational cost since several integrations of the model are required for the evaluation of J. Optimization libraries, e.g., MODULOPT (Gilbert and Lemarechal [1989]), are widely used and are efficient.

For a comprehensive test of powerful large-scale unconstrained minimization methods applied to VDA, see Zou, Navon, Berger, Phua, Schlick and Le Dimet [1993].

3.4. Implementation

A major difficulty encountered in the implementation of this method is the derivation of the adjoint model. A bad solution would be to derive the adjoint model from the continuous direct model, then to discretize it. The convergence of the optimization algorithm requires having the gradient of the cost function with a precision of the order of the computer's roundoff error.
Two steps are carried out for the derivation of the adjoint:

- Differentiation of the direct model. This step serves to determine the linear tangent model. This task is easily carried out by differentiating the direct code line by line.
- Transposition of the linear tangent model. Transposition with respect to time is simply the backward integration. To carry out the transposition, one should start from the last statement of the linear tangent code and transpose each statement. The difficulty stems from the hidden dependencies.

If some rules in the direct code are adhered to, then the derivation of the adjoint model can be made simpler, otherwise it is a long and painful task. Nevertheless, we

can use some automatic differentiation code such as Odyssee (ROSTAING-SCHMIDT and HASSOLD [1994]) (see also TAPENADE, TAMC, FASTOPT). Recent developments on these techniques can be found in MOHAMMADI and PIRONNEAU [2001].

3.5. Remarks

- If the model is nonlinear, then the cost function is not necessarily convex, and the optimization algorithm may converge toward a local minimum. In this case, one can expect convergence toward a global optimum only if the first guess is in the vicinity of the solution. This may occur in meteorology where the former forecast is supposed to be close to the actual state of the atmosphere. In practice, a so-called background term is added to the cost function, measuring the quadratic discrepancy with the prediction. In terms of control, this term could be considered as a regularization in the sense of Tikhonov.
- The optimization algorithm could converge to a correct mathematical solution but would be physically incorrect (e.g., negative humidity).The solution may be far away from the attractor, the regularization term will force the model to verify some additional constraints, e.g., for the solution to remain in the vicinity of the geostrophic equilibrium.
- Regularization terms permit to take into account the statistical information on the error by an adequate choice of the quadratic norm including an error covariance matrix.
- If the control variable U is time dependent, which is the case if boundary conditions are controlled, then we may get problems with a huge dimension. In this case, it will be important to choose an appropriate discretization of the control variable in order to reduce its dimension.
- PUEL [2002] has proposed a new approach to determine the final state, instead of the initial condition, that makes the inverse problem well-posed, thanks to Carleman inequalities.

3.6. Example 1: Saint-Venant's equations

Saint-Venant's equations, also known as shallow water equations, are used for an incompressible fluid for which the depth is small with respect to the horizontal dimensions. General equations of geophysical fluid dynamics are vertically integrated using the hydrostatic hypothesis; therefore, vertical acceleration is neglected. In Cartesian coordinates, they are

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} - fv + \frac{\partial \phi}{\partial x} = 0, \tag{3.6}$$

$$\frac{\partial v}{\partial t} + u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y} + fu + \frac{\partial \phi}{\partial y} = 0, \tag{3.7}$$

$$\frac{\partial \phi}{\partial t} + \frac{\partial u\phi}{\partial x} + \frac{\partial v\phi}{\partial y} = 0. \tag{3.8}$$

In this system, $X = (u, v, \phi)^T$ is the state variable, u and v are the components of the horizontal velocity; ϕ is the geopotential (proportional to the height of the free surface) and f the Coriolis parameter. For sake of simplicity, the following hypotheses are used:

a) The error of the model is neglected. Only the initial condition will be considered as a control variable.

b) Lateral boundary conditions are periodic. This is verified in global models.

c) Observations are supposed to be continuous with respect to time. Of course, this is not the case in practice. $\mathbf{C} \equiv \mathbf{I}$, where $\mathbf{I}$ is the identity operator. If $U_0 = (u_0, v_0, \phi_0)^T$ is the initial condition and if the cost function is given by

$$J(U_0) = \frac{1}{2} \int_0^T [\|u - u_{obs}\|^2 + \|v - v_{obs}\|^2 + \gamma \|\phi - \phi_{obs}\|^2] \mathrm{d}t, \tag{3.9}$$

where γ is a weight function, then the directional derivatives $\overline{X} = (\bar{u}, \bar{v}, \bar{\phi})^T$ in the direction $h = (h_u, h_v, h_\phi)^T$ (in the control space) will be solutions of the linear tangent model:

$$\frac{\partial \bar{u}}{\partial t} + u\frac{\partial \bar{u}}{\partial x} + \bar{u}\frac{\partial u}{\partial x} + v\frac{\partial \bar{u}}{\partial y} + \bar{v}\frac{\partial u}{\partial y} - f\bar{v} + \frac{\partial \bar{\phi}}{\partial x} = 0, \tag{3.10}$$

$$\frac{\partial \bar{v}}{\partial t} + u\frac{\partial \bar{v}}{\partial x} + \bar{u}\frac{\partial v}{\partial x} + v\frac{\partial \bar{v}}{\partial y} + \bar{v}\frac{\partial v}{\partial y} + f\bar{u} + \frac{\partial \bar{\phi}}{\partial y} = 0, \tag{3.11}$$

$$\frac{\partial \bar{\phi}}{\partial t} + \frac{\partial \bar{u}\phi}{\partial x} + \frac{\partial u\bar{\phi}}{\partial x} + \frac{\partial \bar{v}\phi}{\partial y} + \frac{\partial v\bar{\phi}}{\partial y} = 0. \tag{3.12}$$

The adjoint model is obtained by transposition of the linear tangent model. Let $P = (\tilde{u}, \tilde{v}, \tilde{\phi})^T$ be the adjoint variable and after some integrations by parts both in time and space, we see that the adjoint model is defined as being the solution of

$$\frac{\partial \tilde{u}}{\partial t} + u\frac{\partial \tilde{u}}{\partial x} + v\frac{\partial \tilde{u}}{\partial y} + \tilde{u}\frac{\partial v}{\partial y} - \tilde{v}\frac{\partial v}{\partial x} - f\tilde{v} + \phi\frac{\partial \tilde{\phi}}{\partial x} = u - u_{obs}, \tag{3.13}$$

$$\frac{\partial \tilde{v}}{\partial t} - \tilde{u}\frac{\partial u}{\partial y} + u\frac{\partial \tilde{v}}{\partial x} + \tilde{v}\frac{\partial u}{\partial x} + v\frac{\partial \tilde{v}}{\partial y} + f\tilde{u} + \phi\frac{\partial \tilde{\phi}}{\partial y} = v - v_{obs}, \tag{3.14}$$

$$\frac{\partial \tilde{\phi}}{\partial t} + \frac{\partial \tilde{u}}{\partial x} + \frac{\partial \tilde{v}}{\partial y} + u\frac{\partial \tilde{\phi}}{\partial x} + v\frac{\partial \tilde{\phi}}{\partial y} = \gamma(\phi - \phi_{obs}), \tag{3.15}$$

with final conditions equal to 0.

Then the gradient of J is given by

$$\nabla J(U_0) = -P(0) = -\begin{pmatrix} \tilde{u}(0) \\ \tilde{v}(0) \\ \tilde{\phi}(0) \end{pmatrix}. \tag{3.16}$$

In VIDARD [2001], a square oceanic shallow water model (2000 km) is studied, it is discretized with a grid size of 25 km. The period of assimilation lasts one month, and the time-step is 90 mn. Fictitious data provided by the true solution are used after a random perturbation. The optimization code M1QN3 is issued from the MODULOPT optimization library (GILBERT and LEMARECHAL [1989]): it is a quasi Newton algorithm.

3.7. Example 2: a QG model

The oceanic model used in this study is based on the QG approximation obtained by writing the conservation of the potential vorticity (HOLLAND [1978]). The vertical structure of the ocean is divided into N layers. Each one has a constant density ρ_k with a depth H_k ($k = 1, \ldots, N$). We get a coupled system of N equations:

$$\frac{D_k(\theta_k(\Psi) + f)}{Dt} + \delta_{k,N}\, C_1 \Delta \Psi_N - C_3 \Delta^3 \Psi_k \;=\; F_k \quad \begin{array}{l} \text{dans } \Omega \times [0, T], \\ \forall k = 1, \ldots, N, \end{array} \tag{3.17}$$

where

- $\Omega \subset \mathbb{R}^2$ is the oceanic basin, and $[0, T]$ the time interval for the study;
- Ψ_k is the stream function in the layer k;
- $\theta_k(\Psi)$ is the potential vorticity in the layer k, given by

$$\begin{pmatrix} \theta_1(\Psi) \\ \vdots \\ \theta_N(\Psi) \end{pmatrix} = [\Delta - [W]] \begin{pmatrix} \Psi_1 \\ \vdots \\ \Psi_N \end{pmatrix},$$

where $[W]$ is a $N \times N$ tridiagonal matrix, whose entries depend on physical parameters:

$$W_{k,k-1} = -\frac{f_0^2}{H_k g'_{k-\frac{1}{2}}},\; W_{k,k+1} = -\frac{f_0^2}{H_k g'_{k+\frac{1}{2}}},\; W_{k,k} = \frac{f_0^2}{H_k}\Big(\frac{1}{g'_{k-\frac{1}{2}}} + \frac{1}{g'_{k+\frac{1}{2}}}\Big),$$

where f_0 is the value of the Coriolis parameter at the middle latitude of Ω, $g'_{k+\frac{1}{2}} = g\,(\rho_{k+1} - \rho_k)/\rho$ is the reduced gravity at the interface k-$k+1$ (g is the earth gravity and ρ the mean density of the fluid).

- f is the Coriolis force. According to the β-plane approximation, it linearly varies with latitude: $f(x, y) = f_0 + \beta.y$, where (x, y) are the Cartesian coordinates in Ω;
- $\dfrac{D_k.}{Dt}$ is the Lagrangian derivative in layer k, given by

$$\frac{D_k.}{Dt} = \frac{\partial.}{\partial t} - \frac{\partial \Psi_k}{\partial y}\frac{\partial.}{\partial x} + \frac{\partial \Psi_k}{\partial x}\frac{\partial.}{\partial y} = \frac{\partial.}{\partial t} + J(\Psi_k, .),$$

where $J(.,.)$ is the Jacobian operator $J(\varphi, \xi) = \dfrac{\partial \varphi}{\partial x}\dfrac{\partial \xi}{\partial y} - \dfrac{\partial \varphi}{\partial y}\dfrac{\partial \xi}{\partial x}$;

- $C_1\Delta\Psi_N$ is the dissipation on the bottom of the ocean;
- $C_3\Delta^3\Psi_k$ is the parametrization of internal and subgrid dissipation;
- F_k is a forcing term. In this model, only the tension due to the wind, denoted τ, is taken into account. Therefore, we get

$$F_1 = \mathrm{Rot}\tau \text{ and } F_k = 0, \quad \forall k \geq 2.$$

The Eq. (3.17) is written in vector form:

$$\frac{\partial}{\partial t}(\Delta - [W]) \begin{pmatrix} \Psi_1 \\ \vdots \\ \Psi_N \end{pmatrix} = \begin{pmatrix} G_1 \\ \vdots \\ G_N \end{pmatrix}, \tag{3.18}$$

with $G_k = F_k - J(\Psi_k, \theta_k(\Psi) + f) - \delta_{k,N} C_1 \Delta\Psi_N + C_3\Delta^3\Psi_k$.

We are going to consider altimetric measurement of the surface of the ocean given by satellite observations (Topex-Poseidon, Jason). The observed data is the change in the surface of the ocean. According to the QG approximation, it is proportional to the stream function in the surface layer:

$$h^{obs} = \frac{f_0}{g}\Psi_1^{obs}.$$

Therefore, we will assimilate surface data in order to retrieve the fluid circulation especially in the deep ocean layers.

The control vector is the initial state on the N layers:

$$u = \Big(\Psi_k(t=0)\Big)_{k=1,\ldots,N} \in \mathcal{U}_{\mathrm{ad}}.$$

The state vector is

$$\Big(\Psi_k(t)\Big)_{k=1,\ldots,\dot{N}}.$$

We assume that the stream function is observed at each point of the surface layer at discrete times t_j. Then the cost function is defined by

$$\mathcal{J}_\varepsilon(u) = \frac{1}{2}\sum_{j=1}^{n}\int_\Omega \Big(\Psi_1(t_j) - \Psi_1^{obs}(t_j)\Big)^2 \mathrm{d}s + \frac{\varepsilon}{2} \parallel R(u) \parallel_{\mathcal{T}}^2 .$$

The second term in the cost function is the regularization term in the sense of Tikhonov. It renders the inverse problem well posed, by taking into account the square of the potential vorticity of the initial state:

$$\parallel R(u) \parallel_{\mathcal{T}}^2 = \sum_{k=1}^{N} H_k\left[\int_\Omega \Big((\Delta\Psi_k)(0) - [W]_k.(\Psi)(0)\Big)^2 \mathrm{d}s\right].$$

The parameter ε in the cost function is the relative weight of the regularization with respect to the quadratic distance between the observations and the computed state.

The direct and the adjoint models are discretized using finite difference discretization for space and a leap-frog scheme for time.

As above, minimization is carried out with M1QN3. An important point is the choice of the inner product for the space of control. A systematic study was carried out in LUONG, BLUM and VERRON [1998] showing that the best choice is the inner product associated to the natural norm corresponding to the square root of the energy of the system, providing a good preconditioner for the optimization algorithm. Figures 3.1 and 3.2, showing the exact and the identified flow at the beginning and at the end of the assimilation period, prove that the reconstruction method is satisfactory.

The management of various time intervals, in order to improve the penetration of information in the deep layer, is presented in BLUM, LUONG and VERRON [1998]. A main difficulty comes from the dimension of the control space making this method costly from a computational viewpoint. Starting from a statistical analysis of the trajectory of the model, BLAYO, BLUM and VERRON [1998] proposed a method for the reduction of this space using POD vectors, which take into account the dynamics of the system (see Section 6).

For operational purpose, 4D-VAR methods have been implemented on the primitive equations model, by using an incremental method in order to reduce the cost of the resolution of the variational problem (THEPAUT and COURTIER [1987], COURTIER, THEPAUT and HOLLINGSWORTH [1994]).

4. Second-order methods

4.1. Hessian

The optimality system, the Euler-Lagrange equation, provides only a necessary condition for optimality. In the linear case, the solution is unique if the Hessian is definite positive. From a general point of view, the information given by the Hessian is important for theoretical, numerical, and practical issues. For operational models, it is impossible to compute the Hessian itself as it is a square matrix with 10^{16} terms; nevertheless, the most important information can be extracted from the spectrum of the Hessian which can be estimated without an explicit determination of this matrix. This information is of importance for estimating the condition number of the Hessian for preparing an efficient preconditioning.

A general method to get this information is to apply the techniques described above to the couple composed of the direct and adjoint models (LE DIMET, NAVON and DAESCU [2002], WANG, NAVON, LE DIMET and ZOU [1992]), leading to a so-called second-order adjoint. The following steps are carried out:

- Linearization of the direct and adjoint models with respect to the state variable. Since the system is linear with respect to the adjoint variable, no linearization is necessary.

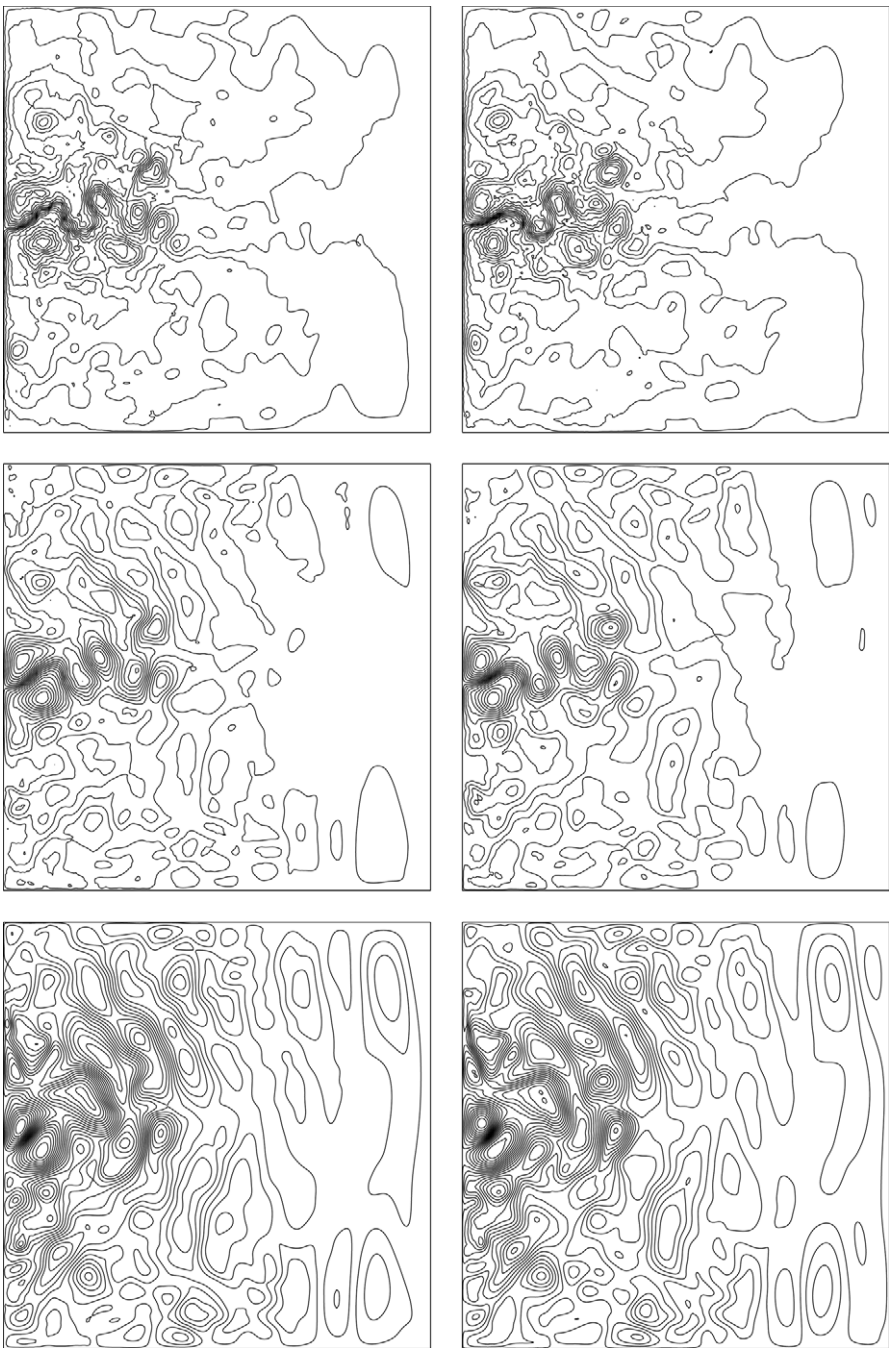


FIG. 3.1 True initial condition (left) and exact solution at the end of the assimilation period (right) for the three (from top to bottom) layers of a quasi-geostrophic model.

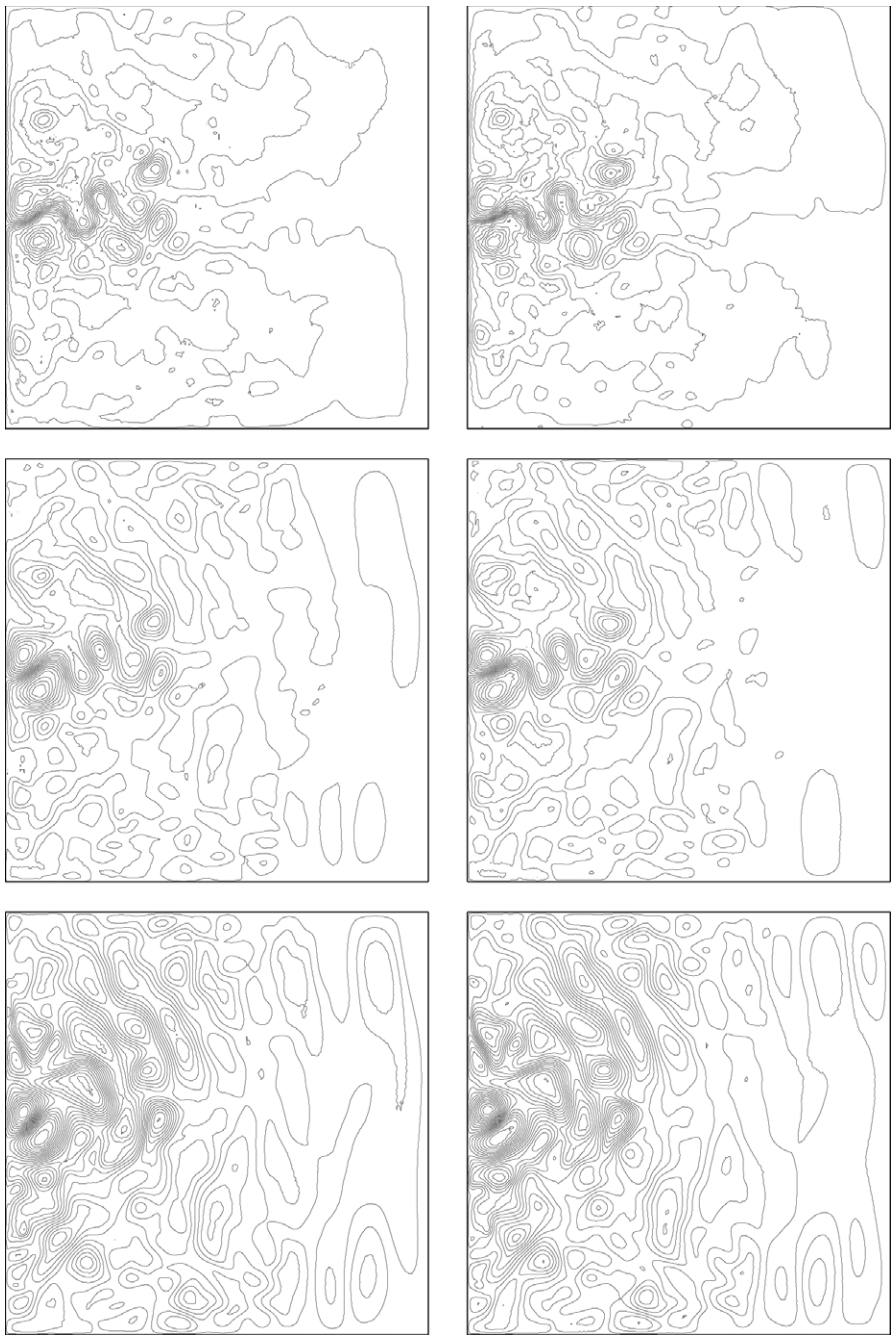


FIG. 3.2 4D-VAR data assimilation results identified initial condition (left) and corresponding solution at the end of the assimilation period (right) for the three (from top to bottom) layers of a quasi-geostrophic model.

- Introducing two second-order adjoint variables.
- Transposition to exhibit the linear dependence with respect to the directions.

If the model (Eq. (3.1)) has the form

$$\begin{cases} \dfrac{\mathrm{d}X}{\mathrm{d}t} = F(X) + B.U \\ X(0) = V \end{cases},$$

U and V being the control variables, and if we consider the cost function defined by Eq. (3.2) and the adjoint equation given by Eq. (3.5), from a backward integration of this adjoint model, the gradient is deduced

$$\nabla J = \begin{pmatrix} \nabla_U J \\ \nabla_V J \end{pmatrix} = \begin{pmatrix} -B^T P \\ -P\,(0) \end{pmatrix}.$$

To calculate the second-order derivative of J with respect to U and V, we have to derive the optimality system (i.e., the model plus the adjoint system). By analogy to the first-order case, we introduce two so-called second-order adjoint variables R and Q as the solution of the system:

$$\begin{cases} \dfrac{\mathrm{d}R}{\mathrm{d}t} = \left[\dfrac{\partial F}{\partial X}\right].R + B.\Lambda \\ \dfrac{\mathrm{d}Q}{\mathrm{d}t} + \left[\dfrac{\partial F}{\partial X}\right]^T.Q = -\left[\dfrac{\partial^2 F}{\partial X^2}.R\right]^T.P + C^T C R, \end{cases} \tag{4.1}$$

where Λ has the dimension of U.

If the Hessian of J is written

$$H(U, V) = \begin{pmatrix} J_{U,U} & J_{U,V} \\ J_{U,V} & J_{V,V} \end{pmatrix}$$

and if system (Eq. (4.1)) is integrated with the conditions:

$$Q(T) = 0$$
$$R(0) = \Theta$$

and $\Lambda = 0$, then

$$J_{V,V}.\Theta = -Q(0),$$
$$J_{V,U}.\Theta = -B^T.Q.$$

Now if the system is integrated with the conditions:

$$Q(T) = 0,$$
$$R(0) = 0,$$

then we obtain

$$J_{U,U}.\Lambda = -B^T.Q.$$

Therefore, without an explicit computation of the Hessian, it is possible to compute the product of the Hessian by any vector and consequently, using classical methods of linear algebra, to evaluate its eigenvalues and eigenvectors and also to carry out Newton-type methods. It is worth pointing out that the R variable is the solution of the linear tangent model (when $\Lambda = 0$), and therefore no extra code has to be written in this case. The left-hand side of the equation verified by Q is the adjoint model, and only the code associated to its right-hand side has to be written.

In the case of the shallow water equations with the initial condition V as unique control vector (no model error), the state variable is $X = (u, v, \phi)$, the adjoint variable is $P = (\tilde{u}, \tilde{v}, \tilde{\phi})$, which is solution of Eqs. (3.13)–(3.15). For the second order, the variable $R = (\bar{u}, \bar{v}, \bar{\phi})$ is the solution of the linear tangent model (Eqs. (3.10)–(3.12)), while the variable $Q = (\hat{u}, \hat{v}, \hat{\phi})$ is the solution of the equations:

$$\begin{aligned}
&\frac{\partial \hat{u}}{\partial t} + u\frac{\partial \hat{u}}{\partial x} + v\frac{\partial \hat{v}}{\partial y} + \hat{u}\frac{\partial v}{\partial y} - \hat{v}\frac{\partial v}{\partial y} - f\hat{v} + \phi\frac{\partial \hat{\phi}}{\partial x} \\
&\quad = \tilde{v}\frac{\partial \bar{v}}{\partial x} - \bar{u}\frac{\partial \tilde{u}}{\partial x} - \bar{v}\frac{\partial \bar{u}}{\partial y} + \tilde{u}\frac{\partial \bar{v}}{\partial y} - \bar{\phi}\frac{\partial \tilde{\phi}}{\partial x} - \bar{u}, \qquad (4.2) \\
&\frac{\partial \hat{v}}{\partial t} + \hat{u}\frac{\partial u}{\partial y} - u\frac{\partial \hat{v}}{\partial x} + \hat{v}\frac{\partial u}{\partial x} + v\frac{\partial \hat{v}}{\partial y} + f\hat{u} + \phi\frac{\partial \hat{\phi}}{\partial y} \\
&\quad = \tilde{u}\frac{\partial \bar{u}}{\partial x} - \bar{u}\frac{\partial \tilde{v}}{\partial x} - \tilde{v}\frac{\partial \bar{u}}{\partial y} + \bar{u}\frac{\partial \tilde{v}}{\partial y} - \bar{\phi}\frac{\partial \tilde{\phi}}{\partial y} - \bar{v}, \qquad (4.3) \\
&\frac{\partial \hat{\phi}}{\partial t} + \frac{\partial \hat{u}}{\partial x} + \frac{\partial \hat{v}}{\partial y} + u\frac{\partial \hat{\phi}}{\partial x} + v\frac{\partial \hat{\phi}}{\partial y} = -\bar{u}\frac{\partial \tilde{\phi}}{\partial x} - \bar{v}\frac{\partial \tilde{\phi}}{\partial x} - \gamma\bar{\phi}. \qquad (4.4)
\end{aligned}$$

From a formal point of view, we see that first-and second-order adjoint models differ by second-order terms which do not take into account the adjoint variable. The computation of second derivatives requires storing both the trajectories of the direct and adjoint models. For very large models, it could be more economical to recompute these trajectories.

The system obtained, i.e., the second-order adjoint, is used to compute the product of the Hessian by any vector. Of course, if we consider all the vectors of the canonical base, then it will be possible to obtain the complete Hessian.

The determination of this product permits access to some information.

- By using Lanczos type methods and deflation, it is possible to compute the eigenvectors and eigenvalues of the Hessian.
- To carry out second-order optimization methods of Newton-type for equations of the form:

$$\nabla J(X) = 0.$$

The iterations are

$$X_{n+1} = X_n - H^{-1}(X_n).\nabla J(X_n),$$

where H is the Hessian of J. At each iteration, a linear system should be solved. This is done by carrying out some iterations of a conjugate gradient method which requires computing the product Hessian-vector.

4.2. *Sensitivity analysis*

In the environmental sciences, the mathematical models contain parameters which cannot be estimated very precisely either because they are difficult to measure or because they represent some subgrid phenomena. Therefore, it is important to be able to estimate the impact of uncertainties on the outputs of the model. Sensitivity analysis is defined as follows:

- X is the state vector of the model and K a vectorial parameter of the model $F(X, K) = 0$.
- $G(X, K)$ the response function: a real value function
- By definition, the sensitivity of the model is the gradient of G with respect to K. The difficulty encountered comes from the implicit dependence of G on K through X solution of the model.

Several methods can be used to estimate the sensitivity:

- By finite differences, we get

$$\frac{\partial G}{\partial e_i} \simeq \frac{G(X(K + \alpha e_i), K + \alpha e_i) - G(X(K), K)}{\alpha}.$$

  The main inconvenience of this method is its computational cost: it requires solving the model as many times as the dimension of K. Furthermore, the determination of the parameter α may be tricky. If it is too large, the variation of G could be nonlinear, while for small values roundoff errors may dominate the variation of G. The main advantage of this method is that it is very easy to implement.
- Sensitivity via an adjoint model. Let $F(X, K) = 0$ be the direct model. We introduce its adjoint:

$$\left[\frac{\partial F}{\partial X}\right]^T .P = \frac{\partial G}{\partial X}.$$

  Then the gradient is given by

$$\nabla G = \frac{\partial G}{\partial K} - \left[\frac{\partial F}{\partial K}\right]^T .P.$$

The advantage of this method is that the sensitivity is obtained by only one run of the adjoint model. The price to be paid is the derivation of the adjoint code.

In geophysics, a usual request is the estimation of the sensitivity with respect to observations. What will be the impact of an uncertainty on the prediction? It is clear that observations are not directly used in the direct model, and they take place only as a forcing term in the adjoint model. Therefore, to apply the general formalism of sensitivity analysis, we should apply it not to the model itself but to the optimality system, i.e., the model plus the adjoint model. A very simple example with a scalar ordinary differential equation is given in LE DIMET, NGODOCK, LUONG and VERRON [1997] showing that the direct model is not sufficient to carry out sensitivity analysis in the presence of data. Deriving the optimality system will introduce second-order derivatives as it has been seen in the previous subsection.

An important problem is the propagation of errors from models and observations toward the predicted fields. Second-order methods provide important tools for this purpose, especially for the estimation of the covariance of errors for the background term (prediction error) and the model error (see LE DIMET, SHUTYAEV and GEJADZE [2006], PARMUZIN, LE DIMET and SHUTYAEV [2006], GEJADZE, LE DIMET and SHUTYAEV [2007]).

5. Nudging method

Nudging is a four-dimensional data assimilation (NDA) method that uses dynamical relaxation to adjust toward observations (observation nudging) or toward an analysis (analysis nudging).

Nudging is accomplished through the inclusion of a forcing term in the model dynamics, with a tunable coefficient that represents the relaxation time scale. Computationally inexpensive nudging is based on both heuristic and physical considerations.

The NDA method relaxes the model state toward the observations during the assimilation period by adding a nonphysical diffusive-type term to the model equations. The nudging terms are defined as the difference between the observation and the model solution multiplied by a nudging coefficient. The size of this coefficient is chosen by numerical experimentation so as to keep the nudging terms small in comparison to the dominating forcing terms in the governing equations, in order to avoid the rebounding effect that slows down the assimilation process, yet large enough to impact the simulation. NDA techniques have been used successfully on the global scale by LYNE, SWINBANK and BIRCH [1982] and KRISHNAMURTI, JISHAN, BEDI, INGLES and OOSTERHOF [1991] and in a wide variety of research applications on mesoscale models (HOKE and ANTHES [1976], RAMAMURTHY and CARR [1987], RAMAMURTHY and CARR [1988], WANG and WARNER [1988], STAUFFER and SEAMAN [1990], VERRON, MOLINES and BLAYO [1992] to cite but a few).

The NDA method is a flexible assimilation technique which is computationally much more economical than the VDA method. However, results from NDA experiments are quite sensitive to the adhoc specification of the nudging relaxation coefficient, and it is not at all clear how to choose a nudging coefficient so as to obtain an optimal solution (LORENC [1986], LORENC [1988]).

5.1. Optimal nudging specification

We assume that the model equations have been discretized in space by a finite difference, finite element, or spectral discretization method. The time continuous model satisfies dynamical equations of the form

$$\frac{\partial X}{\partial t} = F(X), \tag{5.1}$$

$$X(0) = V, \tag{5.2}$$

where X represents the discretized state variable of the model atmosphere, t is time, and V represents the initial condition for the model. Say, for instance, $X^o(t)$ is a given observation, then the objective of VDA is to find model initial conditions that minimize a cost function defined by

$$J(V) = \int_0^T \left\langle W(X - X^o), X - X^o \right\rangle \mathrm{d}t, \tag{5.3}$$

where W is a diagonal weighting matrix. Note that J is only a function of the initial state because X is uniquely defined by the model equations (Eqs. (5.1) and (5.2)). An implicit assumption made in VDA is that the model exactly represents the state of the atmosphere. However, this assumption is not true.

The NDA technique introduced by ANTHES [1974] consists in achieving a compromise between the model and the observations by considering the state of the atmosphere to be defined by

$$\frac{\partial X}{\partial t} = F(X) + G(X^o - X), \tag{5.4}$$

where G is a diagonal matrix.

Together with the initial conditions

$$X(0) = V, \tag{5.5}$$

the system (Eq. (5.1)) has a unique solution $X(V, G)$. The main difficulty in the NDA scheme resides in the estimation of the nudging coefficient G (STAUFFER and SEAMAN [1990]). If G is too large, the fictitious diffusion term will completely dominate the time tendency and will have an effect similar to replacing the model data by the observations at each time-step. Should a particular observation have a large error that prevents obtaining a dynamic balance, an exact fit to the observation is not required since it may lead to a false amplification of observational errors. On the other hand, if G is too small, the observation will have little effect on the solution. In general, G decreases with increasing observation error, increasing horizontal and vertical distance separation, and increasing time separation. In the experiment of ANTHES [1974], a nudging coefficient of 10^{-3} was used for all the fields for a hurricane model and was applied on all the domain of integration. In the experiment of KRISHNAMURTI,

JISHAN, BEDI, INGLES and OOSTERHOF [1991], the relaxation coefficients for the estimated NDA experiment were kept invariant both in space and time, and their values were simply determined by numerical experience. The implicit dynamic constraints of the model then spread the updated information to the other variables (temperature and moisture) resulting eventually in a set of balanced conditions at the end of the nudging period.

In the work of ZOU, NAVON and LE DIMET [1992], a new parameter estimation approach was designed to obtain optimal nudging coefficients. They were optimal in the sense that the difference between the model solution and the observations will be small. For a comprehensive review of parameter estimation addressing issues of identifiability, see NAVON [1998].

5.2. *Parameter estimation of optimal nudging coefficients*

The application of the variational approach to determine model parameters is conceptually similar to that of determining the initial conditions. Here, we present a brief illustration of the method. For the parameter estimation of the nudging coefficients, the cost function J can be defined as

$$J(G) = \int_0^T \left\langle W(X - X^o), X - X^o \right\rangle \mathrm{d}t + \left\langle K(G - \hat{G}), G - \hat{G} \right\rangle, \tag{5.6}$$

where $\hat{G}$ denotes the estimated nudging coefficients, and W and K are specified weighting matrices. The second term plays a double role. On one hand, it ensures that the new value of the nudging parameters is not too far away from the estimated quantity. On the other hand, it enhances the convexity of the cost function since this term contributes a positive term K to the Hessian matrix of J. An optimal NDA procedure can be defined by the optimal nudging coefficients G^* such that

$$J(G^*) \leq J(G), \quad \forall G. \tag{5.7}$$

The problem of extracting the dynamical state from observations is now identified as the mathematical problem of finding initial conditions or external forcing parameters that minimize the cost function. Due to the dynamical coupling of the state variables to the forcing parameters, the dynamics can be enforced through the use of a Lagrange function constructed by appending the model equations to the cost function as constraints in order to avoid the repeated application of the chain rule when differentiating the cost function. The Lagrange function is defined by

$$L(X, G, P) = J + \int_0^T \left\langle P, \frac{\partial X}{\partial t} - F(X) - G(X^o - X) \right\rangle \mathrm{d}t, \tag{5.8}$$

where P is a vector of Lagrange multipliers. The Lagrange multipliers are not specified but computed in determining the best fit. The gradient of the Lagrange function must be

zero at the minimum point. This results in the following first-order conditions:

$$\frac{\partial L}{\partial X} = 0 \sim \text{adjoint model forced by } \ 2\mathbf{W}(X - X^o), \tag{5.9}$$

$$\frac{\partial L}{\partial P} = 0 \sim \text{direct model (Eq. (5.4))}, \tag{5.10}$$

$$\frac{\partial L}{\partial G} = 0 \sim \int_0^T - < P, X^o - X) > \mathrm{d}t + 2K(G - \hat{G}) = 0. \tag{5.11}$$

The solution of equations (Eqs. (5.9)–(5.11)) is called a stationary point of L. Even if the dynamical evolution operator is nonlinear, the equation $\left(\dfrac{\partial L}{\partial X} = 0\right)$ will be the same as those derived by constructing the adjoint of the linear tangent operator; the linearization is automatic due to the Lagrange function L being linear in terms of the Lagrange multipliers P. An important relation between the gradient of the cost function (Eq. (5.7)) with respect to parameters G and the partial derivative of the Lagrange function with respect to the parameters is

$$\nabla_G J(G) = \frac{\partial L}{\partial G} | \text{at stationary point}, \tag{5.12}$$

i.e., the gradient of the cost function with respect to the parameters is equal to the left hand side of Eq. (5.11) which can be obtained in a procedure where the model state P is calculated by integrating the direct model forward and then integrating the adjoint model backward in time with the Lagrange multipliers as adjoint variables. Using this procedure, we can derive the following expressions of the adjoint equation and gradient formulation:

$$\begin{cases} \dfrac{\mathrm{d}P}{\mathrm{d}t} + \left[\dfrac{\partial F}{\partial X}\right]^T . P - G^T . P = 2W(X - X^o) \\ P\,(T) = 0, \end{cases} \tag{5.13}$$

and

$$\nabla_G J = - \int_0^T \left\langle X^o - X,\, P \right\rangle \mathrm{d}t + 2K(G - \hat{G}). \tag{5.14}$$

We see that the adjoint equation of a model with a nudging term added is the same as that without a nudging term except for the additional term $-G^T P$ added to the left hand side of the adjoint equation. Having obtained the value of cost function by integrating the model (Eq. (5.4)) forward and the value of the gradient $\nabla_G J$ by integrating the adjoint equation (Eq. (5.13)) backward in time, any large-scale unconstrained minimization method can be employed to minimize the cost function and to obtain an optimal parameter estimation.

If both the initial condition and the parameter are controlled, the gradient of the cost function for performing the minimization would be

$$\nabla J = (\nabla_V J, \nabla_G J)^T, \tag{5.15}$$

where

$$\nabla_V J = -P(0). \tag{5.16}$$

Zou, Navon and Le Dimet [1992] have shown that estimated NDA, optimal NDA, and KF differ from each other in the choice of the weight matrix often called the gain matrix:

$$G_n^* \equiv W_n^f H_n^T \left(H_n W_n^f H_n^T + R_n\right)^T. \tag{5.17}$$

The VDA, on the other hand, takes both the model forecasts and the observations as perfect. It attempts to obtain an optimal initial condition which minimizes the cost function

$$J^f = E\left(X_n^f - X_n^o\right)^T \left(X_n^f - X_n^o\right). \tag{5.18}$$

The theoretical framework of estimation and control theory provides the foundation of data assimilation techniques. The estimated NDA and the KF are closer to the estimation theory, the VDA to the optimal control aspect while optimal NDA is a combination of both (see also Lorenc [1986]).

See also work of Vidard, Piacentini and Le Dimet [2003] on optimal estimation of nudging coefficients.

5.3. *Back-and-forth nudging*

The backward nudging algorithm consists in solving the state equations of the model backwards in time, starting from the observation of the state of the system at the final instant. A nudging term, with the opposite sign compared with the standard nudging algorithm, is added to the state equations, and the final obtained state is in fact the initial state of the system (Auroux [2008]). The idea is to consider that we have a final condition V_T in Eqs. (5.1) and (5.2) instead of an initial condition V and then to apply nudging to this backward model with the opposite sign of the feedback term (in order to have a well-posed problem). We obtain

$$\begin{cases} \dfrac{\partial \tilde{X}}{\partial t} = F(\tilde{X}) - G'(X^o - \tilde{X}), \quad T > t > 0, \\ \tilde{X}(T) = V_T. \end{cases} \tag{5.19}$$

The back and forth nudging algorithm, introduced in Auroux and Blum [2005], consists in solving first the forward nudging equation and then the direct system backwards in time with a feedback term whose sign is opposite to the one introduced in the forward equation. The "initial" condition of this backward resolution is the final state obtained by the standard nudging method. After resolution of this backward equation, one obtains an estimate of the initial state of the system. We repeat these forward and backward resolutions (with the feedback terms) until convergence of the algorithm:

$$\begin{cases} \dfrac{\partial X_k}{\partial t} = F(X_k) + G(X^o - X_k), \\ X_k(0) = \tilde{X}_{k-1}(0), \end{cases} \tag{5.20}$$

$$\begin{cases} \dfrac{\partial \tilde{X}_k}{\partial t} = F(\tilde{X}_k) - G'(X^o - \tilde{X}_k), \\ \\ \tilde{X}_k(T) = X_k(T), \end{cases} \tag{5.21}$$

with $\tilde{X}_0(0) = V$. Then, $X_1(0) = V$, and a resolution of the direct model gives $X_1(T)$ and hence $\tilde{X}_1(T)$. A resolution of the backward model provides $\tilde{X}_1(0)$, which is equal to $X_2(0)$, the new initial condition of the system, and so on.

This algorithm can be compared with the 4D-VAR algorithm, which consists also in a sequence of forward and backward resolutions. In this algorithm, even for nonlinear problems, it is useless to linearize the system, and the backward system is not the adjoint equation but the direct system with an extra feedback term that stabilizes the resolution of this ill-posed backward resolution.

Auroux and Blum [2005] proved the convergence of the BFN algorithm on a linear model, provided that the feedback term is large enough. Auroux and Blum [2008] discussed the choice of the gain matrices G and G′ and tested the algorithm for Lorenz, Burgers and QG models. This algorithm is hence very promising to obtain a correct initial state, with a very easy implementation because it does not require neither the linearization of the equations to obtain the adjoint model nor any minimization process.

6. POD model reduction methods application to geosciences and 4D-VAR data assimilation

6.1. Introduction

Interest in reduced cost of implementation of 4D-VAR data assimilation in the geosciences motivated research efforts aimed toward reducing dimension of control space without significantly compromising quality of the final solution. POD, also known, when restricted to a finite dimensional case and truncated after a few terms, as equivalent to principal component analysis (PCA) and as empirical orthogonal function (EOF) in oceanography and meteorology, has emerged as a method of choice to be employed in flow control and optimization.

POD is a procedure for extracting a basis for a modal decomposition from an ensemble of signals. POD was introduced in the context of analysis of turbulent flow by Lumley [1967], Berkooz, Holmes and Lumley [1993].

In other disciplines, the same procedure goes by the names of Karhunen-Loeve decomposition or PCA.

The POD method was independently rediscovered several times: Kosambi [1943], Loeve [1945], and Karhunen [1946]. For introductory discussion for POD in fluid mechanics, see Sirovich [1987a,b,c] and Holmes [1990]. The mathematical theory behind it is the spectral theory of compact, self-adjoint operators.

The POD is equivalent to PCA methodology which originated with the work of Pearson [1901], a means of fitting planes by orthogonal least squares also put forward by Hotelling [1933].

If the POD spectrum decays fast enough, practically all the support of the invariant measure is contained in a compact set. Roughly speaking, all the likely realizations in the ensemble can be found in a relatively small set of bounded extent. "Regularity of solutions" is a mathematical property describing, essentially, the rate of decay of the tail of the wave number spectrum of instantaneous solutions of a partial differential equation. The method of snapshots introduced by SIROVICH [1987] is a numerical procedure for saving time in computation of empirical eigenfunctions. KIRBY and SIROVICH [1990] applied the POD procedure directly to the reconstruction of images of human faces. See also KIRBY [2001].

Snapshot bases consist of the flow solution for several flow solutions corresponding to different sets of parameter values evaluated at different time instants of the model evolution. This involves solving the fully discretized model and saving states at various time instants in the time interval under consideration.

POD approximation can be thought of as a Galerkin approximation in the spatial variable, with basis functions corresponding to the solution of the physical system at prespecified time instances. These are called the snapshots.

Due to possible linear dependence or almost linear dependence, the snapshots themselves are not appropriate as a basis. Rather singular value decomposition (SVD) is carried out and the leading generalized eigenfunctions are chosen as a basis, referred to as the POD basis.

6.2. POD: the discrete case

We consider the discrete Karhunen-Loeve expansion to find an optimal representation of the ensemble of snapshots. In general, each sample of snapshots $u_i(\vec{x})$ (defined on a set of m nodal points $\vec{x}$) can be expressed as a dimensional vector as follows:

$$\vec{u}_i = [u_{i1}, \ldots, u_{im}]^T, \tag{6.1}$$

where u_{ij} denotes the jth component of the vector $\vec{u}_i$. The mean vector is given by

$$\bar{u}_k = \frac{1}{n} \sum_{i=1}^{n} u_{ik}, \quad k = 1, \ldots, m. \tag{6.2}$$

We also can form a new ensemble by focusing on deviations from the mean value as follows:

$$v_{ik} = u_{ik} - \bar{u}_k, \quad k = 1, \ldots, m \tag{6.3}$$

Let the matrix A denotes the new ensemble

$$A = \begin{pmatrix} v_{11} & v_{21} & \cdots & vn1 \\ v_{12} & v_{22} & \cdots & v_{n2} \\ \vdots & \vdots & \vdots & \vdots \\ v_{1m} & v_{2m} & \cdots & v_{mm} \end{pmatrix}_{m \times n},$$

where the discrete covariance matrix of the ensemble u may be written as

$$Cy_k = AA^T y_k = \lambda_k y_k. \tag{6.4}$$

Thus, to compute the POD mode, one must solve an eigenvalue problem.

For a discretization of an ocean problem, the dimension often exceeds 10^4 so that a direct solution of this eigenvalue problem is often not feasible. We can transform the $m \times m$ eigenvalue problem into an n eigenvalue problem. In the method of snapshots, one then solves the $n \times n$ eigenvalue problem

$$Dw_k = A^T A w_k = \lambda_k w_k, \quad w_k \in R^n, \tag{6.5}$$

where $1 \leq \lambda_k \leq n$ are the eigenvalues. The eigenvectors w_k may be chosen to be orthonormal, and the POD modes are given by $\phi_k = Aw_k/\sqrt{\lambda_k}$. In matrix form, with $\Phi = [\phi_1, \ldots, \phi_n]$, and $W = [w_1, \ldots, w_n]$, this becomes $\Phi = AW$.

The $n \times n$ eigenvalue problem (Eq. (6.4)) is more efficient than the $m \times m$ eigenvalue problem (Eq. (6.4)) when the number of snapshots n is much smaller than the number of states m.

6.3. POD 4D-VAR

In order to reduce the computational cost of 4D-VAR data assimilation, we consider minimization of the cost functional in a space whose dimension is much smaller than that of the original one. A way to drastically decrease the dimension of the control space without significantly compromising the quality of the final solution but sizably decreasing the cost in memory and CPU time of 4D-VAR motivates us to choose to project the control variable on a basis of characteristic vectors capturing most of the energy and the main directions of variability of the model, i.e., SVD, EOF, Lyapunov, or bred vectors. One would then attempt to control the vector of initial conditions in the reduced space model.

In the 1990s, most efforts of model reduction have been centered on KF and extended Kalman filter (EKF) data assimilation techniques, see Todling and Cohn [1994], Todling, Cohn and Sivakumaran [1998], Pham, Verron and Roubaud [1998], Cane, Kaplan, Miller, Tang, Hackert and Busalacchi [1996], Fukumori and Malanotte-Rizzoli [1995], Verlaan and Heemink [1997] and Hoteit and Pham [2003]. In particular, Cane, Kaplan, Miller, Tang, Hackert and Busalacchi [1996] employed a reduced-order method in which the state space is reduced through the projection onto a linear subspace spanned by a small set of basis functions, using an EOF analysis. This filter is referred to as the reduced rank EKF (see next section).

Some initial efforts aiming at the reduction of the dimension of the control variable, referred to as reduced-order strategy for 4D-VAR ocean data assimilation, were put forward initially by Blayo, Blum and Verron [1998] and Durbiano [2001] and more recently by Hoteit and Kohl [2006] and Robert, Durbiano, Blayo, Verron, Blum and Le Dimet [2005]. They used a low dimension space based on the first few EOFs, which can be computed from a sampling of the model trajectory. Hoteit and Kohl [2006] used the reduced-order model for part of the 4D-VAR assimilation then switched to the full model in a manner done earlier by Peterson [1989].

For a comprehensive description of POD theory and state of the art research, see GUNZBURGER [2003].

At the analysis time $[0, T_N]$, strong constraint 4D-VAR looks to minimize a cost function

$$J(U_0) = (U_0 - U_b)^T B^{-1}(U_0 - U_b) + (HU - y^0)^T O^{-1}(HU - y^0). \tag{6.6}$$

In POD 4D-VAR, we look to minimize the cost function

$$J(c_1(0), \ldots, c_M(0)) = (U_0^{POD} - U_b)B^{-1}(U_0^{POD} - U_b) + (HU^{POD} - y^0)O^{-1}(HU^{POD} - y^0), \tag{6.7}$$

where U_0^{POD} is the control vector, H is an observation operator, B is the background error covariance matrix, and O is the observation error covariance matrix. In Eq. (6.7),

$$U_0^{POD}(x) = U_0^{POD}(0, x) = \overline{U}(x) + \sum_{i=1}^{M} c_i(0)\Phi_i(x), \tag{6.8}$$

$$U^{POD}(x) = U^{POD}(t, x) = \overline{U}(x) + \sum_{i=1}^{M} c_i(t)\Phi_i(x). \tag{6.9}$$

In POD 4D-VAR, the control variables are $c_1(0), \cdots, c_M(0)$. As shown later, the dimension of the POD reduced space could be much smaller than that of the original space. In addition, the forward model is the reduced model which can be very efficiently solved. The adjoint model is used to calculate the gradient of the cost function (Eq. (6.7)) and that will greatly reduce both the computational cost and coding effort.

To establish POD model in POD 4D-VAR, we need first to obtain an ensemble of snapshots, which is taken from the background trajectory, or integrate original model with background initial conditions.

6.4. Adaptive POD 4D-VAR

Since the POD model is based on the solution of the original model for a specified initial condition, it might be a poor model when the new initial condition is significantly different from the one on which the POD model is based upon. Therefore, an adaptive POD 4D-VAR procedure is as follows:

(i) Establish POD model using background initial conditions and then perform optimization iterations to approximate the optimal solution of the cost function (Eq. (6.7)).
(ii) If after a number of iterations, the cost function cannot be reduced significantly as measured by a preset criterion, we generate a new set of snapshots by integrating the original model using the newest initial conditions.
(iii) Establish a new POD model using the new set of snapshots and continue optimization iteration.
(iv) Check if the optimality conditions are reached, if yes, stop; if no, go to step (ii).

6.5. Methods for 4D-VAR POD adaptivity

Working with a low-dimensional model during the computation of a control problem solution has to face the problem that these reduced models are possibly unreliable models if they are not correctly updated during the optimization process.

Consequently, some iterative technique is required, in which the construction of reduced-order models is coupled with the progress of the optimization process. Such an approach leads to the use of reduced-order models that adequately represent the flow dynamics as altered by the control. Crucial at this point is to decide whether or not the reduced-order model has to be adapted to a new flow configuration.

In adaptivity based on a trust-region method (Fahl and Sachs [2003]), the range of validity of a reduced-order model is automatically restricted, and the required update decision for the reduced-order models can be made by employing information that is obtained during the control problem solution.

Ravindran [2002] and Ravindran [2006] propose an adaptive procedure that successively updates the reduced-order model to be used in a sequential quadratic programming constrained optimization algorithm.

6.6. Goal-oriented model-based reduction

Bui-Thanh, Willcox, Ghattas and Van Bloemen Waanders [2007], Willcox, Ghattas, Van Bloemen Waanders and Bader [2005], and Daescu and Navon [2008] proposed an alternative method to determine the reduced-space basis. This method seeks to minimize an error similar in form to Eq. (6.7) to be presented below; however, it will improve upon the POD, first, by minimizing the error in the outputs (as opposed to states) and, second, by imposing additional constraints that $\hat{u}^k(t)$ should result from satisfying the reduced-order governing equations for each parameter instance k.

For a fixed basis size, the POD basis therefore minimizes the error between the original snapshots and their representation in the reduced space defined by

$$E = \sum_{k=1}^{S} \sum_{j=1}^{T} [u^k(t_j) - \tilde{u}^k(t_j)]^T [u^k(t_j) - \tilde{u}^k(t_j)], \tag{6.10}$$

where

$$\tilde{u}^k(t_j) = \Phi \Phi^T u^k(t_j), \tag{6.11}$$

Here, $u^k(t_j)$, $j = 1, \cdots, T$; $k = 1, \ldots, S$ is a snapshot, i.e., the solution of the governing equations at time t_j for parameter instance k. T time instants are considered for each parameter instance, yielding a total of ST snapshots.

The projection matrix $\Phi \in R^{N \times m}$ contains as columns the basis vectors ϕ_i, i.e.,

$$\Phi = [\phi_1, \phi_2, \ldots, \phi_m], \tag{6.12}$$

This error is equal to the sum of the singular values corresponding to those singular vectors (SVs) not included in the POD basis:

$$E = \sum_{i=m+1}^{ST} \sigma_i, \tag{6.13}$$

where σ_i is the ith singular value of U. The POD is an optimal basis in the sense that it minimizes the data reconstruction error given by Eq. (6.7). The goal-oriented, model-based optimization approach presented here provides a general framework for construction of reduced models, and is particularly applicable for optimal design, optimal control, and inverse problems. The optimization approach provides significant advantages over the usual POD by allowing the projection basis to be targeted to output functionals, by providing a framework to consider multiple parameter instances, and by incorporating the reduced-order governing equations as constraints in the basis derivation (see also MEYER and MATTHIES [2003]). Using this method, it is possible to obtain an a priori error estimate for a certain target functional of the solution. This error estimate can be used for adaptively resizing the number of basis vectors and the length of the time-step to satisfy a given error tolerance. It can also be used to form a very efficient low-dimensional basis especially tailored to the target functional of interest. This basis yielded a significantly better approximation of the functional when compared with conventionally chosen bases (see DAESCU and NAVON [2008]).

6.7. State of the art of POD research

ROBERT, DURBIANO, BLAYO, VERRON, BLUM and LE DIMET [2005] apply POD reduced-order modeling in a twin experiment setup for a primitive equation model of the equatorial Pacific Ocean model using an incremental formulation and using a background covariance matrix in the reduced space, obtaining a fast convergence of the minimization of the cost functional.

In a related work, ROBERT, BLAYO and VERRON [2006] applied reduced-order 4D-VAR as a preconditioner to incremental 4D-VAR data assimilation method. They used a low-dimensional space based on first few EOFs chosen from sampling of the model trajectory. See also work of LAWLESS, NICHOLS, BOESS and BUNSE-GERSTNER [2006] using a linear balanced truncation reduced-order modeling as the preconditioner of the inner iteration of an incremental 4D-VAR using a 1D shallow water equation model.

CAO, ZHU, LUO and NAVON [2006, 2007] proposed for the first time a 4D-VAR approach based on POD. Their proposed POD-based 4D-VAR methods are tested and demonstrated using a reduced gravity wave ocean model in Pacific domain in the context of identical twin data assimilation experiments.

LUO, CHEN, ZHU, WANG and NAVON [2007] present an error estimate of a new reduced-order optimizing finite difference system (FDS) model. Numerical examples are presented illustrating that the error between the POD approximate solution and the full FDS solution is consistent with previously obtained theoretical results. The preconditioning aspect of POD for efficient optimization is another topic of active research. DAESCU

and NAVON [2007] use a POD approach to model reduction to identify a reduced-order control space for a 2D global shallow water model. A reduced second-order adjoint model is developed and used to facilitate the implementation of a Hessian-free truncated-Newton (HFTN) minimization algorithm in the POD-based space. The HFTN algorithm benefited most from the order reduction since computational savings were achieved both in the outer and inner iterations of the method.

For use of centroidal Voronoi tessellations (CVT) combined with POD, see the work of BURKARDT, GUNZBURGER and LEE [2006]. Here, POD and CVT approaches to reduced-order modeling are provided, including descriptions of POD and CVT reduced-order bases, their construction from snapshot sets, and their application to the low-cost simulation of a Navier-Stokes system (see also GUNZBURGER, PETERSON and SHADID [2007]). Direct and inverse POD model reduction was applied to a 3D time-dependent finite element adaptive ocean model (Imperial College Ocean Model, ICOM) (FANG, PAIN, NAVON, PIGGOTT, GORMAN and GODDARD [2008]) (see Fig. 6.1).

A novel POD model has been developed for use with an advanced unstructured mesh finite element ocean model, the ICOM, which includes many recent developments in ocean modeling and numerical analysis. The advantages of the POD model developed over existing POD approaches are the ability to increase accuracy when representing geostrophic balance (the balance between the Coriolis terms and the pressure gradient). This is achieved through the use of two sets of geostrophic basis functions where each one is calculated by basis function for velocities u and v.

When adaptive meshes are employed in both the forward and adjoint models, the mesh resolution requirements for each model may be spatially and temporally different as the meshes are adapted according to the flow features of each model. This unavoidably brings to difficulties in the implementation of a POD-based reduced model for an inverse adaptive model. Such challenges include snapshots can be of different length at different time levels and the POD base of the forward model can differ from the POD base of

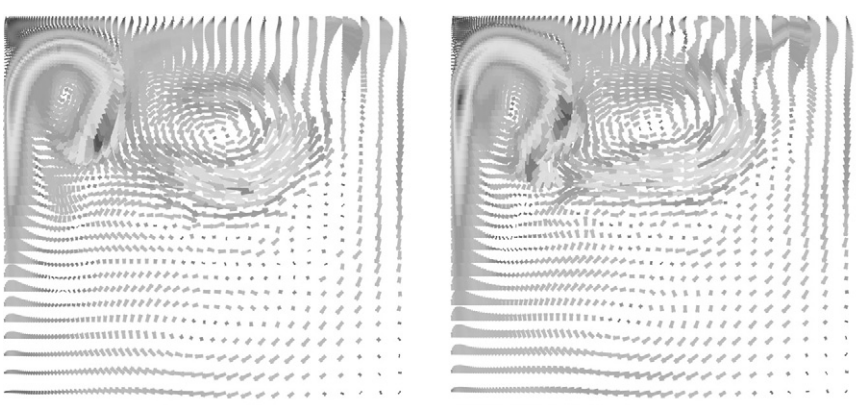


FIG. 6.1 Application of POD model reduction method to Imperial College Ocean Model (ICOM) adaptive finite element ocean Model. Full model and reduced-order model based on first 30 base functions. (See also color insert).

the adjoint model. To overcome these difficulties, a standard reference fixed mesh is adopted for both the forward and adjoint reduced models. The solutions for both are interpolated from their own mesh onto the same reference fixed mesh at each time level. This allows the same number of base modes for both reduced forward and adjoint models. The referenced mesh can also be obtained by superimposing the resolution at each mesh level associated with a goal-based function.

7. Data assimilation with EnKF

7.1. Introduction

In recent years, two trends for operational data assimilation are prevalent for the data assimilation practitioners.

On one hand are the variational methods subdivided between computationally economical 3D-VAR methods which exclude the flow-dependent forecast errors (see PARRISH and DERBER [1992]) while few centers endowed with powerful computing resources adopted 4D-VAR, requiring availability and constant updating of an adjoint model. The latter requires a computationally demanding effort but is significantly more accurate than 3D-VAR (KALNAY, LI, MIYOSHI, YANG and BALLABRERA-POY [2007]) in pretest implementation comparisons. Moreover 4D-VAR allows the assimilation of asynoptic data at their correct observation time along with other advantages such as possibility of inclusion of model error term as a weak constraint 4D-VAR.

On the other hand, in view of the obvious shortcomings of usual KF and EKF, more efficient filter methods have emerged, obviating the prohibitive storage and computational time due to explicit treatment of the state error covariance matrix for KF and EKF such as the EnKF. These new filter algorithms are of special interest due to their simplicity of implementation since no adjoint operators are required, along with their potential for efficient use on parallel computers with large-scale geophysical models (NERGER, HILLER and SCHRATER [2005]). Research on EnKF started with work of EVENSEN [1994], EVENSEN and LEEUWEN [1996], BURGERS, VAN LEEUWEN and EVENSEN [1998], and HOUTEKAMER and MITCHELL [1998]. Their methods can be classified as perturbed observations (or stochastic) EnKF and are essentially a Monte-Carlo approximation of the KF which avoids evolving the covariance matrix of the pdf of the state vector x. A second type of EnKF is a class of square-root (or deterministic) filters (ANDERSON [2003], BISHOP, ETHERTON and MAJUMDAR [2001], WHITAKER and HAMILL [2002], see review of TIPPETT, ANDERSON, BISHOP, HAMILL and WHITAKER [2003]), which consist of a single analysis based on the ensemble mean, and where the analysis perturbations are obtained from the square root of the KF analysis error covariance.

Several variants of the EnKF have been proposed (ANDERSON [2003], BISHOP, ETHERTON and MAJUMDAR [2001], WHITAKER and HAMILL [2002]) which can be interpreted as ensemble square-root KFs. For an improved treatment of nonlinear error evolution in EKF, the singular evolutive interpolated Kalman filter (PHAM, VERRON and ROUBAUD [1998]) was introduced as a variant of the singular evolutive extended Kalman (SEEK) filter. It combines the low-rank approximation with an ensemble representation of the covariance matrix. This idea has also been followed in the concept

of error subspace statistical estimation (LERMUSIAUX and ROBINSON [1999]). Another approach is based on a low-rank approximation of the state covariance matrix of the EKF to reduce the computational costs. Using finite difference approximations for the tangent linear model, these algorithms display better abilities to treat nonlinearity as compared with the EKF. Examples of low-rank filters are the reduced rank square-root algorithm (VERLAAN and HEEMINK [1995]) and the similar SEEK filter (PHAM, VERRON and ROUBAUD [1998]).

We will first present in a short section the basic linear KF (KALMAN [1960]) followed by the EKF and then devote our attention to a brief survey of various flavors of EnKF and its state-of-the-art implementation. See GHIL and MANALOTTE-RIZZOLI [1991] for equivalence between 4D-VAR with strong constraint and the linear KF and LI and NAVON [2001].

Finally, some open issues of advantages of EnKF versus 4D-VAR as have emerged from recent work of KALNAY, LI, MIYOSHI, YANG and BALLABRERA-POY [2007] will be briefly addressed.

7.2. The KF

The KF is an efficient recursive filter that estimates the state of a dynamic system from a series of incomplete and noisy measurements. It was developed by KALMAN [1960]. The KF has been derived in a number of books on control theory, e.g., GELB, KASPER, NASH, PRICE and SUTHERLAND [1974] and JAZWINSKI [1970] to mention but a few. See also early work of DU PLESSIS [1967]. In oceanography, the KF has been used by BUDGELL [1986] to describe nonlinear and linear shallow water wave propagation in branched channels, using one-dimensional (1D) cross-sectionally integrated equations. MILLER [1986] used a 1D linear barotropic QG model to investigate the properties of the KF. He provided a derivation of the KF equations. In meteorology, GHIL [1980] and GHIL et al. [1981] promoted first the use of KF along with COHN, GHIL and ISAACSON [1981] and COHN [1997]. See also MILLER, GHIL and GAUTHIEZ [1994]. GHIL [1989] discussed the KF as a data assimilation method in oceanography and used it with a simple linear barotropic model. The KF for use in meteorology has recently been addressed in work of COHN and PARRISH [1991] who discussed the propagation of error covariances in a 2D linear model.

7.3. The KF formulation

The KF is a recursive filter that estimates the state of a dynamic system from a series of incomplete and noisy measurements.

Consider a linear observation process described by

$$y_k^0 = H_k x_k^t + e_k, \tag{7.1}$$

where k is a multiple of the number of time-steps between two consecutive observations in time. y_k^0 is the vector of observations while the vector e_k is an additive noise representing the error in observations due for instance to instrumental error. Random noise e_k is

assumed white in time with mean 0 and covariance R_k, i.e.,

$$E(e_k e_k^T) = R_k \mathrm{d}_k \mathrm{d}_k^T, \tag{7.2}$$

All the time, we consider a discrete in time stochastic dynamic system

$$x_k^t = M_{k-1} x_{k-1}^t + \eta_{k-1}, \tag{7.3}$$

where M_k represents model dynamics while η_k is model error white in time with mean zero and covariance Q_k,

$$E(h_k h_k^T) = Q_k \mathrm{d}_k \mathrm{d}_k^T, \tag{7.4}$$

One can show that the linear KF (Gelb, Kasper, Nash, Price and Sutherland [1974], Jazwinski [1970], Todling [1999]) consists of following stages:

- Advance in time:

$$\begin{cases} x_k^f = M_{k-1} x_{k-1}^a \\ P_k^f = M_{k-1} P_{k-1}^a M_{k-1}^T + Q_{k-1}, \end{cases} \tag{7.5}$$

  where the forecast and analysis error covariance matrices at time k are given by

$$\begin{cases} P_k^f = E\{(x_k^t - x_k^f)(x_k^t - x_k^f)^T\} \\ P_k^a = E\{(x_k^t - x_k^a)(x_k^t - x_k^a)^T\}. \end{cases} \tag{7.6}$$

  Q_{k-1} is the model error covariance matrix at time $t = t_{k-1}$, and M_{k-1} is the model dynamics. x_{k-1}^a and x_{k-1}^f are the analysis and the forecast at time $t = t_{k-1}$.
- Compute the Kalman gain:

$$K_k = P_k^f H_k^T (H_k P_k^f H_k^T + R_k)^{-1}. \tag{7.7}$$

  The matrix K_k is the optimal weighting matrix known as the Kalman gain matrix.
- Update the state:

$$x_k^a = x_k^f + K_k(y_k^0 - H_k x_k^f), \tag{7.8}$$

  where y_k^0 is the observation at time $t = t_k$, H_k is the observation matrix at time $t = t_k$.
- Update error covariance matrix:

$$P_k^a = (I - K_k H_k) P_k^f. \tag{7.9}$$

7.4. Computational cost of KF

The KF assuming the dynamical model has n unknowns in the state vector then error covariance matrix has n^2 unknowns. The evolution of the error covariance in time requires the cost of $2n$ model integrations. Thus, KF in usual form can only be used for rather low-dimensional dynamical models.

The basic KF is limited to a linear assumption. However, most non-trivial systems are nonlinear. The nonlinearity can be associated either with the process model or with the observation model or with both. In EKF, the state transition and observation models need not be linear functions of the state but may instead be functions.

$$\begin{cases} x_k = f(x_{k-1}, u_k, w_k) \\ z_k = h(x_k, v_k). \end{cases} \tag{7.10}$$

The function f can be used to compute the predicted state from the previous estimate, and similarly, the function h can be used to compute the predicted measurement from the predicted state. However, f and h cannot be applied to the covariance directly. Instead, a matrix of partial derivatives (the Jacobian or Tangent Linear Model) is computed. At each time-step, the Jacobian is evaluated with current predicted states. These matrices can be used in the KF equations. This process essentially linearizes the nonlinear function around the current estimate. This results in the following EKF equations:

- Predict:

$$\begin{cases} x_{k|k-1} = f(\hat{x}_{k-1|k-1}, u_k, 0) \\ P^a_{k|k-1} = F_k P_{k-1|k-1} F_k^T + Q_k, \end{cases} \tag{7.11}$$

- Update:

$$\begin{cases} \tilde{y}_k = z_k - h(\hat{x}_{k|k-1}, 0) \\ S_k = H_k P_{k|k-1} H_k^T + R_k \\ K_k = P_{k|k-1} H_k^T S_K^{-1} \\ x_{k|k-1} = f(\hat{x}_{k-1|k-1}, u_k, 0) \\ P_{k|k} = (I - K_k H_k) P_{k|k-1}, \end{cases} \tag{7.12}$$

  where the state transition and observation matrices are defined to be the following Jacobians

$$\begin{cases} F_k = \dfrac{\partial f}{\partial x}|_{\hat{x}_{k-1|k-1}, u_k} \\ H_k = \dfrac{\partial h}{\partial x}|_{\hat{x}_{k|k-1}}, \end{cases} \tag{7.13}$$

For use in meteorology, see GHIL and MANALOTTE-RIZZOLI [1991], GAUTHIER, COURTIER and MOLL [1993], and BOUTTIER [1994].

7.5. Shortcomings of the EKF

Unlike its linear counterpart, the EKF is not an optimal estimator. In addition, if the initial estimate of the state is wrong, or if the process is modeled incorrectly, the filter may quickly diverge, owing to its linearization. We have really effectuated a closure by discarding moments of third and higher order giving us an approximate equation for the error variance. Usefulness of EKF will depend on properties of the model dynamics. See discussion of MILLER, GHIL and GAUTHIEZ [1994]. EVENSEN [1992] provided the first application of EKF on a nonlinear ocean circulation model. Another problem with the EKF is that the estimated covariance matrix tends to underestimate the true covariance matrix and therefore risks becoming inconsistent in the statistical sense without the addition of "stabilizing noise."

7.6. EnKF

Here, we follow algorithmic presentation of MANDEL [2006]. The EnKF is a Monte Carlo approximation of the KF avoiding evolving the covariance matrix of the pdf of the state vector x. Instead, the probability distribution is represented by a sample

$$X = [x_1, x_2, \ldots, x_N] = [x_i]. \tag{7.14}$$

X is an $n \times N$ matrix whose columns are the ensemble members, and it is called the prior ensemble. Ideally, ensemble members would form a sample from the prior distribution. However, the ensemble members are not in general independent except in the initial ensemble since every EnKF step ties them together. They are deemed to be approximately independent, and all calculations proceed as if they actually were independent. Replicate the data $\mathbf{d}$ into a $m \times N$ matrix

$$D = [\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_N] = [\mathbf{d}_i] \tag{7.15}$$

so that each column consists of the data vector $\mathbf{d}$ plus a random vector from the n-dimensional normal distribution $\mathcal{N}(0, R)$.

Because randomness is introduced in ENKF at every assimilation cycle, the algorithm updates every ensemble member to a different set of observations perturbed by a random noise.

For details, see work of HOUTEKAMER and MITCHELL [1998, 2001], HAMILL and SNYDER [2000, 2002], and more recently HOUTEKAMER, MITCHELL, PELLERIN, BUEHNER, CHARRON, SPACEK and HANSEN [2005]. If, in addition, the columns of X are a sample from the prior probability distribution, then the columns of

$$\hat{X} = X + K(D - HX) \tag{7.16}$$

form a sample from the posterior probability distribution. The EnKF is now obtained simply by replacing the state covariance Q in Kalman gain matrix

$$K = QH^T(HQH^T + R)^{-1} \tag{7.17}$$

by the sample covariance C computed from the ensemble members (called the ensemble covariance).

7.7. Implementation

7.7.1. Basic formulation

Suppose the ensemble matrix X and the data matrix D are as above. The ensemble mean and the covariance are

$$E(X) = \frac{1}{N}\sum_{k=1}^{N} x_k, \quad C = \frac{AA^T}{N-1}, \tag{7.18}$$

where

$$A = X - E(X) = X - \frac{1}{N}(Xe_{N\times 1})e_{1\times N}, \tag{7.19}$$

and e denotes the matrix of all ones of the indicated size. The posterior ensemble X^p is then given by

$$\hat{X} \simeq X^p = X + CH^T(HCH^T + R)^{-1}(D - HX), \tag{7.20}$$

where the perturbed data matrix D is as above. Since C can be written as

$$C = (X - E(X))(X - E(X))^T, \tag{7.21}$$

one can see that the posterior ensemble consists of linear combinations of members of the prior ensemble. Note that since R is a covariance matrix, it is always positive semidefinite and usually positive definite, so the inverse above exists and the formula can be implemented by the Cholesky decomposition. In Evensen [2004], R is replaced by the sample covariance $\dfrac{DD^T}{N-1}$ and the inverse is replaced by a pseudoinverse, computed using the SVD. Since these formulas are matrix operations with dominant Level 3 operations, they are suitable for efficient implementation using software packages such as LAPACK (on serial and shaby it, it is much better (several times cheaper and also more accurate) to compute the Cholesky decomposition of the matrix anred memory computers). Instead of computing the inverse of a matrix and multiplying d treat the multiplication by the inverse as solution of a linear system with many simultaneous right-hand sides. For complex Numerical Weather Prediction (NWP) models, deriving explicitly the background error covariance estimate from original method will be prohibitive

$$\hat{P}^b = \frac{\tilde{X}^b(\tilde{X}^b)^T}{m-1}, \tag{7.22}$$

where $\tilde{X}^b = (\tilde{x}_1^b K, \tilde{x}_m^b)$ and $\tilde{x}_i^b = x_i^b - \bar{x}^b$. Where the ensemble mean is defined by

$$\bar{x}^b = \frac{1}{m}\sum_{i=1}^{m} x_{x_i}^b.$$

Different parallel implementations of parallelized EnKF were proposed by KEPPENNE and RIENECKER [2002] and other method by REICHLE, MCLAUGHLIN and ENTEKHABI [2002], REICHLE, WALKER, KOSTER and HOUSER [2003]. Ensemble data assimilation algorithms with the assumptions of linear error dynamics and Gaussian error statistics will converge as the number of members of the ensemble increases to the state and covariance estimate of those of the EKF (BURGERS, VAN LEEUWEN and EVENSEN [1998]).

7.8. Reduced rank KFs

Due to the expense involved in EKF, many simplified approaches have been proposed attempting to capture only a subset of the flow-dependent error covariances. The covariances evolve with the model dynamics within a specified reduced dimension subspace defined by a fixed set of basis functions (TODLING and COHN [1994]). Possible choices for the basis functions include EOFs (CANE, KAPLAN, MILLER, TANG, HACKERT and BUSALACCHI [1996]), SVs (COHN and TODLING [1996], FISHER [1998]), or a balanced truncation of the Hankel operator (see FARRELL and IOANNOU [2001]). The EOF basis may not be optimal in the sense of providing the best subsequent forecast, which is often the goal of assimilating data. SVs or partially evolved SVs turn out to be more effective. SVs represent the directions that will evolve to optimally account for the error at a future time.

The SEEK filter for data assimilation in oceanography is a variant of the EKF with a low-rank error covariance matrix. It is quite similar in some aspects to reduced rank KF introduced by COHN and TODLING [1996], but differs in some aspects. It is derived from the EKF by approximating the state error covariance matrix by a matrix of reduced rank and evolving this matrix in decomposed form.

7.9. Algorithm of SEEK

For initialization, choose the initial estimate for the model state and an approximate state covariance matrix of low Rank in the decomposed form LUL^T. For forecast, evolve the guessed state with the full nonlinear model and the column vectors L_i with the tangent-linear model. For analysis, compute the updated state covariance matrix by an equation for the matrix U which relates the model state error to the observation error in the spirit of the Riccati equation. With this updated covariance matrix, the state update is given by the analysis step of the EKF. To avoid successive alignment of the vectors L_i, occasionally perform a reorthogonalization of these vectors (see PHAM, VERRON and ROUBAUD [1998], CARME, PHAM and VERRON [2001], HOTEIT, PHAM and BLUM [2002, 2001], HOTEIT and PHAM [2003]).

7.10. Deterministic update ensemble filters

There is also a family of nonstochastic filters (see for instance TIPPETT, ANDERSON, BISHOP, HAMILL and WHITAKER [2003]). These filters do not use perturbed observations, which, it is argued, can be a source of sampling error when ensembles are small. Instead,

they deterministically transform the ensemble of background fields into an ensemble of analyses using

$$P^a(t) = (I - K(t)H(t))\,P^f(t), \tag{7.23}$$

without adding random noise. However, this equation is valid only when the gain K is optimal, which depends, in turn, on Q and R being accurately known. A more general, but more complicated, equation for P^a, which reduces to this equation when the gain is optimal, is given by COHN [1982, Eq. (2.10b)], DALEY [1991, Eq. (13.3.19)], and GHIL and MANALOTTE-RIZZOLI [1991, Eq. (4.13b)]. The performance of stochastic and deterministic filters has been compared in a hierarchy of perfect-model scenarios by LAWSON and HANSEN [2004].

7.11. Ensemble square-root filter

The serial ensemble square-root filter (EnSRF) (WHITAKER and HAMILL [2002]) algorithm has been used for the assimilation at the scale of thunderstorms by SNYDER and ZHANG [2003], ZHANG, SNYDER and SUN [2004], and DOWELL, ZHANG, WICKER, SNYDER and CROOK [2004]. WHITAKER, COMPO, WEI and HAMILL [2004] used the algorithm for the global data assimilation of surface pressure observations. Similar to EnKF, the EnSRF conducts a set of parallel data assimilation cycles. In the EnSRF, one updates the equations for the ensemble mean (denoted by an overbar) and the deviation of the ith member from the mean separately:

$$\begin{cases} \bar{x}^a = \bar{x}^b + \hat{K}(y - H\bar{x}^b) \\ \bar{x}_i^a = \bar{x}_i^b - \tilde{K}H\bar{x}_i^b. \end{cases} \tag{7.24}$$

Here, $\hat{K}$ is the traditional Kalman gain, and $\tilde{K}$ is the reduced gain used to update deviations from the ensemble mean. In the EnSRF, the mean and departures from the mean are updated independently according to Eq. (7.24). If observations are processed one at a time, the EnSRF requires about the same computation as the traditional EnKF with perturbed observations, but for moderately sized ensembles and processes that are generally linear and Gaussian, the EnSRF produces analyses with significantly less error WHITAKER and HAMILL [2002]. Conversely, LAWSON and HANSEN [2004] suggest that if multimodality is typical and ensemble size is large, the EnKF will perform better. For details, see review of HAMILL [2006] in book of PALMER and HAGEDORN [2006].

7.12. Local ensemble Kalman filtering

Local ensemble Kalman filtering proposed by OTT, HUNT, SZUNYOGH, ZIMIN, KOSTELICH, CORAZZA, KALNAY, PATIL and YORKE [2004] and SZUNYOGH, KOSTELICH and GYARMATI et al. [2005] treats local patches surrounding every grid point independently, thus avoiding correlations between distant points. This is preferable in view of the low-rank assumption for the error covariance matrix. These local patches are analyzed independently and are then combined to yield the global analysis.

7.13. Covariance localization

One aspect of ensemble assimilation methods is the requirement of accuracy for covariance matrices. Erroneous representation of error statistics affects the analysis-error covariance, which is propagated forward in time.

The covariance estimate from the ensemble is multiplied point by point with a correlation function that is 1.0 at the observation location and zero beyond some prespecified distance (correlation length).

Two approaches are used : one consists in a cut-off radius so that observations are not assimilated beyond a certain distance from the grid point (see HOUTEKAMER and MITCHELL [1998], EVENSEN [2003]). This may introduce spurious discontinuities.

The second approach is to use a correlation function that decreases monotonically with increasing distance. This results in the Kalman gain

$$K = P^b H^T (HP^b H^T + R)^{-1}, \tag{7.25}$$

being replaced by a modified gain

$$K = (\rho_s \circ P^b) H^T (H(\rho_s \circ P^b) H^T + R)^{-1}, \tag{7.26}$$

where the operation $\rho_s \circ$ denotes a Schur product (an element-by-element multiplication) of a correlation matrix S with local support with the covariance model generated by the ensemble. The Schur product of matrices A and B is a matrix C of the same dimension, where $c_{ij} = a_{ij} b_{ij}$. When covariance localization is applied to smaller ensembles, it may result in more accurate analyses than would be obtained from larger ensembles without localization HOUTEKAMER and MITCHELL [2001]. Localization increases the effective rank of the background error covariances HAMILL, WHITAKER and SNYDER [2001].

Generally, the larger the ensemble, the broader the optimum correlation length scale of the localization function (HOUTEKAMER and MITCHELL [2001], HAMILL, WHITAKER and SNYDER [2001]). See WHITAKER, COMPO, WEI and HAMILL [2004] and HOUTEKAMER, MITCHELL, PELLERIN, BUEHNER, CHARRON, SPACEK and HANSEN [2005] for examples performing ensemble assimilations that also include a vertical covariance localization.

8. Assimilation of images

The observation of the Earth by geostationary or polar-orbiting satellites clearly displays the evolution of some characteristics features such as fronts, the color, or the temperature of the ocean. Figure 8.1 represents the sea surface temperature (SST) of the Black Sea observed by the satellite Moderate Resolution Imaging Spectroradiometer (MODIS), and some geometric features are identified and their temporal evolution has an important informative content. At the present time, this information is used more in a qualitative fashion rather than in a quantitative one. The question arising is: is it possible to couple this information with a numerical model, i.e., how to assimilate images?

Two basic approaches can be considered:

- The first one (HERLIN, LE DIMET, HUOT and BERROIR [2004]) consists, in a first step, to extract from the images some "pseudo" measurements (e.g., surface velocity in

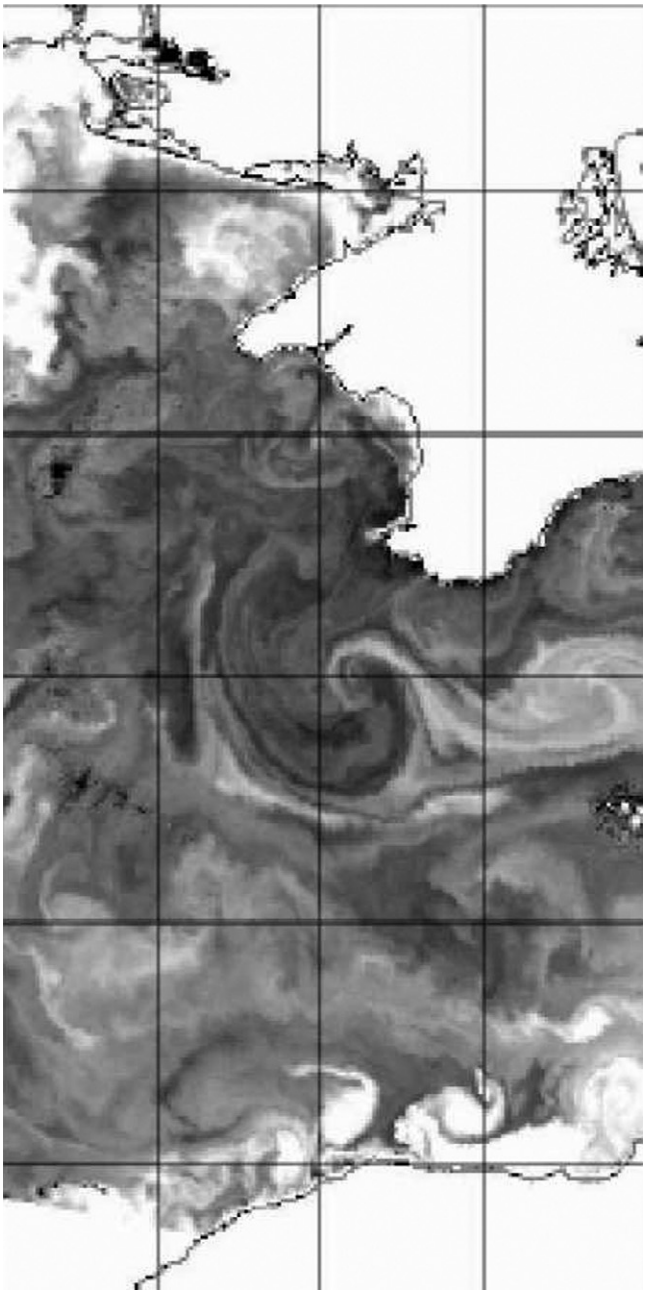


FIG. 8.1 Sea surface temperature in the Black Sea from MODIS.

oceanography, wind components in meteorology), then these measurements will be used in a classical scheme of VDA. For instance, in meteorology some identified clouds can be used as Lagrangian tracers, and assimilated as such. This information makes sense only if the altitudes of the clouds are known, it can be done by an evaluation of the temperature of the cloud and a comparison with the vertical profile of temperature.

- The second approach (MA, ANTONIADIS and LE DIMET [2006]) consists to consider images as objects and insert them directly in the variational analysis.

8.1. Retrieving velocities from images

This approach is a classical one in computer vision, and it is based on the conservation of grey level for individual pixels. Let us consider a pixel of coordinates $(x(t), y(t))$, if I is the luminance of the pixel, this quantity is conservative and its total derivative is equal to 0:

$$\frac{\mathrm{d}I}{\mathrm{d}t}(x(t), y(t), t) = 0.$$

By developing this expression we get

$$\frac{\partial I}{\partial x}.\frac{\mathrm{d}x}{\mathrm{d}t} + \frac{\partial I}{\partial y}.\frac{\partial y}{\partial t} + \frac{\partial I}{\partial t} = \frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v + \frac{\partial I}{\partial t} = 0,$$

u and v are the components of the velocity of the flow and are unknown. This equation is not sufficient to retrieve the velocity field, but it can be included in a variational formulation: we will seek for the velocity field minimizing the functional J defined by

$$J(u, v) = E_1 + E_2,$$

with

$$E_1 = \int_\Omega \left(\frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v + \frac{\partial I}{\partial t}\right)^2 d\Omega,$$

and

$$E_2 = \int_\Omega \|\nabla W\|^2 \, d\Omega,$$

with

$$W = (u, v).$$

E_2 can be considered as a regularization term to smooth the retrieved fields. E_1 is the scalar product of the gradient of luminance with the velocity field, and if these vectors are orthogonal, then the equation does not contain any quantitative information on the

velocity field. Some other conservation laws can be considered according to the nature of the image:

- With an image displaying the color of the ocean, an equation of chlorophyll conservation must be used, and it will have to include sink and source terms, and therefore an equation modeling biological processes must be added to the physical model.
- With an image of the SST, the Boussinesq approximation can be considered.

A selection of points on which the minimization of J is carried out to determine (u, v) has to be done. The structures where the velocity and the gradient of luminance are almost orthogonal must be discarded from the analysis, this being the case of filaments which are elongated structures. These structures are detected by applying operators of mathematical morphology: peak operator detecting brighter areas of maximum width and valley operator detecting darker areas with the same characteristics.

After detection of these structures, a second selection is carried out according to two criteria; the first one consists in removing the filaments having a large elongation – this is done by evaluating the condition number of their inertia matrix, the second selection will discard the quasi-steady state structures. At the pixel level, another selection is done: only points with a significant displacement are considered. Once the pixels have been selected, then an optimization procedure of the function J is performed and a field of velocity is obtained. The result will be considered as pseudo-observation and then included in a VDA scheme. The process is illustrated in Fig. 8.2. An inconvenience of this method is its large number of degrees of freedom: in the choice of the laws of conservation and also in the choice of threshold parameters for the selection of pixels.

8.2. Direct assimilation of images

In the former method, it is necessary to solve several problems of optimization, each iterative algorithm requires the choice of at least one stopping criterion, the accumulation of these quantities being detrimental to the control of the global algorithm. A way to alleviate this difficulty is to consider some characteristic features of the images (e.g., fronts) as objects and they will be assimilated as such in addition to the usual state variables.

In the cost function defining the VDA, an additional term will be added in the form:

$$J_2(X) = \int_0^T \int_\Omega \|DX - I\|^2 \mathrm{d}t\mathrm{d}\Omega.$$

I is the image and D an operator from the space of the state variable of the model toward the space of images. Therefore, the comparison between the images retrieved from the model and the observed images is carried out in the space of images. The questions are

- How to choose the space of images?
- What metric must be used in this space for obtaining an efficient and pertinent comparison?

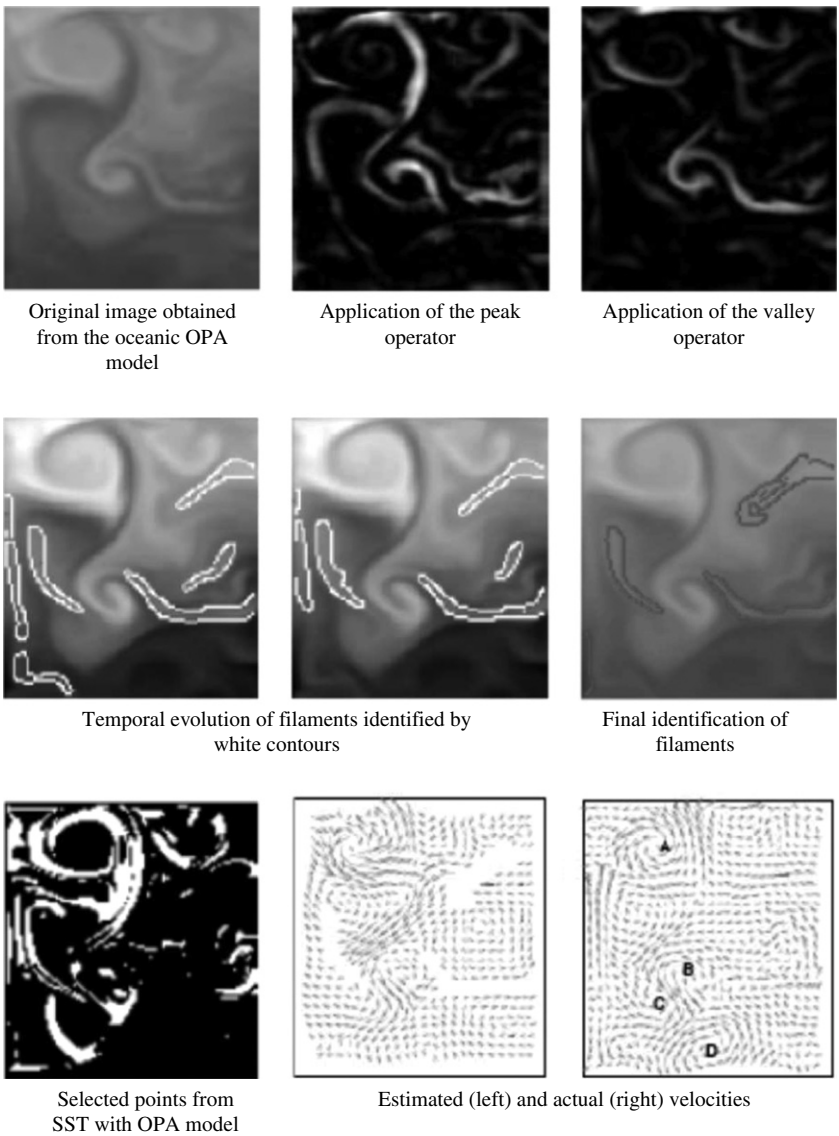
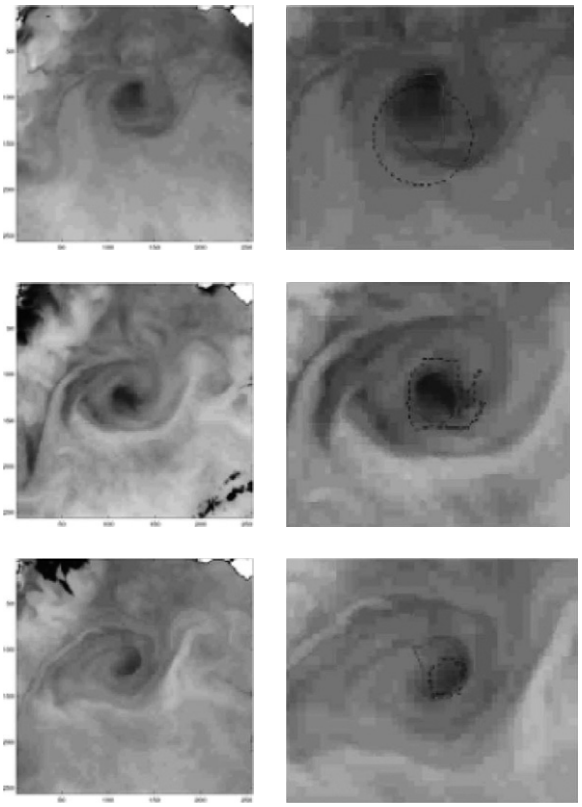


FIG. 8.2 Retrieving velocities from images.

In MA, ANTONIADIS and LE DIMET [2006], the choice has been done to use curvelets (CANDES, DEMANET, DONOHO and YING [2006]). The features of interest in the images are defined by contours: snakes. A snake is a virtual object which can be deformed elastically (thus possessing an internal energy) and which is immersed in a potential field. The main difficulties for applying a snake model to a temporal sequence of images consist in the determination of an initial contour and the design of external forces.



Evolution of a vortex in SST

Tracking in triplet successive frames using the curvelets based method. The initial and final snakes are displayed

FIG. 8.3 Direct assimilation of images. (See also color insert).

The advantage of curvelet-based multiscale methods for snake modeling and tracking is its ability for simultaneously detecting edges and suppressing noise. An example of application of this method is given in Fig. 8.3.

Assimilating images is a generic problem with potential developments not only for geophysical fluids but also in biology and medicine. With respect to the classical data assimilation, due to the multiscale approach, it has the potential to focus on local features such as storms or hurricanes in meteorology. We can expect that further developments will be achieved in the following years.

9. Conclusion

Presently, data assimilation is a very active domain of research with extensions toward several directions.

- The domain of application of these methods has been extended to some other domains in geophysics especially in hydrology for the water cycle surface and underground water. Atmospheric chemistry is an important domain of potential applications.
- From the computational point of view, there is a demand for efficient and fast methods saving both storage and computing time.
- From the theoretical point of view, these methods are not always clearly justified, especially in the nonlinear case. Many problems remain open such as the optimal location of sensors.

Data assimilation has become an essential tool for modeling and prediction of the evolution of geophysical fluids. In many other domains for which data and models are the main sources of information, these methods could be developed in the near future.

10. Acknowledgments

The authors would like to thank Didier Auroux, William Castaings, Mohamed Jardak, and Diana Cheng for their help in the preparation of the manuscript. This paper was partly (assimilation of images) funded by Agence Nationale de la Recherche, France (ANR) in the framework of the ADDISA project. Prof. I. M. Navon would like to acknowledge the support of NSF grants ATM-0201808 and CCF-0635162.

References

ANDERSON, J.L. (2003). A local least squares framework for ensemble filtering. *Mon. Wea. Rev.* **131**, 634–642.

ANDERSON, J.L., ANDERSON, S.L. (1999). A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Mon. Wea. Rev.* **127**, 2741–2758.

ANDERSON, B.D.O., MOORE, J.B. (1979). *Optimal Filtering* (Prentice-Hall, Englewood Cliffs, NJ).

ANTHES, R.A. (1974). Data assimilation and initialization of hurricane prediction models. *J. Atmos. Sci.* **31**, 702–719.

AUROUX, D. (2003). Étude de différentes méthodes d'assimilation de données pour l'environnement, Thesis, Université de Nice Sophia-Antipolis.

AUROUX, D., BLUM, J. (2005). Back and forth nudging algorithm for data assimilation problems. *C. R. Acad. Sci. Paris*. Sér. I **340** (12), 873–878.

AUROUX, D., BLUM, J. (2008). A nudging-based data assimilation method: the Back and Forth Nudging (BFN) algorithm. *Nonlinear Processes in Geophysics*, **15**, 305–319.

BENGTSSON, T., SNYDER, C., NYCHKA, D. (2003). Toward a nonlinear ensemble filter for high dimensional systems. *J. Geophys. Res. Atmos.* **108** (D24).

BERKOOZ, G., HOLMES, P., LUMLEY, J.L. (1993). The proper orthogonal decomposition in the analysis of turbulent flows. *Annu. Rev. Fluid Mech.* **25**, 539–575.

BISHOP, C.H., ETHERTON, B.J., MAJUMDAR, S.J. (2001). Adaptive sampling with the ensemble transform Kalman filter. Part I: theoretical aspects. *Mon. Wea. Rev.* **129** (3), 420–436.

BLAYO, E., BLUM, J., VERRON, J. (1998). Assimilation variationnelle de données en océanographie et réduction de la dimension de l'espace de contrôle. In: *Équations aux dérivées partielles et applications* (Articles dédiés 'a Jacques-Louis Lions, Gauthier-Villars), pp. 199–219.

BLUM, J., LUONG, B., VERRON, J. (1998). Variational assimilation of altimeter data into a non-linear ocean model: temporal strategies, Contrôle et e.d.p. (Luminy, juin 1997). *ESAIM Proc.* **4**, 21–57.

BOUTTIER, F. (1994). A dynamical estimation of forecast error covariances in an assimilation system. *Mon. Wea. Rev.* **122**, 2376–2390.

BUDGELL, W.P. (1986). Nonlinear data assimilation for shallow water equations in branched channels. *J. Geophys. Res.* **91** (C9), 10633–10644.

BUI-THANH, T., WILLCOX, K., GHATTAS, O., VAN BLOEMEN WAANDERS, B. (2007). Goal-oriented model-constrained optimization for reduction of large-scale systems. *J. Comput. Phys.* **224** (2), 880–896.

BURGERS, G., VAN LEEUWEN, P.J., EVENSEN, G. (1998). Analysis scheme in the ensemble Kalman filter. *Mon. Wea. Rev.* **126**, 1719–1724.

BURKARDT, J., GUNZBURGER, M., LEE, H.C. (2006). POD and CVT-based reduced-order modeling of Navier-Stokes flows. *Comput. Methods Appl. Mech. Eng.* **196** (1–3), 337–355.

CANDES, E., DEMANET, D., DONOHO, L., YING, L. (2006). Fast discrete curvelets transforms. *Multiscale Model. Simul.* **5** (3), 861–899.

CANE, M.A., KAPLAN, A., MILLER, R.N., TANG, B., HACKERT, E.C., BUSALACCHI, A.J. (1996). Mapping tropical Pacific sea level: data assimilation via a reduced state space Kalman filter. *J. Geophys. Res.* **101**, 22599–22617.

CAO, Y., ZHU, J., LUO, Z., NAVON, IM. (2006). Reduced-order modeling of the upper tropical Pacific ocean model using proper orthogonal decomposition. *Comput. Math. Appl.* **52** (8–9), 1373–1386.

CAO, Y., ZHU, J., NAVON, I.M., LUO, Z. (2007). A reduced-order approach to four-dimensional variational data assimilation using proper orthogonal decomposition. *Int. J. Numer. Methods Fluids* **53** (10), 1571–1583.

Carme, S., Pham, D.T., Verron, J. (2001). Improving the singular evolutive extended Kalman filter for strongly nonlinear models for use in ocean data assimilation. *Inverse Probl.* **17** (5), 1535–1559.

Chen, Y., Snyder, C. (2007). Assimilating vortex position with an ensemble Kalman filter. *Mon. Wea. Rev.* **135**, 1828–1845.

Cohn, S.E., Ghil, M., Isaacson, E. (1981). Applications of estimation theory to numerical weather prediction. *Bull. Am. Meteorol. Soc.* **62** (6), 899.

Cohn S.E. (1982). Methods of sequential estimation for determining initial data in numerical weather prediction. Thesis (Ph.D.)–New York University (Source: Dissertation Abstracts International, Vol. 43-07, Sec. B, p. 2227).

Cohn, S.E. (1997). An introduction to estimation theory. *J. Meteorol. Soc. Jpn.* **75** (1B), 257–288.

Cohn, S.E., Todling, R. (1996). Approximate data assimilation schemes for stable and unstable dynamics. *J. Meteorol. Soc. Jpn.* **74** (1), 63–75.

Cohn, S.E., Parrish, D.E. (1991). The behavior of forecast error covariances for a kalman filter in two dimensions. *Mon. Wea. Rev.* **119** (8), 1757–1785.

Courtier, Ph., Talagrand, O. (1987). Variational assimilation of meteorological observations with the adjoint equations part 2. Numerical results. *Q. J. Roy. Meteorol. Soc.* **113**, 1329–1347.

Courtier, Ph., Thepaut, J-N., Hollingsworth, A. (1994). A strategy for operational implementation of 4D-Var, using an incremental approach. *Q. J. Roy. Meteorol. Soc.* **120**, 1367–1388.

Daescu, D.N., Navon, I.M. (2007). Efficiency of a POD-based reduced second order adjoint model in 4D-Var data assimilation. *Int. J. Numer. Meth. Fluids.* **53**, 985–1004.

Daescu, D.N., Navon, I.M. (2008). A dual-weighted approach to order reduction in 4D-Var data assimilation. *Monthly Weather Review*. **136** (3), 1026–1041.

Daley, R. (1991). *Atmospheric Data Analysis*, Cambridge University Press, pp. 471.

Dowell, D.C., Zhang, F., Wicker, L.J., Snyder, C., Crook, N.A. (2004). Wind and thermodynamic retrievals in the 17 may 1981 Arcadia, Oklahoma supercell: ensemble Kalman filter experiments. *Mon. Wea. Rev.* **132**, 1982–2005.

Du Plessis, R.M. (1967). *Poor Man's Explanation of Kalman Filters or How I Stopped Worrying and Learned to Love Matrix Inversion*. (Rockwell International Technical Note, Anaheim, CA), pp. 24.

Durbiano, S. (2001). Vecteurs caractéristiques de modéles océaniques pour la réduction d'ordre en assimilation de données, Ph.D. Thesis, Université Joseph Fourier, Laboratoire de Modélisation et calcul. Grenoble, France, pp. 214.

Evensen, G. (1992). Using the extended Kalman filter with a multilayer quasi-geostrophic ocean model. *J. Geophys. Res. Ocean.* **97** (C11), 17905–17924.

Evensen, G. (1994). Sequential data assimilation with a nonlinear quasigeostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.* **99** (C5), 10143–10162.

Evensen, G. (2007). *Data Assimilation: The Ensemble Kalman Filter* (Springer, Berlin, Germany).

Evensen, G., Van Leeuwen, P.J. (1996). Assimilation of Geosat altimeter data for the Agulhas Current using the ensemble Kalman filter with a quasigeostrophic model. *Mon. Wea. Rev.* **124** (1), 85–96.

Evensen, G. (2004). Sampling strategies and square root analysis schemes for the EnKF. *Ocean Dyn.* **54**, 539–560.

Evensen, G. (2003). The ensemble Kalman filter: theoretical formulation and practical implementation. *Ocean Dyn.* **53**, 343–367.

Fahl, M., Sachs, E. (2003). Reduced-order modeling approaches to pdeconstrained optimization based on proper orthogonal decomposition. In: Biegler, L.T., Ghattas, O., Heinkenschloss, M., Van Bloemen Waanders, B. (eds.), *Large Scale PDE-Constrained Optimization* (Springer Verlag), pp. 268–281, 350.

Fang, F., Pain, C., Navon, I.M., Piggott, M.D., Gorman, G.J., Goddard, A.J.H. (2008). Reduced-order modeling of an adaptive mesh ocean model. In Early view with *International Journal for Numerical Methods in Fluids*.

Farrell, B.F., Ioannou, P.J. (2001). Estimation using a reduced-order Kalman filter. *J. Atmos. Sci.* **58** (23), 3666–3680.

Fisher, M. (1998). *Development of a Simplified Kalman Filter. ECMWF Research Department Technical Memorandum 260*. (European Centre for Medium-Range Weather Forecasts), pp. 16. Available from Library, ECMWF, Shinfield Park, Reading, Berkshire, RG29AX England.

Fukumori, I., Malanotte-Rizzoli, P. (1995). An approximate Kalman filter for ocean data assimilation - an example with an idealized gulf-stream model. *J. Geophys. Res. Oceans* **100** (C4), 6777–6793.

Gauthier, P., Courtier, P., Moll, P. (1993). Assimilation of simulated lidar data with a Kalman filter. *Mon. Wea. Rev.* **121**, 1803–1820.

Gejadze, I., Le Dimet, F.-X., Shutyaev, V. (2007). On error covariances in variational data assimilation. *Russ. J. Numer. Anal. Math. Model.* **22** (2), 163.

Gelb, A., Kasper, J.F., Nash, R.A., Price, C.F., Sutherland, A.A. (1974). *Applied Optimal Estimation* (The M.I.T. Press), pp. 374.

Ghil, M., Manalotte-Rizzoli, P. (1991). Data assimilation in meteorology and oceanography. *Adv. Geophys.* **23**, 141–265.

Ghil, M. (1989). Meteorological data assimilation for oceanographers. Part I: description and theoretical framework. *Dyn. Atmos. Oceans* **13**, 171–218.

Ghil, M., Cohn, S., Tavantzis, J., Bube, K., and Isaacson, E. 1981: Applications of estimation theory to numerical weather predication, *Dynamic Meteorology: Data Assimilation Methods*, L. Bengtsson, M. Ghil and E. Källén (eds.), Springer Verlag, pp. 139–224.

Gilbert, J.C., Lemarechal, C. (1989). Some numerical experiments with variable storage quasi-Newton algorithms. *Math. Prog.* **B25**, 407–435.

Golub, G.H., Van Loan, C.F. (1989). *Matrix Computations*, second ed. (Johns Hopkins Univ. Press).

Gunzburger, MD., Peterson, JS., Shadid, JN. (2007). Reduced-order modeling of time-dependent PDEs with multiple parameters in the boundary data. *Comput. Methods Appl. Mech. Eng.* **196** (4–6), 1030–1047.

Gunzburger, M.D. (2003). *Perspectives in Flow Control and Optimization* (Society for Industrial and Applied Mathematics, Philadelphia, PA), pp. 261.

Hager, W.W. (1989). Updating the inverse of a matrix. *SIAM Rev.* **31**, 221–239.

Hamill, T.M. (2006). Ensemble-based atmospheric data assimilation. In: Tim Palmer and Renate Hagedorn (eds.), Predictability of Weather and Climate, Chapter 6, Cambridge University Press, pp. 124–156.

Hamill, T.M., Snyder, C. (2000). A hybrid ensemble Kalman filter/3dvariational analysis scheme. *Mon. Wea. Rev.* **128**, 2905–2919.

Hamill, T.M., Snyder, C. (2002). Using improved background-error covariances from an ensemble Kalman filter for adaptive observations. *Mon. Wea. Rev.* **130**, 1552–1572.

Hamill, T.M., Whitaker, J.S. (2005). Accounting for the error due to unresolved scales in ensemble data assimilation: a comparison of different approaches. *Mon. Wea. Rev.* **133**, 3132–3147.

Hamill, T.M., Whitaker, J.S., Snyder, C. (2001). Distance-dependent filtering of background-error covariance estimates in an ensemble Kalman filter. *Mon. Wea. Rev.* **129**, 2776–2790.

Herlin, I., Le Dimet, F.-X., Huot, E., Berroir, J.-P. (2004). Coupling models and data: which possibilities for remotely-sensed images. In: Prastacos, P., Cortes, U., Diaz, J.L., Murillo, J. (eds.), *e-Environement: Progress and Challenge*, vol. 11 (Research on Computing Sciences, Mexico).

Hoke, J.E., Anthes, R.A. (1976). The initialization of numerical models by a dynamical initialization technique. *Mon. Wea. Rev.* **104**, 1551–1556.

Holland, W.R. (1978). The role of mesoscale eddies in the general circulation of the ocean-numerical experiments using a wind-driven quasigeostrophic model. *J. Phys. Oceanogr.* **8**, 363–392.

Holmes, P., Lumley, J.L., Berkooz, G. (1996). *Turbulence, Coherent Structures, Dynamical Systems and Symmetry* (Cambridge University Press, Cambridge, UK).

Holmes, P.J. (1990). Nonlinear dynamics, chaos and mechanics. *Appl. Mech. Rev.* **43** (5–2), S23–S39.

Hoteit, I., Pham, D.-T. (2003). Evolution of the reduced state space and data assimilation schemes based on the Kalman filter. *J. Meteorol. Soc. Jpn.* **81** (1), 21–39.

Hoteit, I., Kohl, A. (2006). Efficiency of reduced-order, time-dependent adjoint data assimilation approaches. *J. Oceanogr.* **62** (4), 539–550.

Hoteit, I., Pham, D.T., Blum, J. (2002). A simplified reduced-order Kalman filtering and application to altimetric data assimilation in Tropical Pacific. *J. Mar. Syst.* **36** (1–2), 101–127.

Hoteit, I., Pham, D.T., Blum, J. (2001). A semi-evolutive partially local filter for data assimilation. *Mar. Pollut. Bull.* **43** (7–12), 164–174.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **24**, 417–441, 498–520.

HOUTEKAMER, P.L., MITCHELL, H.L. (1998). Data assimilation using an ensemble Kalman filter technique. *Mon. Wea. Rev.* **126**, 796–811.

HOUTEKAMER, P.L. MITCHELL, H.L. (2001). A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon., Wea. Rev.* **129** (1), 123–137.

HOUTEKAMER, P.L., MITCHELL, H.L., PELLERIN, G., BUEHNER, M., CHARRON, M., SPACEK, L., HANSEN, B. (2005). Atmospheric data assimilation with an ensemble Kalman filter: results with real observations. *Mon. Wea. Rev.* **133** (3), 604–620.

HOUTEKAMER, P.L., MITCHELL, H.L. (2005). Ensemble Kalman filtering. *Q. J. Roy. Meteorol. Soc.* **131** (613), 3269–3289, part C.

JAZWINSKI, A.H. (1970). *Stochastic Processes and Filtering Theory* (Academic Press), pp. 376.

KALMAN, R.E. (1960). A new approach to linear filtering and prediction problems. *J. Basic Eng. Trans. ASME* **82**, 35–45.

KALMAN, R.E. (1961). New methods and results in linear filtering and prediction theory. *ASME J. Basic Eng. Ser. D.* **83**.

KALNAY, E. (2003). *Atmospheric Modeling, Data Assimilation and Predictability* (Cambridge University Press, Cambridge, UK).

KALNAY, E., LI, H., MIYOSHI, T., YANG, S.-C., BALLABRERA-POY, J. (2007). 4D-Var or ensemble Kalman filter? *Tellus A* **59** (5), 758–773.

KEPPENNE, C.L., RIENECKER, M.M. (2002). Initial testing of a massively parallel ensemble Kalman filter with the Poseidon isopycnal ocean general circulation model. *Mon. Wea. Rev.* **130**, 2951–2965.

KARHUNEN, K. (1946). Zur spektraltheorie stochastischer prozesse. *Ann. Acad. Sci. Fennicae* 37.

KIRBY, M., SIROVICH, L. (1990). Application of the Karhunen-Loeve procedure for the characterization of human faces. *IEEE Trans. Pattern Anal. Mach. Intell.* **12** (1), 103–108.

KIRBY, M. (2001). *Geometric Data Analysis: An Empirical Approach to Dimensionality Reduction and the Study of Patterns* (John Wiley).

KOSAMBI, D.D. (1943). *J. Indian Math. Soc.* **7**, 76–88.

KRISHNAMURTI, T.N., JISHAN, X., BEDI, H.S., INGLES, K., OOSTERHOF, D. (1991). Physical initialization for numerical weather prediction over the tropics. *Tellus* **43A**, 53–81.

LAWLESS, A.S., NICHOLS, N.K., BOESS, C., BUNSE-GERSTNER, A. (2006). *Using Model Reduction Methods with in Incremental 4D-Var, Numerical Analysis Report 10/06* (Department of Mathematics, The University of Reading).

LAWSON, W.G., HANSEN, J.A. (2004). Implications of stochastic and deterministic filters as ensemble-based data assimilation methods in varying regimes of error growth. *Mon. Wea. Rev.*, **132**, 1966–1981.

LE DIMET, F.-X. (1980). *Une Étude Générale D'analyse Objective Variationnelle Des Champs Mét éorologiques* (Rapport Scientifique LAMP 28, Université de Clermond II, BP 45, 63170 Aubi'ere, France).

LE DIMET, F.-X., TALAGRAND, O. (1986). Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. *Tellus* **38A**, 97–110.

LE DIMET, F.-X., NGODOCK, H.-E., LUONG, B., VERRON, J. (1997). Sensitivity analysis in variational data assimilation. *J. Met. Soc. Jpn.* **75** (1B), 245–255.

LE DIMET, F.-X., NAVON, I.M., DAESCU, D.N. (2002). Second order information in data assimilation. *Mon. Wea. Rev.* **130** (3), 629–648.

LE DIMET, F.-X., SHUTYAEV, V., GEJADZE, I. (2006). On optimal solution error in variational data assimilation: theoretical aspects. *Russ. J. Numer. Anal. Math. Model.* **21** (2), 139.

LIONS, J.L. (1971). *Optimal Control of Systems Governed by Partial Differential Equations* (Springer Verlag).

LOEVE, M. (1945). Fonctions aléatoires de second ordre. *Compte Rend. Acad. Sci. Paris* **220**.

LERMUSIAUX, P.F.J., ROBINSON, A.R. (1999). Data assimilation via error subspace statistical estimation. 1: theory and schemes. *Mon. Wea. Rev.* **127**, 1385–1407.

LI, Z., NAVON, I.M. (2001). Optimality of variational data assimilation and its relationship with the Kalman filter and smoother. *Q. J. Roy. Meteorol. Soc.* **127**, 661–683.

LORENC, A.C. (1986). Analysis methods for numerical weather prediction. *Q. J. Roy. Meteorol. Soc.* **112**, 1177–1194.

LORENC, A.C. (1988). A practical approximation to optimal four-dimensional objective analysis. *Mon. Wea. Rev.* **116**, 730–745.

LUMLEY, J.L. (1967). The structure of inhomogeneous turbulence. In: Yaglom, A.M., Tatarski, V.I. (eds.), *Atmospheric Turbulence and Wave Propagation* (Nauka, Moscow), pp. 166–178.

Luo, Z., Chen, J., Zhu, J., Wang, R., Navon, I.M. (2007). An optimizing reduced-order FDS for the tropical Pacific Ocean reduced gravity model. *Int. J. Numer. Methods Fluids* **55** (2), 143–161.

Luong, B., Blum, J., Verron, J. (1998). A variational method for the resolution of a data assimilation problem in oceanography. *Inverse Probl.* **14**, 979–997.

Lyne, W.H. (1979). *Data Assimilation by Repeated Correction of Model Fields-Theoretical Considerations* (Met O 11 Tech. Note No. 130, Meteorological Office).

Lyne, W.H., Swinbank, R., Birch, N.T. (1982). A data assimilation experiment and the global circulation during the FGGE special observing periods. *Q. J. R. Met. Soc.* **108**, 575–594.

Ma, J., Antoniadis, A., Le Dimet, F.-X. (2006) Curvelet-based snake for multiscale detection and tracking of geophysical fluids. *IEEE Trans. Geosci. Remote Sens.* **14** (12), 3626–3638.

Mandel, J. (2007). *A Brief Tutorial on the Ensemble Kalman Filter* (CCM Report 242, University of Colorado at Denver and Health Sciences Center).

Mandel, J. (2006). *Efficient Implementation of the Ensemble Kalman Filter* (CCM Report 231, University of Colorado at Denver and Health Sciences Center).

Mandel, J., Beezley, J.D. (2007). *Predictor-Corrector and Morphing Ensemble Filters for the Assimilation of Sparse Data into High-Dimensional Nonlinear Systems* CCM Report 239, 11th Symposium on Integrated Observing and Assimilation Systems for the Atmosphere, Oceans, and Land Surface (IOAS-AOLS), CD-ROM, Paper 4.12, 87th American Meteorological Society Annual Meeting, San Antonio, TX, January 2007 presentation, November 2006.

Maybeck, P.S. (1979). *Stochastic Models, Estimation, and Control, vol. 1* (Academic Press).

Meyer, M., Matthies, HG. (2003). Efficient model reduction in non-linear dynamics using the Karhunen-Loeve expansion and dual-weighted-residual methods. *Comput. Mech.* **31**, 179–191.

Miller, R.N., Ghil, M., Gauthiez, F. (1994). Advanced data assimilation in strongly nonlinear dynamical-systems. *J Atmos. Sci.* **51** (8), 1037–1056.

Miller, R.N. (1986). Toward the application of the Kalman filter to regional open ocean modeling. *J. Phys. Oceanogr.* **16** (1), 72–86.

Mohammadi, B., Pironneau, O. (2001). *Applied Shape Optimization for Fluids* (Clarendon Press, Oxford, UK).

Navon, I.M., Zou, X., Derber, J., Sela, J. (1992). Variational data assimilation with an adiabatic version of the NMC spectral model. *Mon. Wea. Rev.* **120** (7), 1433–1446.

Navon, I.M. (1998). Practical and theoretical aspects of adjoint parameter estimation and identifiability in meteorology and oceanography. *Dyn. Atmos. Ocean.* **27** (1–4), 55–79, Special Issue in honor of Richard Pfeffer.

Nerger, L., Hiller, W., Schrater, J. (2005). A comparison of error subspace Kalman filters. *Tellus A* **57** (5), 715–735.

Nodet, M. (2006). Variational assimilation of Lagrangian data in oceanography. *Inverse Probl.* **22**, 245–263.

Ott, E., Hunt, B.R., Szunyogh, I., Zimin, A.V., Kostelich, E.J., Corazza, E., Kalnay, E., Patil, D., Yorke, J.A. (2004). A local ensemble Kalman filter for atmospheric data assimilation. *Tellus Series A - Dynamic Meteorology and Oceanography* **56** (5), 415–428.

Palmer, T., Hagedorn, R. (2006). *Predictability of Weather and Climate* (Cambridge University Press), pp. 718.

Parmuzin, E., Le Dimet, F.-X., Shutyaev, V. (2006). On error analysis in variational data assimilation problem for a nonlinear convection-diffusion model. *Russ. J. Numer. Anal. Math. Model.* **21** (2), 169.

Parrish, D., Derber, J. (1992). The national meteorological centers spectral statistical interpolation analysis system. *Mon. Wea. Rev.* **120**, 1747–1763.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philos. Mag.* **2**, 559–572.

Peterson, J.S. (1989). The reduced basis method for incompressible viscous flow calculations. *SIAM J. Sci. Stat. Comput.* **10** (4), 777–786.

Pham, DT., Verron, J., Roubaud, MC. (1998). A singular evolutive extended Kalman filter for data assimilation in oceanography. *J. Mar. Syst.* **16** (3–4), 323–340.

Puel, J.P. (2002). A non-standard approach to a data assimilation problem. *C. R. Acad. Sci.* Paris, Ser.1 **335**, 161–166.

Rabier, F. (2005). Overview of global data assimilation developments in numerical weather-prediction centres. *Q. J. Roy. Meteorol. Soc.* **131** (613), 3215–3233, part C.

Ramamurthy, M.K., Carr, F.H. (1987). Four-dimensional data assimilation in the monsoon region. Part I: experiments with wind data. *Mon. Wea. Rev.* **115**, 1678–1706.

Ramamurthy, M.K., Carr, F.H. (1988). Four-dimensional data assimilation in the monsoon region. Part II: role of temperature and moisture data. *Mon. Wea. Rev.* **116**, 1896–1913.

Ravindran, S.S. (2002). Adaptive reduced-order controllers for a thermal flow system using proper orthogonal decomposition. *SIAM J. Sci. Comput.* **23** (6), 1924–1942.

Ravindran, S.S. (2006). Reduced-order controllers for control of flow past an airfoil. *Int. J. Numer. Methods Fluids* **50** (5), 531–554.

Ravindran, S.S. (2000). A reduced-order approach for optimal control of fluids using proper orthogonal decomposition. *Int. J Numer. Methods Fluids* **34** (5), 425–448.

Reichle, R.H., McLaughlin, D.B., Entekhabi, D. (2002). Hydrologic data assimilation with the ensemble Kalman filter. *Mon. Wea. Rev.* **130**, 103–114.

Reichle, R.H., Walker, J.P., Koster, R.D., Houser, P.R. (2003). Extended versus ensemble Kalman filtering for land data assimilation. *J. Hydrometeorol.* **3**, 728–740.

Robert, C., Durbiano, S., Blayo, E., Verron, J., Blum, J., Le Dimet, F.X. (2005). A reduced-order strategy for 4D-Var data assimilation. *J. Mar. Syst.* **57** (1–2), 70–82.

Robert, C., Blayo, E., Verron, J. (2006). Reduced-order 4D-Var: a preconditioner for the incremental 4D-Var data assimilation method. *Geophys. Res. Let.* **33**, L18609, doi: 10.1029/2006GL026555.

Robert, C., Blayo, E., Verron, J. (2006). Comparison of reduced-order, sequential and variational data assimilation methods in the tropical Pacific Ocean. *Ocean Dyn.* **56**, 624–633, doi: 10.1007/s10236-006-0079-9.

Rostaing-Schmidt, N., Hassold, E. (1994). Basic function representation of programs for automatic differentiation in the odyssée system. In: Le Dimet, F.-X. (ed.), *High Performance Computing in the Geosciences* (Kluwer Academic Publishers B.V. NATO ASI Series), pp. 207–222.

Sachs, E.W. (2003). Managing POD models in PDE-constrained optimization. Optimization in simulation-based models IMA, Minneapolis. Available at 2003; http://www.ima.umn.edu/talks/workshops/1-9-16.2003/sachs/sachs.pdf.

Sasaki, Y.K. (1958). An objective analysis based on the variational method. *J. Met. Soc. Jpn. II* **36**, 77–88.

Sirovich, L.P. (1987). Turbulence and the dynamics of coherent structures, Part I: coherent structures. *Q. Appl. Math.* **45**, 561–571.

Sirovich, L.P. (1987). Turbulence and the dynamics of coherent structures, Part II: symmetries and transformations. *Q. Appl. Math.* **45**, 573–582.

Sirovich, L.P. (1987). Turbulence and the dynamics of coherent structures, Part III: dynamics and scaling. *Q. Appl. Math.* **45**, 583–590.

Snyder, C., Zhang, F. (2003). Assimilation of simulated Doppler radar observations with an ensemble Kalman filter. *Mon. Wea. Rev.* **131**, 1663–1677.

Stauffer, D.R., Seaman, N.L. (1990). Use of four-dimensional data assimilation in a limited-area mesoscale model. Part I: experiments with synoptic-scale data. *Mon. Wea. Rev.* 118, 1250–1277.

Szunyogh, I., Kostelich, E.J., Gyarmati, G. et al. (2005). Assessing a local ensemble Kalman filter: perfect model experiments with the National Centers for Environmental Prediction global model. *Tellus Series A - Dynamic Meteorology and Oceanography.* **57** (4), 528–545.

Talagrand, O. (1997). Assimilation of observations, an introduction. *J. Met. Soc. Jpn.* **75** (1B), 191–209.

Thepaut, J.N., Courtier, P. (1987). Four-dimensional data assimilation using the adjoint of a multi-level primitive equation model. *Q. J. Roy. Meteorol. Soc.* **117**, 1225–1254.

Tikhonov, A., Arsenin, V. (1977). *Solutions of Ill-posed Problems* (Winston, Washington, DC).

Tippett, M.K., Anderson, J.L., Bishop, C.H., Hamill, T.M., Whitaker, J.S. (2003). Ensemble square root filters. *Mon. Wea. Rev.* **131**, 1485–1490.

Todling, R., Cohn, S.E. (1994). Suboptimal schemes for atmospheric data assimilation based on the Kalman filter. *Mon. Wea. Rev.* **122**, 2530–2557.

Todling, R., Cohn, S.E., Sivakumaran, N.S. (1998). Suboptimal schemes for retrospective data assimilation based on the fixed-lag Kalman smoother. *Mon. Wea. Rev.* **126** (8), 2274–2286.

TODLING, R. (1999). *Estimation Theory and Foundations of Atmospheric Data Assimilation* (DAO Office Note 1999-01).

VAN LEEUWEN, P. (2003). A variance-minimizing filter for large-scale applications. *Mon. Wea. Rev.* **131**, 2071–2084.

VERLAAN, M., HEEMINK, A.W. (1995). Reduced rank square root filters for large scale data assimilation problems. *Proc. Second Int. Symp. On Assimilation of Observations in Meteorology and Oceanography* (WMO TD/-No.651, Tokyo, Japan), pp. 247–252.

VERLAAN, M., HEEMINK, A.W. (1997). Tidal flow forecasting using reduced rank square root filters. *Stoch. Hydrol. Hydraul.* **11** (5), 349–368.

VERRON, J., MOLINES, J.M., BLAYO, E. (1992). Assimilation of Geosat data into a quasigeostrophic model of the North Atlantic between 20 and 50 N: preliminary results. *Oceanol. Acta* **15** (5), 575–583.

VIDARD, P.A. (2001). Vers une prise en compte des erreurs modèle en assimilation de données 4D variationnelle. Application à un modèle réaliste d'océan. Thesis (Université Joseph-Fourier, France).

VIDARD, P.A., PIACENTINI, A., LE DIMET, F.-X. (2003). Determination of optimal nudging coefficients. *Tellus A* **55** (1), 1–15.

WANG, Z., NAVON, I.M., LE DIMET, F.-X., ZOU, X. (1992). The second order adjoint analysis: theory and application. *Met. Atmos. Phys.* **50**, 3–20.

WANG, W., WARNER, T.T. (1988). Use of four-dimensional data assimilation by Newtonian relaxation and latent heat forcing to improve a mesoscale-model precipitation forecast: a case study. *Mon. Wea. Rev.* **116**, 2593–2613.

WHITAKER, J.S., HAMILL, T.M. (2002). Ensemble data assimilation without perturbed observations. *Mon. Wea. Rev.* **130**, 1913–1924.

WHITAKER, J.S., COMPO, G.P., WEI, X., HAMILL, T.M. (2004). Reanalysis without radiosondes using ensemble data assimilation. *Mon. Wea. Rev.* **132**, 1190–1200.

WHITAKER, J.S. (2003). Ensemble square root filters. *Mon. Wea. Rev.* **131**, 1485–1490.

WILLCOX, K., GHATTAS, O., VAN BLOEMEN WAANDERS, B., BADER, B. (2005). An optimization framework for goal-oriented, model-based reduction of large-scale systems. Paper presented at 44th IEEE Conference on Decision and Control and European Control Conference, Seville, 2005.

ZHANG, F., SNYDER, C., SUN, J. (2004). Impacts of initial estimate and observation availability on convective-scale data assimilation with an ensemble Kalman filter. *Mon. Wea. Rev.* **132**, 1238–1253.

ZOU, X., NAVON, I.M., LE DIMET, F.-X. (1992). An optimal nudging data assimilation scheme using parameter estimation. *Q. J. Roy. Met. Soc.* **118**, 1163–1186.

ZOU, X., NAVON, I.M., BERGER, M., PHUA, K.H., SCHLICK,T., LE DIMET, F.-X. (1993). Numerical experience with limited Memory Quasi-Newton and Truncated Newton methods. *SIAM J. Optim.* **3** (3), 582–608.

Energetic Consistency and Coupling of the Mean and Covariance Dynamics

Stephen E. Cohn

Global Modeling and Assimilation Office, NASA Goddard Space Flight Center, Greenbelt, Maryland 20771

1. Introduction

The dynamical state of the ocean and atmosphere is taken to be a large-dimensional random vector in a range of large-scale computational applications, including data assimilation, ensemble prediction, sensitivity analysis, and predictability studies. In each of these applications, numerical evolution of the covariance matrix of the random state plays a central role because this matrix is used to quantify uncertainty in the state of the dynamical system. Since atmospheric and ocean dynamics are nonlinear, there is no closed evolution equation for the covariance matrix or for the mean state. Therefore, approximate evolution equations must be used.

This chapter studies theoretical properties of the evolution equations for the mean state and covariance matrix that arise in the second-moment closure approximation (third- and higher-order moment discard). This approximation was introduced by EPSTEIN [1969] in an early effort to introduce a stochastic element into deterministic weather forecasting and was studied further by EPSTEIN and PITCHER [1972], FLEMING [1971a,b], and PITCHER [1977] also in the context of atmospheric predictability. It has since fallen into disuse, with a simpler one being used in current large-scale applications. The theoretical results of this chapter suggest that this approximation should be reconsidered for use in large-scale applications, however, because the second-moment closure equations possess a property of energetic consistency that the approximate equations now in common use do not possess. A number of properties of solutions of the second-moment closure equations that result from this energetic consistency will be established.

Computational Methods for the Atmosphere and the Oceans
Special Volume (Roger M. Temam and Joseph J. Tribbia, Guest Editors) of
HANDBOOK OF NUMERICAL ANALYSIS, VOL. XIV
P.G. Ciarlet (Editor)
Copyright © 2009 Elsevier B.V.
All rights reserved
ISSN 1570-8659
DOI 10.1016/S1570-8659(08)00210-x

Suppose the dynamics of the state $\mathbf{s} \in \mathbb{R}^N$ are given by a system of nonlinear ordinary differential equations,

$$\frac{\mathrm{d}\mathbf{s}}{\mathrm{d}t} + \mathbf{f}(\mathbf{s}, t) = \mathbf{0}, \tag{1.1}$$

where t is time, $\mathbf{f} : \mathcal{S} \times \mathcal{T} \to \mathbb{R}^N$, $\mathcal{S} \subseteq \mathbb{R}^N$ is a state space appropriate for Eq. (1.1), and $\mathcal{T} = [t_0, T]$ is a closed time interval. The initial condition $\mathbf{s}_{t_0}$ is taken to be a random state, $\mathbf{s}_{t_0} \in \mathcal{S}$ with probability one, so that the problem to be solved is the stochastic initial value problem for Eq. (1.1). Technical assumptions on $\mathbf{f}$ are stated in Sections 2 and 5. In addition, it will be assumed that the dynamics are conservative, and a nonlinear transformation is introduced in Section 3 to ensure that the total energy conserved by solutions of Eq. (1.1), stochastic or deterministic, is just $E = \frac{1}{2}\|\mathbf{s}\|^2$, where $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^N$.

A simple sufficient condition, which is natural for conservative dynamics, under which the stochastic initial value problem for Eq. (1.1) is well posed is stated in Section 4. Under this condition, the solution of the stochastic initial value problem defines a second-order stochastic process, i.e., one that has a mean and covariance matrix at each time t in a closed time interval. Furthermore, it follows immediately from conservation of total energy E for this process that

$$\frac{\mathrm{d}(\|\bar{\mathbf{s}}\|^2 + \operatorname{tr}\mathbf{P})}{\mathrm{d}t} = 0, \tag{1.2}$$

where $\bar{\mathbf{s}} = \bar{\mathbf{s}}_t \in \mathcal{S}$ is the mean state of this process, $\mathbf{P} = \mathbf{P}_t \in \mathbb{R}^{N\times N}$ is the covariance matrix of the process, and $\operatorname{tr}\mathbf{P}$ is the trace, or sum of the diagonal elements, of $\mathbf{P}$. This means that the uncertainty in the random state $\mathbf{s}$, as measured by the total variance $V = \operatorname{tr}\mathbf{P}$, can increase (decrease) only as a result of extracting energy from (inserting energy into) the mean state $\bar{\mathbf{s}}$, with the change in total variance balanced exactly by twice the change in total energy $\frac{1}{2}\|\bar{\mathbf{s}}\|^2$ of the mean state. The mean state and covariance matrix cannot be calculated from Eq. (1.1) without approximation, however, unless $\mathbf{f}$ is linear in $\mathbf{s}$, and one would like to develop approximate evolution equations for $\bar{\mathbf{s}}$ and $\mathbf{P}$ that in the nonlinear case at least preserve this basic conservation property. A closed system of ordinary differential equations for the mean and covariance matrix whose solutions satisfy Eq. (1.2) will be said to be energetically consistent, after Fleming [1971a, p. 872].

The second-moment closure equations for approximate evolution of $\bar{\mathbf{s}}$ and $\mathbf{P}$ are the closed, nonlinearly coupled differential equations

$$\frac{\mathrm{d}\bar{\mathbf{s}}}{\mathrm{d}t} + \mathbf{f}(\bar{\mathbf{s}}, t) + \tfrac{1}{2}\sum_j \sum_k \frac{\partial^2 \mathbf{f}(\bar{\mathbf{s}}, t)}{\partial s_j \partial s_k} P_{jk} = \mathbf{0}, \tag{1.3}$$

$$\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}t} + \mathbf{F}(\bar{\mathbf{s}}, t)\mathbf{P} + \mathbf{P}\mathbf{F}^T(\bar{\mathbf{s}}, t) = \mathbf{0}, \tag{1.4}$$

where P_{jk} is the $(j, k)^{th}$ element of the covariance matrix $\mathbf{P}$ whose evolution is described by Eq. (1.4), $\mathbf{F} = \partial\mathbf{f}/\partial\mathbf{s}$ is the Jacobian matrix of $\mathbf{f}$, and the superscript T denotes transposition. The evolution of $\mathbf{P}$ in Eq. (1.4) depends on that of the mean state $\bar{\mathbf{s}}$ given by

Eq. (1.3), through the dependence of the Jacobian matrix on the mean state, and the evolution of the mean state also depends on that of the covariance matrix, through the double-summation term in Eq. (1.3). If $\mathbf{f}(\mathbf{s}, t)$ is linear in $\mathbf{s}$, then the second partial derivatives of $\mathbf{f}$ with respect to the state variables all vanish so that the double-summation term vanishes. Hence this term is called the nonlinear coupling term. Equations (1.3) and (1.4) are to be solved together for initial conditions $\bar{\mathbf{s}}_{t_0}$ and $\mathbf{P}_{t_0}$, with $\bar{\mathbf{s}}_{t_0}$ being the mean of the random initial condition $\mathbf{s}_{t_0}$ for Eq. (1.1) and $\mathbf{P}_{t_0}$ being the covariance matrix of $\mathbf{s}_{t_0}$.

Conditions under which this initial value problem is well-posed on a closed time interval are given in Section 5, where a stochastic process having the solution $(\bar{\mathbf{s}}, \mathbf{P})$ of the initial value problem as its mean and covariance matrix is also defined. It is shown in Section 6 that the solution satisfies Eq. (1.2). Thus the second-moment closure equations are energetically consistent. For quadratically nonlinear $\mathbf{f}$, energetic consistency of the second-moment closure equations was noted by EPSTEIN [1969] and studied in detail by FLEMING [1971a] who also established energetic consistency of the third-moment closure equations for quadratically nonlinear $\mathbf{f}$. The energetic consistency result established in the present chapter holds for general $\mathbf{f}$.

The derivation of Eq. (1.2) given in Section 6 for the second-moment closure equations shows that the exchange of energy between the mean state and the stochastic perturbations, which balances exactly, occurs solely through the nonlinear coupling term in Eq. (1.3) and the symmetric part $\mathbf{F}^s = \frac{1}{2}(\mathbf{F} + \mathbf{F}^T)$ of the Jacobian matrix in Eq. (1.4). In other words, rewriting Eq. (1.4) as

$$\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}t} + \mathbf{F}^a(\bar{\mathbf{s}}, t)\mathbf{P} - \mathbf{P}\mathbf{F}^a(\bar{\mathbf{s}}, t) + \mathbf{F}^s(\bar{\mathbf{s}}, t)\mathbf{P} + \mathbf{P}\mathbf{F}^s(\bar{\mathbf{s}}, t) = \mathbf{0}, \tag{1.5}$$

where $\mathbf{F}^a = \frac{1}{2}(\mathbf{F} - \mathbf{F}^T)$ is the antisymmetric (skew-symmetric) part of $\mathbf{F}$, one has immediately that

$$\frac{\mathrm{d}\,\mathrm{tr}\,\mathbf{P}}{\mathrm{d}t} + 2\,\mathrm{tr}\,\mathbf{F}^s(\bar{\mathbf{s}}, t)\mathbf{P} = 0,$$

and it is shown in Section 6 that Eq. (1.3) gives

$$\frac{\mathrm{d}\|\bar{\mathbf{s}}\|^2}{\mathrm{d}t} - 2\,\mathrm{tr}\,\mathbf{F}^s(\bar{\mathbf{s}}, t)\mathbf{P} = 0,$$

with contribution only from the nonlinear coupling term, not from the term $\mathbf{f}(\bar{\mathbf{s}}, t)$ in Eq. (1.3). Equation (1.2) then follows.

In case $\mathbf{f}(\mathbf{s}, t)$ is linear in $\mathbf{s}$, then not only does the nonlinear coupling term vanish, but for conservative dynamics, $\mathbf{F}^s$ vanishes as well so that Eqs. (1.3) and (1.4) become simply

$$\frac{\mathrm{d}\bar{\mathbf{s}}}{\mathrm{d}t} + \mathbf{f}(\bar{\mathbf{s}}, t) = \mathbf{0}, \tag{1.6}$$

$$\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}t} + \mathbf{F}^a\mathbf{P} - \mathbf{P}\mathbf{F}^a = \mathbf{0}, \tag{1.7}$$

with $\mathbf{F}^a$ independent of $\bar{\mathbf{s}}$ by linearity. Thus the effect of nonlinearity, to second-moment closure, is to introduce the nonlinear coupling term and the terms $\mathbf{F}^s(\bar{\mathbf{s}}, t)\mathbf{P} + \mathbf{P}\mathbf{F}^s(\bar{\mathbf{s}}, t)$, and these terms together are in energetic balance. Nonlinearity also introduces dependence of $\mathbf{F}^a$ on the mean state, but Eq. (1.7) is energetically neutral, satisfying $\mathrm{d}\,\mathrm{tr}\,\mathbf{P}/\mathrm{d}t = 0$, regardless of any dependence of $\mathbf{F}^a$ on the mean state.

The approximation now widely used in large-scale atmospheric and oceanic applications is to retain Eq. (1.4) as it stands but to neglect the nonlinear coupling term. This is the approximation made for instance in four-dimensional variational data assimilation (COURTIER and TALAGRAND [1987], TALAGRAND and COURTIER [1987], THÉPAUT, COURTIER, BELAUD and LEMAÎTRE [1996]) and in a variety of algorithms based on singular vector calculations (e.g., BUIZZA and PALMER [1995], MOLTENI, BUIZZA, PALMER and PETROLIAGIS [1996], MOORE and KLEEMAN [1997]). This approximation is convenient for computations because the mean state can then be evolved independently of the covariance matrix. Neglecting the nonlinear coupling term, however, destroys energetic consistency. Furthermore, the nonlinear coupling term and the terms in the covariance evolution equation all have formally the same order of magnitude since all are linear in the covariance matrix. Moreover, in the sense of contribution to the total variance, the terms retained in the covariance evolution equation that arise from nonlinearity, $\mathbf{F}^s\mathbf{P} + \mathbf{P}\mathbf{F}^s$, have precisely the same magnitude, and opposite sign, as that of the nonlinear coupling term which is neglected.

The role of $\mathbf{F}^s$ in the energetic coupling of the second-moment closure equations is studied further in Sections 7 and 8. It is shown in Section 8 that if $\mathbf{f}$ is genuinely nonlinear, as defined there, then $\mathbf{F}^s$ has at least one positive and one negative eigenvalue. This means that when the dynamics are genuinely nonlinear, there is always a direction in state space in which uncertainty decays, as well as a direction in which uncertainty grows. One implication is that if $\mathbf{f}$ is genuinely nonlinear, then neglect of the nonlinear coupling term can lead either to increase or decrease of the perceived uncertainty.

When the nonlinear coupling term is neglected, Eq. (1.4) is a linear equation to be solved once the mean state has been calculated, and so its solution can be expressed in the form

$$\mathbf{P}_t = \mathbf{M}_{t,t_0}\mathbf{P}_{t_0}\mathbf{M}_{t,t_0}^T, \tag{1.8}$$

where $\mathbf{M}_{t,t_0}$ is the fundamental matrix (alternatively, solution operator or tangent linear propagator) of the perturbation dynamics corresponding to Eq. (1.1) linearized about the mean state. This expression is particularly convenient for singular value calculations in large-scale applications.

When the nonlinear coupling term is retained, the solution of Eq. (1.4) can still be expressed in the form of Eq. (1.8), but for a matrix $\mathbf{M}_{t,t_0}$ that itself depends on the covariance matrix. In Section 9, Eq. (1.2) is used to establish simple time-independent upper bounds for $\mathrm{tr}\,\mathbf{P}_t/\mathrm{tr}\,\mathbf{P}_{t_0}$, which hold also for the covariance matrix of the original stochastic process. When the nonlinear coupling term is neglected, the largest singular value of $\mathbf{M}_{t,t_0}$ is the least upper bound for $\mathrm{tr}\,\mathbf{P}_t/\mathrm{tr}\,\mathbf{P}_{t_0}$, but there is no time-independent upper bound since $\mathrm{tr}\,\mathbf{P}_t$ can grow unboundedly when Eq. (1.2) does not hold.

A minimum requirement for solutions of Eqs. (1.3) and (1.4) to approximate the mean and covariance matrix of solutions of the stochastic initial value problem for Eq. (1.1) is for $\bar{\mathbf{s}}_{t_0}$ and $\mathbf{P}_{t_0}$ to be the mean and covariance matrix of some random state, $\mathbf{s}_{t_0} \in \mathcal{S}$ with probability one. Because ocean and atmospheric dynamics contain state variables that are constrained to be positive, such as mass and temperature variables, or to satisfy other constraints, this minimum requirement implies that $\mathbf{P}_{t_0}$ cannot be chosen independently of $\bar{\mathbf{s}}_{t_0}$ in general. Thus the initial value problem for Eqs. (1.3) and (1.4) must be posed with some care. Toward this end, the deterministic initial value problem for Eq. (1.1) is reviewed briefly in Section 2, and the stochastic initial value problem for Eq. (1.1) is described in some detail in Section 4. The initial value problem for Eqs. (1.3) and (1.4) is posed in Section 5, and restrictions on the initial covariance matrix are illustrated there using a spatially discretized version of the one-dimensional shallow-water equations.

Brief concluding remarks are given in Section 10.

2. The deterministic initial value problem

Suppose that the evolution of a real vector $\mathbf{s} \in \mathbb{R}^N$ is governed by a nonlinear system of ordinary differential equations,

$$\frac{\mathrm{d}\mathbf{s}}{\mathrm{d}t} + \mathbf{f}(\mathbf{s}) = \mathbf{0}, \tag{2.1}$$

where $\mathbf{f}(\mathbf{s}) = \mathbf{f}(\mathbf{s}, t)$ may depend explicitly on time t. Suppose also that $\mathbf{f} : \mathcal{S} \times \mathcal{T} \to \mathbb{R}^N$, where $\mathcal{S} \subseteq \mathbb{R}^N$ is a given convex open set, i.e., an open set in $\mathbb{R}^N$ such that the line segment between any two points in $\mathcal{S}$ lies entirely in $\mathcal{S}$, and where $\mathcal{T} = [t_0, T]$ is a given closed time interval; $T - t_0$ is finite but may be arbitrarily large. The open set $\mathcal{S}$ is called the state space, an element $\mathbf{s} \in \mathcal{S}$ is called a state, or state vector, and the N components of a state are the state variables. If the actual system under consideration is complex, Eq. (2.1) is obtained by separation into real and imaginary parts.

This section provides a brief review of the deterministic initial value problem for Eq. (2.1), in which one is supposed to find, for each initial state $\mathbf{s}_{t_0} \in \mathcal{S}$, a solution $\mathbf{s} = \mathbf{s}(t) \in \mathcal{S}$ that satisfies $\mathbf{s}(t_0) = \mathbf{s}_{t_0}$. Sufficiently strong hypotheses will be imposed on $\mathbf{f}$ to guarantee existence and uniqueness of solutions on a (possibly short) half-open time interval $\mathcal{T}_* = [t_0, T_*) \subset \mathcal{T}$, with T_* depending in general on $\mathbf{s}_{t_0}$, and also to guarantee that the solution is in the space $C^1(\mathcal{T}_*)$ of functions with one continuous time derivative on $\mathcal{T}_*$. The corresponding stochastic initial value problem, in which $\mathbf{s}_{t_0}$ will depend on a probability variable, is considered in Section 4. It should be noted here that Assumption (2.2) below is stronger than necessary for establishing uniqueness of solutions of Eq. (2.1), but that this assumption will be required later, in Section 5, to ensure that solutions of the covariance evolution equation have one continuous time derivative.

ASSUMPTION 2.1. $\mathbf{f} \in C(\mathcal{S} \times \mathcal{T})$, *the space of continuous functions on* $\mathcal{S} \times \mathcal{T}$.

ASSUMPTION 2.2. $\mathbf{F} = \mathbf{F}(\mathbf{s}, t) \in C(\mathcal{S} \times \mathcal{T})$, *where* $\mathbf{F} = \partial \mathbf{f}/\partial \mathbf{s}$ *denotes the* $N \times N$ *Jacobian matrix of* $\mathbf{f}$, *whose* (j, k)*th element is given by*

$$F_{jk}(\mathbf{s}, t) = \frac{\partial f_j(\mathbf{s}, t)}{\partial s_k}.$$

That is, $F_{jk} \in C(\mathcal{S} \times \mathcal{T})$ *for* $j, k = 1, \ldots, N$.

Assumption (2.1) guarantees that for each $\mathbf{s}_{t_0} \in \mathcal{S}$, there exists a time interval $\mathcal{T}_* = [t_0, T_*)$, with $T_* = T_*(\mathbf{s}_{t_0}) \leq T$, and at least one solution $\mathbf{s}(t) \in \mathcal{S}$ for all $t \in \mathcal{T}_*$, such that $\mathbf{s}(t_0) = \mathbf{s}_{t_0}$. It also guarantees that every such solution is in $C^1(\mathcal{T}_*)$. Furthermore, existence of a solution ceases only if it becomes unbounded or if it hits the boundary of $\mathcal{S}$ (e.g., CODDINGTON and LEVINSON [1955, Chapter 1, Theorem 4.1]), where $\mathbf{f}$ may not even be defined. Assumption (2.2), along with the fact that $\mathcal{S}$ was taken to be convex, implies that $\mathbf{f}$ satisfies on $\mathcal{S} \times \mathcal{T}$, a Lipschitz condition in $\mathbf{s}$, uniformly in t, and this in turn guarantees that there is at most one solution on any time interval.

If the state space $\mathcal{S}$ is all of $\mathbb{R}^N$, and if $\mathbf{f}$ is bounded on $\mathcal{S} \times \mathcal{T}$, then all solutions are bounded and for each $\mathbf{s}_{t_0} \in \mathcal{S}$, there exists a unique solution $\mathbf{s}(t) \in \mathcal{S}$ over the full time interval $\mathcal{T}$, with $\mathbf{s}(t_0) = \mathbf{s}_{t_0}$, and this solution is in $C^1(\mathcal{T})$. If the state space is a bounded set in $\mathbb{R}^N$, then the same is true provided the dynamics (Eq. (2.1)) are such that no solution can ever hit the boundary of $\mathcal{S}$ by time T. It is often the case that the physical problem at hand dictates that the state space cannot be all of $\mathbb{R}^N$, but that there is a choice of state space for which all solutions exist in $\mathcal{S}$ over the full time interval $\mathcal{T}$ as discussed further in Section 3.

Hereafter, the unique solution $\mathbf{s}(t) \in \mathcal{S}$ of Eq. (2.1) on some time interval $\mathcal{T}_* = \mathcal{T}_*(\mathbf{s}_{t_0}) \subset \mathcal{T}$ such that $\mathbf{s}(t_0) = \mathbf{s}_{t_0}$, or on all of $\mathcal{T}$ in case it exists for all time $t \in \mathcal{T}$, will be denoted by $\mathbf{s}_t$. The continuous path through state space traced by $\mathbf{s}_t$ as time progresses is called the trajectory corresponding to $\mathbf{s}_{t_0}$. The bold letter $\mathbf{s}$ without a subscript will usually denote an arbitrary point in the state space or in $\mathbb{R}^N$.

An implication of Assumption (2.2) beyond uniqueness of solutions, which is important for application to the covariance evolution equation, is that solutions depend continuously on parameters such as initial conditions. Regarding each trajectory $\mathbf{s}_t$ as a function of its initial point $\mathbf{s}_{t_0}$, one finds that the $N \times N$ matrix $\mathbf{M} = \mathbf{M}(t) = \partial \mathbf{s}_t/\partial \mathbf{s}_{t_0}$, whose (j, k)th element is given by

$$M_{jk} = \frac{\partial (\mathbf{s}_t)_j}{\partial (\mathbf{s}_{t_0})_k}$$

satisfies the simple *linear* equation

$$\frac{\mathrm{d}\mathbf{M}}{\mathrm{d}t} + \mathbf{F}(\mathbf{s}_t, t)\mathbf{M} = \mathbf{0}, \tag{2.2}$$

with initial condition $\mathbf{M} = \mathbf{I}$, the $N \times N$ identity matrix. Although this equation is linear, it is coupled with Eq. (2.1) through the dependence of the Jacobian matrix $\mathbf{F}$ on the trajectory $\mathbf{s}_t$. Equation (2.2) is obtained by differentiating Eq. (2.1) with respect to $\mathbf{s}_{t_0}$

and applying the chain rule. That there exists a unique solution $\mathbf{M} \in \mathbb{R}^{N\times N}$ of Eq. (2.2), in fact, with one continuous time derivative, for as long as the trajectory $\mathbf{s}_t$ exists in $\mathcal{S}$, is guaranteed by Assumption (2.2) and the linearity of Eq. (2.2). The dependence of the solution $\mathbf{M}$ on the trajectory $\mathbf{s}_t$ can be expressed fully as $\mathbf{M} = \mathbf{M}(t) = \mathbf{M}(t; \mathbf{s}_{t_0})$ since $\mathbf{s}_{t_0}$ determines the trajectory $\mathbf{s}_t$.

Now let $\mathbf{M}_{t,t_0}$ denote the unique solution of Eq. (2.2) in $\mathbb{R}^{N\times N}$, over a finite time interval for which $\mathbf{s}_t$ exists in $\mathcal{S}$, that corresponds to the initial condition $\mathbf{M}_{t_0,t_0} = \mathbf{I}$. From the preceding discussion it follows that for each point $\mathbf{q}_{t_0} \in \mathbb{R}^N$, the linear equation

$$\frac{\mathrm{d}\mathbf{q}}{\mathrm{d}t} + \mathbf{F}(\mathbf{s}_t, t)\mathbf{q} = \mathbf{0} \tag{2.3}$$

has unique solution

$$\mathbf{q}_t = \mathbf{M}_{t,t_0}\mathbf{q}_{t_0} \in \mathbb{R}^N, \tag{2.4}$$

and that $\mathbf{q}_t$ has one continuous time derivative, for as long as $\mathbf{s}_t$ exists in $\mathcal{S}$. The linear equation (Eq. (2.3)) is called the (deterministic) perturbation equation associated with the original nonlinear dynamics (Eq. (2.1)). The matrix $\mathbf{M}_{t,t_0} = \mathbf{M}_{t,t_0}(\mathbf{s}_{t_0})$, which according to Eq. (2.4) expresses the solution of the perturbation equation directly in terms of its initial condition, is called the fundamental matrix, or solution operator, of the perturbation equation. The analogue of Eq. (2.3) for the stochastic initial-value problem, which gives the evolution of stochastic initial perturbations under second-moment closure, is derived in Section 5 along with the corresponding mean and covariance evolution equations. Energetic consistency of the mean and covariance evolution equations is demonstrated in Section 6. The fundamental matrix of the stochastic perturbation equation has special properties due to this energetic consistency, which are described in Section 9.

It is well known (e.g., CODDINGTON and LEVINSON [1955, Chapter 1, Theorem 7.3]) that Eq. (2.2) can be solved explicitly for the determinant of $\mathbf{M}_{t,t_0}$:

$$\det \mathbf{M}_{t,t_0} = \exp\left[-\int_{t_0}^{t} \operatorname{tr} \mathbf{F}(\mathbf{s}_\tau, \tau)\, \mathrm{d}\tau\right], \tag{2.5}$$

where tr $\mathbf{F}$ denotes the trace, or sum of the diagonal elements, of $\mathbf{F}$. Define the symmetric and antisymmetric (skew-symmetric) parts of $\mathbf{F}$ as $\mathbf{F}^s = \frac{1}{2}(\mathbf{F} + \mathbf{F}^T)$ and $\mathbf{F}^a = \frac{1}{2}(\mathbf{F} - \mathbf{F}^T)$, respectively, where the superscript T denotes transposition, so that $\mathbf{F} = \mathbf{F}^s + \mathbf{F}^a$ and $\mathbf{F}^T = \mathbf{F}^s - \mathbf{F}^a$. Since the diagonal elements of a real skew-symmetric matrix are all zero, $\mathbf{F}$ can be replaced in Eq. (2.5) by the symmetric matrix $\mathbf{F}^s$. This is one indication of the important role that the symmetric part of the Jacobian matrix plays in the dependence of trajectories on their initial points. In Section 6 it will be seen that the exchange of energy between the mean state and the stochastic perturbations occurs solely through $\mathbf{F}^s$. This role of $\mathbf{F}^s$ in the energetic coupling of the mean and covariance evolution equations is examined in detail in Sections 7 and 8.

Another immediate consequence of Assumption (2.2) is that if $\partial \mathbf{f}/\partial t \in C(\mathcal{S} \times \mathcal{T})$, then in fact $\mathbf{s}_t$ has two continuous time derivatives, not just one. This follows by differentiating Eq. (2.1) once with respect to time and applying the chain rule:

$$\frac{\mathrm{d}^2\mathbf{s}}{\mathrm{d}t^2} - \mathbf{F}(\mathbf{s}, t)\mathbf{f}(\mathbf{s}, t) + \frac{\partial \mathbf{f}(\mathbf{s}, t)}{\partial t} = \mathbf{0}.$$

Also, if $\partial \mathbf{f}/\partial t \in C(\mathcal{S} \times \mathcal{T})$, then Assumption (2.2) implies Assumption (2.1). It will not be assumed that $\partial \mathbf{f}/\partial t \in C(\mathcal{S} \times \mathcal{T})$ although this often does hold. For instance, in many problems $\mathbf{f}$ does not depend explicitly on time.

3. Conservation of total energy

Suppose now that one is given a nonlinear system

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} + \mathbf{g}(\mathbf{x}) = \mathbf{0}, \tag{3.1}$$

with state space $\mathcal{X}$ and with $\mathbf{g} : \mathcal{X} \times \mathcal{T} \to \mathbb{R}^N$ satisfying the hypotheses (Assumptions (2.1) and (2.2)) imposed earlier on $\mathbf{f}$. Suppose also that there is a known function $\mathbf{s} : \mathcal{X} \to \mathbb{R}^N$ such that the "total energy"

$$E(t) = \tfrac{1}{2}\mathbf{s}^T(\mathbf{x}_t)\mathbf{s}(\mathbf{x}_t) \tag{3.2}$$

is conserved by the solutions of Eq. (3.1), i.e.,

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \mathbf{s}^T(\mathbf{x}_t)\frac{\mathrm{d}\mathbf{s}(\mathbf{x}_t)}{\mathrm{d}t} = 0, \tag{3.3}$$

for each trajectory $\mathbf{x}_t$. Suppose finally that $\mathbf{s} = \mathbf{s}(\mathbf{x})$ defines a continuously differentiable coordinate transformation between the state space $\mathcal{X}$ and the range $\mathbf{s}(\mathcal{X})$, and denote by $\mathbf{A}(\mathbf{x}) = \partial \mathbf{s}/\partial \mathbf{x}$ the Jacobian matrix of this transformation.

Then for $\mathbf{s} = \mathbf{s}(\mathbf{x}_t)$, one has

$$\frac{\mathrm{d}\mathbf{s}}{\mathrm{d}t} = \mathbf{A}(\mathbf{x})\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t},$$

and so from Eq. (3.1) it follows that $\mathbf{s}(\mathbf{x}_t)$ satisfies the nonlinear system (Eq. (2.1)) for $\mathbf{f}$ given by

$$\mathbf{f}(\mathbf{s}, t) = \mathbf{A}\left(\mathbf{x}(\mathbf{s})\right)\mathbf{g}\left(\mathbf{x}(\mathbf{s}), t\right),$$

where $\mathbf{x} = \mathbf{x}(\mathbf{s})$ is the inverse transformation of $\mathbf{s} = \mathbf{s}(\mathbf{x})$. Furthermore, for Eq. (2.1), the total energy becomes simply $E = \frac{1}{2}\mathbf{s}^T\mathbf{s}$, and the statement (Eq. (3.3)) that the total energy is conserved becomes simply $\mathbf{s}^T\mathbf{f} = 0$. Thus, rather than considering Eq. (3.1) directly, for some general energy expression, in this chapter Eq. (2.1) is considered instead under the simple hypothesis that the total energy $E = \frac{1}{2}\mathbf{s}^T\mathbf{s}$ is conserved:

ASSUMPTION 3.1. *For all* $\mathbf{s} \in \mathcal{S}$ *and* $t \in \mathcal{T}$,

$$\mathbf{s}^T \mathbf{f}(\mathbf{s}, t) = 0. \tag{3.4}$$

Transforming a general energy expression out of the problem in this way simplifies substantially the study of energetics of the second-moment closure equations. The cost is potentially a complicated expression for $\mathbf{f}$.

Assumption (3.1) says that the trajectories $\mathbf{s}_t$ of Eq. (2.1) satisfy

$$\mathbf{s}_t^T \mathbf{s}_t = \mathbf{s}_{t_0}^T \mathbf{s}_{t_0},$$

for as long as $t \in \mathcal{T}$ and the trajectory exists in $\mathcal{S}$. In geometrical terms, this means simply that each trajectory remains on the hyperspherical surface $\mathbf{s}^T\mathbf{s} = 2E$ in $\mathbb{R}^N$ on which it originates at time t_0. A trajectory can cease to exist only if this surface intersects the boundary of $\mathcal{S}$, because all solutions with $E < \infty$ initally are bounded for all time.

Since the change of coordinates to "energy variables" $\mathbf{s}$ is central to the results of this chapter, it is worthwhile to consider how it works in simple examples. Consider first the quadratically nonlinear system

$$\frac{\mathrm{d}u}{\mathrm{d}t} + cu = 0,$$
$$\frac{\mathrm{d}\phi}{\mathrm{d}t} - 2cu^2 = 0,$$

with c a nonzero constant. Assumptions (2.1) and (2.2) are satisfied with the state space taken to be all of $\mathbb{R}^2$ and with arbitrarily large final time T for the interval $\mathcal{T} = [t_0, T]$. In addition, for $E = \frac{1}{2}(u^2 + \phi)$, one has $\mathrm{d}E/\mathrm{d}t = 0$. This "energy" expression suggests that one consider also the state space $\mathcal{X}$ consisting of the upper half plane $\phi > 0$ in (u, ϕ)-space so that $E > 0$. The change of variables $s_1 = u$, $s_2 = \phi^{1/2}$ is a continuously differentiable coordinate transformation from $\mathcal{X}$ onto itself, and it yields Eq. (2.1) with

$$\mathbf{f} = c\,\frac{s_1}{s_2}\begin{bmatrix} s_2 \\ -s_1 \end{bmatrix},$$

which is singular along the s_1-axis $s_2 = 0$. Assumptions (2.1), (2.2), and (3.1) are satisfied by this $\mathbf{f}$, with $\mathcal{S}$ being the upper half plane $s_2 > 0$ in (s_1, s_2)-space and again with arbitrarily large final time T for the interval $\mathcal{T}$. The "total energy" $E = \frac{1}{2}\mathbf{s}^T\mathbf{s}$ is conserved on each trajectory of Eq. (2.1), and every trajectory exists either until it hits the s_1-axis, which is the boundary of $\mathcal{S}$, or for all $t \in \mathcal{T}$ if it never hits the s_1-axis.

The solution of Eq. (2.1) in this example, for each $\mathbf{s}_{t_0} \in \mathcal{S}$, is just

$$s_1(t) = s_1(t_0)e^{-c(t-t_0)},$$
$$s_2(t) = \left[2E - s_1^2(t)\right]^{1/2}.$$

There is one stationary solution which is $s_1 = 0$, $s_2 = (2E)^{1/2}$. If $c > 0$ and $s_1(t_0) \neq 0$, then $|s_1(t)|$ decreases monotonically toward zero and $s_2(t)$ increases monotonically

toward $(2E)^{1/2}$: all solutions approach the stationary one and no trajectory can ever hit the s_1-axis. Thus if $c > 0$, then the trajectory corresponding to arbitrary $\mathbf{s}_{t_0} \in \mathcal{S}$ exists in $\mathcal{S}$ over arbitrarily long time intervals. If $c < 0$ on the other hand, and if $s_1(t_0) \neq 0$, then $|s_1(t)|$ increases monotonically toward $(2E)^{1/2}$, $s_2(t)$ decreases monotonically toward zero, and the solution ceases to exist at time T_*,

$$T_* = t_0 - \frac{1}{2c} \log \left[2 \left(1 + \frac{s_2^2(t_0)}{s_1^2(t_0)} \right) \right],$$

when $s_1^2(T_*) = 2E$, $s_2(T_*) = 0$, and $\mathbf{f}$ is no longer defined. However, for the problem in the original variables u and ϕ, as posed on all of $\mathbb{R}^2$, nothing bad happens at this time T_* or at any other finite time. In fact, all that happens is that the trajectory continues downward into the lower half plane $\phi < 0$ in (u, ϕ)-space along the parabola $\phi_t = 2E - u_t^2$. Thus it is possible for longtime existence of solutions to be lost through the transformation to energy variables.

Fortunately, for spatially discretized versions of the main first-order hyperbolic partial differential equations of atmospheric and ocean dynamics, nothing need be lost in the transformation to energy variables since all that is required in the transformation is to take the square root of mass (gravitational potential energy) and temperature (internal energy) variables that are required on physical grounds to remain positive. As a simple example, consider the shallow water equations in one space dimension, taken to be periodic. These are usually written as the momentum equation

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + \frac{\partial \phi}{\partial x} = 0$$

and the continuity equation

$$\frac{\partial \phi}{\partial t} + \frac{\partial u \phi}{\partial x} = 0,$$

where u is the speed, ϕ is the geopotential, and $x \in [0, L]$ is the space variable. From the differential equations, it follows that solutions satisfy the energy equation

$$\frac{\partial (\phi u^2 + \phi^2)}{\partial t} + \frac{\partial (\phi u^3 + 2\phi^2 u)}{\partial x} = 0,$$

which on the periodic domain $[0, L]$ implies conservation of the total energy

$$E = \tfrac{1}{2} \int_0^L \left(\phi u^2 + \phi^2 \right) \mathrm{d}x.$$

Reasonable spatial discretization of the variables u and ϕ gives rise to a system of ordinary differential equations (Eq. (3.1)) that conserves a discretized version (Eq. 3.2)) of this energy integral.

The change of variable from u to $\alpha = \phi^{1/2} u$ transforms the momentum equation to

$$\frac{\partial \alpha}{\partial t} + u \frac{\partial \alpha}{\partial x} + \tfrac{1}{2} \frac{\partial u}{\partial x} \alpha + \phi^{1/2} \frac{\partial \phi}{\partial x} = 0,$$

and substituting $u = \phi^{-1/2}\alpha$ here and in the continuity equation gives the shallow water system in terms of the variables α and ϕ alone. Reasonable spatial discretization of the transformed system, say with $\alpha_n(t) = \alpha(x_n, t)$, $\phi_n(t) = \phi(x_n, t)$, and $x_n = nL/N$, for $n = 1, \ldots, N$, gives a system (Eq. (2.1)) for $\mathbf{s} = [\alpha_1, \ldots, \alpha_N, \phi_1, \ldots, \phi_N]^T$ that conserves the discretized total energy $E = \frac{1}{2}\mathbf{s}^T\mathbf{s}\frac{L}{N}$. One can see that because of the substitution $u = \phi^{-1/2}\alpha$, the function $\mathbf{f}$ will usually be singular on each of the hyperplanes $\phi_n = 0$, $n = 1, \ldots, N$, in $\mathbb{R}^{2N}$. Taking $\mathcal{S}$ to be the convex open set

$$\mathcal{S} = \{\mathbf{s} = [\alpha_1, \ldots, \alpha_N, \phi_1, \ldots, \phi_N]^T : \phi_n > 0, n = 1, \ldots, N\}$$

does not automatically give rise to an initial value problem whose trajectories exist in this $\mathcal{S}$ over arbitrarily long time intervals. Rather, for longtime existence it is necessary to ensure that the spatial discretization also preserves positivity (e.g,. Lin, Chao, Sud and Walker [1994], Lin and Rood [1996]) of the geopotential, so that the boundary of this $\mathcal{S}$ is never hit.

For the one-dimensional shallow water equations, additional state spaces can be defined in a natural way. It follows from the shallow water equations that the characteristic speeds

$$c_+ = u + \phi^{1/2} = \phi^{-1/2}(\alpha + \phi),$$
$$c_- = u - \phi^{1/2} = \phi^{-1/2}(\alpha - \phi),$$

satisfy the coupled advection equations

$$\frac{\partial c_+}{\partial t} + (\tfrac{3}{4}c_+ + \tfrac{1}{4}c_-)\frac{\partial c_+}{\partial x} = 0,$$
$$\frac{\partial c_-}{\partial t} + (\tfrac{1}{4}c_+ + \tfrac{3}{4}c_-)\frac{\partial c_-}{\partial x} = 0.$$

Therefore, if initially $c_+ > 0$ and $c_- < 0$, equivalently $\phi > |\alpha|$, for all $x \in [0, L]$, then this remains true for all time, even though all spatially periodic solutions except the constant one develop discontinuous derivatives in finite time (Lax [1973, Theorem 6.1]). Positivity-preserving (of c_+ and $-c_-$), energy-conserving spatial discretizations of the shallow-water equations have the property that trajectories with initial points in the convex open set

$$\mathcal{S}_0 = \{\mathbf{s} = [\alpha_1, \ldots, \alpha_N, \phi_1, \ldots, \phi_N]^T : \phi_n > |\alpha_n|, n = 1, \ldots, N\}$$

exist in $\mathcal{S}_0$ over arbitrarily long time intervals, although such discretizations must satisfy additional constraints for the trajectories to converge under mesh refinement to solutions of the shallow water equations with discontinuous derivatives. More generally, if there is a constant $c > 0$ such that initially $c_+ > c$ and $c_- < -c$ (equivalently $\phi^{1/2} > c + |u|$, or $\phi^{1/2} > c/2 + (c^2/4 + |\alpha|)^{1/2}$) for all $x \in [0, L]$, the coupled advection equations imply that this also remains true for all time. Therefore, for appropriate spatial discretizations, trajectories with initial points in the convex open set

$$\mathcal{S}_c = \{\mathbf{s} = [\alpha_1, \ldots, \alpha_N, \phi_1, \ldots, \phi_N]^T : \phi_n^{1/2} > c/2 + (c^2/4 + |\alpha_n|)^{1/2}, n = 1, \ldots, N\}$$

exist in $\mathcal{S}_c$ over arbitrarily long time intervals. On the state space $\mathcal{S}_0$, the geopotential can approach zero arbitrarily closely, wherever $u = 0$. On the state space $\mathcal{S}_c$ with $c > 0$, however, the geopotential is guaranteed to be strictly positive everywhere, $\phi > (c + |u|)^2 \geq c^2$, for all $t \in \mathcal{T}$, and therefore the total energy is also strictly positive,

$$E = \tfrac{1}{2} \int_0^L \left(\alpha^2 + \phi^2 \right) \mathrm{d}x \geq \tfrac{1}{2} \int_0^L \phi^2 \, \mathrm{d}x > \tfrac{1}{2} L c^4. \tag{3.5}$$

4. The stochastic initial value problem

Now let Ω be the sample space of a complete probability space and denote by $\mathcal{E}$, the expectation operator on the probability space. A random vector $\mathbf{s} \in \mathbb{R}^N$ is a vector function $\mathbf{s} = \mathbf{s}(\omega)$ of the probability variable $\omega \in \Omega$, i.e., $\mathbf{s} : \Omega \to \mathbb{R}^N$, which is measurable with respect to the probability measure. For fixed ω, $\mathbf{s}(\omega)$ is called a realization of the random vector $\mathbf{s}$.

Consider a second-order random vector, i.e., a random vector $\mathbf{s} \in \mathbb{R}^N$ such that $\mathcal{E}\mathbf{s}^T\mathbf{s} < \infty$. By the Schwarz inequality, one has

$$(\mathcal{E}\mathbf{s})^T(\mathcal{E}\mathbf{s}) \leq (\mathcal{E}|\mathbf{s}|)^T(\mathcal{E}|\mathbf{s}|) \leq \mathcal{E}\mathbf{s}^T\mathbf{s},$$

and therefore the mean $\bar{\mathbf{s}} = \mathcal{E}\mathbf{s}$ exists and is finite. Denote the departure from the mean by $\mathbf{s}' = \mathbf{s} - \bar{\mathbf{s}}$. Since $\mathcal{E}\mathbf{s}' = \mathbf{0}$, one has

$$\mathcal{E}\mathbf{s}^T\mathbf{s} = \bar{\mathbf{s}}^T\bar{\mathbf{s}} + \mathcal{E}\mathbf{s}'^T\mathbf{s}',$$

and therefore, the total variance V defined by

$$V = \mathcal{E}\mathbf{s}'^T\mathbf{s}'$$

is also finite, $V \leq \mathcal{E}\mathbf{s}^T\mathbf{s}$. Since the total variance is finite, it follows by a further application of the Schwarz inequality that the elements of the $N \times N$ covariance matrix $\mathbf{P}$ defined by

$$\mathbf{P} = \mathcal{E}\mathbf{s}'\mathbf{s}'^T$$

also exist and are finite. The covariance matrix is, by definition, symmetric and positive semidefinite. Also, the total variance is just the trace of the covariance matrix, $V = \mathrm{tr}\,\mathbf{P}$. Associating randomness with uncertainty, one can say that the total variance of a second-order random vector is a scalar measure of its uncertainty.

For a random vector $\mathbf{s} \in \mathbb{R}^N$, one writes $\mathbf{s} \in \mathcal{S}$ wp1 (with probability one) if $\mathbf{s}(\omega) \in \mathcal{S}$ for all $\omega \in \Omega$, except possibly for an ω set of probability measure zero. A random vector $\mathbf{s} \in \mathcal{S}$ wp1 will be called a random state. Every second-order random state $\mathbf{s}$ has mean $\bar{\mathbf{s}} \in \mathcal{S}$ because $\mathcal{S}$ was taken to be convex.

In the stochastic initial value problem treated here, Eq. (2.1) is considered for second-order random initial states $\mathbf{s}_{t_0}$, i.e., random initial states with

$$\mathcal{E}E_{t_0} = \tfrac{1}{2}\mathcal{E}\mathbf{s}_{t_0}^T\mathbf{s}_{t_0} < \infty.$$

Denote by $\mathbf{s}_t(\omega)$ the trajectory corresponding to a realization $\mathbf{s}_{t_0}(\omega) \in \mathcal{S}$. From Assumption (3.1), it follows that $\mathbf{s}_t(\omega)$ satisfies the total energy conservation equation

$$\mathbf{s}_t^T(\omega)\mathbf{s}_t(\omega) = \mathbf{s}_{t_0}^T(\omega)\mathbf{s}_{t_0}(\omega) \tag{4.1}$$

for each $\omega \in \Omega$ such that $\mathbf{s}_{t_0}(\omega) \in \mathcal{S}$. If the deterministic initial value problem is such that all trajectories starting in $\mathcal{S}$ exist in $\mathcal{S}$ over arbitrarily long time intervals $\mathcal{T} = [t_0, T]$, for instance if $\mathcal{S} = \mathbb{R}^N$, then the realizations $\mathbf{s}_t(\omega)$ with $\mathbf{s}_{t_0}(\omega) \in \mathcal{S}$ define a random vector $\mathbf{s}_t \in \mathcal{S}$ wp1, for all $t \in \mathcal{T}$, and by taking expectations in Eq. (4.1), it follows that $\mathbf{s}_t$ is also a second-order random vector, for all $t \in \mathcal{T}$. In this case, therefore, $\mathbf{s}_t$ has finite mean $\bar{\mathbf{s}}_t \in \mathcal{S}$ and covariance matrix $\mathbf{P}_t \in \mathbb{R}^{N\times N}$, for all $t \in \mathcal{T}$. The family of random vectors $\{\mathbf{s}_t : t \in \mathcal{T}\}$ is called a (second-order) stochastic process. The trajectories $\mathbf{s}_t(\omega)$, for fixed $\omega \in \Omega$, are called the sample functions, or sample paths, of the process. By Assumption (2.1), each sample path with $\mathbf{s}_{t_0}(\omega) \in \mathcal{S}$ is in $C^1(\mathcal{T})$. In other words, the sample paths are in $C^1(\mathcal{T})$ wp1.

In case the deterministic initial value problem is such that there are trajectories starting in $\mathcal{S}$ that hit the boundary of $\mathcal{S}$ before time T, one can restrict the problem to initial states in some prescribed subset $\mathcal{S}' \subset \mathcal{S}$, not necessarily convex or open, but whose boundary nowhere touches the boundary of $\mathcal{S}$. Since the trajectories are continuous paths in state space, traversed at finite speed, there is then a time $T' > t_0$ such that the trajectory $\mathbf{s}_t(\omega)$ corresponding to an arbitrary point $\mathbf{s}_{t_0}(\omega) \in \mathcal{S}'$ exists in $\mathcal{S}$ for all time t in the closed interval $\mathcal{T}' = [t_0, T']$, with T' independent of $\mathbf{s}_{t_0}(\omega)$. Thus in this case, the stochastic initial value problem is restricted to second-order random initial vectors $\mathbf{s}_{t_0} \in \mathcal{S}'$ wp1. The realizations $\mathbf{s}_t(\omega)$ with $\mathbf{s}_{t_0}(\omega) \in \mathcal{S}'$ still satisfy Eq. (4.1) but now only for $t \in \mathcal{T}'$. Thus they define a second-order stochastic process $\{\mathbf{s}_t : t \in \mathcal{T}'\}$, with $\mathbf{s}_t \in \mathcal{S}$ wp1, and with finite mean $\bar{\mathbf{s}}_t \in \mathcal{S}$ and covariance matrix $\mathbf{P}_t \in \mathbb{R}^{N\times N}$, for all $t \in \mathcal{T}'$. The sample paths are in $C^1(\mathcal{T}')$ wp1.

Thus two cases have been distinguished for the stochastic initial value problem. In the first one, Eq. (2.1) defines a second-order stochastic process $\{\mathbf{s}_t : t \in \mathcal{T}\}$ for each second-order random initial state. In the second case, Eq. (2.1) defines a second-order stochastic process $\{\mathbf{s}_t : t \in \mathcal{T}'\}$ for each second-order random initial vector $\mathbf{s}_{t_0}$ such that $\mathbf{s}_{t_0} \in \mathcal{S}'$ wp1. Denote by V_t the total variance

$$V_t = \mathcal{E}\mathbf{s}_t'^T\mathbf{s}_t' = \operatorname{tr}\mathbf{P}_t,$$

for $t \in \mathcal{T}$ in the first case and for $t \in \mathcal{T}'$ in the second. Taking expectations in Eq. (4.1) gives

$$\bar{\mathbf{s}}_t^T\bar{\mathbf{s}}_t + V_t = \bar{\mathbf{s}}_{t_0}^T\bar{\mathbf{s}}_{t_0} + V_{t_0}, \tag{4.2}$$

for $t \in \mathcal{T}$ in the first case and for $t \in \mathcal{T}'$ in the second. The total variance V_t is a scalar measure of the uncertainty present in the random state $\mathbf{s}_t$ due to uncertainty in the random initial state $\mathbf{s}_{t_0}$.

Equation (4.2) says that the uncertainty in solutions of Eq. (2.1) due to uncertainty in the initial condition, as measured by the total variance, can increase (decrease) only as a result of extracting energy from (inserting energy into) the mean state $\bar{\mathbf{s}}$, with the

change in total variance balanced exactly by twice the change in total energy $\frac{1}{2}\bar{\mathbf{s}}^T\bar{\mathbf{s}}$ of the mean state. This is purely a consequence of conservation of total energy for the nonlinear dynamics, and it holds regardless of any assumptions one might make on the form of the probability distribution of the random initial state $\mathbf{s}_{t_0}$, apart from existence of its first two moments. In fact, Eq. (4.2) holds even if no moments beyond the first two exist at any time. It is simply a statement about second-order stochastic processes whose realizations satisfy Eq. (4.1) with probability one.

Equation (4.2) implies in particular the bound

$$V_t \leq \bar{\mathbf{s}}_{t_0}^T \bar{\mathbf{s}}_{t_0} + V_{t_0}, \tag{4.3}$$

valid for all time $t \in \mathcal{T}$ in the first case, $t \in \mathcal{T}'$ in the second, with equality holding at some particular time τ if, and only if, all the energy of the mean state has been extracted at that time, $\bar{\mathbf{s}}_\tau = \mathbf{0}$. This is a simple, general statement of the maximum level of uncertainty that can occur in solutions of Eq. (2.1). There is no implication, however, that the uncertainty actually saturates, i.e., becomes or approaches a constant in time, either at the level $\bar{\mathbf{s}}_{t_0}^T \bar{\mathbf{s}}_{t_0} + V_{t_0}$ or at any other level. Saturation of uncertainty will be discussed briefly at the end of Section 8.

For many dynamical systems, the state space is naturally bounded away from the origin of coordinates $\mathbf{s} = \mathbf{0}$ in $\mathbb{R}^N$ so that in fact the mean state cannot vanish at any time and the bound (Eq. (4.3)) can be improved upon. This is the case when there are mass-like and/or temperature-like state variables, for instance, that are constrained by the physics of the problem to be bounded from below by positive constants. If every point $\mathbf{s}$ in the state space $\mathcal{S}$ of the problem satisfies an inequality $\mathbf{s}^T\mathbf{s} > 2E_{\min} \geq 0$, then since $\bar{\mathbf{s}}_t \in \mathcal{S}$, one has $\bar{\mathbf{s}}_t^T \bar{\mathbf{s}}_t > 2E_{\min} \geq 0$; here, $2E_{\min}$ is just the minimum Euclidean distance from the origin to the boundary of $\mathcal{S}$. In this case, Eq. (4.2) implies the stronger bound

$$V_t < \bar{\mathbf{s}}_{t_0}^T \bar{\mathbf{s}}_{t_0} + V_{t_0} - 2E_{\min}. \tag{4.4}$$

For example, in Section 3, it was shown that for appropriately discretized versions of the one-dimensional shallow water equations with state space $\mathcal{S}_c$, $c \geq 0$, the geopotential satisfies $\phi > c^2$ for all $t \in \mathcal{T}$, and since the discretized total energy was $E = \frac{1}{2}\mathbf{s}^T\mathbf{s}\frac{L}{N}$, Eq. (3.5) gives $\bar{\mathbf{s}}_t^T \bar{\mathbf{s}}_t > Nc^4$. Thus for appropriate discrete shallow water dynamics on $\mathcal{S}_c$, one has

$$V_t < \bar{\mathbf{s}}_{t_0}^T \bar{\mathbf{s}}_{t_0} + V_{t_0} - Nc^4, \tag{4.5}$$

for all $t \in \mathcal{T}$.

Since Eq. (4.2) holds without any assumptions on moments beyond the first two, even without their existence, one expects to be able to find approximate evolution equations for just the mean and covariance matrix such that their solutions are guaranteed to satisfy Eq. (4.2) and therefore also the bounds (Eqs. (4.3) and (4.4)). A closed system of differential equations for the mean and covariance matrix that has this property will be said to be energetically consistent. The second-moment closure equations derived in Section 5 constitute a closed, nonlinearly coupled system of differential equations for the

mean and covariance matrix. In Section 6, it will be shown that the nonlinear coupling term in the second-moment closure equation for the mean makes them energetically consistent. First, the exact equations for the mean and covariance matrix, which are not closed unless the dynamics are linear, will be derived.

If E represents a physical total energy, then for both the deterministic and stochastic initial value problems there seems little reason to consider initial points in $\mathbb{R}^N$ with total energy greater than or equal to some prescribed, fixed amount, say $E \geq E_{\max}$. Such points will be eliminated from consideration by making the following simple hypothesis on $\mathcal{S}$:

ASSUMPTION 4.1. $\mathcal{S} \subseteq \mathcal{S}_{\max}$, *where* $\mathcal{S}_{\max}$ *denotes the interior of the hypersphere* $\mathbf{s}^T\mathbf{s} = 2E_{\max}$ *in* $\mathbb{R}^N$*:*

$$\mathcal{S}_{\max} = \{\mathbf{s} \in \mathbb{R}^N : \mathbf{s}^T\mathbf{s} < 2E_{\max}\}.$$

This assumption imposes no restriction on existence of solutions since no trajectory starting in $\mathcal{S} \subseteq \mathcal{S}_{\max}$ can ever hit the boundary of $\mathcal{S}_{\max}$ at any time, by conservation of total energy. It is, however, a restriction on the initial probability distribution, and therefore on the probability distribution at any time, because it implies that if $\mathbf{s}$ is a random state, then $\mathbf{s}^T\mathbf{s} < 2E_{\max}$ wp1. In particular, no random state is (multivariate) normally distributed, and the marginal distributions of a random state also cannot be normal. This may seem a significant restriction. However, it is suggested by the physical problem at hand. Moreover, as discussed already, for atmospheric and ocean dynamics there are also usually state variables that are constrained to be positive, in fact often bounded from below by positive constants, and these cannot be normally distributed.

An immediate consequence of Assumption (4.1) is that every random state $\mathbf{s}$ is a second-order random vector, in fact, $\mathcal{E}\mathbf{s}^T\mathbf{s} < 2E_{\max}$. Also, since $\mathbf{f} \in C(\mathcal{S} \times \mathcal{T})$ by Assumption (2.1), it follows from Assumption (4.1) that $\mathbf{f}^T\mathbf{f} < \infty$ on $\mathcal{S} \times \mathcal{T}$. Therefore, if $\mathbf{s}$ is a random state, then $\mathbf{f}(\mathbf{s})$ is a second-order random vector, $\mathcal{E}\mathbf{f}^T(\mathbf{s})\mathbf{f}(\mathbf{s}) < \infty$.

Again let $\mathbf{s}_t$ (either for $t \in \mathcal{T}$ or $t \in \mathcal{T}'$, depending on the case) denote the solution of the stochastic initial value problem for Eq. (2.1). Then $\mathcal{E}|\mathbf{f}(\mathbf{s}_t, t)| < \infty$ since $\mathbf{f}(\mathbf{s}_t, t)$ is a second-order random vector, and therefore

$$\int_{t_0}^{t} \mathcal{E}|\mathbf{f}(\mathbf{s}_\tau, \tau)|\, \mathrm{d}\tau < \infty.$$

It follows (e.g., DOOB [1953, Theorem (2.7), p. 62]) that

$$\mathcal{E}\int_{t_0}^{t} \mathbf{f}(\mathbf{s}_\tau, \tau)\, \mathrm{d}\tau = \int_{t_0}^{t} \mathcal{E}\mathbf{f}(\mathbf{s}_\tau, \tau)\, \mathrm{d}\tau, \tag{4.6}$$

where both integrals exist and are finite. Now write Eq. (2.1) as the integral equation

$$\mathbf{s}_t(\omega) = \mathbf{s}_{t_0}(\omega) - \int_{t_0}^{t} \mathbf{f}(\mathbf{s}_\tau(\omega), \tau)\, \mathrm{d}\tau. \tag{4.7}$$

Taking expectations and using Eq. (4.6) gives

$$\bar{\mathbf{s}}_t = \bar{\mathbf{s}}_{t_0} - \int_{t_0}^{t} \bar{\mathbf{f}}(\mathbf{s}_\tau, \tau)\, \mathrm{d}\tau, \tag{4.8}$$

where $\bar{\mathbf{f}}(\mathbf{s}_\tau, \tau) = \mathcal{E}\mathbf{f}(\mathbf{s}_\tau, \tau)$, which shows that the mean state $\bar{\mathbf{s}}_t$ is a continuous function of time. Differentiating in Eq. (4.8) gives an equation for the mean state in differential form,

$$\frac{\mathrm{d}\bar{\mathbf{s}}_t}{\mathrm{d}t} + \bar{\mathbf{f}}(\mathbf{s}_t, t) = \mathbf{0}, \tag{4.9}$$

which is satisfied almost everywhere in $\mathcal{T}$ in the first case, i.e., except possibly on a subset of $\mathcal{T}$ of Lebesgue measure zero, and almost everywhere in $\mathcal{T}'$ in the second. This exact equation for the mean state is not a differential equation unless $\mathbf{f}(\mathbf{s}, t)$ is linear in $\mathbf{s}$, in which case $\bar{\mathbf{f}}(\mathbf{s}_t, t) = \mathbf{f}(\bar{\mathbf{s}}_t, t)$. Also, $\bar{\mathbf{f}}(\mathbf{s}_t, t)$ is not necessarily continuous in time, and so $\bar{\mathbf{s}}_t$ is not necessarily continuously differentiable. Comparing Eqs. (2.1) and (4.9) shows that the commutation

$$\mathcal{E}\frac{\mathrm{d}\mathbf{s}_t}{\mathrm{d}t} = \frac{\mathrm{d}\bar{\mathbf{s}}_t}{\mathrm{d}t} \tag{4.10}$$

holds, almost everywhere in time.

To derive the equation for the covariance matrix, first subtract Eq. (4.8) from Eq. (4.7) to obtain

$$\mathbf{s}'_t(\omega) = \mathbf{s}'_{t_0}(\omega) - \int_{t_0}^{t} \mathbf{f}'(\mathbf{s}_\tau(\omega), \tau)\, \mathrm{d}\tau, \tag{4.11}$$

where $\mathbf{s}' = \mathbf{s} - \bar{\mathbf{s}}$ and $\mathbf{f}' = \mathbf{f} - \bar{\mathbf{f}}$. Postmultiplying this equation by the transpose of itself, then taking expectations, and then exchanging the order of expectation and integration, which again can be justified by Assumption (4.1), gives

$$\begin{aligned}\mathbf{P}_t = \mathbf{P}_{t_0} &- \int_{t_0}^{t} \mathcal{E}\left[\mathbf{f}'(\mathbf{s}_\tau, \tau)\mathbf{s}'^{T}_{t_0}\right] \mathrm{d}\tau - \int_{t_0}^{t} \mathcal{E}\left[\mathbf{s}'_{t_0}\mathbf{f}'^{T}(\mathbf{s}_\tau, \tau)\right] \mathrm{d}\tau \\ &+ \int_{t_0}^{t}\int_{t_0}^{t} \mathcal{E}\left[\mathbf{f}'(\mathbf{s}_{\tau_1}, \tau_1)\mathbf{f}'^{T}(\mathbf{s}_{\tau_2}, \tau_2)\right] \mathrm{d}\tau_1\, \mathrm{d}\tau_2,\end{aligned}$$

which shows that the covariance matrix $\mathbf{P}_t$ is a continuous function of time. Differentiating this result and using Eq. (4.11) gives an equation for the covariance matrix in differential form,

$$\frac{\mathrm{d}\mathbf{P}_t}{\mathrm{d}t} + \mathcal{E}\left[\mathbf{f}'(\mathbf{s}_t, t)\mathbf{s}'^{T}_t\right] + \mathcal{E}\left[\mathbf{f}'(\mathbf{s}_t, t)\mathbf{s}'^{T}_t\right]^T = 0, \tag{4.12}$$

satisfied almost everywhere in $\mathcal{T}$ or in $\mathcal{T}'$. Again, this is not a differential equation unless $\mathbf{f}$ is linear, and the elements of $\mathbf{P}_t$ are not necessarily continuously differentiable. Equation (4.12) can be used to show that the commutation

$$\mathcal{E}\frac{\mathrm{d}\mathbf{s}_t'\mathbf{s}_t'^T}{\mathrm{d}t} = \frac{d\mathbf{P}_t}{\mathrm{d}t} \tag{4.13}$$

holds, almost everywhere in time.

Equations (4.9) and (4.12) are usually derived under an hypothesis much weaker than Assumption (4.1) which, like Assumption (4.1), also implies that $\mathbf{f}(\mathbf{s}_t, t)$ is a second-order random vector (e.g., Doob [1953, p. 277, hypothesis H_2], Jazwinski [1970, p. 105, hypothesis H_1]). However, Assumption (4.1) makes sense for conservative dynamics, and it greatly simplifies the derivation of Eqs. (4.9) and (4.12). It also makes for little difference between the formulation of the stochastic initial value problem and that of the deterministic initial value problem since it makes every random state a second-order random vector. All that has been necessary for the stochastic problem was to restrict initial random vectors to lie in some set $\mathcal{S}'$ (wp1) contained wholly in the interior of $\mathcal{S}$ in case solutions of the deterministic initial value problem do not exist over arbitrarily long time intervals, to ensure that all the sample paths exist for some minimum amount of time $T' - t_0 > 0$ wp1.

5. The second-moment closure equations

Two final hypotheses on $\mathbf{f}$ will now be made:

Assumption 5.1. *$\partial^2\mathbf{f}(\mathbf{s}, t)/\partial s_j \partial s_k \in C(\mathcal{S} \times \mathcal{T})$, for $j, k = 1, \ldots, N$.*

Assumption 5.2. *The second partial derivatives of $\mathbf{f}$ are Lipschitz continuous in $\mathbf{s}$ on $\mathcal{S} \times \mathcal{T}$, uniformly in t. That is, there are constants K_{jk} such that*

$$\left\| \frac{\partial^2\mathbf{f}(\mathbf{s}_1, t)}{\partial s_j \partial s_k} - \frac{\partial^2\mathbf{f}(\mathbf{s}_2, t)}{\partial s_j \partial s_k} \right\| \leq K_{jk} \|\mathbf{s}_1 - \mathbf{s}_2\|,$$

where $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^N$, for each $\mathbf{s}_1, \mathbf{s}_2 \in \mathcal{S}$, for each $t \in \mathcal{T}$ and for $j, k = 1, \ldots, N$.

Let $\mathbf{s}$ be a random state, i.e., $\mathbf{s} \in \mathcal{S}$ wp1. Since $\bar{\mathbf{s}} = \mathcal{E}\mathbf{s} \in \mathcal{S}$, it follows from Assumption (5.1) that $\mathbf{f}(\mathbf{s}) = \mathbf{f}(\mathbf{s}, t)$ has a Taylor expansion about $\bar{\mathbf{s}}$ up to second order,

$$\mathbf{f}(\mathbf{s}) = \mathbf{f}(\bar{\mathbf{s}}) + \mathbf{F}(\bar{\mathbf{s}})\mathbf{s}' + \tfrac{1}{2}\sum_j \sum_k \frac{\partial^2\mathbf{f}(\bar{\mathbf{s}})}{\partial s_j \, \partial s_k} s_j' s_k' + \mathbf{g}(\mathbf{s}', \bar{\mathbf{s}}), \tag{5.1}$$

where $\mathbf{F} = \mathbf{F}(\mathbf{s}) = \mathbf{F}(\mathbf{s}, t)$ is the Jacobian matrix of $\mathbf{f}$ introduced in Section 2, and where s_j' is the jth element of the random vector $\mathbf{s}' = \mathbf{s} - \bar{\mathbf{s}}$. It follows from Assumption (5.2) that the remainder term $\mathbf{g}$ is $\mathcal{O}((\mathbf{s}')^3)$ for each fixed $\bar{\mathbf{s}} \in \mathcal{S}$. It followed from Assumption (4.1) that $\mathbf{s}$ is a second-order random vector and along with Assumption (2.1) that $\mathbf{f}(\mathbf{s})$ is also a

second-order random vector so that $\overline{\mathbf{f}} = \mathcal{E}\mathbf{f}$ exists and is finite. Taking expectations in the Taylor expansion then shows that $\overline{\mathbf{g}} = \mathcal{E}\mathbf{g}$ also exists and is finite, and that

$$\overline{\mathbf{f}} = \mathbf{f}(\overline{\mathbf{s}}) + \tfrac{1}{2} \sum_j \sum_k \frac{\partial^2 \mathbf{f}(\overline{\mathbf{s}})}{\partial s_j \partial s_k} P_{jk} + \overline{\mathbf{g}}, \tag{5.2}$$

where $P_{jk} = \mathcal{E} s'_j s'_k$ is the (j, k)th element of the covariance matrix $\mathbf{P} = \mathcal{E}\mathbf{s}'\mathbf{s}'^T$.

The *mean equation* is obtained by substituting Eq. (5.2) into Eq. (4.9) and neglecting the remainder term $\overline{\mathbf{g}}$ to yield

$$\frac{\mathrm{d}\overline{\mathbf{s}}}{\mathrm{d}t} + \mathbf{f}(\overline{\mathbf{s}}) + \tfrac{1}{2} \sum_j \sum_k \frac{\partial^2 \mathbf{f}(\overline{\mathbf{s}})}{\partial s_j \partial s_k} P_{jk} = \mathbf{0}, \tag{5.3}$$

where the time subscripts have been omitted because the mean equation is only an approximate equation for the evolution of $\overline{\mathbf{s}}_t$. The *covariance evolution equation* is obtained by using Eqs. (5.1) and (5.2) to approximate $\mathbf{f}' = \mathbf{f}'(\mathbf{s}) = \mathbf{f}(\mathbf{s}) - \overline{\mathbf{f}}$ by $\mathbf{F}(\overline{\mathbf{s}})\mathbf{s}'$, and substituting this into Eq. (4.12) to yield

$$\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}t} + \mathbf{F}(\overline{\mathbf{s}})\mathbf{P} + \mathbf{P}\mathbf{F}^T(\overline{\mathbf{s}}) = \mathbf{0}, \tag{5.4}$$

where the time subscripts have again been omitted because the covariance evolution equation is only an approximate equation for the evolution of $\mathbf{P}_t$.

Equations (5.3) and (5.4) together constitute the second-moment closure equations for the evolution of the mean state and covariance matrix. In case $\mathbf{f} = \mathbf{f}(\mathbf{s}, t)$ is linear in its first argument, the second-moment closure equations decouple and they are exact; the Jacobian matrix $\mathbf{F} = \mathbf{F}(\mathbf{s}, t)$ is independent of its first argument so that the covariance evolution equation decouples from the mean equation, and the second partial derivatives of $\mathbf{f}(\mathbf{s}, t)$ with respect to their first argument all vanish so that the mean equation likewise decouples from the covariance evolution equation. In case $\mathbf{f}$ is nonlinear, the mean and covariance evolution equations are fully coupled in both directions.

Let $\mathcal{P} \subseteq \mathbb{R}^{N \times N}$ be an open set, and suppose that $\overline{\mathbf{s}}_{t_0} \in \mathcal{S}$ and $\mathbf{P}_{t_0} \in \mathcal{P}$. Then there is a half-open time interval $\mathcal{T}_0 = [t_0, T_0)$, with T_0 depending in general on $\overline{\mathbf{s}}_{t_0}$ and $\mathbf{P}_{t_0}$, such that there exists a unique solution $(\overline{\mathbf{s}}(t) \in \mathcal{S}, \mathbf{P}(t) \in \mathcal{P})$ of this coupled system for all $t \in \mathcal{T}_0$, satisfying initial condition $(\overline{\mathbf{s}}(t_0), \mathbf{P}(t_0)) = \left(\overline{\mathbf{s}}_{t_0}, \mathbf{P}_{t_0}\right)$. Assumptions (2.1), (2.2), and (5.1) guarantee that there exists at least one solution on $\mathcal{T}_0$ satisfying the initial condition, and also that every solution on $\mathcal{T}_0$ is in $C^1(\mathcal{T}_0)$. Assumptions (2.2) and (5.2) guarantee that there exists at most one solution on $\mathcal{T}_0$ satisfying the initial condition. As in Section 4, if $\overline{\mathbf{s}}_{t_0}$ and $\mathbf{P}_{t_0}$ are restricted to lie in some subsets $\overline{\mathcal{S}} \subset \mathcal{S}$ and $\overline{\mathcal{P}} \subset \mathcal{P}$, respectively, whose boundaries nowhere touch the boundaries of $\mathcal{S}$ and $\mathcal{P}$, then existence and uniqueness of solutions are assured over some closed time interval $\overline{\mathcal{T}} = [t_0, \overline{T}]$ with $\overline{T} > t_0$. Thus if one is interested in solving Eqs. (5.3) and (5.4) with some particular $\overline{\mathbf{s}}_{t_0} \in \mathcal{S}$ given, as is often the case, then existence and uniqueness on some closed time interval $\overline{\mathcal{T}} = [t_0, \overline{T}]$ are assured for any $\mathbf{P}_{t_0} \in \overline{\mathcal{P}}$, simply by taking $\overline{\mathcal{S}} = \{\overline{\mathbf{s}}_{t_0}\}$.

In fact, Eqs. (5.3) and (5.4) are supposed to be solved for initial condition $(\overline{\mathbf{s}}_{t_0}, \mathbf{P}_{t_0})$ being the mean state and covariance matrix of a random state $\mathbf{s}_{t_0}$, if the equations are to

approximate the evolution of the mean and covariance matrix of the random state $\mathbf{s}_t$. In particular, $\mathbf{P}_{t_0}$ is supposed to be symmetric positive semidefinite. Moreover, this means that $\mathbf{P}_{t_0}$ cannot be specified independently of $\bar{\mathbf{s}}_{t_0}$. For instance, under Assumption (4.1) one has the simple restriction

$$\bar{\mathbf{s}}_{t_0}^T \bar{\mathbf{s}}_{t_0} + \operatorname{tr} \mathbf{P}_{t_0} < 2E_{\max}. \tag{5.5}$$

The particular geometry of $\mathcal{S}$ for a given problem can restrict $\mathbf{P}_{t_0}$ in terms of $\bar{\mathbf{s}}_{t_0}$ much more severely than this, with the restriction being the strongest when $\bar{\mathbf{s}}_{t_0}$ is near the boundary of $\mathcal{S}$ as discussed further below. Since $\mathbf{P}_{t_0}$ is not supposed to be specified independently of $\bar{\mathbf{s}}_{t_0}$, it is simplest to pose the initial value problem for a given fixed $\bar{\mathbf{s}}_{t_0} \in \mathcal{S}$, then to define the set $\mathcal{P} = \mathcal{P}(\bar{\mathbf{s}}_{t_0})$ of symmetric positive semidefinite matrices over which $\mathbf{P}_{t_0}$ is allowed to vary due to the particular geometry of $\mathcal{S}$, and finally to define a subset $\overline{\mathcal{P}} = \overline{\mathcal{P}}(\bar{\mathbf{s}}_{t_0}) \subset \mathcal{P}(\bar{\mathbf{s}}_{t_0})$ whose boundary nowhere touches that of $\mathcal{P}(\bar{\mathbf{s}}_{t_0})$. In this way, one has existence and uniqueness on a closed time interval $\overline{\mathcal{T}} = [t_0, \overline{T}]$, for the given $\bar{\mathbf{s}}_{t_0} \in \mathcal{S}$ and for all $\mathbf{P}_{t_0} \in \overline{\mathcal{P}}(\bar{\mathbf{s}}_{t_0})$, along with the assurance that every $\mathbf{P}_{t_0} \in \overline{\mathcal{P}}(\bar{\mathbf{s}}_{t_0})$ is the covariance matrix of a random state $\mathbf{s}_{t_0}$ with the given mean $\bar{\mathbf{s}}_{t_0} \in \mathcal{S}$. Then one can conclude that a physically meaningful problem has been posed. If one is given both $\bar{\mathbf{s}}_{t_0} \in \mathcal{S}$ and any particular $\mathbf{P}_{t_0} \in \mathcal{P}(\bar{\mathbf{s}}_{t_0})$, then existence and uniqueness are guaranteed on a closed time interval $\overline{\mathcal{T}} = [t_0, \overline{T}]$ by taking $\overline{\mathcal{P}} = \{\mathbf{P}_{t_0}\}$.

As a simple example of how the requirement that $\mathbf{P}_{t_0}$ is the covariance matrix of a random state with mean $\bar{\mathbf{s}}_{t_0}$ makes $\mathbf{P}_{t_0}$ depend on $\bar{\mathbf{s}}_{t_0}$, consider again the shallow water system described at the end of Section 3. There it was shown that on the state space $\mathcal{S}_c$, solutions exist over arbitrarily long time intervals for appropriate spatial discretizations. In terms of the variable u rather than α, the space $\mathcal{S}_c$ is described by the inequality

$$u_n^2 + 2c|u_n| + c^2 < \phi_n,$$

for $n = 1, \ldots, N$. Taking expectations gives

$$\mathcal{E}\left(u_n'\right)^2 + \overline{u}_n^2 + 2c\mathcal{E}|u_n| + c^2 < \overline{\phi}_n,$$

and since $|\overline{u}_n| < \mathcal{E}|u_n|$, this implies that

$$\mathcal{E}\left(u_n'\right)^2 < \overline{\phi}_n - (|\overline{u}_n| + c)^2. \tag{5.6}$$

Since the mean state is in $\mathcal{S}_c$, one has $(|\overline{u}_n| + c)^2 < \overline{\phi}_n$ so that the right-hand side of inequality (Eq. 5.6)) is indeed positive. This inequality is a restriction on the variance $\mathcal{E}\left(u_n'\right)^2$ in terms of $|\overline{u}_n|$ and $\overline{\phi}_n$ for every random state on $\mathcal{S}_c$, and the restriction becomes stronger as the mean state approaches the boundary of $\mathcal{S}_c$, i.e., as the right-hand side of inequality (Eq. (5.6)) becomes small. For given $\overline{\phi}_n$, $n = 1, \ldots, N$, the restriction is mildest when the mean state is a state of rest, $\overline{u}_n = 0$ for $n = 1, \ldots, N$.

Returning to the general problem, suppose now that the set $\mathcal{P} = \mathcal{P}(\bar{\mathbf{s}}_{t_0})$ restricted by positive semidefiniteness and the geometry of $\mathcal{S}$ has been defined for each $\bar{\mathbf{s}}_{t_0} \in \mathcal{S}$. From inequality (Eq. (5.5)) and the fact that $\mathcal{S}$ is an open set, it follows that $\mathcal{P}(\bar{\mathbf{s}}_{t_0})$ is

an open set, for each $\bar{\mathbf{s}}_{t_0} \in \mathcal{S}$. A simple choice for the set $\overline{\mathcal{P}} = \overline{\mathcal{P}}(\bar{\mathbf{s}}_{t_0}) \subset \mathcal{P}(\bar{\mathbf{s}}_{t_0})$ of initial covariance matrices $\mathbf{P}_{t_0}$, is the set:

$$\overline{\mathcal{P}} = \mathcal{P}_\mu = \{\mathbf{P} \in \mathcal{P} : \tfrac{1}{\mu}\mathbf{P} \in \mathcal{P}\}, \tag{5.7}$$

for any μ with $0 < \mu < 1$, which will be made for the sake of definiteness in Section 9, where a specific choice of $\overline{\mathcal{P}}$ is needed. Since $\mathcal{P}(\bar{\mathbf{s}}_{t_0})$ is an open set, $\mathcal{P}_\mu(\bar{\mathbf{s}}_{t_0})$ is also an open set. Note that both $\mathcal{P}(\bar{\mathbf{s}}_{t_0})$ and $\mathcal{P}_\mu(\bar{\mathbf{s}}_{t_0})$ contain the origin in $\mathbb{R}^{N\times N}$, for every $\bar{\mathbf{s}}_{t_0} \in \mathcal{S}$, since $\mathcal{S}$ is an open set.

In case the solution of the deterministic initial value problem for Eq. (2.1) exists over arbitrarily long time intervals, one would like to know whether or not the solution of the initial value problem for Eqs. (5.3) and (5.4) also exists over arbitrarily long time intervals, for each $\bar{\mathbf{s}}_{t_0} \in \mathcal{S}$ and $\mathbf{P}_{t_0} \in \mathcal{P}(\bar{\mathbf{s}}_{t_0})$. This important question is not addressed in the present chapter. Also, the question of how well the solution of Eqs. (5.3) and (5.4) approximates the mean and covariance matrix of the stochastic process defined in Section 4 is not considered here.

For given $\bar{\mathbf{s}}_{t_0} \in \mathcal{S}$ and all $\mathbf{P}_{t_0} \in \overline{\mathcal{P}}(\bar{\mathbf{s}}_{t_0})$, there exists on some closed time interval $\overline{\mathcal{T}} = [t_0, \overline{T}]$ a unique solution of Eqs. (5.3) and (5.4) satisfying the initial conditions. From this point onwards, $\bar{\mathbf{s}}_t$ and $\mathbf{P}_t$ (defined originally in Section 4) are redefined as this solution. The random vector $\mathbf{s}'_t = \mathbf{s}'_t(\omega)$ is also redefined, for each $\omega \in \Omega$ such that $\mathbf{s}_{t_0}(\omega) = \bar{\mathbf{s}}_{t_0} + \mathbf{s}'_{t_0}(\omega) \in \mathcal{S}$, as the unique solution on $\overline{\mathcal{T}}$ of the *(stochastic) perturbation equation*

$$\frac{\mathrm{d}\mathbf{s}'}{\mathrm{d}t} + \mathbf{F}(\bar{\mathbf{s}})\mathbf{s}' = \mathbf{0}, \tag{5.8}$$

corresponding to given initial condition $\mathbf{s}'_{t_0}(\omega) = \mathbf{s}_{t_0}(\omega) - \bar{\mathbf{s}}_{t_0}$. That there exists a unique solution on $\overline{\mathcal{T}}$ for each such ω is guaranteed by Assumption (2.2) since the perturbation equation is linear. Finally the stochastic process $\{\mathbf{s}_t : t \in \overline{\mathcal{T}}\}$ is defined as the one whose sample paths are given by $\mathbf{s}_t(\omega) = \bar{\mathbf{s}}_t + \mathbf{s}'_t(\omega)$. The mean state of this second-order process is $\bar{\mathbf{s}}_t$, and $\mathbf{P}_t$ is its covariance matrix. Note that for this process,

$$\frac{\mathrm{d}\mathbf{s}_t}{\mathrm{d}t} = \frac{\mathrm{d}\bar{\mathbf{s}}_t}{\mathrm{d}t} + \frac{\mathrm{d}\mathbf{s}'_t}{\mathrm{d}t} \tag{5.9}$$

is continuous on $\overline{\mathcal{T}}$ wp1 since $\mathrm{d}\bar{\mathbf{s}}_t/\mathrm{d}t$ is continuous on $\overline{\mathcal{T}}$ and $\mathrm{d}\mathbf{s}'_t/\mathrm{d}t$ is continuous on $\overline{\mathcal{T}}$ wp1.

The stochastic perturbation equation has the form of the deterministic perturbation equation (Eq. (2.3)) for the original nonlinear dynamics. When $\mathbf{f}$ is nonlinear, the evolution of $\mathbf{s}'_t$ according to the stochastic perturbation equation depends through the Jacobian matrix on the mean state, and therefore also on the covariance matrix, since Eqs. (5.3) and (5.4) are fully coupled. Thus in the second-moment closure framework, to simulate the evolution of individual sample paths $\mathbf{s}_t(\omega)$ requires first solving for not only the mean state but for the covariance matrix as well. This is one consequence of the presence of the nonlinear coupling term in the mean equation.

To close this section, it will be established that Eqs. (4.10) and (4.13) hold for the newly defined stochastic process. First, because the stochastic perturbation equation is linear, it has a fundamental matrix $\mathbf{M}_{t,t_0}$, not to be confused with that of the deterministic perturbation equation defined by Eq. (2.2). The fundamental matrix of the stochastic perturbation equation is defined for all $t \in \overline{\mathcal{T}}$ as the solution of the deterministic linear equation

$$\frac{\mathrm{d}\mathbf{M}_{t,t_0}}{\mathrm{d}t} + \mathbf{F}(\bar{\mathbf{s}})\mathbf{M}_{t,t_0} = \mathbf{0}, \tag{5.10}$$

corresponding to initial condition $\mathbf{M}_{t_0,t_0} = \mathbf{I}$, the $N \times N$ identity matrix. Therefore, it expresses the solution of the stochastic perturbation equation directly in terms of random initial condition $\mathbf{s}'_{t_0}$:

$$\mathbf{s}'_t = \mathbf{M}_{t,t_0}\mathbf{s}'_{t_0}. \tag{5.11}$$

This fundamental matrix depends in general on the mean state, and therefore also on the covariance matrix, due to the nonlinear coupling term in the mean equation. This dependence can be expressed fully as $\mathbf{M}_{t,t_0} = \mathbf{M}_{t,t_0}(\bar{\mathbf{s}}_{t_0}, \mathbf{P}_{t_0})$ since $\bar{\mathbf{s}}_{t_0}$ and $\mathbf{P}_{t_0}$ determine $\bar{\mathbf{s}}_t$ and $\mathbf{P}_t$ uniquely, for all $t \in \overline{\mathcal{T}}$. The fundamental matrix does not depend on the probability variable ω. Taking expectations in Eq. (5.11), therefore, gives $\mathcal{E}\mathbf{s}'_t = \mathbf{0}$ since $\mathcal{E}\mathbf{s}'_{t_0} = \mathbf{0}$, and then taking expectations in Eq. (5.8) gives

$$\mathcal{E}\frac{\mathrm{d}\mathbf{s}'}{\mathrm{d}t} = \mathbf{0}$$

since the Jacobian matrix does not depend on ω. Taking expectations in Eq. (5.9) then yields Eq. (4.10).

It can be verified that the same fundamental matrix can be used to express the solution of the covariance evolution equation (Eq. (5.4)) as

$$\mathbf{P}_t = \mathbf{M}_{t,t_0}\mathbf{P}_{t_0}\mathbf{M}^T_{t,t_0}, \tag{5.12}$$

which is the *operator form* of the covariance evolution equation. Since $\mathbf{P}_{t_0} = \mathcal{E}\mathbf{s}'_{t_0}\mathbf{s}'^T_{t_0}$ is symmetric positive semidefinite, it follows from Eq. (5.12) that $\mathbf{P}_t$ is also symmetric positive semidefinite. From Eq. (5.11), one has

$$\mathbf{s}'_t\mathbf{s}'^T_t = \mathbf{M}_{t,t_0}\mathbf{s}'_{t_0}\mathbf{s}'^T_{t_0}\mathbf{M}^T_{t,t_0},$$

and on taking expectations, it follows that

$$\mathbf{P}_t = \mathcal{E}\mathbf{s}'_t\,\mathbf{s}'^T_t. \tag{5.13}$$

Postmultiplying Eq. (5.8) by $\mathbf{s}'^T$, then adding the transpose of the result to itself, then taking expectations and using Eq. (5.13) leads to

$$\mathcal{E}\frac{\mathrm{d}\mathbf{s}'\mathbf{s}'^T}{\mathrm{d}t} + \mathbf{F}(\bar{\mathbf{s}})\mathbf{P} + \mathbf{P}\mathbf{F}^T(\bar{\mathbf{s}}) = \mathbf{0}.$$

Comparing this result with Eq. (5.4) gives Eq. (4.13).

To summarize, attention will now be focused on the mean equation (5.3), the covariance evolution equation (5.4) which can be written equivalently in operator form as Eq. (5.12), and the stochastic perturbation equation (5.8) which can be written equivalently in operator form as Eq. (5.11). The covariance matrix can be taken to be defined by Eq. (5.13). The fundamental matrix is defined by Eq. (5.10).

6. Energetic consistency of the second-moment closure equations

The second-moment closure equations were derived without reference to Assumption (3.1) that total energy is conserved by the nonlinear dynamics (Eq. (2.1)). Assumption (3.1) will now be used to show that the second-moment closure equations are energetically consistent.

By Assumption (5.1), Eq. (3.4) can be differentiated twice with respect to each of the N state variables and then evaluated for any $\mathbf{s} \in \mathcal{S}$ and $t \in \mathcal{T}$. Doing so once gives

$$\mathbf{s}^T \frac{\partial \mathbf{f}}{\partial s_k} + \mathbf{e}^{kT}\mathbf{f} = 0, \tag{6.1}$$

for each $\mathbf{s} \in \mathcal{S}$ and $t \in \mathcal{T}$ and for $k = 1, \ldots, N$, where $\mathbf{e}^k$ denotes the kth column of the $N \times N$ identity matrix. This can be written equivalently as

$$\mathbf{f}(\mathbf{s}) = -\mathbf{F}^T(\mathbf{s})\mathbf{s}, \tag{6.2}$$

which is a special relationship between $\mathbf{f}$ and its Jacobian matrix. Equation (6.2) implies in particular that if $\mathbf{0} \in \mathcal{S}$, then

$$\mathbf{f}(\mathbf{0}) = \mathbf{0}, \tag{6.3}$$

which means simply that the nonlinear dynamics are not externally forced, and that $\mathbf{s} = \mathbf{0}$ is a steady state solution of the nonlinear dynamics. Recall, however, that typically $\mathbf{0} \notin \mathcal{S}$ for geophysical dynamics since there are usually state variables that must be positive on physical grounds.

Differentiating Eq. (6.1) gives

$$\mathbf{s}^T \frac{\partial^2 \mathbf{f}}{\partial s_j \partial s_k} + \mathbf{e}^{jT} \frac{\partial \mathbf{f}}{\partial s_k} + \mathbf{e}^{kT} \frac{\partial \mathbf{f}}{\partial s_j} = 0, \tag{6.4}$$

for each $\mathbf{s} \in \mathcal{S}$ and $t \in \mathcal{T}$ and for $j, k = 1, \ldots, N$. The symmetric and anti-symmetric (skew-symmetric) parts of the Jacobian matrix, respectively, $\mathbf{F}^s$ and $\mathbf{F}^a$, were defined following Eq. (2.5). Equation (6.4) can be rewritten in terms of $\mathbf{F}^s$ as

$$\mathbf{F}^s_{jk}(\mathbf{s}) = -\tfrac{1}{2}\mathbf{s}^T \frac{\partial^2 \mathbf{f}(\mathbf{s})}{\partial s_j \partial s_k}, \tag{6.5}$$

for each $\mathbf{s} \in \mathcal{S}$ and $t \in \mathcal{T}$ and for $j, k = 1, \ldots, N$. Energetic consistency of the second-moment closure equations is an immediate consequence of this special relationship

between the symmetric part of the Jacobian matrix and the second partial derivatives of $\mathbf{f}$.

To see this, premultiply the mean equation (Eq. (5.3)) by $\bar{\mathbf{s}}^T$ and evaluate Eqs. (3.4) and (6.5) at $\mathbf{s} = \bar{\mathbf{s}} \in \mathcal{S}$ to find that

$$\tfrac{1}{2}\frac{\mathrm{d}\bar{\mathbf{s}}^T\,\bar{\mathbf{s}}}{\mathrm{d}t} - \sum_j \sum_k \mathbf{F}^s{}_{jk}(\bar{\mathbf{s}})\,P_{jk} = 0. \tag{6.6}$$

Recall that the total variance V is the trace of the covariance matrix,

$$V = \operatorname{tr}\mathbf{P} = \mathcal{E}\mathbf{s}'^T\mathbf{s}',$$

where the second equality follows from Eq. (5.13). Applying the trace operator to the covariance evolution equation (Eq. (5.4)), and using the property that $\operatorname{tr}\mathbf{P}\mathbf{F}^T = \operatorname{tr}\mathbf{F}^T\mathbf{P}$, gives

$$\tfrac{1}{2}\frac{\mathrm{d}V}{\mathrm{d}t} + \operatorname{tr}\mathbf{F}^s(\bar{\mathbf{s}})\mathbf{P} = 0. \tag{6.7}$$

This result follows also from the stochastic perturbation equation, as it must, by premultiplying Eq. (5.8) by $\mathbf{s}'^T$, then using the fact that $\mathbf{s}'^T(\omega)\mathbf{F}^a(\bar{\mathbf{s}})\mathbf{s}'(\omega) = 0$ for each $\omega \in \Omega$ since $\mathbf{F}^a$ is skew-symmetric and $\mathbf{s}'$ is real, and then taking expectations. Adding Eqs. (6.6) and (6.7) gives

$$\frac{\mathrm{d}(\bar{\mathbf{s}}^T\,\bar{\mathbf{s}} + V)}{\mathrm{d}t} = 0, \tag{6.8}$$

which is the statement (Eq. (4.2)) of energetic consistency for the second-moment closure equations. It implies in particular the bounds (Eqs. (4.3) and (4.4)) on their solutions.

7. The role of the symmetric part of the Jacobian matrix

Consider for a moment the case of linear, conservative dynamics. If $\mathbf{f}(\mathbf{s}, t)$ is linear in $\mathbf{s}$, then the Jacobian matrix $\mathbf{F} = \mathbf{F}(\mathbf{s}, t)$ is independent of $\mathbf{s}$, and the second partial derivatives of $\mathbf{f}$ with respect to the state variables all vanish. Thus from Eq. (6.5), it follows that $\mathbf{F}^s = \mathbf{0}$, and therefore from Eq. (6.2) one has simply $\mathbf{f}(\mathbf{s}) = \mathbf{F}^a\mathbf{s}$, with $\mathbf{F}^a$ independent of $\mathbf{s}$. The mean and covariance evolution equations, already simple for linear dynamics in general, simplify still further for conservative dynamics.

That the Jacobian matrix has a nonzero symmetric part $\mathbf{F}^s$ is one consequence of nonlinearity. Moreover, Eqs. (6.6) and (6.7) show that the exchange of energy between the mean state and the stochastic perturbations, which leads to the exact balance (6.8), occurs solely through the symmetric part of the Jacobian matrix. Equation (6.7) shows also that $\mathbf{F}^s$ directly controls the growth and/or decay of the uncertainty measured by the total variance V.

This section gives an overview of the role that $\mathbf{F}^s$ plays in controlling the behavior of the total variance and therefore in determining how the amount of energy exchanged

with the mean state changes through time. Section 8 will then examine in more depth some of the special properties that $\mathbf{F}^s$ has in this role, which are properties that result from conservation of total energy.

According to Eq. (5.12), the solution $\mathbf{P}$ of the covariance evolution equation is symmetric positive semidefinite, with rank not exceeding that of $\mathbf{P}_{t_0}$. Therefore, $\mathbf{P}$ has eigendecomposition

$$\mathbf{P} = \mathbf{W}\Sigma^2\mathbf{W}^T,$$

where Σ^2 is the diagonal matrix of nonnegative eigenvalues $\sigma_1^2, \ldots, \sigma_L^2$, where rank $\mathbf{P} \leq L \leq N$, and where $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_L]$ is the $N \times L$ matrix of normalized real eigenvectors $\mathbf{w}_l$, $\mathbf{w}_l^T\mathbf{w}_l = 1$ for $l = 1, \ldots, L$. The eigenvalues and eigenvectors depend on time. It follows from the eigendecomposition that

$$V = \operatorname{tr}\mathbf{P} = \sum_{l=1}^{L} \sigma_l^2, \tag{7.1}$$

and that

$$\operatorname{tr}\mathbf{F}^s(\bar{\mathbf{s}})\mathbf{P} = \sum_{l=1}^{L} \sigma_l^2 \mathbf{w}_l^T \mathbf{F}^s(\bar{\mathbf{s}})\mathbf{w}_l. \tag{7.2}$$

Substituting into Eq. (6.7) gives

$$\sum_{l=1}^{L} \left[\tfrac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t} + \mathbf{w}_l^T\mathbf{F}^s(\bar{\mathbf{s}})\mathbf{w}_l \right] \sigma_l^2 = 0. \tag{7.3}$$

Equation (7.3) says that the presence in $\mathbf{P}$ of an eigenvector $\mathbf{w}_l$ with nonzero eigenvalue acts to increase (decrease) uncertainty if $\mathbf{w}_l^T\mathbf{F}^s(\bar{\mathbf{s}})\mathbf{w}_l < 0$ ($\mathbf{w}_l^T\mathbf{F}^s(\bar{\mathbf{s}})\mathbf{w}_l > 0$). In particular, if $\mathbf{F}^s(\mathbf{s}, t)$ happens to be negative semidefinite, for all $\mathbf{s} \in \mathcal{S}$ and $\mathbf{t} \in \mathcal{T}$, then the total variance V is monotone nondecreasing, $\mathrm{d}V/\mathrm{d}t \geq 0$, and hence the total energy of the mean state is monotone nonincreasing, $\mathrm{d}\bar{\mathbf{s}}^T\bar{\mathbf{s}}/\mathrm{d}t \leq 0$, independently of the initial condition $(\bar{\mathbf{s}}_{t_0}, \mathbf{P}_{t_0})$, for as long as the solution $(\bar{\mathbf{s}}, \mathbf{P})$ exists. Thus dynamics with negative semidefinite $\mathbf{F}^s$ constitute a worst case for predictability; the uncertainty V can only increase with time or at best hold constant at times. That this should be a worst case is to be expected already from the deterministic initial value problem: the deterministic perturbation equation (Eq. (2.3)) gives

$$\tfrac{1}{2}\frac{\mathrm{d}\mathbf{q}^T\mathbf{q}}{\mathrm{d}t} + \mathbf{q}^T\mathbf{F}^s(\mathbf{s}_t, t)\mathbf{q} = 0$$

for perturbations of the deterministic trajectory $\mathbf{s}_t$, and so the size $\mathbf{q}^T\mathbf{q}$ of the perturbation is monotone nondecreasing, regardless of the initial perturbation, if $\mathbf{F}^s$ is negative semidefinite on $\mathcal{S} \times \mathcal{T}$. Similarly, if $\mathbf{F}^s$ is positive semidefinite on $\mathcal{S} \times \mathcal{T}$, then $\mathrm{d}V/\mathrm{d}t \leq 0$ and $\mathrm{d}\bar{\mathbf{s}}^T\bar{\mathbf{s}}/\mathrm{d}t \geq 0$, independently of initial condition $(\bar{\mathbf{s}}_{t_0}, \mathbf{P}_{t_0})$, for as long as the solution $(\bar{\mathbf{s}}, \mathbf{P})$ exists. The dynamics in this case would be eminently predictable.

By Assumptions (2.2) and (4.1), $\mathbf{F}^s$ is bounded on $\mathcal{S} \times \mathcal{T}$, and therefore the eigenvalues of $\mathbf{F}^s$ are bounded on $\mathcal{S} \times \mathcal{T}$. The eigenvalues are real since $\mathbf{F}^s$ is symmetric. Denote the largest and smallest eigenvalues of $\mathbf{F}^s$ by $\lambda_{\max}(\mathbf{F}^s)$ and $\lambda_{\min}(\mathbf{F}^s)$, respectively. In the next section, it will be shown that by Assumption (3.1),

$$\lambda_{\min}(\mathbf{F}^s(\mathbf{s}, t)) \leq 0 \leq \lambda_{\max}(\mathbf{F}^s(\mathbf{s}, t)) \tag{7.4}$$

at each point $\mathbf{s} \in \mathcal{S}$ and $t \in \mathcal{T}$. This means that $\mathbf{F}^s$ cannot be either positive or negative definite, anywhere on $\mathcal{S} \times \mathcal{T}$, although it does not rule out positive or negative semidefiniteness. Also, it will be shown that at each point $\mathbf{s}$ in any open set $\mathcal{S}_*$ on which $\mathbf{f}$ is genuinely nonlinear at time t, as defined there, the inequalities in Eq. (7.4) are strict at time t:

$$\lambda_{\min}(\mathbf{F}^s(\mathbf{s}, t)) < 0 < \lambda_{\max}(\mathbf{F}^s(\mathbf{s}, t)), \tag{7.5}$$

which means that $\mathbf{F}^s(\mathbf{s}, t)$ has at least one positive and one negative eigenvalue at time t, at every point $\mathbf{s} \in \mathcal{S}_*$. Thus, if $\mathbf{F}^s$ is genuinely nonlinear on all of $\mathcal{S}$, for all $t \in \mathcal{T}$, then $\mathbf{F}^s$ cannot be either positive or negative semidefinite, anywhere on $\mathcal{S} \times \mathcal{T}$; there is potential for both growth and decay of uncertainty, at all times, regardless of where the mean state happens to be in $\mathcal{S}$. Equation (7.3) shows in the genuinely nonlinear case that whether growth, decay, or neither actually occurs at a particular time depends on the eigenstructure of $\mathbf{P}$, relative to that of $\mathbf{F}^s(\bar{\mathbf{s}})$, at that time.

Now denote by $\mathbf{u}_{\max} = \mathbf{u}_{\max}(\mathbf{F}^s(\bar{\mathbf{s}}, t))$ and $\mathbf{u}_{\min} = \mathbf{u}_{\min}(\mathbf{F}^s(\bar{\mathbf{s}}, t))$, respectively, the eigenvectors of $\mathbf{F}^s$ corresponding to eigenvalues $\lambda_{\max} = \lambda_{\max}(\mathbf{F}^s(\bar{\mathbf{s}}, t))$ and $\lambda_{\min} = \lambda_{\min}(\mathbf{F}^s(\bar{\mathbf{s}}, t))$. It follows from Eqs. (7.1) and (7.2) that

$$\min_{\mathbf{P}} \frac{\operatorname{tr} \mathbf{F}^s(\bar{\mathbf{s}}, t)\mathbf{P}}{\operatorname{tr} \mathbf{P}} = \lambda_{\min},$$

where the minimization is over all symmetric positive semidefinite matrices $\mathbf{P}$, and that the minimum is attained for $\mathbf{P} = \mathbf{P}_{\min} = \gamma \mathbf{u}_{\min} \mathbf{u}_{\min}^T$, with γ an arbitrary positive constant. Furthermore, $\mathbf{P}_{\min} \in \mathcal{P}(\bar{\mathbf{s}})$ for all γ small enough, where $\mathcal{P}$ is the set defined in Section 5, since $\mathcal{P}$ is an open set containing the origin in $\mathbb{R}^{N \times N}$. Also, $\mathbf{P}_{\min} \in \mathcal{P}_\mu(\bar{\mathbf{s}})$ for each $\mu \in (0, 1)$, by taking γ still smaller, where $\mathcal{P}_\mu$ is the set defined in Eq. (5.7). Thus the minimum is achieved for matrices in $\mathcal{P}(\bar{\mathbf{s}})$ and in $\mathcal{P}_\mu(\bar{\mathbf{s}})$. Similarly,

$$\max_{\mathbf{P}} \frac{\operatorname{tr} \mathbf{F}^s(\bar{\mathbf{s}}, t)\mathbf{P}}{\operatorname{tr} \mathbf{P}} = \lambda_{\max},$$

and the maximum is attained for $\mathbf{P} = \mathbf{P}_{\max} = \delta \mathbf{u}_{\max} \mathbf{u}_{\max}^T$, with δ an arbitrary positive constant. Rewriting Eq. (6.7) as

$$\tfrac{1}{2} \frac{1}{V} \frac{\mathrm{d}V}{\mathrm{d}t} + \frac{\operatorname{tr} \mathbf{F}^s(\bar{\mathbf{s}}, t)\mathbf{P}}{\operatorname{tr} \mathbf{P}} = 0,$$

then shows that $-2\lambda_{\min}$ ($+2\lambda_{\max}$) is the maximum instantaneous relative rate of increase (decrease) of uncertainty and is attained for $\mathbf{P} = \mathbf{P}_{\min}$ ($\mathbf{P} = \mathbf{P}_{\max}$).

The presence of both positive and negative eigenvalues of $\mathbf{F}^s$ can lead to complex behavior for the total variance V as a function of time and therefore also for the behavior of the total energy of the mean state as a function of time, which according to Eq. (6.8) mirrors precisely that of $V/2$. In general, both growth and decay of total variance can occur, and therefore $\mathrm{d}V/\mathrm{d}t$ can also vanish instantaneously. For instance, one can check that $\mathrm{d}V/\mathrm{d}t = 0$ at

$$\mathbf{P} = \lambda_{\max}\mathbf{u}_{\min}\mathbf{u}_{\min}^T - \lambda_{\min}\mathbf{u}_{\max}\mathbf{u}_{\max}^T.$$

8. Genuine nonlinearity and essential linearity

Inequality (Eq. (7.4)) is readily established. First observe that Eqs. (3.4) and (6.2) together imply that

$$\mathbf{s}^T\mathbf{F}^s(\mathbf{s})\mathbf{s} = 0, \tag{8.1}$$

for every $\mathbf{s} \in \mathcal{S}$ and $t \in \mathcal{T}$, since $\mathbf{F}^a$ is skew-symmetric. Also note for later reference that if $\mathbf{0} \in \mathcal{S}$, then Eq. (6.5) gives

$$\mathbf{F}^s(\mathbf{0}) = \mathbf{0} \tag{8.2}$$

although again one should recall that typically $\mathbf{0} \notin \mathcal{S}$ for models of ocean and atmospheric dynamics.

Equation (8.1) implies that at each point $\mathbf{s} \in \mathcal{S}$, one has the following alternative at each fixed time $t \in \mathcal{T}$: either $\mathbf{F}^s(\mathbf{s})$ has at least one positive and one negative eigenvalue or else $\mathbf{s}$ is a null vector of $\mathbf{F}^s(\mathbf{s})$,

$$\mathbf{F}^s(\mathbf{s})\mathbf{s} = \mathbf{0}. \tag{8.3}$$

To see this, let

$$\mathbf{F}^s(\mathbf{s}) = \mathbf{U}(\mathbf{s})\Lambda(\mathbf{s})\mathbf{U}^T(\mathbf{s})$$

be the eigendecomposition of $\mathbf{F}^s$, where Λ is the diagonal matrix of eigenvalues $\lambda_1, \ldots, \lambda_N$, all of which are real since $\mathbf{F}^s$ is real and symmetric and where $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_N]$ is the matrix of normalized real eigenvectors $\mathbf{u}_l$, $\mathbf{u}_l^T\mathbf{u}_l = 1$ for $l = 1, \ldots, N$. Then Eq. (8.1) can be rewritten as

$$\sum_{l=1}^{N} \lambda_l(\mathbf{s})[\mathbf{s}^T\mathbf{u}_l(\mathbf{s})]^2 = 0.$$

Either all the terms in this sum vanish or there is at least one positive and one negative term, equivalently, at least one positive and one negative eigenvalue. The condition that all the terms vanish can be expressed as $\Lambda(\mathbf{s})\mathbf{U}^T(\mathbf{s})\mathbf{s} = \mathbf{0}$, which is equivalent to Eq. (8.3) since $\mathbf{U}(\mathbf{s})$ is nonsingular. Thus the statement of alternatives has been demonstrated.

Inequality (Eq. (7.4)) holds at each point $\mathbf{s} \in \mathcal{S}, t \in \mathcal{T}$, where $\mathbf{F}^s(\mathbf{s}, t)$ has a null vector. Therefore, it holds at each $\mathbf{s} \in \mathcal{S}, t \in \mathcal{T}$ where the second alternative condition, Eq. (8.3), holds. Furthermore, inequality (Eq. (7.5)) holds at each point $\mathbf{s} \in \mathcal{S}, t \in \mathcal{T}$ where the first alternative condition, that $\mathbf{F}^s(\mathbf{s}, t)$ has at least one positive and one negative eigenvalue, holds. Therefore, inequality (Eq. (7.4)) holds for all $\mathbf{s} \in \mathcal{S}$ and $t \in \mathcal{T}$. The rest of this section will distinguish the two alternatives more tangibly and examples will be given at the end of the section.

The condition expressed by Eq. (8.3) can be expressed equivalently in a way that clarifies when it occurs and also allows one to check for its occurrence essentially by inspection of $\mathbf{f}$, without even calculating $\mathbf{F}^s$. This is done by first introducing the polar coordinate $\rho = (\mathbf{s}^T\mathbf{s})^{1/2}$ so that for each $\mathbf{s} \in \mathcal{S}$, one has $\mathbf{s} = \rho\mathbf{c}$ with $\mathbf{c}$ on the unit hypersphere $\mathbf{c}^T\mathbf{c} = 1$ in $\mathbb{R}^N$. Any scalar, vector, or matrix function $\phi = \phi(\mathbf{s}) = \phi(\mathbf{s}, t)$ that is continuously differentiable with respect to $\mathbf{s} \in \mathcal{S}$ for all $t \in \mathcal{T}$ also has a continuous derivative with respect to ρ, for all $\mathbf{s} \in \mathcal{S}$ and $t \in \mathcal{T}$, and

$$\frac{\partial \phi}{\partial \rho} = \sum_{l=1}^{N} \frac{\partial \phi}{\partial s_l} \frac{\partial s_l}{\partial \rho} = \sum_{l=1}^{N} \frac{\partial \phi}{\partial s_l} c_l$$

so that

$$\rho \frac{\partial \phi}{\partial \rho} = \sum_{l=1}^{N} \frac{\partial \phi}{\partial s_l} s_l.$$

By Assumption (2.2), ϕ can be taken to be $\mathbf{f}$, in which case this relationship reads

$$\rho \frac{\partial \mathbf{f}(\mathbf{s})}{\partial \rho} = \mathbf{F}(\mathbf{s})\mathbf{s}$$

since $\partial \mathbf{f}(\mathbf{s})/\partial s_l$ is by definition the lth column of the Jacobian matrix $\mathbf{F}(\mathbf{s})$. Using Eq. (6.2), one then has

$$\mathbf{F}^s(\mathbf{s})\mathbf{s} = \tfrac{1}{2}\left[\mathbf{F}(\mathbf{s}) + \mathbf{F}^T(\mathbf{s})\right]\mathbf{s} = \tfrac{1}{2}\left[\rho \frac{\partial \mathbf{f}(\mathbf{s})}{\partial \rho} - \mathbf{f}(\mathbf{s})\right] = \tfrac{1}{2}\rho^2 \frac{\partial \rho^{-1}\mathbf{f}(\mathbf{s})}{\partial \rho}, \tag{8.4}$$

for every $\mathbf{s} \in \mathcal{S}$ and $t \in \mathcal{T}$.

Thus the second alternative condition, Eq. (8.3), is equivalent to

$$\mathbf{f}(\mathbf{s}) = \rho \frac{\partial \mathbf{f}(\mathbf{s})}{\partial \rho}. \tag{8.5}$$

It is always the second alternative that holds at the origin if $\mathbf{0} \in \mathcal{S}$, as seen either by setting $\mathbf{s} = \mathbf{0}$ in Eq. (8.3) or by setting $\rho = 0$ in Eq. (8.5) and using Eq. (6.3). Away from the origin, Eq. (8.4) says that the second alternative condition is equivalent to

$$\frac{\partial \rho^{-1}\mathbf{f}(\mathbf{s})}{\partial \rho} = \mathbf{0}. \tag{8.6}$$

Substituting Eq. (8.5) into Eq. (8.6) shows that, away from the origin, the second alternative condition is equivalent simply to

$$\frac{\partial^2 \mathbf{f}(\mathbf{s})}{\partial \rho^2} = \mathbf{0}. \tag{8.7}$$

If Eq. (8.7) holds not just at a point $\mathbf{s} \in \mathcal{S}$, but for all points in an open set $\mathcal{S}_* \subseteq \mathcal{S}$, then $\mathbf{f}$ is linear as a function of ρ for all $\mathbf{s} \in \mathcal{S}_*$. By Assumption (5.1), $\partial^2 \mathbf{f}/\partial \rho^2$ is continuous on the state space $\mathcal{S}$ for each $t \in \mathcal{T}$, and therefore if Eq. (8.7) holds on an open set $\mathcal{S}_* \subseteq \mathcal{S}$, then it holds also on $\overline{\mathcal{S}_*} \cap \mathcal{S}$, where $\overline{\mathcal{S}_*}$ denotes the closure of $\mathcal{S}_*$ in $\mathbb{R}^N$. If $\mathbf{f}$ is linear in all the state variables, on an open set $\mathcal{S}_* \subseteq \mathcal{S}$, then certainly $\mathbf{f}$ is linear in ρ throughout $\mathcal{S}_*$, and not only does the second alternative, therefore, hold on $\mathcal{S}_*$ but it was seen at the beginning of Section 7 that then in fact $\mathbf{F}^s = \mathbf{0}$ on $\mathcal{S}_*$. It is possible for $\mathbf{f}$ to be linear in ρ on an open set $\mathcal{S}_* \subseteq \mathcal{S}$ without being linear in any of the state variables there. This is the case when $\mathbf{f}(\mathbf{s}) = \rho \mathbf{g}(\mathbf{c})$, for $\mathbf{s} \in \mathcal{S}_*$, for some function $\mathbf{g}$ whose first partial derivatives with respect to all the c_l, $l = 1, \ldots, N$, vanish nowhere for $\mathbf{s} = \rho \mathbf{c} \in \mathcal{S}_*$.

In case condition (Eq. (8.7)) holds at every point $\mathbf{s}$ in an open set $\mathcal{S}_* = \mathcal{S}_*(t) \subseteq \mathcal{S}$ at a particular time $t \in \mathcal{T}$, then $\mathbf{f} = \mathbf{f}(\mathbf{s}, t)$ will be said to be *essentially linear* on $\mathcal{S}_*(t)$. The equivalence between conditions (Eqs. (8.3) and (8.7)), together with Eq. (6.2), implies that $\mathbf{f}(\mathbf{s}, t)$ is essentially linear on an open set $\mathcal{S}_*(t)$ if, and only if,

$$\mathbf{f}(\mathbf{s}, t) = \mathbf{F}^a(\mathbf{s}, t)\mathbf{s}, \tag{8.8}$$

for all $\mathbf{s} \in \mathcal{S}_*(t)$. The discussion at the beginning of Section 7 shows that if $\mathbf{f}(\mathbf{s}, t)$ is linear in all the state variables on an open set $\mathcal{S}_*(t) \subseteq \mathcal{S}$, then Eq. (8.8) holds with $\mathbf{F}^a(\mathbf{s}, t)$ independent of $\mathbf{s}$ on $\mathcal{S}_*(t)$. Thus essential linearity, as defined here, amounts to generalizing the notion of linearity in such a way that one still has $\mathbf{f} = \mathbf{F}^a\mathbf{s}$ but with $\mathbf{F}^a$ depending on $\mathbf{s}$.

In case condition Eq. (8.7) holds at *no* point $\mathbf{s}$ in an open set $\mathcal{S}_* = \mathcal{S}_*(t) \subseteq \mathcal{S}$ at a particular time $t \in \mathcal{T}$, then $\mathbf{f} = \mathbf{f}(\mathbf{s}, t)$ will be said to be *genuinely nonlinear* on $\mathcal{S}_*(t)$. With this definition, the original statement of alternatives can finally be rephrased, for open sets instead of points, as follows: $\mathbf{F}^s(\mathbf{s}, t)$ has at least one positive and one negative eigenvalue at each point $\mathbf{s}$ in an open set $\mathcal{S}_*(t) \subseteq \mathcal{S}$ if, and only if, $\mathbf{f}(\mathbf{s}, t)$ is genuinely nonlinear on $\mathcal{S}_*(t)$. It has also been shown that an open set $\mathcal{S}_*(t) \subseteq \mathcal{S}$ on which $\mathbf{f}(\mathbf{s}, t)$ is genuinely nonlinear cannot contain the origin $\mathbf{s} = \mathbf{0}$ since Eq. (8.3) holds at the origin. If $\mathbf{0} \notin \mathcal{S}$, as is typical for geophysical problems, then it is possible for $\mathbf{f}$ to be genuinely nonlinear on all of $\mathcal{S}$, for all time $t \in \mathcal{T}$. It is not difficult to show that for reasonable discretizations of the one-dimensional shallow water system considered in Sections 3–5, $\mathbf{f}$ is genuinely nonlinear on all of the state space $\mathcal{S}_c$, for arbitrarily long time intervals $\mathcal{T}$.

To illustrate the ideas of genuine nonlinearity and essential linearity in the simplest setting, consider the case $N = 2$. It follows from Eq. (3.4) that $\mathbf{f}$ has the form

$$\mathbf{f}(\mathbf{s}, t) = \beta(\mathbf{s}, t) \begin{bmatrix} s_2 \\ -s_1 \end{bmatrix} = \rho \beta(\mathbf{s}, t) \begin{bmatrix} c_2 \\ -c_1 \end{bmatrix},$$

for some scalar function β, thus generalizing the first example of Section 3. It is immediate that, away from the origin, $\mathbf{f}$ is essentially linear precisely on those open sets $\mathcal{S}_*(t)$

on which β is independent of ρ at time t, i.e., on which β is a function only of the ratio s_1/s_2 at time t. That Eq. (8.8) does indeed hold on every such set can be verified directly by calculating $\mathbf{F}^a$. One finds that

$$\mathbf{F}^a = \left(\beta + \tfrac{1}{2}\rho\frac{\partial\beta}{\partial\rho}\right)\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

so that

$$\mathbf{F}^a\mathbf{s} = \left(\beta + \tfrac{1}{2}\rho\frac{\partial\beta}{\partial\rho}\right)\begin{bmatrix} s_2 \\ -s_1 \end{bmatrix},$$

and so Eq. (8.8) does hold where $\partial\beta/\partial\rho = 0$ as well as at the origin. Similarly, it is immediate that, away from the origin, $\mathbf{f}$ is genuinely nonlinear precisely on those open sets $\mathcal{S}_*(t)$ on which $\partial\beta/\partial\rho$ vanishes nowhere at time t. The statement of alternatives says that on such a set, $\mathbf{F}^s$ must have at least one positive and one negative eigenvalue, which for $N = 2$ means that $\mathbf{F}^s$ must have exactly one positive and one negative eigenvalue and therefore must have a negative determinant, $\det \mathbf{F}^s < 0$. One can calculate directly that

$$\det \mathbf{F}^s = -\tfrac{1}{4}\rho^2\left(\frac{\partial\beta}{\partial\rho}\right)^2,$$

which verifies the statement.

Finally, although it is not usually the case that $\mathbf{0} \in \mathcal{S}$ for geophysical problems, it is instructive to consider dynamical behavior near the origin in case $\mathbf{0} \in \mathcal{S}$, particularly in light of the ideas of genuine nonlinearity and essential linearity. Recall from Eq. (6.3) that $\mathbf{s} = \mathbf{0}$ is a steady state solution of the original, deterministic nonlinear dynamics (Eq. (2.1)).

Consider first the case in which $\mathbf{f}$ is essentially linear on an open set $\mathcal{S}_*(t) \subseteq \mathcal{S}$ with $\mathbf{0} \in \mathcal{S}_*(t)$. Thus, $\mathbf{f}$ is linear in ρ on $\mathcal{S}_*(t)$, and hence, $\mathbf{f}$ is linear in ρ near the origin along each coordinate axis. Therefore,

$$f_k(\epsilon\mathbf{e}^l, t) = \alpha_k^l(t)|\epsilon|, \tag{8.9}$$

for some scalars α_k^l, $k, l = 1, \ldots, N$ and for all ϵ small enough, where $\mathbf{e}^l$ denotes the lth column of the $N \times N$ identity matrix. This follows from the fact that $\rho = (\mathbf{s}^T\mathbf{s})^{1/2} = |\epsilon|$ for $\mathbf{s} = \epsilon\mathbf{e}^l$. By Assumption (2.2), $\partial f_k/\partial s_l$ exists and is continuous at the origin, and since $|\epsilon|$ is not differentiable at $\epsilon = 0$, it follows from Eq. (8.9) that $\alpha_k^l(t) = 0$ for $k, l = 1, \ldots, N$. Therefore, $\mathbf{f}(\mathbf{s}, t) = \mathbf{0}$ for all $\mathbf{s} \in \mathcal{S}_*(t)$: essentially linear dynamics near the origin are trivial dynamics. Further, if $\mathbf{f}$ is essentially linear on all of $\mathcal{S}$ for all $t \in \mathcal{T}$, and if $\mathbf{0} \in \mathcal{S}$, what has just been shown is that then $\mathbf{f} = \mathbf{0}$ on $\mathcal{S} \times \mathcal{T}$.

Now consider dynamics near the origin for general $\mathbf{f}$. If $\mathbf{0} \in \mathcal{S}$, one has that $\epsilon\mathbf{e}^l \in \mathcal{S}$, for $l = 1, \ldots, N$ and for all ϵ small enough. Evaluating Eq. (6.5) at $\mathbf{s} = \epsilon\mathbf{e}^l$ and using Eq. (8.2) gives

$$\frac{1}{\epsilon}\left[\mathbf{F}_{jk}^s(\epsilon\mathbf{e}^l, t) - \mathbf{F}_{jk}^s(\mathbf{0}, t)\right] = -\tfrac{1}{2}\mathbf{e}^{lT}\frac{\partial^2\mathbf{f}(\epsilon\mathbf{e}^l, t)}{\partial s_j \partial s_k},$$

and taking the limit here as $\epsilon \to 0$ then gives

$$\frac{\partial \mathbf{F}^s_{jk}(\mathbf{0}, t)}{\partial s_l} = -\frac{1}{2}\frac{\partial^2 f_l(\mathbf{0}, t)}{\partial s_j \partial s_k}.$$

In view of Eq. (6.3), at $\bar{\mathbf{s}} = \mathbf{0}$ the mean Eq. (5.3) therefore reads

$$\frac{\mathrm{d}\bar{s}_l}{\mathrm{d}t} - \sum_j \sum_k \frac{\partial \mathbf{F}^s_{jk}(\mathbf{0}, t)}{\partial s_l} P_{jk} = \mathbf{0}, \tag{8.10}$$

for $l = 1, \ldots, N$. Thus the behavior of the mean state $\bar{\mathbf{s}}$ at $\bar{\mathbf{s}} = \mathbf{0}$ depends on the symmetric matrices $\partial \mathbf{F}^s(\mathbf{0}, t)/\partial s_l$, $l = 1, \ldots, N$. In particular, even though $\mathbf{s} = \mathbf{0}$ is a steady state solution of the deterministic nonlinear dynamics, $\bar{\mathbf{s}} = \mathbf{0}$ is not necessarily a steady state solution of the mean equation. The reason for this is the presence of uncertainty in the stochastic dynamics, which is manifested through the covariance matrix in the nonlinear coupling term in Eq. (8.10).

In view of Eq. (8.2), at $\bar{\mathbf{s}} = \mathbf{0}$ the covariance evolution Eq. (5.4) simplifies to

$$\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}t} + \mathbf{F}^a(\mathbf{0}, t)\mathbf{P} - \mathbf{P}\mathbf{F}^a(\mathbf{0}, t) = \mathbf{0}. \tag{8.11}$$

Thus at $\bar{\mathbf{s}} = \mathbf{0}$, the covariance matrix evolves in an energetically neutral way, $\mathrm{d}V/\mathrm{d}t = 0$, and the evolution of the mean state at $\bar{\mathbf{s}} = \mathbf{0}$ still depends on this covariance evolution through the nonlinear coupling term in Eq. (8.10). In particular, it is possible for the mean state to remain zero for a period of time, not just instantaneously, and subsequently to regain energy. That is to say, although it is possible for all of the energy of the mean state to have been extracted at some particular time $t = \tau$, and therefore for the upper bound in Eq. (4.3) actually to be attained at time τ, it is also possible for the mean state to regain energy after time τ through interaction with the covariance matrix in the nonlinear coupling term. This effect is not present if the nonlinear coupling term is neglected.

9. Bounds on the growth of relative uncertainty

It was shown in Section 7 that the instantaneous relative rate of change of the total variance V_t satisfies the bounds

$$-2\lambda_{\max}(\mathbf{F}^s(\bar{\mathbf{s}}_t, t)) \leq \frac{1}{V_t}\frac{\mathrm{d}V_t}{\mathrm{d}t} \leq -2\lambda_{\min}(\mathbf{F}^s(\bar{\mathbf{s}}_t, t)),$$

with each bound attainable by rank one matrices $\mathbf{P}_t \in \mathcal{P}_\mu(\bar{\mathbf{s}}_t) \subset \mathcal{P}(\bar{\mathbf{s}}_t)$, for any given $\mu \in (0, 1)$, where $\mathcal{P}$ and $\mathcal{P}_\mu$ are the sets defined in Section 5. It was shown in Section 8 that $\lambda_{\min} \leq 0$ and $\lambda_{\max} \geq 0$, and furthermore that if $\bar{\mathbf{s}}_t$ is in an open set on which $\mathbf{f}(\bar{\mathbf{s}}_t, t)$ is genuinely nonlinear, then $\lambda_{\min} < 0$ and $\lambda_{\max} > 0$. Both $\lambda_{\max}$ and $|\lambda_{\min}|$ can be large, although finite, since no assumption has been introduced that would otherwise limit these values. Thus not only is it possible to have both growth and decay of total

variance, at any instant of time it is also possible for the relative rate of growth or decay to be large.

On the other hand, inequalities (Eqs. (4.3) and (4.4)) show that V_t cannot grow unboundedly. That is to say, although the total variance can grow rapidly at particular instants of time, growth over any interval of time is strictly limited due to the energetic consistency of the mean and covariance evolution equations, which was demonstrated in Section 6. The present section examines the growth of V_t, relative to V_{t_0}, over every interval of time for which the solution of the second-moment closure equations exists, thus providing time-independent bounds on the relative uncertainty V_t/V_{t_0}. The latter part of the discussion will concern inequalities (Eqs. (4.3) and (4.4)) only, for the given state space $\mathcal{S}$, without specific reference to the particular problem whose solutions satisfy the inequalities. Therefore, the time-independent bounds obtained for V_t/V_{t_0} apply as well to the original stochastic process defined in Section 4 for as long as it exists.

Consider first the behavior of solutions of the second-moment closure equations when the nonlinear coupling term in the mean equation is neglected, so that the fundamental matrix $\mathbf{M}_{t,t_0} = \mathbf{M}_{t,t_0}(\bar{\mathbf{s}}_{t_0}, \mathbf{P}_{t_0})$ introduced in Section 5 becomes a function of $\bar{\mathbf{s}}_{t_0}$ alone, $\mathbf{M}_{t,t_0} = \mathbf{M}_{t,t_0}(\bar{\mathbf{s}}_{t_0})$. Denote by $\sigma_{t,t_0} = \sigma_{t,t_0}(\bar{\mathbf{s}}_{t_0})$ the largest singular value of $\mathbf{M}_{t,t_0}$, and denote by $\mathbf{v}_{t_0} = \mathbf{v}_{t_0}(\bar{\mathbf{s}}_{t_0})$ the corresponding right singular vector, normalized so that $\mathbf{v}_{t_0}^T\mathbf{v}_{t_0} = 1$. The corresponding normalized left singular vector, $\mathbf{u}_t = \mathbf{u}_t(\bar{\mathbf{s}}_{t_0})$, is then given by

$$\mathbf{u}_t = \frac{1}{\sigma_{t,t_0}}\mathbf{M}_{t,t_0}\mathbf{v}_{t_0}.$$

Now take

$$\mathbf{P}_{t_0} = V_{t_0}\mathbf{v}_{t_0}\mathbf{v}_{t_0}^T \tag{9.1}$$

so that $V_{t_0} = \operatorname{tr}\mathbf{P}_{t_0}$, and use Eq. (5.12) to obtain

$$\frac{V_t}{V_{t_0}} = \frac{\operatorname{tr}\mathbf{M}_{t,t_0}\mathbf{P}_{t_0}\mathbf{M}_{t,t_0}^T}{\operatorname{tr}\mathbf{P}_{t_0}} = \frac{\operatorname{tr}\sigma_{t,t_0}^2 V_{t_0}\mathbf{u}_t\mathbf{u}_t^T}{\operatorname{tr}\mathbf{P}_{t_0}} = \sigma_{t,t_0}^2.$$

The largest singular value σ_{t,t_0} of any matrix $\mathbf{M}_{t,t_0}$ also has the property that

$$\frac{\operatorname{tr}\mathbf{M}_{t,t_0}\mathbf{P}\mathbf{M}_{t,t_0}^T}{\operatorname{tr}\mathbf{P}} \leq \sigma_{t,t_0}^2 \tag{9.2}$$

for every symmetric positive semidefinite matrix $\mathbf{P}$. This leads to the usual result that

$$\frac{V_t}{V_{t_0}} \leq \sigma_{t,t_0}^2(\bar{\mathbf{s}}_{t_0}), \tag{9.3}$$

with equality holding for the choice of $\mathbf{P}_{t_0}$ given in Eq. (9.1). For this choice, it is guaranteed that $\mathbf{P}_{t_0} \in \mathcal{P}_\mu(\bar{\mathbf{s}}_{t_0}) \subset \mathcal{P}(\bar{\mathbf{s}}_{t_0})$, for any given $\mu \in (0, 1)$, by taking V_{t_0} small enough. The bound in Eq. (9.3) depends on the initial mean state $\bar{\mathbf{s}}_{t_0}$. Also, there is no general upper bound for the largest singular value $\sigma_{t,t_0}(\bar{\mathbf{s}}_{t_0})$ itself.

Now consider the behavior of solutions of the second-moment closure equations with the nonlinear coupling term retained, so that the fundamental matrix is fully a function of $\mathbf{P}_{t_0}$, $\mathbf{M}_{t,t_0} = \mathbf{M}_{t,t_0}(\bar{\mathbf{s}}_{t_0}, \mathbf{P}_{t_0})$. The largest singular value of $\mathbf{M}_{t,t_0}$ is in general then a function of $\mathbf{P}_{t_0}$ as well, $\sigma_{t,t_0} = \sigma_{t,t_0}(\bar{\mathbf{s}}_{t_0}, \mathbf{P}_{t_0})$, as are the corresponding right and left singular vectors. For any particular choice of $\mathbf{P}_{t_0}$, for instance the one described in the preceding paragraph, inequality (Eq. (9.2)) still holds for every symmetric positive semidefinite matrix $\mathbf{P}$. Therefore, it holds for $\mathbf{P} = \mathbf{P}_{t_0}$, and it follows that

$$\frac{V_t}{V_{t_0}} = \frac{\operatorname{tr} \mathbf{M}_{t,t_0} \mathbf{P}_{t_0} \mathbf{M}_{t,t_0}^T}{\operatorname{tr} \mathbf{P}_{t_0}} \leq \sigma_{t,t_0}^2(\bar{\mathbf{s}}_{t_0}, \mathbf{P}_{t_0}).$$

Thus σ_{t,t_0}^2 is still an upper bound for V_t / V_{t_0}, but the property that it is necessarily attained for some choice of $\mathbf{P}_{t_0}$ has been lost.

What is actually desired is the *least* upper bound for V_t / V_{t_0}, call it $\hat{\sigma}_{t,t_0}^2(\bar{\mathbf{s}}_{t_0})$:

$$\hat{\sigma}_{t,t_0}^2(\bar{\mathbf{s}}_{t_0}) = \sup_{\mathbf{P}_{t_0}} \frac{\operatorname{tr} \mathbf{M}_{t,t_0}(\bar{\mathbf{s}}_{t_0}, \mathbf{P}_{t_0}) \mathbf{P}_{t_0} \mathbf{M}_{t,t_0}^T(\bar{\mathbf{s}}_{t_0}, \mathbf{P}_{t_0})}{\operatorname{tr} \mathbf{P}_{t_0}}, \tag{9.4}$$

where the supremum is taken over all initial covariance matrices $\mathbf{P}_{t_0} \in \mathcal{P}_\mu(\bar{\mathbf{s}}_{t_0})$, for some given $\mu \in (0, 1)$, or over all $\mathbf{P}_{t_0}$ in some chosen subset of $\mathcal{P}_\mu(\bar{\mathbf{s}}_{t_0})$. It will be convenient to take the supremum over all initial covariance matrices in the open set $\mathcal{P}_\mu(\bar{\mathbf{s}}_{t_0}; V_{\min}, V_{\max})$ for some given μ, $V_{\min}$, and $V_{\max}$, where this set is defined for $0 < \mu < 1$ and $0 \leq V_{\min} < V_{\max} \leq 2(E_{\max} - E_{\min})$ by

$$\mathcal{P}_\mu(\bar{\mathbf{s}}_{t_0}; V_{\min}, V_{\max}) = \{\mathbf{P} \in \mathcal{P}_\mu(\bar{\mathbf{s}}_{t_0}) : V_{\min} < \operatorname{tr} \mathbf{P} < V_{\max}\}. \tag{9.5}$$

This set is never empty since $\mathcal{P}_\mu(\bar{\mathbf{s}}_{t_0})$ is an open set. Taking the supremum over all $\mathbf{P}_{t_0} \in \mathcal{P}_\mu(\bar{\mathbf{s}}_{t_0}; V_{\min}, V_{\max})$ makes $\hat{\sigma}_{t,t_0}(\bar{\mathbf{s}}_{t_0})$ defined in Eq. (9.4) depend also on $V_{\min}$ and $V_{\max}$, i.e., $\hat{\sigma}_{t,t_0}(\bar{\mathbf{s}}_{t_0}) = \hat{\sigma}_{t,t_0}(\bar{\mathbf{s}}_{t_0}; V_{\min}, V_{\max})$. Then one has

$$V_t / V_{t_0} \leq \hat{\sigma}_{t,t_0}^2(\bar{\mathbf{s}}_{t_0}; V_{\min}, V_{\max}),$$

with equality either holding or holding arbitrarily closely, for some $\mathbf{P}_{t_0} \in \mathcal{P}_\mu(\bar{\mathbf{s}}_{t_0}; V_{\min}, V_{\max})$. One can also define

$$\hat{\sigma}_{\overline{\mathcal{T}}}(\bar{\mathbf{s}}_{t_0}; V_{\min}, V_{\max}) = \sup_{t \in \overline{\mathcal{T}}} \hat{\sigma}_{t,t_0}(\bar{\mathbf{s}}_{t_0}; V_{\min}, V_{\max}),$$

where $\overline{\mathcal{T}} = [t_0, \overline{T}] \subseteq \mathcal{T}$ is an interval of existence of solutions for all $\mathbf{P}_{t_0} \in \mathcal{P}_\mu(\bar{\mathbf{s}}_{t_0}; V_{\min}, V_{\max})$. This yields

$$V_t / V_{t_0} \leq \hat{\sigma}_{\overline{\mathcal{T}}}^2(\bar{\mathbf{s}}_{t_0}; V_{\min}, V_{\max}), \tag{9.6}$$

for all $t \in \overline{\mathcal{T}}$, with equality either holding or holding arbitrarily closely, at some time $t \in \overline{\mathcal{T}}$ and for some $\mathbf{P}_{t_0} \in \mathcal{P}_\mu(\bar{\mathbf{s}}_{t_0}; V_{\min}, V_{\max})$.

The maximization problem of Eq. (9.4) is nonlinear, and in particular one cannot expect its solution to be independent of V_{t_0} as it was seen to be in the linear case, in which the nonlinear coupling term in the mean equation was neglected. It is for this reason that the supremum is to be taken over the set defined in Eq. (9.5), which allows a range $V_{t_0} \in (V_{\min}, V_{\max})$ for the initial total variance. Inequalities (Eqs. (4.3) and (4.4)) easily imply upper bounds for $\hat{\sigma}^2_{\overline{\mathcal{T}}}(\bar{\mathbf{s}}_{t_0}; V_{\min}, V_{\max})$, whose dependence on $\bar{\mathbf{s}}_{t_0}$, $V_{\min}$, and $V_{\max}$ are explicit, as will be seen next.

Consider first some cases in which $\mathbf{0} \notin \mathcal{S}$ so that inequality (Eq. (4.4)) holds. Rewriting it for V_t / V_{t_0} gives

$$\frac{V_t}{V_{t_0}} < 1 + \frac{\bar{\mathbf{s}}_{t_0}^T \bar{\mathbf{s}}_{t_0} - 2E_{\min}}{V_{t_0}}. \tag{9.7}$$

Recall that $\bar{\mathbf{s}}_{t_0}$ and V_{t_0} must also satisfy inequality (Eq. (5.5)), rewritten here as

$$V_{t_0} < 2E_{\max} - \bar{\mathbf{s}}_{t_0}^T \bar{\mathbf{s}}_{t_0}. \tag{9.8}$$

The most unpredictable (largest possible V_t / V_{t_0}) case occurs when $\bar{\mathbf{s}}_{t_0}$ is near the outer boundary of $\mathcal{S}$, where the total energy is $E_{\max}$, for then V_{t_0} in inequality (Eq. (9.7)) must be small, according to inequality (Eq. (9.8)). Therefore, let $\bar{\mathbf{s}}_{t_0}^T \bar{\mathbf{s}}_{t_0} = 2E_{\max} - \epsilon$, with $\epsilon > 0$ fixed and small. Then inequality (Eq. (9.8)) holds if $V_{t_0} < \epsilon$, so take the supremum over all $\mathbf{P}_{t_0} \in \mathcal{P}_\mu(\bar{\mathbf{s}}_{t_0}; \alpha\epsilon, \epsilon)$, for some given $\mu \in (0, 1)$ and $\alpha \in (0, 1)$. Setting $\bar{\mathbf{s}}_{t_0}^T \bar{\mathbf{s}}_{t_0} = 2E_{\max} - \epsilon$ and $V_{t_0} > \alpha\epsilon$ in inequality (Eq. (9.7)), and using inequality (Eq. (9.6)), then gives

$$\frac{V_t}{V_{t_0}} \leq \hat{\sigma}^2_{\overline{\mathcal{T}}}(\bar{\mathbf{s}}_{t_0}; \alpha\epsilon, \epsilon) < 1 - \frac{1}{\alpha} + \frac{2(E_{\max} - E_{\min})}{\alpha\epsilon}, \tag{9.9}$$

for all $t \in \overline{\mathcal{T}}$. This shows that V_t / V_{t_0} can be large if ϵ is small compared with $2(E_{\max} - E_{\min})$ even if α is taken to be close to one, i.e., even if one allows only a small range $V_{t_0} \in (\alpha\epsilon, \epsilon)$ for the initial total variance. The upper bound (Eq. (9.9)) for the supremum $\hat{\sigma}^2_{\overline{\mathcal{T}}}(\bar{\mathbf{s}}_{t_0}; \alpha\epsilon, \epsilon)$ depends on the initial mean state $\bar{\mathbf{s}}_{t_0}$ only through its total energy $E_{\max} - \epsilon/2$. Recall from the derivation of inequality (Eq. (4.4)) in Section 4 that the bound (Eq. (9.7)) is tight if the mean state $\bar{\mathbf{s}}_t$ corresponding to $\bar{\mathbf{s}}_{t_0}$ has energy near the minimum energy level $E_{\min}$. Therefore, the bound (Eq. (9.9)) for $\hat{\sigma}^2_{\overline{\mathcal{T}}}(\bar{\mathbf{s}}_{t_0}; \alpha\epsilon, \epsilon)$ is tight if some mean state with large initial energy $E_{\max} - \epsilon/2$ approaches the minimum energy level $E_{\min}$ at any time $t \in \overline{\mathcal{T}}$.

The most predictable (smallest possible V_t / V_{t_0}) case occurs when $\bar{\mathbf{s}}_{t_0}$ is near the inner boundary of $\mathcal{S}$, where the total energy is $E_{\min}$, for then inequality (Eq. (9.8)) implies that V_{t_0} need not be small. Let $\bar{\mathbf{s}}_{t_0}^T \bar{\mathbf{s}}_{t_0} = 2E_{\min} + \epsilon$, with $\epsilon > 0$ fixed and small. Then inequality (Eq. (9.8)) holds if $V_{t_0} < V_{\max} = 2(E_{\max} - E_{\min}) - \epsilon$, so take the supremum over all $\mathbf{P}_{t_0} \in \mathcal{P}_\mu(\bar{\mathbf{s}}_{t_0}; \alpha V_{\max}, V_{\max})$, for some given $\mu \in (0, 1)$ and $\alpha \in (0, 1)$. Substitution into inequality (Eq. (9.7)) then gives

$$\frac{V_t}{V_{t_0}} \leq \hat{\sigma}^2_{\overline{\mathcal{T}}}(\bar{\mathbf{s}}_{t_0}; \alpha V_{\max}, V_{\max}) < \frac{2(E_{\max} - E_{\min}) + (\frac{1}{\alpha} - 1)\epsilon}{2(E_{\max} - E_{\min}) - \epsilon}, \tag{9.10}$$

for all $t \in \overline{\mathcal{T}}$. This shows that V_t/V_{t_0} can never be much larger than one if ϵ is small compared with $2(E_{\max} - E_{\min})$ unless α is also taken to be small, much smaller than $\epsilon/2(E_{\max} - E_{\min})$. The explanation for this is simply that $V_t < 2(E_{\max} - E_{\min})$ for all $t \in \overline{\mathcal{T}}$ by Assumption (4.1), and so if V_{t_0} is already close to the value $2(E_{\max} - E_{\min})$, then there is no room for growth of the relative uncertainty V_t/V_{t_0}. This behavior is far different than what occurs when the nonlinear coupling term is neglected, for then the initial total variance V_{t_0} simply scales out of the problem.

It is expected that in many applications $E_{\min}$ and $E_{\max}$ are both known reasonably well, with $E_{\max}/E_{\min}$ not much larger than one. This still leaves plenty of room for growth of the relative uncertainty. Suppose the total energy of the initial mean state is the average of the minimum and maximum energy levels so that $\bar{\mathbf{s}}_{t_0}^T \bar{\mathbf{s}}_{t_0} = E_{\min} + E_{\max}$. Then inequality (Eq. (9.8)) holds if $V_{t_0} < V_{\max} = E_{\max} - E_{\min}$. Again taking the supremum over all $\mathbf{P}_{t_0} \in \mathcal{P}_\mu(\bar{\mathbf{s}}_{t_0}; \alpha V_{\max}, V_{\max})$, for some given $\mu \in (0, 1)$ and $\alpha \in (0, 1)$, then gives simply

$$\frac{V_t}{V_{t_0}} \leq \hat{\sigma}^2_{\overline{\mathcal{T}}}(\bar{\mathbf{s}}_{t_0}; \alpha V_{\max}, V_{\max}) < 1 + \frac{1}{\alpha}, \tag{9.11}$$

for all $t \in \overline{\mathcal{T}}$. If the value of α is taken to be modest, say $\alpha = 1/3$, then little growth of uncertainty can occur. However, if V_{t_0} is allowed to be small, say with $\alpha = 1/100$, then considerable growth of uncertainty can occur, relative to this small value.

In case $\mathbf{0} \in \mathcal{S}$, the strict inequality (Eq. (9.7)) simply becomes nonstrict, with $E_{\min} = 0$ also. The bounds in inequalities (Eqs. (9.9)–(9.11)) remain strict, however, because they were obtained from the strict inequality (Eq. (9.8)). Thus all that is necessary is to set $E_{\min} = 0$ everywhere following inequality (Eq. (9.7)). To see that the discussion from Eq. (9.4) onwards applies equally well to the stochastic process defined in Section 4, replace the numerator in Eq. (9.4) by $\operatorname{tr} \mathbf{P}_t$, where $\mathbf{P}_t$ is the covariance matrix of that process, and then reinterpret the various other quantities following Eq. (9.4) accordingly.

10. Conclusions

The problem of finding a system of approximate evolution equations for the mean and covariance matrix of second-order stochastic processes defined by unforced, nonlinear, conservative systems of ordinary differential equations with random initial conditions has been examined in this chapter from the viewpoint of energetics. A brief treatment of the stochastic initial value problem for conservative nonlinear systems of ordinary differential equations was given, and it was used to show that the mean and covariance matrix of the resulting stochastic process are dynamically linked through an energy relationship.

The second-moment closure equations are a nonlinearly coupled system of approximate evolution equations for the mean and covariance matrix of this process. An existence and uniqueness theory was given for these equations based largely on existence and uniqueness theory for the stochastic initial value problem. This theory was then used to show that, under appropriate hypotheses, the mean and covariance matrix whose evolution is given by the second-moment closure equations are the mean and covariance matrix of an additional, well-defined stochastic process. It was shown further that the

second-moment closure equations are energetically consistent; the mean and covariance matrix whose evolution they define are dynamically linked through precisely the same energy relationship as that of the mean and covariance matrix of the original stochastic process.

Several implications followed from this energetic consistency. One is that the total variance V_t of each of the two covariance matrices, that of the original stochastic process and that of the one whose evolution is given by the second-moment closure equations, is strictly and identically bounded in time t. It was shown further that energetic consistency implies simple, identical, time-independent bounds on the ratio V_t / V_{t_0}. Also it was shown that when the original conservative system is genuinely nonlinear, as defined in this chapter, total variance for the second-moment closure equations may be increasing, decreasing, or stationary at times.

Essential to the results of this chapter is that no assumption was made on the initial probability distribution, apart from the existence of two moments. It was also argued that the normal distribution is not appropriate in general for conservative dynamics because it requires the existence of realizations, with nonzero probability, having total energy larger than any given amount. Furthermore, for atmospheric and ocean dynamics, there are mass-like and/or temperature-like state variables that are constrained by the dynamics to be positive. With this motivation, an hypothesis was introduced that requires the realizations to have bounded energy, with probability one, and appropriate state spaces were introduced to handle state variables that are bounded from below. Many of the results were illustrated with examples.

The behavior of solutions of the second-moment closure equations was contrasted with the behavior of solutions of the approximate system obtained by neglecting the nonlinear coupling term in the mean equation. This approximate system, usually derived under an assumption of normality, is at the heart of many current large-scale computational applications in atmospheric and ocean dynamics. It was shown that this approximate system is not energetically consistent because the nonlinear coupling term is crucial for energetic consistency.

The results of this article have left a number of open questions. For instance, it will be important to establish hypotheses under which solutions of the second-moment closure equations exist over arbitrarily long time intervals, in case solutions of the nonlinear dynamical system from which they are derived also have this property. It will also be important to develop efficient computational algorithms for implementing the second-moment closure equations in large-scale oceanic and atmospheric applications. Addressing these issues successfully will require the continued strong collaboration among physical scientists, computational scientists, and mathematicians that has been so fruitful in recent years as this volume attests.

11. Acknowledgments

The author would like to thank the editors, Roger Temam and Joe Tribbia, and Lulu Stader of Elsevier, for their enthusiasm and patience during the preparation of this manuscript. The generous support of NASA's Modeling, Analysis and Prediction program, managed by Don Anderson, is also gratefully acknowledged.

References

BUIZZA, R., PALMER, T.N. (1995). The singular vector structure of the atmospheric global circulation. *J. Atmos. Sci.* **52**, 1434–1456.

CODDINGTON, E.A., LEVINSON, N. (1955). *Theory of Ordinary Differential Equations* (McGraw-Hill, New York).

COURTIER, P., TALAGRAND, O. (1987). Variational assimilation of meteorological observations with the adjoint vorticity equation. Part II: numerical results. *Q. J. R. Meteorol. Soc.* **113**, 1329–1368.

DOOB, J.L. (1953). *Stochastic Processes* (Wiley, New York).

EPSTEIN, E.S. (1969). Stochastic dynamic prediction. *Tellus* **21**, 739–759.

EPSTEIN, E.S., PITCHER, E.J. (1972). Stochastic analysis of meteorological fields. *J. Atmos. Sci.* **29**, 244–257.

FLEMING, R.J. (1971a). On stochastic dynamic prediction I. The energetics of uncertainty and the question of closure. *Mon. Wea. Rev.* **99**, 851–872.

FLEMING, R.J. (1971b). On stochastic dynamic prediction II. Predictability and utility. *Mon. Wea. Rev.* **99**, 927–938.

JAZWINSKI, A.H. (1970). *Stochastic Processes and Filtering Theory* (Academic Press, New York).

LAX, P.D. (1973). *Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves* (Society for Industrial and Applied Mathematics, Philadelphia).

LIN, S.-J., CHAO, W.C., SUD, Y.C., WALKER, G.K. (1994). A class of the van Leer-type transport schemes and its application to the moisture transport in a general circulation model. *Mon. Wea. Rev.* **122**, 1575–1593.

LIN, S.-J., ROOD, R.B. (1996). Multidimensional flux-form semi-Lagrangian transport schemes. *Mon. Wea. Rev.* **124**, 2046–2070.

MOLTENI, F., BUIZZA, R., PALMER, T.N., PETROLIAGIS, T. (1996). The ECMWF ensemble prediction system: methodology and validation. *Q. J. R. Meteorol. Soc.* **122**, 73–119.

MOORE, A.M., KLEEMAN, R. (1997). The singular vectors of a coupled ocean-atmosphere model of ENSO. I: thermodynamics, energetics and error growth. *Q. J. R. Meteorol. Soc.* **123**, 953–981.

PITCHER, E.J. (1977). Application of stochastic dynamic prediction to real data. *J. Atmos. Sci.* **34**, 3–21.

TALAGRAND, O., COURTIER, P. (1987). Variational assimilation of meteorological observations with the adjoint vorticity equation. Part I: theory. *Q. J. R. Meteorol. Soc.* **113**, 1311–1328.

THÉPAUT, J.-N., COURTIER, P., BELAUD, G., LEMAÎTRE, G. (1996). Dynamical structure functions in a four-dimensional variational assimilation: a case study. *Q. J. R. Meteorol. Soc.* **122**, 535–561.

Boundary Value Problems for the Inviscid Primitive Equations in Limited Domains

Antoine Rousseau

Institut National de Recherche en Informatique et Automatique (INRIA), Laboratoire Jean Kuntzmann, Grenoble, France

Roger M. Temam

The Institute for Scientific Computing and Applied Mathematics, Indiana University, Bloomington, IN, USA

Joseph J. Tribbia

National Center for Atmospheric Research, Boulder, Colorado, USA

Abstract

This work aims to contribute to what is considered as a major computational issue for the geophysical fluid dynamics (GFD) for the coming years, i.e., the boundary conditions for numerical computations in a limited domain, with a boundary that has (at least partly) no physical justification. Numerical computations in limited domains in ocean and atmosphere are constantly required (and sometimes lead to commercial softwares) in order to provide forecasts for agriculture, tourism industry, insurances, aircraft navigation, etc. This chapter focuses on the nonviscous primitive equations in a limited domain in space dimensions 2, 2.5, and 3 and provides in each case a set of boundary conditions which is shown to lead to a well-posed problem. The suitability of these new boundary conditions is also computationally evidenced in space dimension 2.

1. Introduction

Limited area models (LAM) are constantly used for numerical simulations of geophysical flows. Indeed when refined information are needed, it would be far too costly and in

Computational Methods for the Atmosphere and the Oceans
Special Volume (Roger M. Temam and Joseph J. Tribbia, Guest Editors) of
HANDBOOK OF NUMERICAL ANALYSIS, VOL. XIV
P.G. Ciarlet (Editor)
Copyright © 2009 Elsevier B.V.
All rights reserved
ISSN 1570-8659
DOI 10.1016/S1570-8659(08)00211-1

fact not computationally feasible at all, to perform simulations for the whole atmosphere, or for the whole ocean. Numerical computations in limited domains in ocean and atmosphere are hence constantly required (and sometimes lead to commercial softwares) in order to provide forecasts for agriculture, tourism industry, insurances, aircraft navigation, etc. LAM are used in an essential way for strongly perturbed systems, such as tropical storms, squall lines, or mid-latitude cyclones, to overcome the error due to the parametrization schemes and numerical truncation errors introduced purely by insufficient model resolutions for large areas or global models.

The penalty for using a LAM is the appearance of a domain with nonphysical boundaries where no physical law will provide natural boundary conditions. Hence, beside the usual difficulty of writing boundary conditions on top or bottom of the ocean or atmosphere, now appears the difficulty of writing boundary conditions on the nonphysical lateral boundary. Such a difficulty has been known since early works of J. von Neumann and J.G. Charney, and various remedies have been proposed and implemented over the years (see e.g., Bennett and Chua [1999], Bennett and Kloeden [1978], Charney, Fjörtoft and von Neumann [1950]). However, it is expected that, for the high-resolution models which will be used in coming years with the increase of computing powers and computer memory, the remedies which have been used will lead to spurious modes which will damage the whole computation; see for example in Temam and Tribbia [2003, Figs. 3 and 4], the effect of spurious modes and their resolution. The tutorial article by Warner, Peterson and Treadon [1997] describes the motivations and the computational difficulties for this problem. The present work aims to contribute to what is considered as a major computational issue for the geophysical fluid dynamics (GFD) for the coming years, i.e., the boundary conditions for numerical computations in a limited domain, with a boundary that has (at least partly) no physical justification.

Beside the computational difficulty, a mathematical difficulty arises. Indeed, as we explain below, for the equations that we consider (the primitive equations without viscosity), there is no set of local boundary conditions which produces a well-posed problem. Hence, for the boundary conditions that we propose, we need also to address questions of well-posedness, which we do in the context of the linearized equations.

The equations concerned by this kind of applications are well known: primitive equations (PEs) and shallow water equations (SWEs). This work is fully dedicated to the PEs, which we will recall later on. More precisely, we only consider *inviscid* PEs since the viscosity effects mainly appear after a few days, which is beyond the forecast period that we are interested in. For the case of viscous PEs, for which the mathematical aspects have been widely studied from the initial works of Lions, Temam and Wang [1992a,b] to the most recent results of Cao and Titi [2007], Kobelkov [2006], Kukavica and Ziane [2007], and numerous authors in between, the reader is referred to the review papers Temam and Ziane [2004] and its updated form Petcu, Temam and Ziane [2008] in this volume. To the best of our knowledge, the inviscid case has been left unexplored for years since the negative result of Oliger and Sundström [1978], which showed that inviscid PEs could not be well-posed for any set of boundary conditions of local type. Recent needs in GFD put these problems back on the frontstage, and the issue of open boundary conditions has been recently studied in Rousseau, Temam and Tribbia

[2005b, 2007], TEMAM and TRIBBIA [2003], and BLAYO and DEBREU [2005]. The use of such boundary conditions is now taken into account in realistic numerical simulations, at least for the so-called barotropic mode (see e.g., DUMAS and LAZURE [2007], MADEC, DELECLUSE, IMBARD, and LÉVY [1998]).

We focus our study on the linearized PEs since the boundary condition difficulty is already fully present in the linear case. However, the numerical simulations presented in the sequel relate as well to nonlinear simulations. They seem to indicate that the nonlinear system behaves like the linear one, at least for sufficiently small times and initial data, as far as the boundary conditions are concerned.

Before recalling the equations and the negative result OLIGER and SUNDSTRÖM [1978] quoted above, let us draw the outline of this article. In Section 2, we both provide the mathematical and numerical results obtained in ROUSSEAU, TEMAM and TRIBBIA [2005b, 2007], and that concern the two-dimensional $(x - z)$ case. Then, before going into the full 3D model, we propose in Section 3 a simplified 2.5D model, in view of performing (in dimension 2) *computations of physical significance*; see the introduction to Section 3 for a detailed description of the motivations for this model. Finally, we end this overview with the most recent mathematical results on the full three-dimensional case presented in Section 4.

In summary, our contributions in this chapter are as follows:

1. In space dimensions 2, 2.5, and 3, we propose sets of boundary conditions which lead to well-posed initial and boundary value problems for the linearized PEs without viscosity.
2. Numerical simulations performed in space dimension 2 support the conjecture that the nonlinear analog of these boundary conditions produce a well-posed nonlinear problem.
3. Numerical simulations performed in space dimension 2 show that the boundary conditions that we propose satisfactorily solve the problem of lateral boundary conditions for the LAM, with a precision of a few percents.

The rest of this section is devoted to describing the PEs, their linearization around a stratified state, and the normal mode expansion in the vertical direction.

1.1. The inviscid primitive equations

We now recall the primitive equations (PEs); the emphasis will be on the case of the ocean. The case of the atmosphere can be studied similarly with minor changes, as well as the coupled atmosphere and ocean (see e.g., PETCU, TEMAM and ZIANE [2008] in this volume). The equations are derived from the Boussinesq equations by making the hydrostatic assumption which amounts to replacing the conservation of momentum in the vertical direction by the hydrostatic equation, hence the equations

$$\frac{\partial \widetilde{\mathbf{v}}}{\partial t} + (\widetilde{\mathbf{v}}.\nabla)\,\widetilde{\mathbf{v}} + \widetilde{w}\,\frac{\partial \widetilde{\mathbf{v}}}{\partial z} + f\,k \times \widetilde{\mathbf{v}} + \frac{1}{\rho_0}\nabla \widetilde{p} = F_{\widetilde{\mathrm{v}}}, \tag{1.1a}$$

$$\frac{\partial \widetilde{p}}{\partial z} = -\,\widetilde{\rho} g, \tag{1.1b}$$

$$\nabla \widetilde{\mathbf{v}} + \frac{\partial \widetilde{w}}{\partial z} = 0, \tag{1.1c}$$

$$\frac{\partial \widetilde{T}}{\partial t} + (\widetilde{\mathbf{v}}.\nabla)\,\widetilde{T} + \widetilde{w}\,\frac{\partial \widetilde{T}}{\partial z} = Q_{\widetilde{T}}, \tag{1.1d}$$

$$\widetilde{\rho} = \rho_0\,(1 - \alpha\,(\widetilde{T} - T_0)). \tag{1.1e}$$

In these equations, $\widetilde{\mathbf{v}} = (\widetilde{u}, \widetilde{v})$ is the horizontal velocity, $\widetilde{w}$ the vertical velocity, $\widetilde{p}$ the pressure, $\widetilde{\rho}$ the density, $\widetilde{T}$ the temperature, g the gravitational acceleration, and f the Coriolis parameter. The horizontal gradient is denoted by ∇. Equation (1.1e) is the equation of state of the fluid, ρ_0 and T_0 are constant reference values of $\widetilde{\rho}$ and $\widetilde{T}$, and $\alpha > 0$ is constant; this equation of state is linear although more involved nonlinear state equations could be considered.

Equation (1.1b) is the so-called hydrostatic equation. The other equations correspond to the Boussinesq approximation (see, e.g., PEDLOSKY [1987], WASHINGTON and PARKINSON [1986], and SALMON [1998] for more details).

In the physical context, the forcing terms $F_{\widetilde{\mathbf{v}}} = (F_{\widetilde{u}}, F_{\widetilde{v}})$ and $Q_{\widetilde{T}}$ do not exist, but we introduce them here for mathematical generality and to study the case of nonhomogeneous boundary conditions by homogenization of the boundary conditions.

We now consider a reference stratified flow with constant velocity $\overline{\mathbf{v}}_0 = (\overline{U}_0, 0) = \overline{U}_0\,e_x$, and density, temperature, and pressure of the form $\rho_0 + \overline{\rho}$, $T_0 + \overline{T}$, $p_0 + \overline{p}$ with $\mathrm{d}\overline{p}/\mathrm{d}z$ constant and thus

$$\overline{T}(z) = \frac{N^2}{\alpha\,g}\,z, \tag{1.2}$$

$$\overline{\rho}(z) = -\,\rho_0\,\alpha\,\overline{T}(z) = -\,\frac{\rho_0\,N^2}{g}\,z, \tag{1.3}$$

$$\frac{\mathrm{d}\overline{T}}{\mathrm{d}z}(z) = \frac{N^2}{\alpha\,g}, \tag{1.4}$$

$$\frac{\mathrm{d}\overline{\rho}}{\mathrm{d}z}(z) = -\,\frac{\rho_0}{g}\,N^2, \tag{1.5}$$

$$\frac{\mathrm{d}\overline{p}}{\mathrm{d}z}(z) = -\,(\rho_0 + \overline{\rho})\,g. \tag{1.6}$$

Here, N is the buoyancy frequency or Brunt–Väisälä frequency, assumed to be constant.

We then decompose the unknown functions $\widetilde{\mathbf{v}}$, $\widetilde{\rho}$, $\widetilde{T}$, $\widetilde{p}$ in the following way:

$$\begin{cases} \widetilde{\mathbf{v}} = \overline{U}_0\,e_x + \mathbf{v}(x, y, z, t), \\ \widetilde{\rho} = \rho_0 \quad + \overline{\rho}(z) \quad + \rho(x, y, z, t), \\ \widetilde{T} = T_0 \quad + \overline{T}(z) \quad + T(x, y, z, t), \\ \widetilde{p} = p_0 \quad + \overline{p}(z) \quad + p(x, y, z, t). \end{cases} \tag{1.7}$$

Equations (1.1b), (1.1d), and (1.1e) become

$$\frac{\partial p}{\partial z} = -\rho\, g, \tag{1.8}$$

$$\frac{\partial T}{\partial t} + (\widetilde{\mathbf{v}}.\nabla)\, T + w\, \frac{\partial T}{\partial z} + \frac{N^2}{\alpha\, g}\, w = F_T, \tag{1.9}$$

$$\rho = -\rho_0\, \alpha\, T. \tag{1.10}$$

We infer from Eqs. (1.1) and (1.8)–(1.10) the following equations for u, v, w, $\phi = p/\rho_0$, and $\psi = \phi_z = \alpha\, g\, T$:

$$\frac{\partial u}{\partial t} + \overline{U}_0\, \frac{\partial u}{\partial x} + u\, \frac{\partial u}{\partial x} + v\, \frac{\partial u}{\partial y} + w\, \frac{\partial u}{\partial z} - fv + \frac{\partial \phi}{\partial x} = F_u, \tag{1.11a}$$

$$\frac{\partial v}{\partial t} + \overline{U}_0\, \frac{\partial v}{\partial x} + u\, \frac{\partial v}{\partial x} + v\, \frac{\partial v}{\partial y} + w\, \frac{\partial v}{\partial z} + fu + \frac{\partial \phi}{\partial y} = F_v - f\, \overline{U}_0, \tag{1.11b}$$

$$\frac{\partial \psi}{\partial t} + \overline{U}_0\, \frac{\partial \psi}{\partial x} + u\, \frac{\partial \psi}{\partial x} + v\, \frac{\partial \psi}{\partial y} + w\, \frac{\partial \psi}{\partial z} + N^2\, w = F_\psi, \tag{1.11c}$$

$$\frac{\partial \phi}{\partial z} = -\frac{\rho}{\rho_0}\, g = \psi, \tag{1.11d}$$

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0. \tag{1.11e}$$

From Eq. (1.11e), we find:

$$w(x, y, z) = \int_z^0 \left(\frac{\partial u}{\partial x}(x, y, z') + \frac{\partial v}{\partial y}(x, y, z') \right) \mathrm{d}z', \tag{1.12}$$

which makes the vertical velocity w a diagnostic variable, whereas it is a prognostic one in the Navier–Stokes equations.

The PEs Eqs. (1.11a)–(1.11e), linearized around the stratified flow $\overline{\mathbf{v}}_0 = \overline{U}_0\, e_x$, $\overline{\rho}$, $\overline{T}$, $\overline{p}$, read:

$$\frac{\partial u}{\partial t} + \overline{U}_0\, \frac{\partial u}{\partial x} - fv + \frac{\partial \phi}{\partial x} = F_u, \tag{1.13a}$$

$$\frac{\partial v}{\partial t} + \overline{U}_0\, \frac{\partial v}{\partial x} + fu + \frac{\partial \phi}{\partial y} = F_v - f\, \overline{U}_0, \tag{1.13b}$$

$$\frac{\partial \psi}{\partial t} + \overline{U}_0\, \frac{\partial \psi}{\partial x} + N^2\, w = F_\psi, \tag{1.13c}$$

$$\frac{\partial \phi}{\partial z} = -\frac{\rho}{\rho_0}\, g = \psi, \tag{1.13d}$$

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0. \tag{1.13e}$$

We will consider the flow in the three-dimensional domain $\mathcal{M} = \mathcal{M}' \times (-L_3, 0)$, where $\mathcal{M}'$ is the interface atmosphere/ocean, $\mathcal{M}' = (0, L_1) \times (0, L_2)$. Naturally, we supplement Eq. (1.13) with the following top and bottom boundary conditions (just imposed by kinematics):

$$w(x, y, z = -L_3, t) = w(x, y, z = 0, t) = 0, \quad \forall (x, y) \in \mathcal{M}', t > 0. \tag{1.14}$$

The aim of this work is to introduce some lateral boundary conditions at $x = 0, L_1$ and $y = 0, L_2$, which are both physically reasonable and computationnally satisfying,[1] and that lead to the well-posedness of the problem (Eq. (1.13)).

1.2. Normal modes expansion

The first step of the analysis of Eq. (1.13) consists, by separation of variables, in looking for solutions of the form

$$\begin{cases} u(x, y, z, t) = \mathcal{U}(z)\, \hat{u}(x, y, t), \;\; v(x, y, z, t) = \mathcal{V}(z)\, \hat{v}(x, y, t), \\ \psi(x, y, z, t) = \Psi(z)\, \hat{\psi}(x, y, t), \\ w(x, y, z, t) = \mathcal{W}(z)\, \hat{w}(x, y, t), \;\; \phi(x, y, z, t) = \Phi(z)\, \hat{\phi}(x, y, t). \end{cases} \tag{1.15}$$

Substituting these expressions into Eq. (1.13), we find that $\mathcal{U}$, $\mathcal{V}$, and Φ must be proportional and $\mathcal{W}$ proportional to Ψ. So we just take $\mathcal{V} = \Phi = \mathcal{U}$, and $\Psi = \mathcal{W}$. Indeed, the third equation of Eq. (1.13) implies that

$$-\frac{\hat{\psi}_t + \overline{U}_0 \hat{\psi}_x}{N^2 \hat{w}} = \frac{\mathcal{W}}{\Psi} \; (= c_1'),$$

and these quantities are constant since the left-hand side of the last equation depends[2] on x, y, and t and the right-hand side depends on z only. For the sake of simplicity, we can take this constant c_1' equal to one, i.e., $\mathcal{W} = \Psi$. Similarly, applying the operator $\partial/\partial t + \overline{U}_0\, \partial/\partial x$ to the first and second equations of Eq. (1.13), we obtain that $\mathcal{U}$, $\mathcal{V}$, and Φ must be proportional, and so we can take $\mathcal{U} = \mathcal{V} = \Phi$. Finally, the fourth and fifth equations of Eq. (1.13) imply that

$$-\frac{\hat{u}_x + \hat{v}_y}{\hat{w}} = \frac{\mathcal{W}'}{\mathcal{U}} = c_2', \quad \frac{\hat{\phi}}{\hat{\psi}} = \frac{\Psi}{\Phi'} = c_3',$$

where c_2', c_3' are constant; hence, $\mathcal{W} = c_2'\, \mathcal{U}'$ and

$$\mathcal{U}'' + \lambda^2 \mathcal{U} = 0, \quad \mathcal{W}'' + \lambda^2 \mathcal{W} = 0, \tag{1.16}$$

[1] Assuming that we are willing to pay the price of a nonlocal (mode by mode) boundary condition for increased accuracy. The necessity of nonlocal boundary conditions appears below.

[2] We recall that the buoyancy frequency N is assumed to be constant.

with $\lambda^2 = -c_2'/c_3'$. By Eq. (1.14), the natural boundary conditions for w and $\mathcal{W}$ are $\mathcal{W} = 0$ at $z = 0$ and $-L_3$; thus, $\mathcal{U}$ and $\mathcal{W}$ are solutions of the two-point boundary value problems consisting of Eq. (1.16) and

$$\mathcal{U}'(0) = \mathcal{U}'(-L_3) = \mathcal{W}(0) = \mathcal{W}(-L_3) = 0. \tag{1.17}$$

We denote by λ_n^2 the corresponding eigenvalues and write

$$\begin{cases} \lambda_n = \dfrac{n\pi}{L_3}, \quad \lambda_n^2 = \dfrac{1}{gH_n}, \quad \text{i.e.,} \quad H_n = \dfrac{L_3^2}{gn^2\pi^2}, \\ \\ \mathcal{W}_n = \sqrt{\dfrac{2}{L_3}}\sin(\lambda_n z), \quad \mathcal{U}_n = \sqrt{\dfrac{2}{L_3}}\cos(\lambda_n z), \quad n \geq 1, \quad \mathcal{U}_0 = \dfrac{1}{\sqrt{L_3}}. \end{cases} \tag{1.18}$$

As usual, the functions $\mathcal{U}_n$ and $\mathcal{W}_n$ have been chosen to form an orthonormal set in $L^2(-L_3, 0)$.

The equations satisfied by $\hat{u}$, $\hat{v}$, etc., will appear below. Indeed, having found these special solutions to Eq. (1.13), we now look for the general solution in the form

$$\begin{cases} (u, v, \phi) = \displaystyle\sum_{n\geq 0} \mathcal{U}_n(z)(u_n, v_n, \phi_n)(x, y, t), \\ \\ (w, \psi) = \displaystyle\sum_{n\geq 1} \mathcal{W}_n(z)(w_n, \psi_n)(x, y, t). \end{cases} \tag{1.19}$$

Substituting these expressions in Eq. (1.13), we arrive at the following systems, for $n \geq 1$,

$$\begin{cases} \dfrac{\partial u_n}{\partial t} + \overline{U}_0\dfrac{\partial u_n}{\partial x} - fv_n + \dfrac{\partial \phi_n}{\partial x} = 0, \\ \\ \dfrac{\partial v_n}{\partial t} + \overline{U}_0\dfrac{\partial v_n}{\partial x} + fu_n + \dfrac{\partial \phi_n}{\partial y} = 0, \\ \\ \dfrac{\partial \psi_n}{\partial t} + \overline{U}_0\dfrac{\partial \psi_n}{\partial x} + N^2 w_n = 0, \\ \\ \phi_n = -\dfrac{1}{\lambda_n}\psi_n, \quad w_n = -\dfrac{1}{\lambda_n}\left(\dfrac{\partial u_n}{\partial x} + \dfrac{\partial v_n}{\partial y}\right). \end{cases} \tag{1.20}$$

For $n = 0$, $w_0 = \psi_0 = 0$ and there remains

$$\begin{cases} \dfrac{\partial u_0}{\partial t} + \overline{U}_0\dfrac{\partial u_0}{\partial x} - fv_0 + \dfrac{\partial \phi_0}{\partial x} = 0, \\ \\ \dfrac{\partial v_0}{\partial t} + \overline{U}_0\dfrac{\partial v_0}{\partial x} + fu_0 + \dfrac{\partial \phi_0}{\partial y} = 0, \\ \\ \dfrac{\partial u_0}{\partial x} + \dfrac{\partial v_0}{\partial y} = 0. \end{cases} \tag{1.21}$$

Note that, since the considered problem is linear, there is no coupling between the equations of modes m and n for $m \neq n$.

In the sequel, we will always study the barotropic mode ($n = 0$) separately, and for $n \geq 1$, we use the last two equations (Eq. (1.20)) and rewrite the first three equations in the form

$$\begin{cases} \dfrac{\partial u_n}{\partial t} + \overline{U}_0 \dfrac{\partial u_n}{\partial x} - f v_n - \dfrac{1}{\lambda_n} \dfrac{\partial \psi_n}{\partial x} = 0, \\ \dfrac{\partial v_n}{\partial t} + \overline{U}_0 \dfrac{\partial v_n}{\partial x} + f u_n - \dfrac{1}{\lambda_n} \dfrac{\partial \psi_n}{\partial y} = 0, \\ \dfrac{\partial \psi_n}{\partial t} + \overline{U}_0 \dfrac{\partial \psi_n}{\partial x} - \dfrac{N^2}{\lambda_n} \left(\dfrac{\partial u_n}{\partial x} + \dfrac{\partial v_n}{\partial y} \right) = 0. \end{cases} \tag{1.22}$$

As indicated before, our aim is to propose boundary conditions for Eqs. (1.20)–(1.22) which make these equations well-posed and consequently the Eq. (1.13) also. As we shall see in Sections 2–4 below (see also Rousseau, Temam and Tribbia [2005b]), the boundary conditions are different depending on whether

$$1 \leq n \leq n_c, \quad \text{or} \quad n > n_c,$$

where n_c and λ_{n_c} are such that

$$\frac{n_c \pi}{L_3} = \lambda_{n_c} < \frac{N}{\overline{U}_0} < \lambda_{n_c+1} = \frac{(n_c+1)\pi}{L_3}. \tag{1.23}$$

We will not study the nongeneric case, where $L_3 N / \pi \overline{U}_0$ is an *integer*, i.e., we will assume throughout that

$$\overline{U}_0 \neq \frac{N}{\lambda_n}, \ \forall n \geq 1, \text{ or equivalently } \frac{N L_3}{\pi \overline{U}_0} \notin \mathbb{N}. \tag{1.24}$$

The modes $0 \leq n \leq n_c$ are called *subcritical*, and the modes $n > n_c$ are called *supercritical*. It is convenient to introduce the sub- and supercritical components of the functions defined by

$$u^0 = P_0 u = \mathcal{U}_0 u_0, u^I = P_I u = \sum_{n=1}^{n_c} \mathcal{U}_n u_n, \ \ u^{II} = P_{II} u = \sum_{n>n_c}^{\infty} \mathcal{U}_n u_n, \tag{1.25}$$

and similarly for all the other functions; of course, the zero mode u^0 is a subcritical mode, but as we will see, we need to treat it separately. With these notations, the Eqs. (1.13), (1.20), and (1.22) are equivalent to the following system:

$$\begin{cases} u^0_t + \bar{U}_0 u^0_x - f v^0 + \phi^0_x = 0, \\ v^0_t + \bar{U}_0 v^0_x + f u^0 + \phi^0_y = 0, \\ u^0_x + v^0_y = 0, \end{cases} \tag{1.26}$$

$$\begin{cases} u_t^I + \overline{U}_0 u_x^I - f v^I + \phi_x^I = 0, \\ v_t^I + \overline{U}_0 v_x^I + f u^I + \phi_y^I = 0, \\ \psi_t^I + \overline{U}_0 \psi_x^I + N^2 w^I = 0, \end{cases} \tag{1.27}$$

$$\begin{cases} u_t^{II} + \overline{U}_0 u_x^{II} - f v^{II} + \phi_x^{II} = 0, \\ v_t^{II} + \overline{U}_0 u_x^{II} + f u^{II} + \phi_y^{II} = 0, \\ \psi_t^{II} + \overline{U}_0 \psi_x^{II} + N^2 w^{II} = 0, \end{cases} \tag{1.28}$$

with the additional relations $\phi = \phi(\psi)$, $w = w(u, v)$:

$$\begin{cases} \phi^I = -\sum_{n=1}^{n_c} \dfrac{1}{\lambda_n} \psi_n \mathcal{U}_n, \quad w^I = -\sum_{n=1}^{n_c} \dfrac{1}{\lambda_n} (u_{nx} + v_{ny}) \mathcal{W}_n, \\ \phi^{II} = -\sum_{n>n_c} \dfrac{1}{\lambda_n} \psi_n \mathcal{U}_n, \quad w^{II} = -\sum_{n>n_c} \dfrac{1}{\lambda_n} (u_{nx} + v_{ny}) \mathcal{W}_n. \end{cases} \tag{1.29}$$

We will also set $U = (u, v, \psi)$, $U^0 = P_0 U$, $U^I = P_I U$, $U^{II} = P_{II} U$.

1.3. Position of the problem

Let us focus on Eq. (1.22) for the moment. Considering for the sake of simplicity that the functions do not depend on y (see Section 2 below), one can notice that the characteristic values of the resulting system are $(\overline{U}_0, \overline{U}_0 - N/\lambda_n, \overline{U}_0 + N/\lambda_n)$. Since $\overline{U}_0 > 0$, $N/\lambda_n > 0$, we always have at least two positive eigenvalues. But, $\overline{U}_0 - N/\lambda_n$ can be either positive or negative.[3] We say that the corresponding mode is supercritical in the first case and subcritical in the second case (see above); it appears then that the subcritical modes require two boundary conditions on the left of the domain ($x = 0$) and one boundary condition on the right ($x = L_1$), whereas the supercritical modes require three boundary conditions at $x = 0$. Based on this remark, OLIGER and SUNDSTRÖM [1978] concluded that the boundary value problem associated with Eq. (1.13) is ill-posed for any set of local boundary conditions (see also TEMAM and TRIBBIA [2003]).

Hereafter, our aim will be to study separately the subcritical and supercritical modes, proposing suitable boundary conditions for them, and to combine them and obtain existence, uniqueness, and regularity of solutions for the whole linearized problem. In each case, we will study one (subcritical/supercritical) mode separately and then combine them for the whole subcritical and supercritical components.

[3] In the ocean, taking the following values $\overline{U}_0 = 1\, ms^{-1}$, $N = 0.001\, s^{-1}$, and $L_3 = 1000\, m$, we end up with three subcritical modes. Realistic physical situations usually lead to a number n_c of subcritical modes between one and five (see TEMAM and TRIBBIA [2003] for more details).

2. Two-dimensional $x - z$ case

We start our study with the 2D (x, z) model for which both theoretical and numerical results have been established. We first rewrite the equations, assuming that the functions do not depend on the y variable, both in their nonlinear and linear formulations. Then, a well-posedness theorem is established in Section 2.3 for the linearized PEs. After introducing the numerical scheme that is used, we end this study of the 2D case with some numerical simulations for the linear and nonlinear cases that achieve two objectives. On the one hand, the absence of blow-up in these computations indicates that the nonlinear inviscid PEs are well-posed when supplemented with the boundary conditions that we propose. On the other hand, they show a very good coincidence on the subdomain Ω_1 of the two solutions, thus showing also the computational relevance of these new boundary conditions.[4]

2.1. Linear and nonlinear 2D primitive equations

Let us consider the nonlinear PEs without viscosity (Eq. (1.11)), without any dependance on the y variable:

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + w\frac{\partial u}{\partial z} - fv + \frac{\partial \phi}{\partial x} = F_u, \tag{2.1a}$$

$$\frac{\partial v}{\partial t} + u\frac{\partial v}{\partial x} + w\frac{\partial v}{\partial z} + fu = F_v, \tag{2.1b}$$

$$\frac{\partial \psi}{\partial t} + u\frac{\partial \psi}{\partial x} + (N^2 + \frac{\partial \psi}{\partial z})\, w = F_\psi, \tag{2.1c}$$

$$\frac{\partial \phi}{\partial z} = -\frac{\rho}{\rho_0}\, g = \psi, \tag{2.1d}$$

$$\frac{\partial u}{\partial x} + \frac{\partial w}{\partial z} = 0. \tag{2.1e}$$

We will consider the flow in the 2D domain $\mathcal{M} = (0, L_1) \times (-L_3, 0)$, and supplement Eq. (2.1) with an initial data u_0, v_0, ψ_0. The top and bottom boundary conditions (just imposed by kinematics) are the same as for the complete 3D problem (see Section 1 above):

$$w(x, z = -L_3, t) = w(x, z = 0, t) = 0, \quad \forall x \in (0, L_1), t > 0, \tag{2.2}$$

and we also have

$$\frac{\partial \phi}{\partial x}(x, z) = \phi_s'(x) - \int_z^0 \frac{\partial \psi}{\partial x}(x, z')\, \mathrm{d}z', \tag{2.3}$$

$$w(x, z) = \int_z^0 \frac{\partial u}{\partial x}(x, z')\, \mathrm{d}z', \tag{2.4}$$

[4] See Section 2.4 for the full description of this numerical test.

where $\phi_s(x, t) = \phi(x, z = 0, t)$ is the surface pressure divided by ρ_0, and ϕ_s' its derivative with respect to x.

The 2D PEs (Eq. (2.1)), linearized around the stratified flow $\overline{\mathbf{v}}_0 = \overline{U}_0\, e_x, \overline{\rho}, \overline{T}$, $\overline{p}$, read:

$$\frac{\partial u}{\partial t} + \overline{U}_0 \frac{\partial u}{\partial x} - fv + \frac{\partial \phi}{\partial x} = F_u, \tag{2.5a}$$

$$\frac{\partial v}{\partial t} + \overline{U}_0 \frac{\partial v}{\partial x} + fu = F_v - f\,\overline{U}_0, \tag{2.5b}$$

$$\frac{\partial \psi}{\partial t} + \overline{U}_0 \frac{\partial \psi}{\partial x} + N^2\, w = F_\psi, \tag{2.5c}$$

$$\frac{\partial \phi}{\partial z} = -\frac{\rho}{\rho_0}\, g = \psi, \tag{2.5d}$$

$$\frac{\partial u}{\partial x} + \frac{\partial w}{\partial z} = 0. \tag{2.5e}$$

The aim of this section is to consider some lateral boundary conditions at $x = 0$ and $x = L_1$ that are both physically reasonable and computationally satisfying, and that lead to the well-posedness of the problem (Eq. (2.5)).

2.2. The modal equations and boundary conditions

We consider a normal mode decomposition of the solution of the following form (see Section 1 above for the details and the justifications):

$$(u, v, \phi) = \sum_{n \geq 0} \mathcal{U}_n(z)\, (u_n, v_n, \phi_n)\, (x, t), \tag{2.6}$$

$$(w, \psi) = \sum_{n \geq 1} \mathcal{W}_n(z)\, (w_n, \psi_n)\, (x, t). \tag{2.7}$$

We now introduce the expansion in Eqs. (2.6) and (2.7) into Eq. (2.5). We multiply Eqs. (2.5a), (2.5b), and (2.5e) by $\mathcal{U}_n$ and Eqs. (2.5c) and (2.5d) by $\mathcal{W}_n$ and integrate on $(-L_3, 0)$, and we find the same Eq. (1.20) but with no dependence on the y variable, namely, for $n \geq 1$:

$$\begin{cases} \dfrac{\partial u_n}{\partial t} + \overline{U}_0 \dfrac{\partial u_n}{\partial x} - fv_n + \dfrac{\partial \phi_n}{\partial x} = F_{u,n}, \\ \dfrac{\partial v_n}{\partial t} + \overline{U}_0 \dfrac{\partial v_n}{\partial x} + fu_n = F_{v,n}, \\ \dfrac{\partial \psi_n}{\partial t} + \overline{U}_0 \dfrac{\partial \psi_n}{\partial x} + N^2\, w_n = F_{\psi,n}, \\ \phi_n = -\dfrac{1}{\lambda_n}\psi_n, \\ w_n = -\dfrac{1}{\lambda_n}\dfrac{\partial u_n}{\partial x}. \end{cases} \tag{2.8}$$

The constant mode in $z(n = 0)$ is different (simpler), and we postpone its study to Section 2.3.5 below. Taking into account the last two equations of Eq. (2.8), the first three become:

$$\begin{cases} \dfrac{\partial u_n}{\partial t} + \overline{U}_0 \dfrac{\partial u_n}{\partial x} - f\, v_n - \dfrac{1}{\lambda_n} \dfrac{\partial \psi_n}{\partial x} = F_{u,n}, \\ \dfrac{\partial v_n}{\partial t} + \overline{U}_0 \dfrac{\partial v_n}{\partial x} + f\, u_n = F_{v,n}, \\ \dfrac{\partial \psi_n}{\partial t} + \overline{U}_0 \dfrac{\partial \psi_n}{\partial x} - \dfrac{N^2}{\lambda_n} \dfrac{\partial u_n}{\partial x} = F_{\psi,n}. \end{cases} \tag{2.9}$$

For the nonlinear PEs, the normal modes decomposition reads

$$\frac{\partial u_n}{\partial t} + \overline{U}_0 \frac{\partial u_n}{\partial x} - f v_n + \frac{\partial \phi_n}{\partial x} + B_{u,n}(U) = F_{u,n}, \tag{2.10a}$$

$$\frac{\partial v_n}{\partial t} + \overline{U}_0 \frac{\partial v_n}{\partial x} + f u_n + B_{v,n}(U) = F_{v,n}, \tag{2.10b}$$

$$\frac{\partial \psi_n}{\partial t} + \overline{U}_0 \frac{\partial \psi_n}{\partial x} + N^2 w_n + B_{\psi,n}(U) = F_{\psi,n}, \tag{2.10c}$$

where $B_{u,n}$, $B_{v,n}$, and $B_{\psi,n}$ are the following modal parts of the nonlinearities:

$$B_{u,n} = \int_{-L_3}^{0} \left(u \frac{\partial u}{\partial x} + w \frac{\partial u}{\partial z} \right) \mathcal{U}_n \mathrm{d}z, \tag{2.11a}$$

$$B_{v,n} = \int_{-L_3}^{0} \left(u \frac{\partial v}{\partial x} + w \frac{\partial v}{\partial z} \right) \mathcal{U}_n \mathrm{d}z, \tag{2.11b}$$

$$B_{\psi,n} = \int_{-L_3}^{0} \left(u \frac{\partial \psi}{\partial x} + w \frac{\partial \psi}{\partial z} \right) \mathcal{W}_n \mathrm{d}z. \tag{2.11c}$$

with u, v, ψ, w truncated to M modes.

Let us now introduce the lateral boundary conditions which, for each $n \geq 1$, will supplement this system. We recall (see Section 1.3 above) that lateral boundary conditions at $x = 0$ and $x = L_1$ cannot be imposed without separating the subcritical and supercritical modes.

The boundary conditions for the subcritical modes were discussed in Rousseau, Temam and Tribbia [2004], Rousseau, Temam and Tribbia [2005a], they are recalled below. The boundary conditions for the supercritical modes are less problematic, we now present them. For $n > n_c$, a set of natural boundary conditions for system (Eq. (2.9)) is

$$\begin{cases} u_n(0, t) = 0, \\ v_n(0, t) = 0, \\ \psi_n(0, t) = 0. \end{cases} \tag{2.12}$$

In Eqs. (2.12) and (2.14), we chose, for simplicity, homogeneous boundary conditions, but we discuss in Section 2.3.4 below the case of nonzero boundary values.

For $1 \le n \le n_c, \overline{U}_0 - N/\lambda_n < 0$, the corresponding eigenvector is $\eta_n = u_n + \psi_n/N$. The eigenvectors related to $\overline{U}_0$ and $\overline{U}_0 + N/\lambda_n$ are, respectively, v_n and $\xi_n = u_n - \psi_n/N$. Thanks to (Eq. (1.20)), we have, for $n \ge 1$, $(\xi_n, \eta_n) = (u_n + \lambda_n \phi_n, u_n - \lambda_n \phi_n)$.

Using the variables ξ_n, v_n, η_n, we rewrite Eq. (2.9) as follows:

$$\begin{cases} \dfrac{\partial \xi_n}{\partial t} + (\overline{U}_0 + \dfrac{N}{\lambda_n}) \dfrac{\partial \xi_n}{\partial x} - f v_n = F_{\xi,n}, \\ \dfrac{\partial v_n}{\partial t} + \overline{U}_0 \dfrac{\partial v_n}{\partial x} + \dfrac{1}{2} f(\xi_n + \eta_n) = F_{v,n}, \\ \dfrac{\partial \eta_n}{\partial t} + (\overline{U}_0 - \dfrac{N}{\lambda_n}) \dfrac{\partial \eta_n}{\partial x} = F_{\eta,n}. \end{cases} \tag{2.13}$$

Hence, for these subcritical modes ($n \le n_c$), a set of natural and nonreflective boundary conditions is the following

$$\begin{cases} \xi_n(0, t) = 0, \\ v_n(0, t) = 0, \\ \eta_n(L_1, t) = 0. \end{cases} \tag{2.14}$$

In Section 2.3, we will prove the well-posedness of the linear PEs (Eq. (2.5)) (equivalent mode by mode to Eq. (2.8)) with the modal boundary conditions in Eq. (2.12) and (2.14).

For the nonlinear case, we perform the same change of variables $\xi_n = u_n - \psi_n/N$ $\eta_n = u_n + \psi_n/N$, and obtain the nonlinear version of Eq. (2.13), namely:

$$\frac{\partial \xi_n}{\partial t} + (\overline{U}_0 + \frac{N}{\lambda_n}) \frac{\partial \xi_n}{\partial x} - f v_n + B_{\xi,n}(U) = F_{\xi,n}, \tag{2.15a}$$

$$\frac{\partial v_n}{\partial t} + \overline{U}_0 \frac{\partial v_n}{\partial x} + f \frac{\xi_n + \eta_n}{2} + B_{v,n}(U) = F_{v,n}, \tag{2.15b}$$

$$\frac{\partial \eta_n}{\partial t} + (\overline{U}_0 - \frac{N}{\lambda_n}) \frac{\partial \eta_n}{\partial x} - f v_n + B_{\eta,n}(U) = F_{\eta,n}, \tag{2.15c}$$

where $B_{\xi,n} = B_{u,n} - B_{\psi,n}/N$ and $B_{\eta,n} = B_{u,n} + B_{\psi,n}/N$.

We assume in the following that the initial data is such that the nonlinear part is small as compared to the stratified flow $(\overline{U}_0, 0, 0)$ so that the characteristic values do not change sign, at least during a certain period of time. Assuming so, we conjecture that the boundary conditions provided for the linearized system will give a well-posed problem for the nonlinear equations, at least for some time. We leave the theoretical analysis to subsequent studies, and perform hereafter the corresponding numerical simulations based on this hypothesis, which is conforted by the lack of numerical blow-up. Hence, the boundary conditions that we consider for the nonlinear case are also Eqs. (2.12)–(2.14).

2.3. Well-posedness results

We aim to implement the boundary conditions in Eqs. (2.12) and (2.14) in the linear case, and we first set the functional framework appropriate for these boundary conditions.

2.3.1. Theoretical framework

We aim to write the initial value problem under consideration as a functional evolution in an appropriate Hilbert space H:

$$\begin{cases} \dfrac{\mathrm{d}U}{\mathrm{d}t} + A\,U = F, \\ U(0) = U_0. \end{cases} \tag{2.16}$$

Here A is an unbounded operator with domain $D(A) \subset H$, the forcing F taking its values in H and the initial data $U_0 \in D(A)$ are given.

We define H by setting

$$H = H_u \times H_v \times H_\psi, \tag{2.17}$$

$$H_u = \left\{ u \in L^2(\mathcal{M}) \;/\; \int_{-L_3}^{0} u(x,z)\,\mathrm{d}z = 0 \text{ a.e. on } (0, L_1) \right\},$$

$$H_v = H_\psi = L^2(\mathcal{M}),$$

where $\mathcal{M}$ is the 2D domain $(0, L_1) \times (-L_3, 0)$. We endow H with the scalar product[5]

$$(U, \widetilde{U})_H = \int_{\mathcal{M}} (u\,\widetilde{u} + v\,\widetilde{v} + \frac{1}{N^2}\,\psi\,\widetilde{\psi})\,\mathrm{d}\mathcal{M}, \qquad \forall (U, \widetilde{U}) \in H^2, \tag{2.18}$$

and the associated norm

$$|\,U\,|_H = \{(U,U)_H\}^{1/2}, \qquad \forall U \in H. \tag{2.19}$$

The space H_u is clearly closed in $L^2(\mathcal{M})$, and $H = H_u \times H_v \times H_\psi$ is a closed subspace of $(L^2(\mathcal{M}))^3$, which we endow with the scalar product and norm derived from Eq. (2.18) and equivalent to those of $(L^2(\mathcal{M}))^3$. We denote by P the orthogonal projector from $L^2(\mathcal{M})$ onto H_u. For every $g \in L^2(\mathcal{M})$,

$$P(g)(x,z) = g(x,z) - \frac{1}{L_3} \int_{-L_3}^{0} g(x,z')\,\mathrm{d}z', \tag{2.20}$$

$$(I-P)(g)(x,z) = \frac{1}{L_3} \int_{-L_3}^{0} g(x,z')\,\mathrm{d}z'. \tag{2.21}$$

It is easily checked that $Pg \in H_u$ and $(I-P)g \perp Pg$. Finally, $H_u^\perp$ is identical to $L_x^2(0, L_1)$. Indeed, for $g \in H_u^\perp$, $(I-P)\,g = g$ so that g does not depend on z and belongs to $L_x^2(0, L_1)$. Conversely, if $h \in L_x^2(0, L_1)$, then for every $u \in H_u$, $(u,h)_{L^2(\mathcal{M})} = \int_0^{L_1} h(x) \int_{-L_3}^{0} u(x,z)dz\,dx = 0$ and $h \in H_u^\perp$.

[5]It is not surprising to have $1/N^2$ as a multiplicative coefficient in front of the last term of $(U, \widetilde{U})_H$, since $\int_{\mathcal{M}} (u^2 + v^2) d\mathcal{M}$ represents the kinetic energy whereas $N^{-2} \int_{\mathcal{M}} \psi^2\,d\mathcal{M}$ is the available potential energy so that the square of the norm in H represents the total energy.

We are now in position to define the operator A; its domain $D(A)$ is defined by

$$D(A) = \left\{ U = (u, v, \psi) \in H \ / \ \left| \begin{array}{l} (u_x, v_x, \psi_x) \in L^2(\mathcal{M}) \\ (u, v, \psi) \text{ verify (2.23) and (2.24)} \end{array} \right. \right\}. \quad (2.22)$$

Here and in the sequel, u_x, u_z denote the partial derivatives $\partial u/\partial x$, $\partial u/\partial z$ of a function u.

The boundary conditions in Eqs. (2.23) and (2.24), identical to Eqs. (2.12) and (2.14), are written in the following form.[6] For the subcritical modes ($1 \leq n \leq n_c$):

$$\left\{ \begin{array}{l} \displaystyle\int_{-L_3}^{0} u(0, z)\, \mathcal{U}_n(z) \mathrm{d}z - \frac{1}{N} \int_{-L_3}^{0} \psi(0, z)\, \mathcal{W}_n(z)\, \mathrm{d}z = 0, \\ \displaystyle\int_{-L_3}^{0} v(0, z)\, \mathcal{U}_n(z)\, \mathrm{d}z = 0, \\ \displaystyle\int_{-L_3}^{0} u(L_1, z)\, \mathcal{U}_n(z)\, \mathrm{d}z + \frac{1}{N} \int_{-L_3}^{0} \psi(L_1, z)\, \mathcal{W}_n(z)\, \mathrm{d}z = 0, \end{array} \right. \quad (2.23)$$

and for the supercritical ones ($n > n_c$):

$$\left\{ \begin{array}{l} \displaystyle\int_{-L_3}^{0} u(0, z)\, \mathcal{U}_n(z)\, \mathrm{d}z = 0, \\ \displaystyle\int_{-L_3}^{0} v(0, z)\, \mathcal{U}_n(z)\, \mathrm{d}z = 0, \\ \displaystyle\int_{-L_3}^{0} \psi(0, z)\, \mathcal{W}_n(z)\, \mathrm{d}z = 0, \end{array} \right. \quad (2.24)$$

For every $U = (u, v, \psi) \in D(A)$, AU is given by

$$AU = \begin{pmatrix} P\big[\overline{U}_0\, u_x - fv - \displaystyle\int_z^0 \psi_x(x, z')\, \mathrm{d}z' \big] \\ \overline{U}_0\, v_x + fu \\ \overline{U}_0\, \psi_x + N^2\, w \end{pmatrix}, \quad (2.25)$$

where $w = w(u)$ is given by Eq. (2.4).

We now intend to prove the well-posedness of Eq. (2.16), corresponding to the linearized PEs supplemented with the boundary conditions in Eqs. (2.23) and (2.24), in the context of the linear semi-group theory.

[6] We note that the boundary conditions on v do not depend on the modes (see also the boundary condition on the constant mode v_0 in Section 2.3.5 below); hence, they could be written in the form $v(0, z) = 0, \ \forall z \in (-L_3, 0)$. However, we keep the modal notation by analogy with the other functions u and ψ, and because this is the way this boundary condition is actually implemented in numerical simulations, see Section 2.5 below.

2.3.2. Main theorem

To prove the well-posedness of the initial value problem (Eq. (2.16)), we will use the following version of the Hille–Yosida theorem borrowed from Burq and Gérard [2003] (see also Brézis [1973], Henry [1981], Lions [1965], Pazy [1983], Yosida [1980]):

Theorem 2.1. *(Hille–Yosida theorem) Let H be a Hilbert space and let A: $D(A) \longrightarrow H$ be a linear unbounded operator, with domain $D(A) \subset H$. Assume the following:*

(i) $D(A)$ *is dense in* H *and* A *is closed,*
(ii) A *is* ≥ 0, *i.e.,* $(AU, U)_H \geq 0, \quad \forall U \in D(A)$,
(iii) $\exists \mu_0 > 0$, *such that* $A + \mu_0 I$ *is onto.*

Then, $-A$ is infinitesimal generator of a semigroup of contractions $\{S(t)\}_{t\geq 0}$ in H, and for every $U_0 \in H$ and $F \in L^1(0, T; H)$, there exists a unique solution $U \in \mathcal{C}([0, T]; H)$ of Eq. (2.16)*,*

$$U(t) = S(t)\, U_0 + \int_0^t S(t-s)\, F(s)\, \mathrm{d}s. \tag{2.26}$$

If, furthermore, $U_0 \in D(A)$ and $F' = \mathrm{d}F/\mathrm{d}t \in L^1(0, T; H)$, then U satisfies (2.16) *and*

$$U \in \mathcal{C}([0, T]; H) \cap L^\infty(0, T; D(A)), \quad \frac{\mathrm{d}U}{\mathrm{d}t} \in L^\infty(0, T; H). \tag{2.27}$$

We now state and prove the main result for the homogeneous boundary conditions in 2D.

Theorem 2.2. *Let H be the Hilbert space defined in Eq. (2.17) and A be the linear operator defined in Eq. (2.25) corresponding to the linearized PEs with vanishing viscosity and homogeneous modal boundary conditions.*

Then, the initial value problem Eq. (2.16), corresponding to Eq. (2.5) supplemented with the boundary conditions in Eqs. (2.23) and (2.24), is well-posed, i.e., for every initial data $U_0 \in D(A)$ and forcing $F \in L^1(0, T; H)$, there exists a unique solution $U \in \mathcal{C}([0, T]; H)$ of Eq. (2.16).

2.3.3. Proof of Theorem 2.2

We want to apply Theorem 2.1 to Eq. (2.16). To this aim, we verify the hypotheses (i), (ii), and (iii) of the Hille–Yosida theorem (Theorem 2.1); we start with (ii) and (iii) and postpone the proof of (i) to Lemma 2.3 below. We start with the proof of (ii):

Lemma 2.1. *For every $U \in D(A)$, $(A\,U, U)_H \geq 0$.*

PROOF. For any $U \in H$, let us compute the scalar product $(A\,U, U)_H$:

$$(A\,U, U)_H = \int_{\mathcal{M}} P(\overline{U}_0\, u_x - fv - \int_z^0 \psi_x(x, z')\,\mathrm{d}z')\, u\mathrm{d}\mathcal{M} + \int_{\mathcal{M}} (\overline{U}_0\, v_x + fu)\, v\mathrm{d}\mathcal{M} + \int_{\mathcal{M}} (\overline{U}_0\, \psi_x + N^2\, w)\, \frac{\psi}{N^2}\,\mathrm{d}\mathcal{M}.$$

Since $u \in H_u$, we have, using Eq. (2.4):

$$\begin{aligned}(A\,U, U)_H &= \int_{\mathcal{M}} (\overline{U}_0\, u_x - fv - \int_z^0 \psi_x(x, z')\,\mathrm{d}z')\, u\mathrm{d}\mathcal{M} \\ &\quad + \int_{\mathcal{M}} (\overline{U}_0\, v_x + fu)\, v\,\mathrm{d}\mathcal{M} \\ &\quad + \int_{\mathcal{M}} (\overline{U}_0\, \psi_x + N^2\, w)\, \frac{\psi}{N^2}\mathrm{d}\mathcal{M} \\ &= \int_{-L_3}^0 \frac{\overline{U}_0}{2} \Big(u^2(L_1) - u^2(0) + v^2(L_1) - v^2(0) \\ &\quad + \frac{1}{N^2}\,\psi^2(L_1) - \frac{1}{N^2}\,\psi^2(0)\Big)\,\mathrm{d}z \\ &\quad - \int_{\mathcal{M}} \Big\{u(x, z) \int_z^0 \psi_x(x, z')\,\mathrm{d}z' \\ &\quad - \psi(x, z) \int_z^0 u_x(x, z')\,\mathrm{d}z'\Big\}\,\mathrm{d}x\mathrm{d}z.\end{aligned}$$

Here $u(L_1)$, $u(0)$ stand for $u(L_1, z)$, $u(0, z)$, etc. Using the expansion in Eqs. (2.6), (2.7) with Eq. (1.18), it is easy to check that

$$\begin{cases} -\displaystyle\int_z^0 \psi_x(x, z')\,\mathrm{d}z' = \sum_{n\geq 1} \frac{\psi_{n\,x}(x)}{\lambda_n}\,(1 - \mathcal{U}_n(z)) \\ \qquad\qquad = \theta(x) - \displaystyle\sum_{n\geq 1} \frac{\psi_{n\,x}(x)}{\lambda_n}\,\mathcal{U}_n(z), \\ \displaystyle\int_z^0 u_x(x, z')\,\mathrm{d}z' = -\sum_{n\geq 1} \frac{u_{n\,x}(x)}{\lambda_n}\,\mathcal{W}_n(z), \end{cases} \tag{2.28}$$

where $\theta = \theta(x)$ is an L^2-function depending only on x.

Using again the expansion in Eqs. (2.6), (2.7) and remembering that $u \in H_u$, the integral $\int_{\mathcal{M}} u\,\theta\,\mathrm{d}\mathcal{M}$ vanishes, and we find:

$$
\begin{aligned}
(A\,U,U)_H = \sum_{n\geq 1} \frac{\overline{U}_0}{2} \Big(& u_n^2(L_1) - u_n^2(0) + v_n^2(L_1) - v_n^2(0) \\
& + \frac{1}{N^2}\,\psi_n^2(L_1) - \frac{1}{N^2}\,\psi_n^2(0) \Big) + \frac{\overline{U}_0}{2} \Big(v_0^2(L_1) - v_0^2(0) \Big) \\
& - \sum_{n\geq 1} \frac{1}{\lambda_n} \int_0^{L_1} (\psi_{n\,x}\,u_n + \psi_n\,u_{n\,x})\,\mathrm{d}x.
\end{aligned}
$$

Using the boundary conditions in Eq. (2.23) for the subcritical modes and Eq. (2.24) for the supercritical ones, we find:

$$
\begin{aligned}
(A\,U,U)_H = & \sum_{1\leq n\leq n_c} \frac{\overline{U}_0}{2} \Big(u_n^2(L_1) - u_n^2(0) + v_n^2(L_1) + u_n^2(L_1) - u_n^2(0) \Big) \\
& + \frac{\overline{U}_0}{2}\,v_0^2(L_1) + \sum_{1\leq n\leq n_c} \frac{N}{\lambda_n} \Big(u_n^2(L_1) + u_n^2(0) \Big) \\
& + \sum_{n>n_c} \frac{\overline{U}_0}{2} \Big(u_n^2(L_1) + v_n^2(L_1) + \frac{1}{N^2}\,\psi_n^2(L_1) \Big) \\
& - \sum_{n>n_c} \frac{1}{\lambda_n}\,u_n(L_1)\,\psi_n(L_1).
\end{aligned}
$$

For every subcritical mode (when $n \leq n_c$):

$$
\begin{aligned}
& \overline{U}_0 \Big(u_n^2(L_1) - u_n^2(0) + \frac{1}{2}\,v_n^2(L_1) \Big) + \frac{N}{\lambda_n} \Big(u_n^2(L_1) + u_n^2(0) \Big) \\
& \qquad = (\overline{U}_0 + \frac{N}{\lambda_n})\,u_n^2(L_1) + \frac{\overline{U}_0}{2}\,v_n^2(L_1) + (\frac{N}{\lambda_n} - \overline{U}_0)\,u_n^2(0) \geq 0,
\end{aligned}
$$

the latter quantity is nonnegative, thanks to the definition of n_c. For every supercritical mode (when $n > n_c$):

$$
\begin{aligned}
& \frac{\overline{U}_0}{2} \Big(u_n^2(L_1) + v_n^2(L_1) + \frac{1}{N^2}\,\psi_n^2(L_1) \Big) - \frac{1}{\lambda_n}\,u_n(L_1)\,\psi_n(L_1) \\
& = \frac{\overline{U}_0}{2}\,v_n^2(L_1) + \frac{\overline{U}_0}{2} \Big(u_n(L_1) - \frac{N}{\overline{U}_0\,\lambda_n}\,\psi_n(L_1) \Big)^2 \\
& \qquad + \frac{\overline{U}_0}{2\,N^2} \Big(1 - \frac{N^2}{\overline{U}_0^2\,\lambda_n^2} \Big)\,\psi_n^2(L_1) \geq 0.
\end{aligned}
$$

This quantity is also nonnegative, which achieves the proof of Lemma 2.1. □

Due to Eq. (1.24), which is assumed throughout, we can choose μ_0 such that

$$\mu_0 \neq f^2\,(1 - \overline{U}_0^2\,\lambda_n^2/N^2), \quad \forall n \geq 1, \tag{2.29}$$

$$\mu_0 \neq f^2\,\overline{U}_0^2\,\lambda_n^2/N, \quad \forall n \geq 1. \tag{2.30}$$

With this choice of μ_0, we can prove the following lemma:

LEMMA 2.2. *The operator $A + \mu_0 I$ is onto from $D(A)$ onto H, where μ_o satisfies Eqs.* (2.29) *and* (2.30).

PROOF. For μ_0 as indicated, we are given $F = (F_u, F_v, F_\psi)$ in H, and we look for $U = (u, v, \psi)$ in $D(A)$ such that $(A + \mu_0\, I)\, U = F$. Writing this equation componentwise, we find:

$$\begin{cases} \overline{U}_0\, u_x(x,z) - f\, v(x,z) + \mu_0\, u(x,z) \\ \qquad\qquad - \displaystyle\int_z^0 \psi_x(x,z')\,\mathrm{d}z' + \phi_s'(x) = F_u(x,z), \\ \overline{U}_0\, v_x(x,z) + f\, u(x,z) + \mu_0\, v(x,z) = F_v(x,z), \\ \overline{U}_0\, \psi_x(x,z) + N^2\, w(x,z) + \mu_0\, \psi(x,z) = F_\psi(x,z). \end{cases} \tag{2.31}$$

To obtain the modal equations corresponding to Eq. (2.31), we multiply the three equations by $\mathcal{U}_n, \mathcal{U}_n$, and $\mathcal{W}_n$, respectively, and integrate on $(-L_3, 0)$.

Of course, since $F = (F_u, F_v, F_\psi) \in H$, we also have the following modal decompositions:

$$\begin{cases} F_u(x,z) = \displaystyle\sum_{n\geq 1} \mathcal{U}_n(z)\, F_{u,n}(x), \\ F_v(x,z) = \displaystyle\sum_{n\geq 0} \mathcal{U}_n(z)\, F_{v,n}(x), \\ F_\psi(x,z) = \displaystyle\sum_{n\geq 1} \mathcal{W}_n(z)\, F_{\psi,n}(x). \end{cases} \tag{2.32}$$

Note that for F as for U, since $F_u \in H_u \subset L^2(\mathcal{M})$, $F_{u,0} = 0$ and the decomposition of F_u starts from $n = 1$.

For the barotropic mode $n = 0$ (constant in the variable z), we only consider the first two equations since multiplying the third one by $\mathcal{W}_0 = 0$ would be useless. Integrating the equation for v and reporting in the equation for u (in which $u_0 = 0$, see above), we find v_0 and the surface pressure ϕ_s, up to an additive constant $\phi_s(0)$:

$$\begin{cases} v_0(x) = \dfrac{1}{\overline{U}_0} \displaystyle\int_0^x F_{v\,0}(x')\, e^{(x'-x)\,\mu_0/\overline{U}_0}\mathrm{d}x', \\ \phi_s(x) = \phi_s(0) + \displaystyle\int_0^x \Big(f\, v_0(x') - \frac{L_3^2}{\pi} \sum_{n\geq 1} \psi_{n\,x}(x') \Big)\mathrm{d}x'. \end{cases} \tag{2.33}$$

Naturally, the above expression of ϕ_s depends on the other modes ($n \geq 1$). We now write the equations, derived from Eq. (2.31) mode by mode:

$$\begin{cases} \overline{U}_0\, u_{n\,x} - f v_n + \mu_0\, u_n - \dfrac{1}{\lambda_n}\, \psi_{n\,x} = F_{u,n}, \\ \overline{U}_0\, v_{n\,x} + f u_n + \mu_0\, v_n = F_{v,n}, \\ \overline{U}_0\, \psi_{n\,x} - \dfrac{N^2}{\lambda_n}\, u_{n\,x} + \mu_0\, \psi_n = F_{\psi,n}. \end{cases} \tag{2.34}$$

We recall that the functions (u_n, v_n, ψ_n) only depend on the x variable. Hence, Eq. (2.34) is just a linear system of ordinary differential equations for u_n, v_n, ψ_n.

As usual, to solve Eq. (2.34), we first consider the corresponding homogeneous system. Dropping the subscripts n for the moment, we write:

$$\begin{cases} \overline{U}_0\, \dfrac{\partial^2 u}{\partial x^2} - f v - \dfrac{1}{\lambda}\, \dfrac{\partial^2 \psi}{\partial x^2} + \mu_0\, u = 0, \\ \overline{U}_0\, \dfrac{\partial^2 v}{\partial x^2} + f u + \mu_0\, v = 0, \\ \overline{U}_0\, \dfrac{\partial^2 \psi}{\partial x^2} - \dfrac{N^2}{\lambda}\, \dfrac{\partial^2 u}{\partial x^2} + \mu_0\, \psi = 0. \end{cases} \tag{2.35}$$

The general solution of this linear system is of the form

$$(u, v, \psi) = \sum_{i=1}^{3} (A_i, B_i, C_i)\, e^{R_i\, x}, \tag{2.36}$$

where the coefficients R_i are as follows:

$$\begin{cases} R_1 = -\dfrac{\mu_0}{\overline{U}_0}, \\ R_2 = \dfrac{-\mu_0\, \overline{U}_0 + \dfrac{N}{\lambda}\Big(\mu_0^2 - f^2\, (\overline{U}_0^2\, \lambda^2/N^2 - 1)\Big)^{1/2}}{\overline{U}_0^2 - \dfrac{N^2}{\lambda^2}}, \\ R_3 = \dfrac{-\mu_0\, \overline{U}_0 - \dfrac{N}{\lambda}\Big(\mu_0^2 - f^2\, (\overline{U}_0^2\, \lambda^2/N^2 - 1)\Big)^{1/2}}{\overline{U}_0^2 - \dfrac{N^2}{\lambda^2}}. \end{cases} \tag{2.37}$$

The coefficients $(A_i, B_i, C_i)_{1 \leq i \leq 3}$ satisfy the equations:

$$\begin{cases} A_i = a_i\, B_i, \\ C_i = c_i\, B_i, \end{cases} \tag{2.38}$$

with

$$\begin{cases} a_1 = 0, \\ c_1 = -\dfrac{f\, \lambda}{R_1}, \end{cases} \tag{2.39}$$

and, for $i = 2, 3$:

$$\begin{cases} a_i = -\dfrac{\overline{U}_0\, R_i + \mu_0}{f}, \\ c_i = \dfrac{N^2\, R_i}{\lambda\,(\overline{U}_0\, R_i + \mu_0)}. \end{cases} \tag{2.40}$$

Now, returning to the nonhomogeneous system Eq. (2.34), we look for a solution $(u_n, v_n, \psi_n) = (u, v, \psi)$ of the form:

$$Y = (u, v, \psi)^T = \sum_{i=1}^{3} (a_i, 1, c_i,)^T\, B_i(x)\, e^{R_i\, x}, \tag{2.41}$$

where the (a_i, c_i) and R_i have been defined above. Equation (2.34) reads then:

$$M\, Y' + N\, Y = F, \tag{2.42}$$

where

$$M = \begin{pmatrix} \overline{U}_0 & 0 & -\frac{1}{\lambda} \\ 0 & \overline{U}_0 & 0 \\ -\frac{N^2}{\lambda} & 0 & \overline{U}_0 \end{pmatrix}, \quad N = \begin{pmatrix} \mu_0 & -f & 0 \\ f & \mu_0 & 0 \\ 0 & 0 & \mu_0 \end{pmatrix}, \tag{2.43}$$

$$F = (F_u, F_v, F_\psi)^T. \tag{2.44}$$

Thanks to assumption Eq. (1.24), $\overline{U}_0 \neq N/\lambda_n$, the matrix M is regular and it can be inverted. Equation (2.42) then implies:

$$\sum_{i=1}^{3} (a_i, 1, c_i)^T\, B_i'(x)\, e^{R_i\, x} = M^{-1} F =: \widetilde{F}. \tag{2.45}$$

We now write the latter equation component by component. We find:

$$\Lambda(x).(B_1'(x), B_2'(x), B_3'(x))^T = (\widetilde{F}_1(x), \widetilde{F}_2(x), \widetilde{F}_3(x))^T, \tag{2.46}$$

with

$$\Lambda(x) = \begin{pmatrix} 0 & a_2\, e^{R_2\, x} & a_3\, e^{R_3\, x} \\ e^{R_1\, x} & e^{R_2\, x} & e^{R_3\, x} \\ c_1\, e^{R_1\, x} & c_2\, e^{R_2\, x} & c_3\, e^{R_3\, x} \end{pmatrix}. \tag{2.47}$$

Let us check that the matrix $\Lambda(x)$ is regular for every $x \in \mathbb{R}$; it is clearly sufficient to do so for $x = 0$, for which

$$\Lambda(0) = \begin{pmatrix} 0 & a_2 & a_3 \\ 1 & 1 & 1 \\ c_1 & c_2 & c_3 \end{pmatrix}.$$

We call $\mathbf{L}_1$, $\mathbf{L}_2$, and $\mathbf{L}_3$ the lines of $\Lambda(0)$. It is clear that $\mathbf{L}_1$ and $\mathbf{L}_2$ are linearly independent vectors. Then, if $\Lambda(0)$ were not regular, there would exist $(\alpha, \beta) \in \mathbb{R}^2$ such that $\mathbf{L}_3 = \alpha\,\mathbf{L}_1 + \beta\,\mathbf{L}_2$. After some easy computations, we would find that necessarily:

$$a_3\,(c_2 - c_1) = a_2\,(c_3 - c_1), \tag{2.48}$$

which leads (see Eqs. (2.39) and (2.40)) to

$$\overline{U}_0\,(R_3 - R_2)\,f^2\,\lambda^2/N^2 = -\mu_0\,R_1\,(R_3 - R_2). \tag{2.49}$$

From Eq. (2.29) we find that $R_2 \neq R_3$, and thanks to the definition of R_1, Eq. (2.49) becomes:

$$\overline{U}_0^2\,f^2\,\lambda^2/N^2 = \mu_0^2, \tag{2.50}$$

which contradicts Eq. (2.30). Thus, the matrix $\Lambda(x)$ is regular for every $x \in \mathbb{R}$.

Back to Eq. (2.46), and thanks to the latter result, the functions $B_i'(x)$ are uniquely determined for $i = 1, 2, 3$. It remains to use the modal boundary conditions to determine the constants $B_i(0)$ and thus the functions $B_i(x)$.

At this point, it is desirable to reintroduce the indices n, i.e., to return to the notation (u_n, v_n, ψ_n) since the boundary conditions depend on the mode considered. For the supercritical modes ($n > n_c$), the modal boundary condition is that in Eq. (2.12). We thus look for the $B_i(0)$ satisfying:

$$\begin{cases} a_2\,B_2(0) + a_3\,B_3(0) = 0, \\ B_1(0) + B_2(0) + B_3(0) = 0, \\ c_1\,B_1(0) + c_2\,B_2(0) + c_3\,B_3(0) = 0. \end{cases} \tag{2.51}$$

The matrix of this system is again $\Lambda(0)$, which was shown to be regular (see above). We conclude that the constants $B_i(0)$ are uniquely determined by Eq. (2.51) and equal to zero. The functions $B_i(x)$ for the supercritical modes ($n > n_c$) are now fully determined.

If $n \leq n_c$, the mode is subcritical and we consider the boundary condition in Eq. (2.14). We thus want to solve the following system:

$$\begin{cases} -N\,c_1\,B_1(0) + (a_2 - N\,c_2)\,B_2(0) + (a_3 - N\,c_3)\,B_3(0) = 0, \\ B_1(0) + B_2(0) + B_3(0) = 0, \\ N\,c_1\,B_1(0) + (a_2 + N\,c_2)\,B_2(0) + (a_3 + N\,c_3)\,B_3(0) = \Gamma, \end{cases} \tag{2.52}$$

where

$$\Gamma = -\sum_{i=1}^{3} \int_0^{L_1} (a_i + N\,c_i)\,B_i'(x)\,dx. \tag{2.53}$$

The quantity Γ depends only on the data and on the B_i'; hence, it is known at this stage. After some computations and using hypotheses (Eqs. (2.29) and (2.30)), we check that the matrix of the linear system (Eq. (2.52)) is regular (same proof exactly as for

$\Lambda(0)$). This achieves the determination of the B_i in the subcritical case, and the lemma is proved. □

REMARK 2.1. The case when there exists $n \geq 1$ such that $\overline{U}_0 = N/\lambda_n$ is slightly different and actually simpler since the third equation of Eq. (2.13) becomes $\partial \eta_n(x,t)/\partial t = F_{\eta,n}(t)$, which can be integrated directly. We note that no boundary condition (neither in the subcritical case nor in the supercritical one) would then be required for η_n so that Eqs. (2.23) and (2.24) would have to be modified. But, we do not want to go into the details since this nongeneric situation seldom occurs in numerical simulations.

To conclude, there remains to verify the hypothesis (i) of the Hille–Yosida theorem.

LEMMA 2.3. *The domain $D(A)$ of A is dense in H, and the operator A is closed.*

PROOF. We first verify that the orthogonal in H of $D(A)$, $D(A)^{\perp}$, is reduced to $\{0\}$.

Let v be an element of $D(A)^{\perp}$. Since $A + \mu_0 I$ is onto, there exists $u \in D(A)$ such that $(A + \mu_0 I)\, u = v$. Then,

$$0 = (v,u)_H = \Big((A + \mu_0 I)\, u, u \Big)_H \geq \mu_0 \mid u \mid_H^2;$$

hence, $u = v = 0$, which implies that $D(A)^{\perp} = \{0\}$, and $D(A)$ is dense in H.

To show that A is closed, we consider a sequence $(u_j, v_j, \psi_j) = U_j$ of $D(A)$ such that:

$$U_j \longrightarrow U \text{ in } H, \tag{2.54}$$

$$A\, U_j = F_j \longrightarrow F \text{ in } H, \tag{2.55}$$

and we want to verify that $U = (u, v, \psi) \in D(A)$ and $F = A\, U$ so that the graph of A is closed.

Thanks to Eq. (2.54), we know that

$$u_j \longrightarrow u \text{ in } H_u \subset L^2(\mathcal{M}), \tag{2.56}$$

$$v_j \longrightarrow v \text{ in } L^2(\mathcal{M}). \tag{2.57}$$

We also find from Eqs. (2.25) and (2.55) that

$$\overline{U}_0 \frac{\partial^2 v_j}{\partial x^2} + f u_j \longrightarrow F_2 \text{ in } L^2(\mathcal{M}). \tag{2.58}$$

Hence the sequence $(\partial v_j/\partial x)_{j \in \mathbb{N}}$ is bounded in $L^2(\mathcal{M})$, and thanks to Eq. (2.57), we obtain that $v_x \in L^2(\mathcal{M})$.

In view of proving that $(u_x, \psi_x) \in L^2(\mathcal{M})$, we consider the decomposition in normal modes. Thanks to Eq. (2.54), we have for every $n \geq 1$:

$$u_{j,n} \longrightarrow u_n \text{ in } L^2(0, L_1), \tag{2.59}$$

$$v_{j,n} \longrightarrow v_n \text{ in } L^2(0, L_1), \tag{2.60}$$

$$\psi_{j,n} \longrightarrow \psi_n \text{ in } L^2(0, L_1), \tag{2.61}$$

and the quantities $\sum_{n\geq 1} |u_{jn}|^2$, $\sum_{n\geq 1} |v_{jn}|^2$, and $\sum_{n\geq 1} |\psi_{jn}|^2$ are bounded uniformly in j.

Similarly, we infer from Eq. (2.55) that for every $n \geq 1$ the following convergences in $L^2(0, L_1)$:

$$\overline{U}_0 \frac{\partial^2 u_{j,n}}{\partial x^2} - f v_n - \frac{1}{\lambda_n} \frac{\partial^2 \psi_{j,n}}{\partial x^2} = F^j_{u,n} \longrightarrow F_{u,n}, \tag{2.62}$$

$$\overline{U}_0 \frac{\partial^2 v_{j,n}}{\partial x^2} + f u_n = F^j_{v,n} \longrightarrow F_{v,n}, \tag{2.63}$$

$$\overline{U}_0 \frac{\partial^2 \psi_{j,n}}{\partial x^2} - \frac{N^2}{\lambda_n} \frac{\partial^2 u_{j,n}}{\partial x^2} = F^j_{\psi,n} \longrightarrow F_{\psi,n}, \tag{2.64}$$

and the quantities $\sum_{n\geq 1} |F^j_{u,n}|^2$, $\sum_{n\geq 1} |F^j_{v,n}|^2$, and $\sum_{n\geq 1} |F^j_{\psi,n}|^2$ are bounded uniformly in j.

Combining Eqs. (2.62) and (2.64), we find that:

$$\frac{\partial^2 u_{j,n}}{\partial x^2} = \frac{1}{\overline{U}_0^2 - N^2/\lambda_n^2} \left(\overline{U}_0 F^j_{u,n} + f \overline{U}_0 v_{j,n} + \frac{F^j_{\psi,n}}{\lambda_n}\right); \tag{2.65}$$

hence, the $(\partial u_{j,n}/\partial x)_{j\geq 1}$ is bounded in $L^2(0, L_1)$ and $(du_n/dx) \in L^2(0, L_1)$. Moreover, we find that[7]

$$\sum_{n\geq 1} \left| \frac{\partial^2 u_{j,n}}{\partial x^2} \right|^2 \leq \frac{4}{\min\limits_{n\geq 1} |\overline{U}_0^2 - N^2/\lambda_n^2|^2} \sum_{n\geq 1} \left(\overline{U}_0^2 |F^j_{u,n}|^2 + f^2 \overline{U}_0^2 |v_{j,n}|^2 + \left| \frac{F^j_{\psi,n}}{\lambda_n} \right|^2\right) \tag{2.66}$$

so that the latter quantity is bounded uniformly in j. This guarantees that $u_x \in L^2(\mathcal{M})$. Following the same idea, and using either Eq. (2.62) or (2.64), we also prove that $\psi_x \in L^2(\mathcal{M})$.

To insure that $U \in D(A)$, we need to verify that the modal boundary conditions in Eqs. (2.12) and (2.14) are satisfied by U. This is clear since the convergence of $(u_{j,n}, v_{j,n}, \psi_{j,n})$ to (u_n, v_n, ψ_n) is in fact in $H^1(0, L_1)$ so that the boundary conditions pass to the limit.

Finally, let us show that $A\,U = F$. Thanks to Eq. (2.54), we find that $A\,U_j \to A\,U$ in the distribution sense in $\mathcal{M}$, hence $A\,U = F$ in the sense of distributions on $\mathcal{M}$. We infer from $U \in D(A)$ that $A\,U \in L^2(\mathcal{M})$, and conclude that $A\,U = F$ in $L^2(\mathcal{M})$, which ends the proof of Lemma 2.3. □

[7] Thanks to Eq. (1.24), we know that $\min\limits_{n\geq 1} |\overline{U}_0^2 - N^2/\lambda_n^2| > 0$.

2.3.4. The case of nonhomogeneous boundary conditions

In practical simulations, we want to solve the PEs with nonhomogeneous boundary conditions on U at $x = 0$ and $x = L_1$, i.e., U given respectively equal to $U^{g,l}$ and $U^{g,r}$. We assume that these boundary values are derived from a solution $\widetilde{U}$ given or computed on a domain $\widetilde{\mathcal{M}}$ larger than $\mathcal{M}$.[8]

We discussed in Section 2.3.3 above the case when $U^{g,l} = U^{g,r} = 0$. The issue is now to determine which components of $U^{g,l}$ and $U^{g,r}$ are needed to obtain a well-posed problem. In this context, all components of $U^{g,l}$ and $U^{g,r}$ are available, but we know (or surmise at this point) that they will not be all used, those used depending on the mode that we consider.

Based on the data $U^{g,l}$,$U^{g,r}$, let us now construct the following function $U^g = (u^g, v^g, \psi^g)$ depending on z and t and defined by

$$(u^g, v^g, \psi^g)(z,t) = \sum_{n\geq 1} \Big(u_n^g(t)\,\mathcal{U}_n(z),\, v_n^g(t)\mathcal{U}_n(z),\, \psi_n^g(t)\mathcal{W}_n(z) \Big), \tag{2.67}$$

where (u_n^g, v_n^g, ψ_n^g) are found using the boundary values $U^{g,l}$ and $U^{g,r}$ by

$$\begin{cases} u_n^g(t) - \dfrac{1}{N}\,\psi_n^g(t) = u_n^{g,l}(t) - \dfrac{1}{N}\,\psi_n^{g,l}(t), \\ \qquad\qquad v_n^g(t) = v_n^{g,l}(t), \qquad\qquad 1 \leq n \leq n_c, \\ u_n^g(t) + \dfrac{1}{N}\,\psi_n^g(t) = u_n^{g,r}(t) + \dfrac{1}{N}\,\psi_n^{g,r}(t), \end{cases} \tag{2.68}$$

$$\begin{cases} u_n^g(t) = u_n^{g,l}(t), \\ v_n^g(t) = v_n^{g,l}(t), \qquad n > n_c. \\ \psi_n^g(t) = \psi_n^{g,l}(t), \end{cases} \tag{2.69}$$

We note that U^g is a function of z and t only, and hence it does not depend on the horizontal coordinate x. Setting $F^\# = F - \mathrm{d}U^g/\mathrm{d}t$ and $U_0^\# = U_0 - U_0^g$, where $U_0^g = U^g(t=0)$, we will look for $U^\#$ solution of

$$\begin{cases} \dfrac{\mathrm{d}U^\#}{\mathrm{d}t} + A\,U^\# = F^\#, \\ U^\#(t=0) = U_0^\#. \end{cases} \tag{2.70}$$

Like Eq. (2.16), this equation corresponds to the case with homogeneous boundary conditions. In order to apply Theorem 2.2 to Eq. (2.70), we would need to have

$$U_0^\# = U_0 - U_{t=0}^g \in D(A), \tag{2.71}$$

and

$$F^\#,\, \frac{\mathrm{d}F^\#}{\mathrm{d}t} \in L^1(O,T;H). \tag{2.72}$$

[8] Assuming periodical boundary conditions for $\widetilde{\mathcal{M}}$.

We will state in Theorem 2.3 below some assumptions on $U^{g,l}$ and $U^{g,r}$, which guarantee that $U_0^{\#}$ and $F^{\#}$ satisfy Eqs. (2.72) and (2.71). Writting $U = U^{\#} + U^g$, we find that U is solution of Eq. (2.1), and the boundary conditions of U at $x = 0$ and $x = L_1$ are those of U^g, i.e., for the subcritical modes ($1 \le n \le n_c$):

$$\begin{cases} \displaystyle\int_{-L_3}^{0} u(0,z,t)\,\mathcal{U}_n(z)\,\mathrm{d}z - \frac{1}{N}\int_{-L_3}^{0} \psi(0,z,t)\,\mathcal{W}_n(z)\,\mathrm{d}z \\ \qquad = \displaystyle\int_{-L_3}^{0} u^{g,l}(z,t)\,\mathcal{U}_n(z)\,\mathrm{d}z - \frac{1}{N}\int_{-L_3}^{0} \psi^{g,l}(z,t)\,\mathcal{W}_n(z)\,\mathrm{d}z, \\ \displaystyle\int_{-L_3}^{0} v(0,z,t)\,\mathcal{U}_n(z)\,\mathrm{d}z = \int_{-L_3}^{0} v^{g,l}(z,t)\,\mathcal{U}_n(z)\,\mathrm{d}z, \\ \displaystyle\int_{-L_3}^{0} u(L_1,z,t)\,\mathcal{U}_n(z)\,\mathrm{d}z + \frac{1}{N}\int_{-L_3}^{0} \psi(L_1,z,t)\,\mathcal{W}_n(z)\,\mathrm{d}z \\ \qquad = \displaystyle\int_{-L_3}^{0} u^{g,r}(z,t)\,\mathcal{U}_n(z)\mathrm{d}z + \frac{1}{N}\int_{-L_3}^{0} \psi^{g,r}(z,t)\,\mathcal{W}_n(z)\,\mathrm{d}z, \end{cases} \tag{2.73}$$

and for the supercritical ones ($n > n_c$):

$$\begin{cases} \displaystyle\int_{-L_3}^{0} u(0,z,t)\,\mathcal{U}_n(z)\,\mathrm{d}z = \int_{-L_3}^{0} u^{g,l}(z,t)\,\mathcal{U}_n(z)\,\mathrm{d}z, \\ \displaystyle\int_{-L_3}^{0} v(0,z,t)\,\mathcal{U}_n(z)\,\mathrm{d}z = \int_{-L_3}^{0} v^{g,l}(z,t)\,\mathcal{U}_n(z)\,\mathrm{d}z, \\ \displaystyle\int_{-L_3}^{0} \psi(0,z,t)\,\mathcal{W}_n(z)\,\mathrm{d}z = \int_{-L_3}^{0} \psi^{g,l}(z,t)\,\mathcal{W}_n(z)\,\mathrm{d}z. \end{cases} \tag{2.74}$$

Thus, we have established the following result:

THEOREM 2.3. *Let H be the Hilbert space defined in Eq. (2.17) and A be the linear operator defined in Eq. (2.25) corresponding to the two-dimensional linearized PEs with vanishing viscosity. We are given the boundary values $U^{g,l}$ and $U^{g,r}$ which are in $L^1\big(0,T;L^2(-L_3,0)^3\big)$, together with their first time derivatives, and F and $F' = \mathrm{d}F/\mathrm{d}t \in L^1(0,T;L^2(\mathcal{M})^3)$.*

Then, the initial value problem corresponding to Eq. (2.1)*, supplemented with the boundary conditions in Eqs.* (2.73) *and* (2.74)*, is well-posed, i.e., for every initial data $U_0 \in U_0^g + D(A)$,*[9] *there exists a unique solution $U \in \mathcal{C}([0,T];H)$ of Eq.* (2.1) *verifying Eqs.* (2.73) *and* (2.74)*, and $U(0) = U_0$.*

2.3.5. The barotropic mode

We now return to the mode constant in z, when $n = 0$. This mode does not raise any mathematical difficulty, but it is fundamental in the numerical simulations, since it carries much energy.

[9] This means that U_0 has the same smoothness as a function of $D(A)$ and Eqs. (2.73), (2.74) are satisfied at $t = 0$.

Integrating Eqs. (2.5a), (2.5b), and (2.5e), on $(-L_3, 0)$, we find:

$$\frac{\partial u_0}{\partial t} + \overline{U}_0 \frac{\partial u_0}{\partial x} - f v_0 + \frac{\partial \phi_0}{\partial x} = F_{u,0}, \tag{2.75}$$

$$\frac{\partial v_0}{\partial t} + \overline{U}_0 \frac{\partial v_0}{\partial x} + f u_0 = F_{v,0}, \tag{2.76}$$

$$\frac{\partial u_0}{\partial x} = 0. \tag{2.77}$$

We propose to supplement this system with the following boundary conditions:

$$u_0(0, t) = u_l(t), \tag{2.78}$$

$$v_0(0, t) = v_l(t), \tag{2.79}$$

with u_l, v_l given (not necessarily zero, as in Section 2.3.4).

Then, since $\partial u_0/\partial x = 0$, u_0 does not depend on x, and it is thus equal to $u_l(t)$ everywhere so that Eq. (2.78) means in fact that

$$u_0(x, t) = u_l(t), \quad \forall (x, t) \in (0, L_1) \times \mathbb{R}_+^*. \tag{2.80}$$

Introducing Eq. (2.80) in Eq. (2.76), we find that:

$$\frac{\partial v_0}{\partial t} + \overline{U}_0 \frac{\partial v_0}{\partial x} = F_{v,0} - f\,(\overline{U}_0 + u_l). \tag{2.81}$$

When we supplement Eq. (2.81) with the boundary condition in Eq. (2.79), we have a simple well-posed problem, and v_0 is given in terms of the data by integration along the characteristics.

Finally, once both u_0 and v_0 are known, Eq. (2.75) gives ϕ_0, up to an additive constant (as expected):

$$\begin{aligned} \phi_0(x, t) &= \phi_0(0, t) + \int_0^x \left\{ f\, v_0(x', t) - \frac{\partial u_0}{\partial t}(x', t) \right\} \mathrm{d}x' \\ &= \phi_0(0, t) - x\, u_l'(t) + f \int_0^x v_0(x', t)\, \mathrm{d}x'. \end{aligned} \tag{2.82}$$

2.4. Numerical simulations

In this section and the next one, we describe the numerical simulations performed in the 2D linear and nonlinear cases. We start by presenting the numerical scheme.

2.4.1. Vertical decomposition

In the vertical direction, we proceed by normal modes decomposition as in Eqs. (2.6), (2.7). From the numerical point of view, we will need to transform some grid-data into

modal coefficients in the $\mathcal{U}_m$ or $\mathcal{W}_m$ bases of $L^2(-L_3, 0)$, and vice versa, using Fourier and inverse Fourier transforms for example.

Given a function f represented by its values f_l on a grid $z_l = -L_3 + l\Delta z$, $0 \leq l \leq l_{max}$, $\Delta z = L_3 / l_{max}$, we want to find its coefficients f_m in the modal decomposition (Eqs. (2.6) and (2.7)), limited to $0 \leq m \leq M$. To this aim, we use the second-order central point integration method, with the z_l as collocation points. For the functions u, v, and ϕ, we decompose them in the $\mathcal{U}_m$ basis of $L^2(-L_3, 0)$. For $0 \leq m \leq M$,

$$\{u_m, v_m, \phi_m\} = \Delta z \sum_{l=0}^{l_{\max}-1} \frac{\mathcal{U}_m(z_l) \cdot \{u, v, \phi\}(z_l) + \mathcal{U}_m(z_{l+1}) \cdot \{u, v, \phi\}(z_{l+1})}{2}, \tag{2.83}$$

and for w and ψ, $1 \leq m \leq M$:

$$\{w_m, \psi_m\} = \Delta z \sum_{l=0}^{l_{\max}-1} \frac{\mathcal{W}_m(z_l) \cdot \{w, \psi\}(z_l) + \mathcal{W}_m(z_{l+1}) \cdot \{w, \psi\}(z_{l+1})}{2}. \tag{2.84}$$

This approach which is proposed by the physicists is different from the more mathematical approach to spectral and pseudo-spectral methods as in e.g., Bernardi and Maday [1997], Gottlieb and Hesthaven [2001]. The advantage of such a choice is that the orthogonality relations (see Section 1.2) are satisfied from the numerical point of view. Further studies and comparisons of the two approaches will be needed in the future.

On the contrary, if the function is given by its modal coefficients, the values on the z-grid z_l, $0 \leq l \leq l_{\max}$ is simply given by

$$(u, v, \phi)(z_l) = \sum_{m=0}^{M} (u_m, v_m, \phi_m)\, \mathcal{U}_m(z_l), \tag{2.85}$$

$$(w, \psi)(z_l) = \sum_{m=0}^{M} (w_m, \psi_m)\, \mathcal{W}_m(z_l). \tag{2.86}$$

In the numerical simulations, we are given some initial data on the physical grid $(z_l)_{0 \leq l \leq l_{max}}$. We transform them into modal coefficients, thanks to Eq. (2.83) or (2.84), and if the problem is linear, we keep them all along the computations, except for graphic purposes, for which we use the inverse formulas Eqs. (2.85) and (2.86) to return to the physical space. Naturally, in the nonlinear case, we will need to operate Eqs. (2.83)–(2.86) once at every time step, in order to avoid the computation of a convolution product, that would cost too much in term of CPU time and is not considered an appropriate (stable) numerical procedure. We compute the nonlinear terms of the equations in the physical space (x, z) thanks to Fourier and inverse Fourier transforms.

2.4.2. Finite differences in time and space

Looking at the form of Eq. (2.13), we choose to discretize these equations in the horizontal direction with a finite differences method. Naturally, care has to be taken to the sign of the

characteristic values, in order to take an upwind (hence stable) spatial discretization of the x-derivative. Whereas $\overline{U}_0$ and $\overline{U}_0 + N/\lambda_m$ are always positive, the third characteristic value of the *mth* mode, in the linear case, is $\overline{U}_0 - N/\lambda_m$ and can either be positive or negative for the actual physical values that we consider.

With this in mind, for every subcritical mode $m \leq n_c$, we discretize Eq. (2.15) as follows:

$$\frac{\xi^{n+1}_{m,j} - \xi^n_{m,j}}{\Delta t^n} + \left(\overline{U}_0 + \frac{N}{\lambda_m}\right) \frac{\xi^n_{m,j} - \xi^n_{m,j-1}}{\Delta x} - f\, v^n_{m,j} = F^n_{\xi,m,j} - B^n_{\xi,m,j}, \quad (2.87a)$$

$$\frac{v^{n+1}_{m,j} - v^n_{m,j}}{\Delta t^n} + \overline{U}_0 \frac{v^n_{m,j} - v^n_{m,j-1}}{\Delta x} + f \frac{\xi^n_{m,j} + \eta^n_{m,j}}{2} = F^n_{v,m,j} - B^n_{v,m,j}, \quad (2.87b)$$

$$\frac{\eta^{n+1}_{m,j} - \eta^n_{m,j}}{\Delta t^n} + \left(\overline{U}_0 - \frac{N}{\lambda_m}\right) \frac{\eta^n_{m,j+1} - \eta^n_{m,j}}{\Delta x} - f v^n_{m,j} = F^n_{\eta,m,j} - B^n_{\eta,m,j}. \quad (2.87c)$$

where the right-hand side of Eq. (2.87) contains the nonlinear terms, computed explicitly thanks to an Adams–Bashforth scheme.

Equations (2.87a) and (2.87b) hold for $1 \leq j \leq J$, whereas (2.87c) is written for $0 \leq j \leq J-1$. There are no equations for $\xi^{n+1}_{m,0}$, $v^{n+1}_{m,0}$, and $\eta^{n+1}_{m,J}$, these quantities being given by the boundary conditions as required in Eqs. (2.23), (2.24) (see also Rousseau, Temam and Tribbia [2005b]). On the contrary, if $m > n_c$ (supercritical case), we propose for $1 \leq j \leq J$ the discretized equations

$$\frac{\xi^{n+1}_{m,j} - \xi^n_{m,j}}{\Delta t^n} + \left(\overline{U}_0 + \frac{N}{\lambda_m}\right) \frac{\xi^n_{m,j} - \xi^n_{m,j-1}}{\Delta x} - f v^n_{m,j} = F^n_{\xi,m,j} - B^n_{\xi,m,j}, \quad (2.88a)$$

$$\frac{v^{n+1}_{m,j} - v^n_{m,j}}{\Delta t^n} + \overline{U}_0 \frac{v^n_{m,j} - v^n_{m,j-1}}{\Delta x} + f \frac{\xi^n_{m,j} + \eta^n_{m,j}}{2} = F^n_{v,m,j} - B^n_{v,m,j}, \quad (2.88b)$$

$$\frac{\eta^{n+1}_{m,j} - \eta^n_{m,j}}{\Delta t^n} + \left(\overline{U}_0 - \frac{N}{\lambda_m}\right) \frac{\eta^n_{m,j} - \eta^n_{m,j-1}}{\Delta x} - f v^n_{m,j} = F^n_{\eta,m,j} - B^n_{\eta,m,j}. \quad (2.88c)$$

Either $\xi^{n+1}_{m,0}$, $v^{n+1}_{m,0}$, and $\eta^{n+1}_{m,0}$ are given by the boundary conditions defined as in Eqs. (2.23), (2.24) (see also Rousseau, Temam and Tribbia [2005b], transparent boundary conditions case), or they satisfy the periodicity conditions in Eq. (2.95) below (periodical case).

For every function $f(x, z, t)$, $f^n_{m,j}$ represents $f_m(x_j, t_n)$ for $0 \leq j \leq J, 0 \leq n \leq n_{\max}$, with

$$0 = x_0 < x_1 < \ldots < x_j < \ldots < x_J = L, \quad (2.89)$$

$$0 = t_0 < t_1 < \ldots < t_n < \ldots < t_{n_{\max}} = T, \quad (2.90)$$

$$\Delta x = x_{j+1} - x_j = \frac{L}{J}, \quad (2.91)$$

$$\Delta t^n = t_{n+1} - t_n .n. \quad (2.92)$$

In the numerical experiments, we choose an homogeneous space discretization ($\Delta x = const = L/J$). For the sake of simplicity, we choose an explicit time scheme, with a constant time step Δt, which will be restricted by the well-known CFL condition to guarantee stability in the linear case:

$$\Delta t \leq \frac{\Delta x}{\max\limits_{1\leq m\leq M}\left(\overline{U}_0, \overline{U}_0 + \frac{N}{\lambda_m}, |\overline{U}_0 - \frac{N}{\lambda_m}|\right)} = \frac{\Delta x}{\overline{U}_0 + \frac{N}{\lambda_1}}. \tag{2.93}$$

When the equations are nonlinear, the characteristic values depend on time since $\overline{U}_0$ has to be replaced by $u + \overline{U}_0$, but we assume that the initial data is such that $|u_0| << \overline{U}_0$, which is physically relevant (Temam and Tribbia [2003]). We actually base our computations on the data in the already quoted article Temam and Tribbia [2003], i.e., the initial data is such that the ratio between the perturbation and the reference flow $\overline{U}_0\,\mathbf{e_x}$ is less than 10%, which is physically relevant. In the case of numerical simulations with periodic boundary conditions, we multiply the initial data of Temam and Tribbia [2003] by $sin(\pi\, x/L)$ to make it periodic and avoid any boundary layer at $t = 0$.

2.5. Numerical results

We present hereafter two different sets of numerical results. In Section 2.5.1, we resolve the nonlinear PEs in a domain $\Omega_0 = (0, L) \times (-L_3, 0)$ with periodic boundary conditions in the horizontal (x) direction, and $w = 0$ at $z = -L_3, 0$ (see Figure 2.1). These numerical results will provide the boundary conditions needed for the computations of Section 2.5.2, below. In Section 2.5.3, the transparent properties of the boundary conditions are introduced, and we compare the solutions computed in Section 2.5.1 and those of Section 2.5.2; this allows us to confirm the computational suitability and relevance of the boundary conditions that we have introduced.

The computations are done as follows. We fix M (the chosen number of modes) and compute $(u_m^0, v_m^0, \psi_m^0)_{0<m\leq M}$ from the given data u^0, v^0, ψ^0, thanks to Eqs. (2.83) and (2.84).

Then, for every mode $m \leq M$, we consider the modal equations (Eq. (2.15)) and their discretization (Eqs. (2.87) and (2.88)), and supplement them with the appropriate boundary conditions, either Eq. (2.96) for the periodical case or Eqs. (2.100) and (2.101) for the case of transparent boundary conditions. We recall here that for every m, $(\xi_m, \eta_m) = (u_m - N\,\psi_m/\lambda_m, u_m + N\,\psi_m/\lambda_m)$ will be the numerical unknowns to be computed so that $(u_m, w_m, \psi_m, \phi_m)$ can be obtained with

$$u_m(x,t) = \frac{\xi_m + \eta_m}{2}(x,t), \tag{2.94a}$$

$$w_m(x,t) = -\frac{u_{m_x}}{\lambda_m}(x,t), \tag{2.94b}$$

$$\psi_m(x,t) = \frac{N\,(\eta_m - \xi_m)}{2}(x,t), \tag{2.94c}$$

$$\phi_m(x,t) = -\frac{\psi_m}{\lambda_m}(x,t). \tag{2.94d}$$

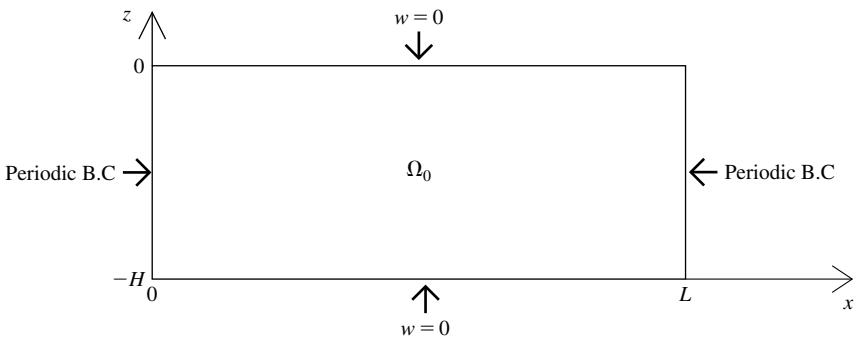


FIG. 2.1 Domain Ω_0.

As a consequence, we will only consider the quantities (ξ_m, v_m, η_m) in the sequel, the other physical quantities being easily computed, thanks to Eq. (2.94).

2.5.1. Periodic boundary conditions for the large domain Ω_0
In the periodical case, we consider the following modal boundary conditions:

$$\xi_m(0,t) = \xi_m(L,t), \tag{2.95a}$$

$$v_m(0,t) = v_m(L,t), \tag{2.95b}$$

$$\eta_m(0,t) = \eta_m(L,t). \tag{2.95c}$$

For each time step $\Delta t^n = \Delta t$ satisfying Eq. (2.93) we compute the unknown functions $(\xi_m^{n+1}, v_m^{n+1}, \eta_m^{n+1})$, thanks to Eqs. (2.87) and (2.88), with the numerical boundary conditions:

$$\xi_{m,0}^{n+1} = \xi_{m,J}^{n+1}, \tag{2.96a}$$

$$v_{m,0}^{n+1} = v_{m,J}^{n+1}, \tag{2.96b}$$

$$\eta_{m,0}^{n+1} = \eta_{m,J}^{n+1}. \tag{2.96c}$$

The following figures plot u, v, and ψ in the domain Ω_0 at two different times. Figures 2.2, 2.3, and 2.4 represent the initial data ($t = 0$) for these three quantities, whereas Figs. 2.5, 2.6 and 2.7 represent u, v, and ψ at $t = t_1 > 0$.

2.5.2. Transparent boundary conditions for the subdomain $\Omega_1 \subset \Omega_0$
We now intend to simulate the PEs in the subdomain $\Omega_1 = (a, b) \times (-L_3, 0)$, and the boundary conditions at $x = a$ and b are those considered in the linear case, see Eqs. (2.23) and (2.24) (see also ROUSSEAU, TEMAM and TRIBBIA [2005b]). In the numerical simulations below, we will consider a domain $\Omega_1 = (a, b) \times (-L_3, 0)$ such that $0 < a < b < L$, so that Ω_1 is fully imbedded in $\Omega_0 = (0, L) \times (-L_3, 0)$ see Figure 2.8. The space discretization points are now changed to $x_j = a + j(b - a)/J$, $0 \le j \le J$.

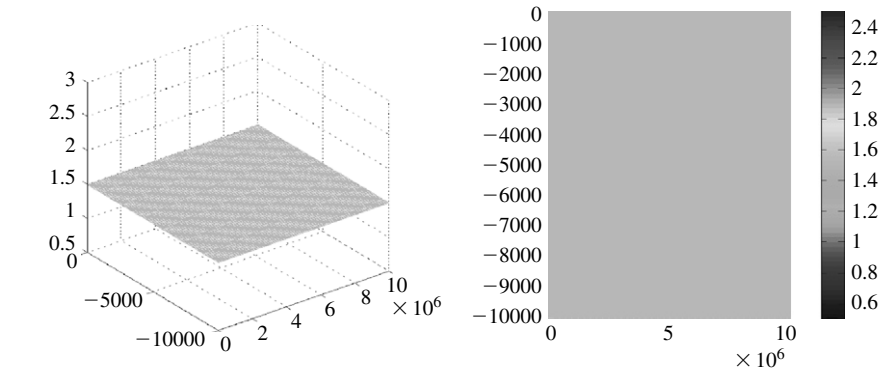


FIG. 2.2 Periodic boundary condition. Initial data u_0. (See also color insert).

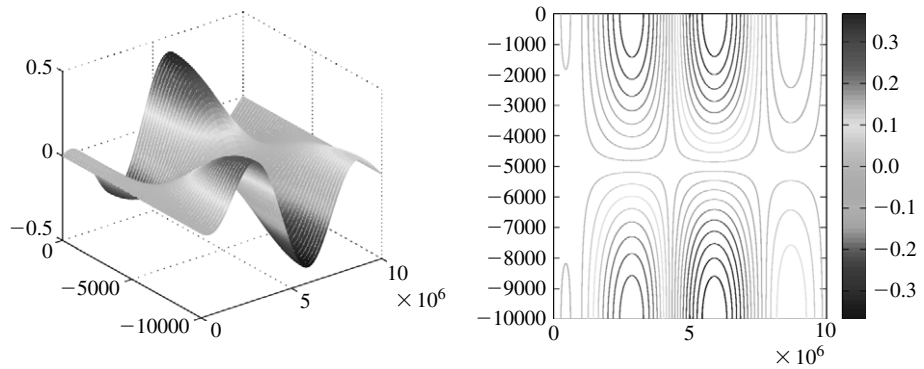


FIG. 2.3 Periodic boundary condition. Initial data v_0. (See also color insert).

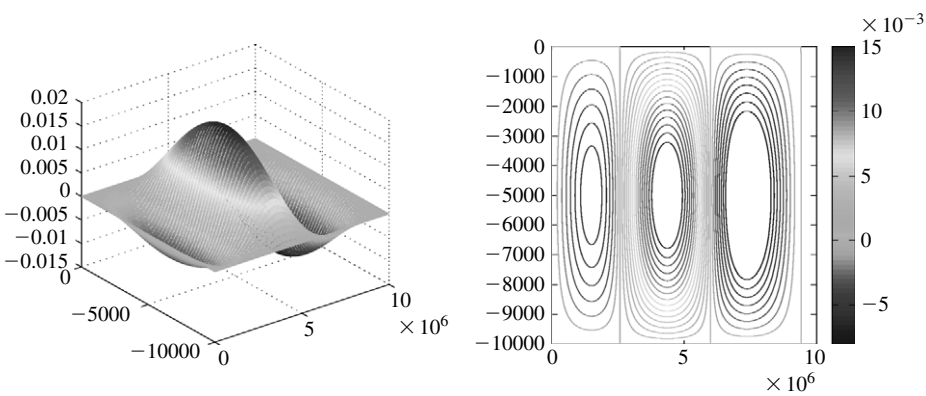


FIG. 2.4 Periodic boundary condition. Initial data ψ_0. (See also color insert).

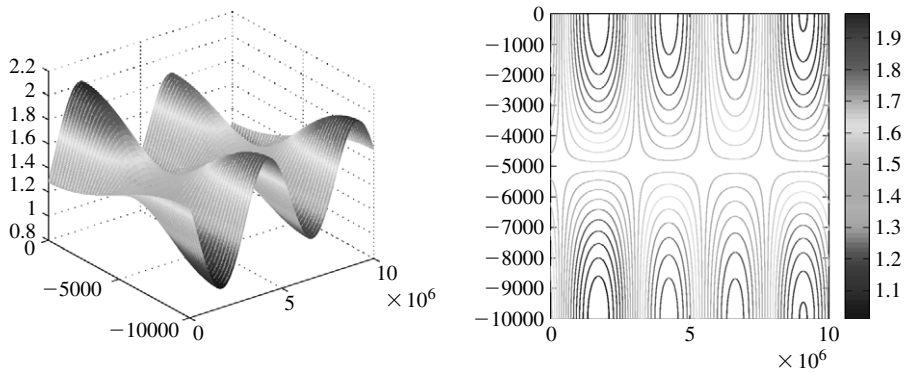


FIG. 2.5 Periodic boundary condition. Values of u at $t = t_1$. (See also color insert).

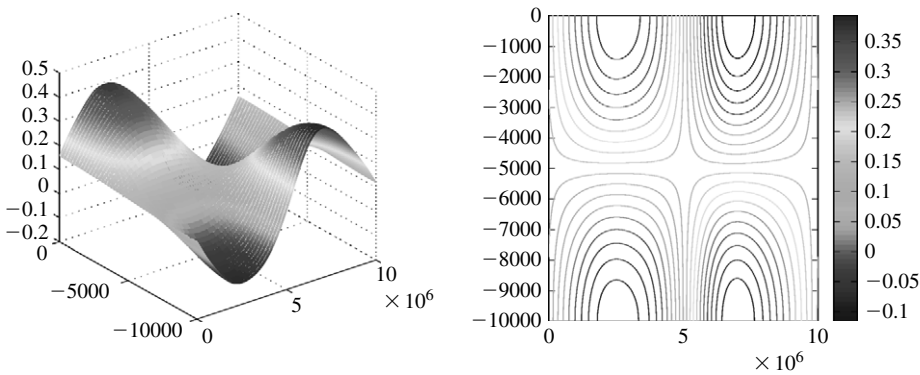


FIG. 2.6 Periodic boundary condition. Values of v at $t = t_1$. (See also color insert).

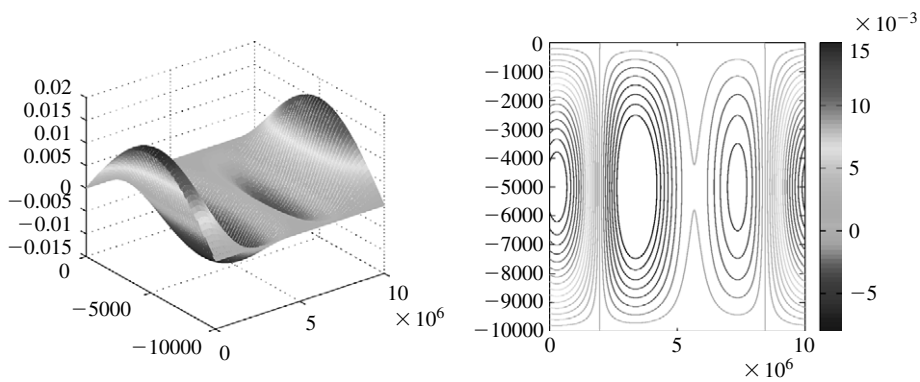


FIG. 2.7 Periodic boundary condition. Values of ψ at $t = t_1$. (See also color insert).

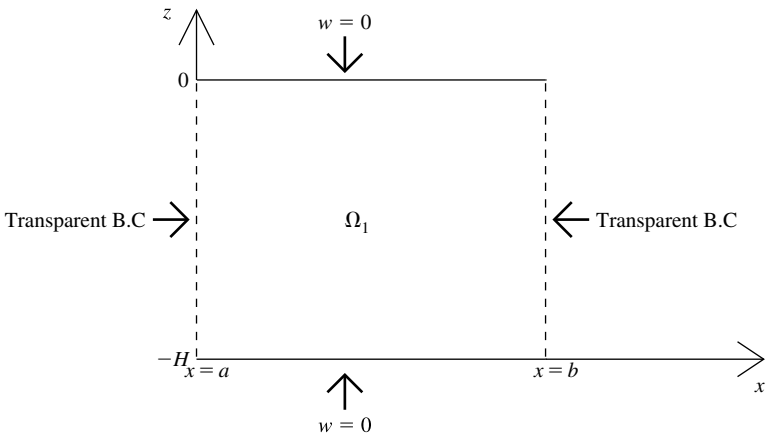


FIG. 2.8 Subdomain Ω_1.

At the boundaries $x = a$ and $x = b$, we will consider the nonhomogeneous form of the transparent boundary conditions of Eqs. (2.23), (2.24). We use the computations of Section 2.5.1 above to provide the right-hand side of the boundary conditions in Eqs. (2.98) and (2.99) below, and afterwards use them for comparison in the whole subdomain Ω_1. These boundary conditions, expressed in a general way, are given in Eqs. (2.23), (2.24). They consist in *an infinite set of integral boundary conditions*. For example:

$$\int_{-L_3}^{0} v(a,z,t)\,\mathcal{U}_m(z)\mathrm{d}z = \int_{-L_3}^{0} \widetilde{v}(a,z,t)\,\mathcal{U}_m(z)\mathrm{d}z, \quad \forall m \le M, \tag{2.97}$$

where $\widetilde{U} = (\widetilde{u}, \widetilde{v}, \widetilde{w}, \widetilde{\psi}, \widetilde{\phi})$ are known functions, computed in the domain Ω_0 using the periodic boundary conditions (see Section 2.5.1 above).

Hence, for every subcritical mode ($m \le m_c$) and every time $t > 0$, we have

$$\xi_m(a,t) = \widetilde{\xi}_m(a,t), \tag{2.98a}$$

$$v_m(a,t) = \widetilde{v}_m(a,t), \tag{2.98b}$$

$$\eta_m(b,t) = \widetilde{\eta}_m(b,t), \tag{2.98c}$$

where $\widetilde{\xi}_m$ and $\widetilde{\eta}_m$ are defined as usual.

For the supercritical modes, we set for every $t > 0$:

$$\xi_m(a,t) = \widetilde{\xi}_m(a,t), \tag{2.99a}$$

$$v_m(a,t) = \widetilde{v}_m(a,t), \tag{2.99b}$$

$$\eta_m(a,t) = \widetilde{\eta}_m(a,t). \tag{2.99c}$$

To implement these boundary conditions, we discretize Eq. (2.15) with the finite differences method, taking into account the sign of $\overline{U}_0 - N/\lambda_m$ for the discretization of the first x-derivative of η_m in Eq. (2.15c) (see Eqs. (2.87) and (2.88) of Section 2.2 above).

For each time step $\Delta t^n = \Delta t$ satisfying Eq. (2.93) we compute the unknown functions $(\xi_m^{n+1}, v_m^{n+1}, \eta_m^{n+1})$, thanks to Eqs. (2.87) and (2.88), with the numerical boundary conditions:

$$\xi_{m,0}^{n+1} = \widetilde{\xi}_m(a, t_{n+1}), \tag{2.100a}$$

$$v_{m,0}^{n+1} = \widetilde{v}_m(a, t_{n+1}), \tag{2.100b}$$

$$\eta_{m,J}^{n+1} = \widetilde{\eta}_m(b, t_{n+1}), \tag{2.100c}$$

if m is subcritical ($m \leq m_c$). If m is supercritical ($m > m_c$), we set

$$\xi_{m,0}^{n+1} = \widetilde{\xi}_m(a, t_{n+1}), \tag{2.101a}$$

$$v_{m,0}^{n+1} = \widetilde{v}_m(a, t_{n+1}), \tag{2.101b}$$

$$\eta_{m,0}^{n+1} = \widetilde{\eta}_m(a, t_{n+1}). \tag{2.101c}$$

The following figures plot u, v, and ψ in the domain Ω_1 at two different times. Figures 2.9, 2.10, and 2.11 represent the initial data ($t = 0$) for these three quantities, whereas Figs. 2.12, 2.13 and 2.14 represent u, v, and ψ at $t = t_1 > 0$.

Here, one can see that Figures 2.12, 2.13 and 2.14, respectively, match well with Figures 2.5, 2.6 and 2.7 in the domain Ω_1.

2.5.3. Comparisons

In order to quantitatively confirm what can be observed, we finally choose an interior point $(x_0, z_0) = (5.8 \times 10^6, -4.0 \times 10^3) \in \Omega_1$, and plot in Fig. 2.16 the values of $(u, v, \psi)(x_0, z_0, t)$ computed in Ω_1 with transparent boundary conditions compared to the same quantities computed in Ω_0 with periodic boundary conditions. The results are similar if one considers another choice of (x_0, z_0); this shows the transparency property of the boundary conditions in Eqs. (2.98) and (2.99).

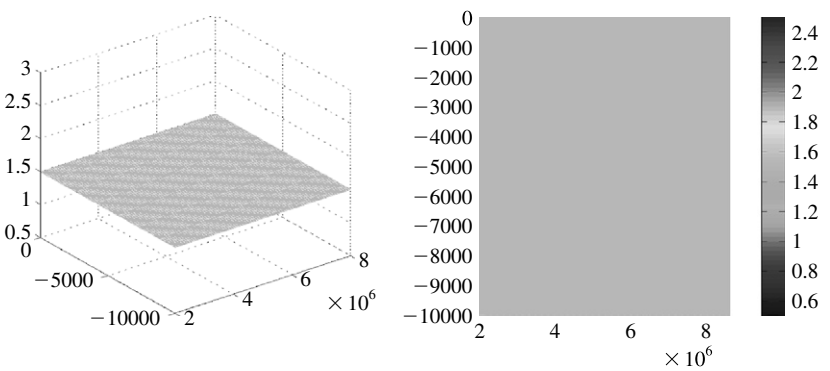


FIG. 2.9 Transparent boundary condition. Initial data u_0. (See also color insert).

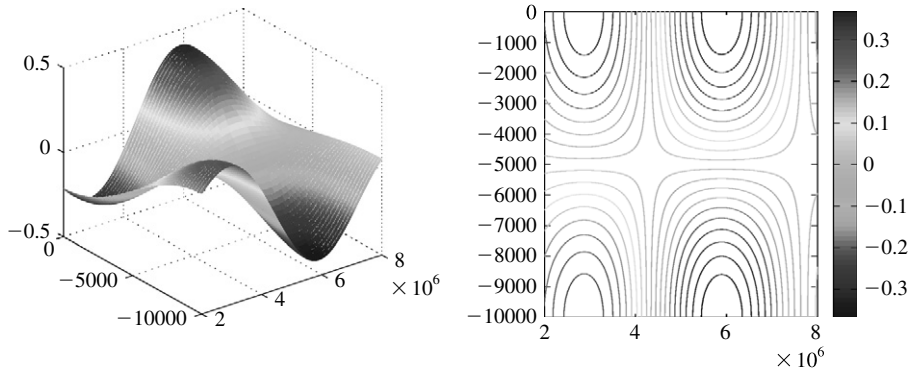


FIG. 2.10 Transparent boundary condition. Initial data v_0. (See also color insert).

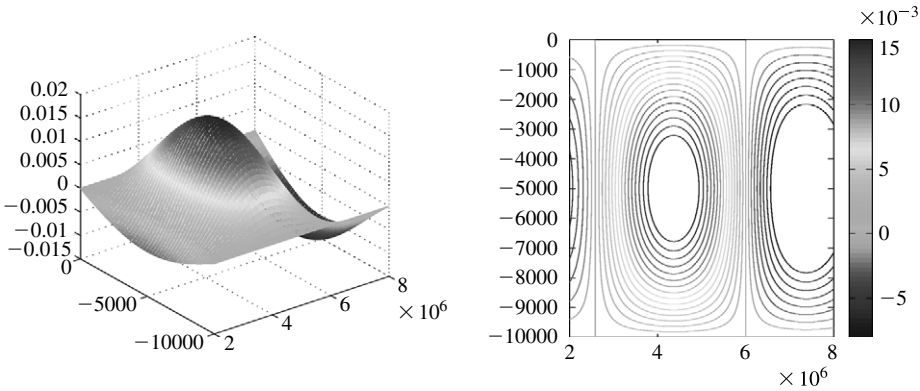


FIG. 2.11 Transparent boundary condition. Initial data ψ_0. (See also color insert).

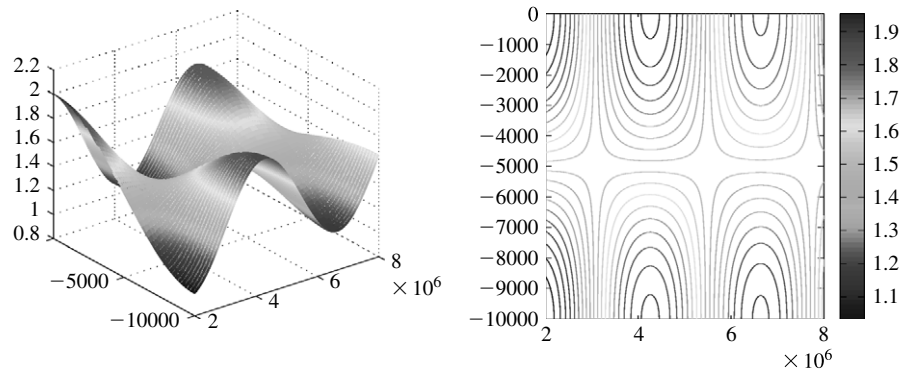


FIG. 2.12 Transparent boundary condition. Values of u at $t = t_1$. (See also color insert).

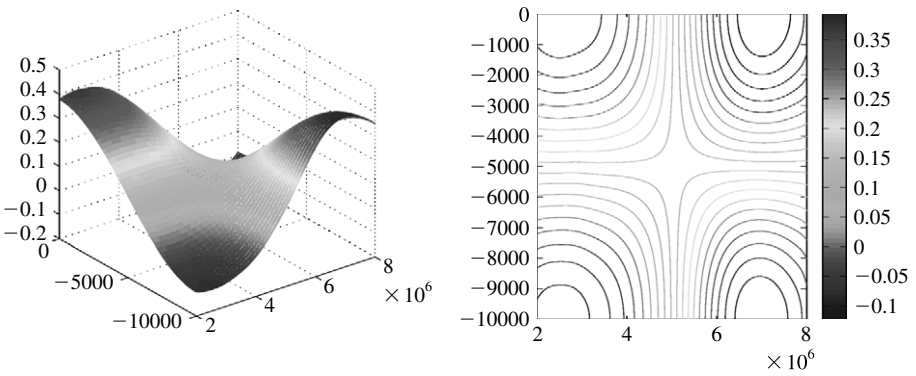


FIG. 2.13 Transparent boundary condition. Values of v at $t = t_1$. (See also color insert).

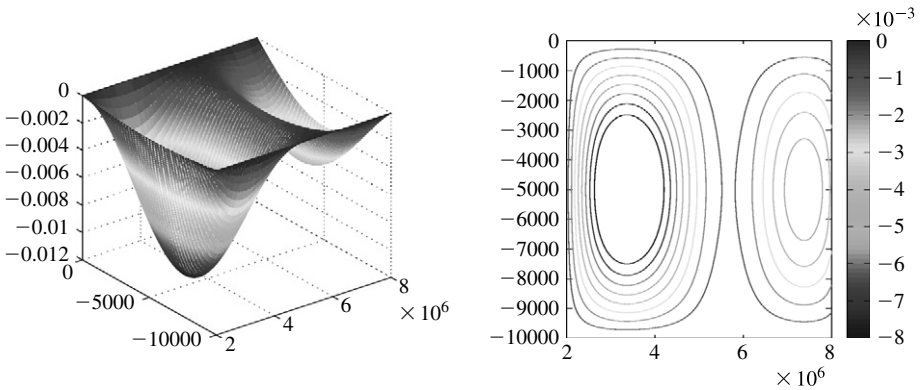


FIG. 2.14 Transparent boundary condition. Values of ψ at $t = t_1$. (See also color insert).

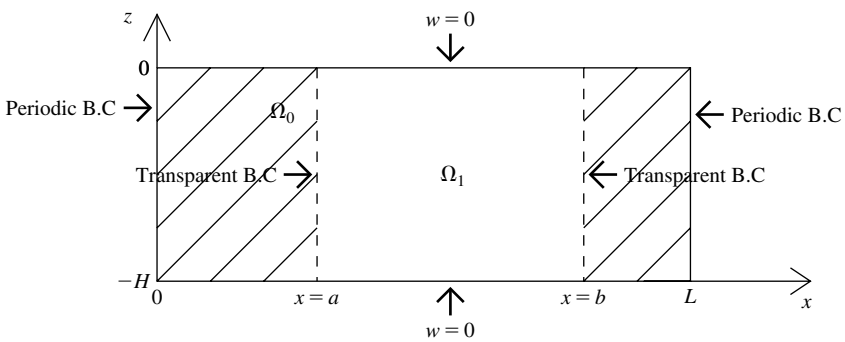


FIG. 2.15 Subdomains Ω_0 and Ω_1.

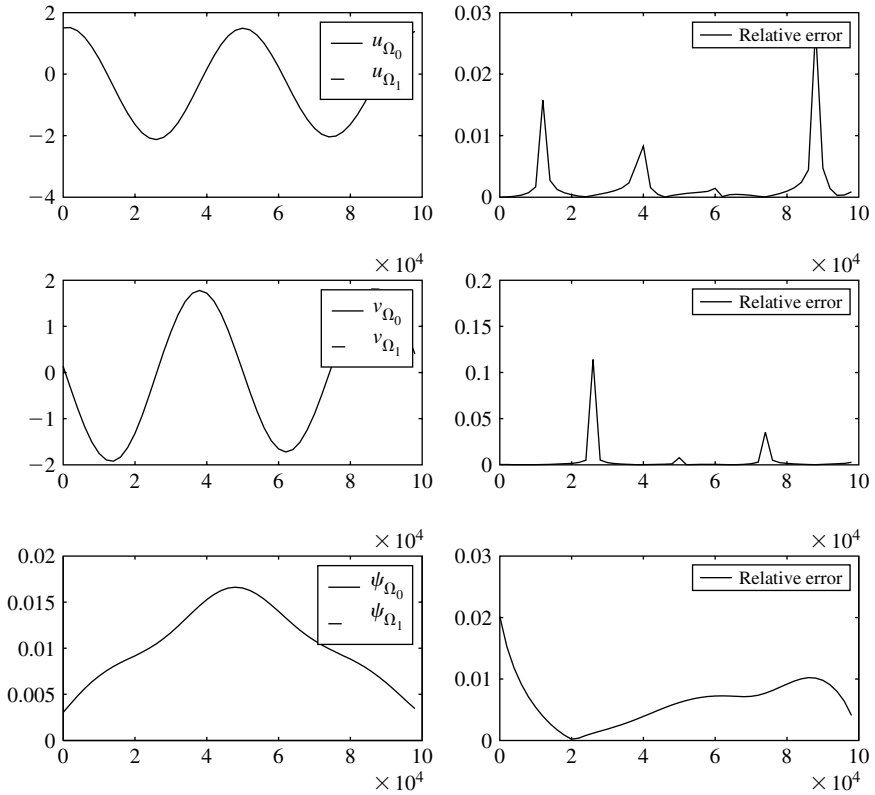


FIG. 2.16 Computations of $(u, v, \psi)(x_0, z_0, t)$ with two different types of boundary conditions (left). Relative errors (right).

In the left column of Fig. 2.16, we plot $(u, v, \psi)(x_0, z_0, t)$, these quantities being computed with the two types of boundary conditions. In the right column, we plot the corresponding relative errors $|f_{\Omega_0} - f_{\Omega_1}|/|f_{\Omega_0}|$, where f is successively u, v, and ψ. The reader might think that the relative error reaches some local high values, but this is actually due to the fact that the quantity u_{Ω_0} (or $v_{\Omega_0}, \psi_{\Omega_0}$) vanishes; these local maximum are not meaningful.

3. Space dimension 2.5

3.1. Motivations

We now pursue our study with a simplified 3D case, in which we allow the unknown functions to mildly depend on the y variable. This case, whose motivations are given below, is called the 2.5 dimension case.

The numerical simulations performed in Section 2 were mainly motivated by computational preoccupations and the need to support the idea that the proposed boundary conditions are computationally feasible and lead indeed to well-posedness. In view of performing (in dimension 2) *computations of physical significance*, the last author expressed the wish that the flow were a perturbation of a geostrophic flow (which is not the case in Section 2). Now, the geostrophic equation

$$p_y = -\rho f u, \tag{3.1}$$

implies that there does not exist any geostrophic solution depending only on x and z.[10] In this context, it is then necessary, even in dimension 2, to introduce some y-dependence. A number of natural choices had to be abandoned, in particular the use of a few Fourier modes in y for a Lorentz type model. Indeed this model would produce undesirable Gibbs phenomena when we approximate the periodic extension of the function $\sigma(y) = y$ on $[0, L_2]$, this function being introduced in the model by Eq. (3.1). In this way, we were led to choose, for the y-direction, a three-mode linear finite element model. In this section, we present the full derivation of the model and study the well-posedness of the linearized equations, leaving for further studies the nonlinear case and the numerical studies.

This section is organized as follows. The model is derived in Section 3.2. We first derive the Galerkin finite element approximation based on the use of three piecewise linear elements in the direction y; we thus arrive at three coupled systems, each one similar to the 2D primitive equations in the variables x and z (and t). We then perform the normal mode decomposition of these equations in the direction z as in Section 2, the normal modes in z being either sines or cosines (depending on the functions), and these sines and cosines are the eigenfunctions of the two-point boundary value Sturm–Liouville problem in Eqs. (1.16) and (1.17) (Temam and Tribbia [2003]). At this stage, each mode consists of three coupled equations for the functions of the variables x and t (Section 3.2.2). We finally introduce, in Section 3.2.3, the boundary conditions for the latest systems in x and t, the boundary conditions depending on the nature of the mode (subcritical or supercritical), the subcritical modes being the mathematically most challenging and physically most relevant ones. In Section 3.3, the objectives are as follows: we first establish, in the absence of the zero mode, the well-posedness of the linearized PEs, all the non-zero modes taken into account. We then pay special attention to the mode zero (barotropic part), and finally consider the case of nonhomogeneous boundary conditions; we classically reduce this case to the homogeneous case by homogenization of the boundary conditions. We consider in Section 3.4 a related model, physically interesting but with fewer degrees of freedom. The well-posedness of this model is addressed in a similar way as in Sections 3.2 and 3.3 for the first model. For this model, we only emphasize the parts of the proof and the discussions which are different from the first model. The actual numerical simulations will be performed and discussed in a separate work.

[10] Ox is the local west–east direction, Oy is the local south–north direction, and Oz is the ascendant vertical.

3.2. The 2.5D primitive equations

We rewrite the 3D PEs for the ocean and the atmosphere without viscosity with the same notations as in Eq. (1.1):

$$\begin{cases} \widetilde{\mathrm{v}}_t + (\widetilde{\mathrm{v}} \cdot \nabla)\widetilde{\mathrm{v}} + \widetilde{w}\mathrm{v}_z + f\, k \times \widetilde{\mathrm{v}} + \dfrac{1}{\rho_0} \nabla \widetilde{p} = F_{\widetilde{\mathrm{v}}}, \\ \widetilde{p}_z = -\widetilde{\rho} g, \\ \nabla \cdot \widetilde{\mathrm{v}} + \widetilde{w}_z = 0, \\ \widetilde{T}_t + (\widetilde{\mathrm{v}} \cdot \nabla)\widetilde{T} + \widetilde{w}\widetilde{T}_z = Q_{\widetilde{T}}, \\ \widetilde{\rho} = \rho_0(1 - \alpha(\widetilde{T} - T_0)). \end{cases} \tag{3.2}$$

Here $\widetilde{\mathrm{v}} = (\widetilde{u}, \widetilde{v})$ is the horizontal velocity, $\widetilde{w}$ the vertical velocity, $\widetilde{\rho}$ the density, $\widetilde{p}$ the pressure, and $\widetilde{T}$ the temperature; ∇ denotes the horizontal gradient operator; $\widetilde{\mathrm{v}}_t = \partial \mathrm{v}/\partial t$, etc. The independent variables are $(x, y, z) \in \mathcal{M} = (0, L_1) \times (0, L_2) \times (-L_3, 0)$, and $t > 0$.

As said before, in the physical context, the forcing terms $F_{\widetilde{\mathrm{v}}} = (F_{\widetilde{u}}, F_{\widetilde{v}})$ and $Q_{\widetilde{T}}$ vanish, but we introduce them here for mathematical generality and, below, to study the case of nonhomogeneous boundary conditions by homogenization of the boundary conditions.

After introducing the basic stratified flow, following the steps of Section 1, we reach the following system with five equations and five unknowns:

$$\begin{cases} u_t + \overline{U}_0 u_x - fv + \phi_x = 0, \\ v_t + \overline{U}_0 v_x + f(\overline{U}_0 + u) + \phi_y = 0, \\ \psi_t + \overline{U}_0 \psi_x + N^2 w = 0, \\ u_x + v_y + w_z = 0, \\ \phi_z = \psi. \end{cases} \tag{3.3}$$

The functions $u, v, \ldots$, are related to $\widetilde{u}, \widetilde{v}$ by Eq. (1.7).

3.2.1. Finite element expansion in the y-direction

The aim is to find (and numerically study) a 2D version of Eq. (3.3), which is physically interesting. For that purpose, we want the flow to be close to geostrophic equilibrium so that $u = u^g + u'$, or $\widetilde{u} = \overline{U} + u^g + u'$,[11] etc., where u^g, etc. (and as well $\overline{U} + u^g$, etc.), are geostrophic, and u', v', etc., are small compared to u^g, v^g, etc., which are themselves small compared to $\overline{u}$, $\overline{v}$, etc.

The geostrophic equation

$$p^g_y = -\rho_0 f u^g \tag{3.4}$$

[11] The notations u', v', etc., are not used in the sequel.

prevents us from taking functions $\widetilde{u} = u^g + u$, $\widetilde{p} = p^g + p, \ldots,$ independent of y. Indeed, if we consider a space periodic approximation with two or three modes of the Fourier series in y, Eq. (3.4) will introduce the Fourier series expansion of

$$h(y) = y, \quad 0 < y < L_2, \tag{3.5}$$

and, as is well-known, the discontinuity of (the periodic extension of) h leads to numerical oscillations.

Hence, our 2.5D model will allow linear variations in y, and in view of Eq. (3.4), it is then natural to introduce piecewise linear finite elements in the y direction. We introduce one middle point $0.5\,L_2$ in the middle of the interval $(0,\, L_2)$: $0,\, L_2$ play the role of boundaries, and values at $0.5\,L_2$ play the role of the flow "independent of y." We introduce three hat functions (finite elements) h_1, h_2, and h_3 (see Fig. 3.1) corresponding to the points $0,\, 0.5\,L_2$, and L_2. Instead of the usual hat function $\widetilde{h}_2$ (see Fig. 3.1), we use h_2 such that h_1, h_2, and h_3 are *orthogonal*.

We now look for approximate solutions of Eq. (3.3) of the form of

$$\begin{cases} u \simeq u_1(x,z,t)h_1(y) + u_2(x,z,t)h_2(y) + u_3(x,z,t)h_3(y), \\ v \simeq v_1(x,z,t)h_1(y) + v_2(x,z,t)h_2(y) + v_3(x,z,t)h_3(y), \\ w \simeq w_1(x,z,t)h_1(y) + w_2(x,z,t)h_2(y) + w_3(x,z,t)h_3(y), \\ \phi \simeq \phi_1(x,z,t)h_1(y) + \phi_2(x,z,t)h_2(y) + \phi_3(x,z,t)h_3(y), \\ \psi \simeq \psi_1(x,z,t)h_1(y) + \psi_2(x,z,t)h_2(y) + \psi_3(x,z,t)h_3(y), \end{cases} \tag{3.6}$$

and consider the corresponding finite elements (Galerkin) approximation of Eq. (3.3). We then introduce the expressions in Eq. (3.6) for u, v, w, ϕ, and ψ into the system in Eq. (3.3), multiply each equation by h_1, h_2, and h_3, respectively, and integrate over

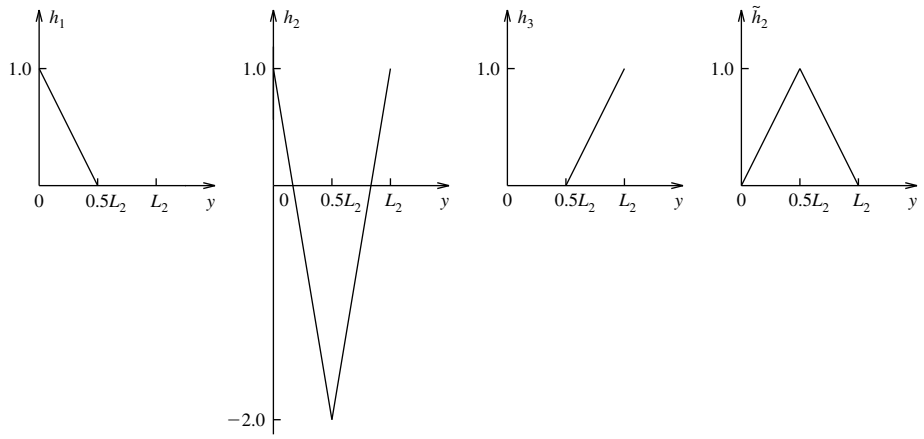


FIG. 3.1 The hat functions h_1, h_2, h_3 and $\widetilde{h}_2$.

(0, 1). Thanks to the orthogonality of h_1, h_2, and h_3, we obtain the following system:

$$\begin{cases} \boldsymbol{u}_t + \overline{U}_0 \boldsymbol{u}_x + \boldsymbol{\phi}_x - f\boldsymbol{v} = 0, \\ \boldsymbol{v}_t + \overline{U}_0 \boldsymbol{v}_x + f\boldsymbol{u} + \Lambda\boldsymbol{\phi} + \boldsymbol{f} = 0, \\ \boldsymbol{\psi}_t + \overline{U}_0 \boldsymbol{\psi}_x + N^2 \boldsymbol{w} = 0, \\ \boldsymbol{u}_x + \Lambda\boldsymbol{v} + \boldsymbol{w}_z = 0, \\ \boldsymbol{\psi} = \boldsymbol{\phi}_z. \end{cases} \tag{3.7}$$

Here

$$\boldsymbol{u} = (u_1, u_2, u_3)^T, \quad \boldsymbol{v} = (v_1, v_2, v_3)^T, \quad \boldsymbol{\phi} = (\phi_1, \phi_2, \phi_3)^T,$$
$$\boldsymbol{\psi} = (\psi_1, \psi_2, \psi_3)^T, \quad \boldsymbol{w} = (w_1, w_2, w_3)^T,$$

and

$$\Lambda = \frac{1}{L_2} \begin{pmatrix} -3 & -9 & 0 \\ \frac{1}{2} & 0 & -\frac{1}{2} \\ 0 & 9 & 3 \end{pmatrix} \quad \boldsymbol{f} = \frac{1}{L_2} \begin{pmatrix} \frac{3}{2} f\overline{U}_0 \\ -\frac{1}{2} f\overline{U}_0 \\ \frac{3}{2} f\overline{U}_0 \end{pmatrix}. \tag{3.8}$$

Note that the matrix Λ which has the physical dimension of $length^{-1}$ can be basically seen as the discretized form, in the context of this Galerkin procedure, of the differential operator $\partial/\partial y$.

We denote by $\mathcal{M}' = (0, L_1) \times (-L_3, 0)$ the two-dimensional spatial domain for the system in Eq. (3.7).

3.2.2. The normal mode expansion

As in Section 2 for the 2D case, we consider a normal mode expansion of the solutions of system in Eq. (3.7). That is, we look for solutions of this system in the form:

$$\begin{cases} (\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{\phi}) = \sum_{n \geq 0} \mathcal{U}_n(z)(\boldsymbol{u}_n, \boldsymbol{v}_n, \boldsymbol{\phi}_n)(x, t), \\ (\boldsymbol{w}, \boldsymbol{\psi}) = \sum_{n \geq 1} \mathcal{W}_n(z)(\boldsymbol{w}_n, \boldsymbol{\psi}_n)(x, t). \end{cases} \tag{3.9}$$

Here, $\boldsymbol{u}_n$, $\boldsymbol{v}_n$, etc., are vector functions as $\boldsymbol{u}$, $\boldsymbol{v}$, etc., but are independent of z. We refer the reader to Section 1 (or Temam and Tribbia [2003]) for the justification of the normal mode expansion. The specifications of the eigenfunctions $\mathcal{U}_n$ and $\mathcal{W}_n$ given in Section 2 are repeated here for convenience of the reader:

$$\begin{cases} \mathcal{U}_0 = \sqrt{\frac{1}{L_3}}, \text{ and } \mathcal{U}_n = \sqrt{\frac{2}{L_3}} \cos(\lambda_n z) \quad \text{for } n \geq 1, \\ \mathcal{W}_n = \sqrt{\frac{2}{L_3}} \sin(\lambda_n z) \quad \text{for } n \geq 1, \end{cases} \tag{3.10}$$

where $\lambda_n = n\pi/L_3$. We observe that, for $n, m \geq 1$,

$$\begin{cases} \displaystyle\int_{-L_3}^{0} \mathcal{U}_n(z)\mathcal{U}_m(z)\,\mathrm{d}z = \delta_{n,m}, \\ \displaystyle\int_{-L_3}^{0} \mathcal{W}_n(z)\mathcal{W}_m(z)\,\mathrm{d}z = \delta_{n,m}, \\ \displaystyle\int_{-L_3}^{0} \mathcal{U}_n(z)\mathcal{W}_m(z)\,\mathrm{d}z = 0, \\ \mathcal{U}_n'(z) = -\lambda_n \mathcal{W}_n(z), \\ \mathcal{W}_n'(z) = \lambda_n \mathcal{U}_n(z). \end{cases} \tag{3.11}$$

We then introduce Eq. (3.9) into the system in Eq. (3.7). For each $n \geq 0$, we multiply each equation by $\mathcal{U}_n$ (or $\mathcal{W}_n$ for the third and fifth equations) and integrate over $(-L_3, 0)$. When $n = 0$, we obtain a system for $\boldsymbol{u}_0$, $\boldsymbol{v}_0$, and $\boldsymbol{\phi}_0$ only:

$$\begin{cases} \boldsymbol{u}_{0t} + \overline{U}_0 \boldsymbol{u}_{0x} + \boldsymbol{\phi}_{0x} - f\boldsymbol{v}_0 = 0, \\ \boldsymbol{v}_{0t} + \overline{U}_0 \boldsymbol{v}_{0x} + f\boldsymbol{u}_0 + \Lambda\boldsymbol{\phi}_0 + \boldsymbol{f}_0 = 0, \\ \boldsymbol{u}_{0x} + \Lambda\boldsymbol{v}_0 = 0. \end{cases} \tag{3.12}$$

Here, Λ is the same as in Eq. (3.8), and

$$\boldsymbol{f}_0 = \sqrt{L_3}\boldsymbol{f}. \tag{3.13}$$

When $n \geq 1$, the corresponding system for each mode has the same form:

$$\begin{cases} \boldsymbol{u}_{nt} + \overline{U}_0 \boldsymbol{u}_{nx} + \boldsymbol{\phi}_{nx} - f\boldsymbol{v}_n = 0, \\ \boldsymbol{v}_{nt} + \overline{U}_0 \boldsymbol{v}_{nx} + f\boldsymbol{u}_n + \Lambda\boldsymbol{\phi}_n = 0, \\ \boldsymbol{\psi}_{nt} + \overline{U}_0 \boldsymbol{\psi}_{nx} + N^2 \boldsymbol{w}_n = 0, \\ \boldsymbol{u}_{nx} + \Lambda\boldsymbol{v}_n + \lambda_n \boldsymbol{w}_n = 0, \\ -\lambda_n \boldsymbol{\phi}_n = \boldsymbol{\psi}_n. \end{cases} \tag{3.14}$$

From the last two equations, we notice that

$$\boldsymbol{\phi}_n = -\frac{1}{\lambda_n}\boldsymbol{\psi}_n, \quad \boldsymbol{w}_n = -\frac{1}{\lambda_n}(\boldsymbol{u}_{nx} + \Lambda\boldsymbol{v}_n), \tag{3.15}$$

which means that $\boldsymbol{\phi}_n$ and $\boldsymbol{w}_n$ are determined by the other three unknowns, they are *diagnostic variables*. Then, we can eliminate $\boldsymbol{\phi}_n$ and $\boldsymbol{w}_n$ in Eq. (3.14) and obtain a

system for $\boldsymbol{u}_n$, $\boldsymbol{v}_n$, and $\boldsymbol{\psi}_n$, for each $n \geq 1$:

$$\begin{cases} \boldsymbol{u}_{nt} + \overline{U}_0 \boldsymbol{u}_{nx} - \dfrac{1}{\lambda_n} \boldsymbol{\psi}_{nx} - f \boldsymbol{v}_n = 0, \\ \boldsymbol{v}_{nt} + \overline{U}_0 \boldsymbol{v}_{nx} + f \boldsymbol{u}_n - \dfrac{1}{\lambda_n} \Lambda \boldsymbol{\psi}_n = 0, \\ \boldsymbol{\psi}_{nt} - \dfrac{N^2}{\lambda_n} \boldsymbol{u}_{nx} + \overline{U}_0 \boldsymbol{\psi}_{nx} - \dfrac{N^2}{\lambda_n} \Lambda \boldsymbol{v}_n = 0. \end{cases} \tag{3.16}$$

In Section 3.2.3, we will present the boundary conditions at $x = 0$ and L_1 for the modes $n \geq 1$. These boundary conditions will ensure the well-posedness of the system in Eq. (3.7), which we are going to establish in Section 3.3. Due to its different and somehow irregular form, the system in Eq. (3.12) of the zero mode will be treated separately at the end of Section 3.3.

3.2.3. Boundary conditions at $x = 0$, L_1

The analysis by which we determine the boundary conditions for the systems in Eq. (3.16), and ultimately for Eq. (3.7), is similar to that in the 2D case. We will review the spirit of the analysis here for the sake of completeness, and then list the boundary conditions that we propose for the systems in Eq. (3.16).

The matrix associated with the coefficients of the first-order terms with respect to x reads:

$$\begin{pmatrix} \overline{U}_0 & 0 & -\frac{1}{\lambda_n} \\ 0 & \overline{U}_0 & 0 \\ -\frac{N^2}{\lambda_n} & 0 & \overline{U}_0 \end{pmatrix}.$$

There are three eigenvalues to this matrix, namely $\overline{U}_0 + N/\lambda_n$, $\overline{U}_0$, and $\overline{U}_0 - N/\lambda_n$. Because $\overline{U}_0$ and λ_n are positive, each mode has at least two positive eigenvalues. However, as before and depending on n, the third eigenvalue $\overline{U}_0 - 1/\lambda_n$ can be either positive or negative for the actual (physical) values of $\overline{U}_0$. We say that the corresponding mode is supercritical in the first case and subcritical in the second case. The supercritical modes require three boundary conditions at $x = 0$, while the subcritical modes require two boundary conditions at $x = 0$ and one at $x = L_1$. This mandates that we impose different boundary conditions according to the type of the modes.

We first note that the sequence $\{ \lambda_n \}$ is monotone and $\lambda_n \longrightarrow \infty$ as $n \longrightarrow \infty$. Therefore, there are only a finite number of subcritical modes, which, however, are the most challenging and also the most important ones as they carry much energy.

We continue to denote by n_c the number of subcritical modes, which is defined as in Eq. (1.23):

$$\lambda_{n_c} = \frac{n_c \pi}{L_3} < \frac{1}{\overline{U}_0},$$

$$\lambda_{n_c+1} = \frac{(n_c + 1)\pi}{L_3} > \frac{1}{\overline{U}_0}.$$

REMARK 3.1. When $\overline{U}_0 - N/\lambda_n = 0$ for some n, the mode will be neither subcritical nor supercritical. But, this can be easily avoided by modifying as necessary the velocity of the reference flow. For this reason, we assume as in Eq. (1.24) that $\overline{U}_0 \neq N/\lambda_n$ for all n's.

For the supercritical modes, i.e., when $n > n_c$, we take the natural boundary conditions:

$$\begin{cases} \boldsymbol{u}_n(0,t) = 0, \\ \boldsymbol{v}_n(0,t) = 0, \\ \boldsymbol{\psi}_n(0,t) = 0. \end{cases} \tag{3.17}$$

For the subcritical modes, i.e., when $1 \leq n \leq n_c$, we impose the boundary conditions in the following way:

$$\begin{cases} \boldsymbol{\xi}_n(0,t) = 0, \\ \boldsymbol{v}_n(0,t) = 0, \\ \boldsymbol{\eta}_n(L_1,t) = 0. \end{cases} \tag{3.18}$$

Here, $\boldsymbol{\xi}_n = \boldsymbol{u}_n - \boldsymbol{\psi}_n/N$, $\boldsymbol{v}_n = \boldsymbol{v}_n$, and $\boldsymbol{\eta}_n = \boldsymbol{u}_n + \boldsymbol{\psi}_n/N$ are the three eigenvectors corresponding to $\overline{U}_0 + N/\lambda_n$, $\overline{U}_0$, and $\overline{U}_0 - N/\lambda_n$, respectively.

REMARK 3.2. In this subsection, the boundary conditions are given for each mode. The boundary conditions for the system in Eq. (3.7) will come directly from Eqs. (3.17) and (3.18), and will be presented later on (see Eqs. (3.22) and (3.23)).

REMARK 3.3. For most of this section, the boundary conditions will be homogeneous. But, at the end, we will explain how to handle the nonhomogeneous case. Some technicalities related to the so-called compatibility conditions will appear.

3.3. Well-posedness of the linear system

3.3.1. The functional setting

We want to write Eq. (3.7) (the zero mode excluded, see Remark 3.4 below) as an initial value problem of the form

$$\begin{cases} \dfrac{\mathrm{d}U}{\mathrm{d}t} + AU = F, \\ U(0) = U_0. \end{cases} \tag{3.19}$$

Here, $U = U(t)$ stands for $(\boldsymbol{u}(t), \boldsymbol{v}(t), \boldsymbol{\psi}(t))$, A is an unbounded operator in H with domain $D(A) \subset H$, and $U_0 \in D(A)$, $F \in H$. The space H is defined as follows:

$$H = H_{\boldsymbol{u}} \times H_{\boldsymbol{v}} \times H_{\boldsymbol{\psi}}, \tag{3.20}$$

where

$$H_{\boldsymbol{u}} = H_{\boldsymbol{v}} = \left\{ \boldsymbol{u} \in \mathbf{L}^2(\mathcal{M}') \;\middle|\; \int_{-L_3}^{0} \boldsymbol{u}(x, z)\, \mathrm{d}z = \mathbf{0}, \text{ for a.e. } x \in (0, L_1) \right\},$$

$$H_{\boldsymbol{\psi}} = \mathbf{L}^2(\mathcal{M}').$$

In the definitions above, $\mathcal{M}'$ is the two-dimensional domain $(0, L_1) \times (-L_3, 0)$ (see at the end of Section 3.2.1). The convention that $\mathbf{L}^2(\mathcal{M}') = (L^2(\mathcal{M}'))^3$ has been used. Similarly, later in this section, we will use $\mathbf{H}^1(\mathcal{M}')$, $\boldsymbol{\mathcal{D}}(\mathcal{M}')$, etc., for the corresponding vector function spaces; H is endowed with the following scalar product:

$$(U, \widetilde{U})_H = \int_{\mathcal{M}'} \left(\boldsymbol{u} \cdot \widetilde{\boldsymbol{u}} + \boldsymbol{v} \cdot \widetilde{\boldsymbol{v}} + \frac{1}{N^2} \boldsymbol{\psi} \cdot \widetilde{\boldsymbol{\psi}} \right) \mathrm{d}\mathcal{M}'.$$

Clearly, H is a closed subspace of $(\mathbf{L}^2(\mathcal{M}'))^3$, and the norm of H derived from the scalar product $(\cdot, \cdot)_H$ is equivalent to that of $(\mathbf{L}^2(\mathcal{M}'))^3$.

We denote by P the orthogonal projector from $\mathbf{L}^2(\mathcal{M}')$ onto $H_{\boldsymbol{u}}$ (and also onto $H_{\boldsymbol{v}}$, since $H_{\boldsymbol{u}}$ and $H_{\boldsymbol{v}}$ are identical.) Hence, for each $\boldsymbol{g} \in \mathbf{L}^2(\mathcal{M}')$,

$$\begin{cases} P(\boldsymbol{g}) = \boldsymbol{g} - \dfrac{1}{L_3} \displaystyle\int_{-L_3}^{0} \boldsymbol{g}(x, z)\, \mathrm{d}z, \\ \\ (I - P)(\boldsymbol{g}) = \dfrac{1}{L_3} \displaystyle\int_{-L_3}^{0} \boldsymbol{g}(x, z)\, \mathrm{d}z. \end{cases} \tag{3.21}$$

We can easily check that $P(\boldsymbol{g}) \in H_{\boldsymbol{u}}$ and $(I - P)(\boldsymbol{g}) \in H_{\boldsymbol{u}}^{\perp}$. We can also show that $H_{\boldsymbol{u}}^{\perp} = \mathbf{L}_x^2(0, L_1)$. Indeed, for each $\boldsymbol{f} \in \mathbf{L}_x^2(0, L_1)$, $P(\boldsymbol{f}) = \mathbf{0}$, and so $\boldsymbol{f} \in H_{\boldsymbol{u}}^{\perp}$. If, on the other hand, $\boldsymbol{f} \in H_{\boldsymbol{u}}^{\perp}$, then $(I - P)\boldsymbol{f} = \boldsymbol{f}$. Hence, $\boldsymbol{f}$ is independent of z, and $\boldsymbol{f} \in \mathbf{L}_x^2(0, L_1)$.

The unknown U is subjected to modal boundary conditions, which are listed below. The parallelism between the modal boundary conditions for U and the boundary conditions for each mode (see Eqs. (3.17) and (3.18)) is obvious.

For $n > n_c$ (i.e., *for the supercritical modes*),

$$\begin{cases} \displaystyle\int_{-L_3}^{0} \boldsymbol{u}(0, z) \mathcal{U}_n(z)\, \mathrm{d}z = \mathbf{0}, \\ \\ \displaystyle\int_{-L_3}^{0} \boldsymbol{v}(0, z) \mathcal{U}_n(z)\, \mathrm{d}z = \mathbf{0}, \\ \\ \displaystyle\int_{-L_3}^{0} \boldsymbol{\psi}(0, z) \mathcal{W}_n(z)\, \mathrm{d}z = \mathbf{0}. \end{cases} \tag{3.22}$$

For $1 \le n \le n_c$ (i.e., *for the subcritical modes*),

$$\begin{cases} \displaystyle\int_{-L_3}^{0} \boldsymbol{u}(0,z)\mathcal{U}_n(z) - \frac{1}{N}\int_{-L_3}^{0} \boldsymbol{\psi}(0,z)\mathcal{W}_n(z)\,\mathrm{d}z = \mathbf{0}, \\ \displaystyle\int_{-L_3}^{0} \boldsymbol{v}(0,z)\mathcal{U}_n(z) = \mathbf{0}, \\ \displaystyle\int_{-L_3}^{0} \boldsymbol{u}(L_1,z)\mathcal{U}_n(z) + \frac{1}{N}\int_{-L_3}^{0} \boldsymbol{\psi}(L_1,z)\mathcal{W}_n(z)\,\mathrm{d}z = \mathbf{0}. \end{cases} \tag{3.23}$$

We now define $D(A)$ as follows:

$$D(A) = \{\, U \in H \mid U_x \in (\mathbf{L}^2(\mathcal{M}'))^3, \tag{3.24}$$

and U verifies the BCs Eqs. (3.22) and (3.23) $\}$.

Then, for every $U \in D(A)$, AU is defined by

$$AU = \begin{pmatrix} \overline{U}_0 \boldsymbol{u}_x - f\boldsymbol{v} - P[\int_z^0 \boldsymbol{\psi}_x(x,z')\,\mathrm{d}z'] \\ \overline{U}_0 \boldsymbol{v}_x + f\boldsymbol{u} - P[\int_z^0 \Lambda\boldsymbol{\psi}(x,z')\,\mathrm{d}z'] \\ \overline{U}_0 \boldsymbol{\psi}_x + N^2 \int_z^0 (\boldsymbol{u}_x + \Lambda \boldsymbol{v})\,\mathrm{d}z' \end{pmatrix}. \tag{3.25}$$

REMARK 3.4. By the definition of the spaces H and $D(A)$, we include in the system Eq. (3.19) all the modes with $n \ge 1$. The zero mode ($n = 0$) is excluded from the system, and will be treated separately.

3.3.2. Main result

We will prove the well-posedness of the system in Eq. (3.19) with the help of the Hille–Yosida theorem, Theorem 2.1. The main result of this section concerning Eq. (3.19) is as follows:

THEOREM 3.1. *Let H, A, and $D(A)$ be defined as in Section 3.3.1. Then, the initial value problem in Eq. (3.19) is well-posed. That is, for every $t_1 > 0$, and for every $U_0 \in D(A)$, $F \in L^1(0,t_1;H)$, with $F' \in L^1(0,t_1;H)$, Eq. (3.19) has a unique solution U such that*

$$U \in C([0,t_1];H) \cap L^\infty(0,t_1;D(A)), \quad \frac{\mathrm{d}U}{\mathrm{d}t} \in L^\infty(0,t_1;H). \tag{3.26}$$

3.3.3. Proof of Theorem 3.1

We first want to rewrite AU in another form, and we also want to introduce the adjoint A^* of A and its domain $D(A^*)$, which are needed in the course of the proof.

We want to express AU in terms of $\boldsymbol{u}_n$, $\boldsymbol{v}_n$, and $\boldsymbol{\psi}_n$. This form is more convenient for the calculations. To this end, we simply introduce the normal mode expansions in Eq. (3.9) of $\boldsymbol{u}$, $\boldsymbol{v}$, and $\boldsymbol{\psi}$ into Eq. (3.25). Note that since $\boldsymbol{u} \in H_{\boldsymbol{u}}$ and $\boldsymbol{v} \in H_{\boldsymbol{v}}$, $\boldsymbol{u}_0$, and $\boldsymbol{v}_0$ vanish.

After working out the integrations and grouping the coefficients of the eigenfunctions, we obtain

$$AU = \begin{pmatrix} \displaystyle\sum_{n\geq 1} (\overline{U}_0 \boldsymbol{u}_{nx} - \frac{1}{\lambda_n}\boldsymbol{\psi}_{nx} - f\boldsymbol{v}_n)\,\mathcal{U}_n \\ \displaystyle\sum_{n\geq 1} (\overline{U}_0 \boldsymbol{v}_{nx} + f\boldsymbol{u}_n - \frac{1}{\lambda_n}\Lambda\boldsymbol{\psi}_n)\,\mathcal{U}_n \\ \displaystyle\sum_{n\geq 1} \left(-\frac{N^2}{\lambda_n}\boldsymbol{u}_{nx} + \overline{U}_0\boldsymbol{\psi}_{nx} - \frac{N^2}{\lambda_n}\Lambda\boldsymbol{v}_n\right)\mathcal{W}_n \end{pmatrix}. \tag{3.27}$$

We recall (see, e.g., Rudin [1991]) that given an unbounded operator A from $D(A)$ into H, the domain of its adjoint consists in the U in H such that $U \longrightarrow (AU, \widetilde{U})$ is a linear functional K on $D(A)$, continuous for the norm of H, in which case $A^*\widetilde{U} = K$. The determination of A^* introduces the following boundary conditions for $\widetilde{U}$:
For the supercritical modes, i.e., $n > n_c$,

$$\begin{cases} \widetilde{\boldsymbol{u}}_n(L_1, t) = 0, \\ \widetilde{\boldsymbol{v}}_n(L_1, t) = 0, \\ \widetilde{\boldsymbol{\psi}}_n(L_1, t) = 0. \end{cases} \tag{3.28}$$

For the the subcritical modes, i.e., $1 \leq n \leq n_c$,

$$\begin{cases} \widetilde{\boldsymbol{u}}_n(L_1, t) - \dfrac{1}{N}\widetilde{\boldsymbol{\psi}}_n(L_1, t) = 0, \\ \widetilde{\boldsymbol{v}}_n(L_1, t) = 0, \\ \widetilde{\boldsymbol{u}}_n(0, t) + \dfrac{1}{N}\widetilde{\boldsymbol{\psi}}_n(0, t) = 0. \end{cases} \tag{3.29}$$

A simple analysis, which we skip,[12] shows that the domain $D(A^*)$ of A^* is as follows:

$$D(A^*) = \{\, \widetilde{U} \in H \mid \widetilde{U}_x \in (\mathbf{L}^2(\mathcal{M}'))^3, \tag{3.30}$$

and $\widetilde{U}$ verifies the BCs in Eqs. (3.28) and (3.29) $\}$.

Furthermore, for $\widetilde{U} \in D(A^*)$,

$$A^*\widetilde{U} = \begin{pmatrix} \displaystyle\sum_{n\geq 1} \Big(-\overline{U}_0\widetilde{\boldsymbol{u}}_{nx} + \frac{1}{\lambda_n}\widetilde{\boldsymbol{\psi}}_{nx} + f\,\widetilde{\boldsymbol{v}}_n\Big)\mathcal{U}_n \\ \displaystyle\sum_{n\geq 1} \Big(-\overline{U}_0\widetilde{\boldsymbol{v}}_{nx} - f\widetilde{\boldsymbol{u}}_n - \frac{1}{\lambda_n}\Lambda^T\widetilde{\boldsymbol{\psi}}_n\Big)\mathcal{U}_n \\ \displaystyle\sum_{n\geq 1} \left(\frac{N^2}{\lambda_n}\widetilde{\boldsymbol{u}}_{nx} - \overline{U}_0\widetilde{\boldsymbol{\psi}}_{nx} - \frac{N^2}{\lambda_n}\Lambda^T\widetilde{\boldsymbol{v}}_n\right)\mathcal{W}_n \end{pmatrix}. \tag{3.31}$$

[12] See in Section 4 a more involved analysis in dimension 3.

The coefficient matrix Λ in Eq. (3.31) is the same as in Eq. (3.8), and Λ^T is its transpose.

The proof of Theorem 3.1 essentially consists of the verification of the hypotheses of Theorem 2.1. We will do it in the Lemmas 3.1 to 3.5. We will then summarize the proof of Theorem 3.1 at the end of this section.

LEMMA 3.1. *There exists* $\delta > 0$ *such that* $A + \delta I \geq 0$ *and* $A^* + \delta I \geq 0$, *i.e.,* $((A + \delta I)U,\ U)_H \geq 0$ *for each* $U \in D(A)$ *and* $((A^* + \delta I)\widetilde{U},\ \widetilde{U})_H \geq 0$ *for each* $\widetilde{U} \in D(A^*)$.

PROOF. Let $U \in D(A)$. Then U has the normal mode expansion

$$U = \Big(\sum_{n\geq 1} \boldsymbol{u}_n \mathcal{U}_n, \sum_{n\geq 1} \boldsymbol{v}_n \mathcal{U}_n, \sum_{n\geq 1} \boldsymbol{\psi}_n \mathcal{W}_n \Big).$$

Using the expression in Eq. (3.27) for AU, we compute

$$\begin{aligned}
(AU, U)_H &= \int_0^{L_1} \int_{-L_3}^{0} AU \cdot U \, \mathrm{d}z \, \mathrm{d}x \\
&= \int_0^{L_1} \Bigg[\sum_{n\geq 1} \Big(\overline{U}_0 \boldsymbol{u}_{nx} \cdot \boldsymbol{u}_n - \frac{1}{\lambda_n} \boldsymbol{\psi}_{nx} \cdot \boldsymbol{u}_n - f \boldsymbol{v}_n \cdot \boldsymbol{u}_n \Big) \\
&\quad + \sum_{n\geq 1} \Big(\overline{U}_0 \boldsymbol{v}_{nx} \cdot \boldsymbol{v}_n + f \boldsymbol{u}_n \cdot \boldsymbol{v}_n - \frac{1}{\lambda_n} \Lambda \boldsymbol{\psi}_n \cdot \boldsymbol{v}_n \Big) \\
&\quad + \sum_{n\geq 1} \Big(- \frac{1}{\lambda_n} \boldsymbol{u}_{nx} \cdot \boldsymbol{\psi}_n + \frac{\overline{U}_0}{N^2} \boldsymbol{\psi}_{nx} \cdot \boldsymbol{\psi}_n - \frac{1}{\lambda_n} \Lambda \boldsymbol{v}_n \cdot \boldsymbol{\psi}_n \Big) \Bigg] \mathrm{d}x. \\
&= \sum_{n\geq 1} \Bigg[\frac{\overline{U}_0}{2} {\boldsymbol{u}_n}^2(L_1) - \frac{\overline{U}_0}{2} {\boldsymbol{u}_n}^2(0) + \frac{\overline{U}_0}{2} {\boldsymbol{v}_n}^2(L_1) - \frac{\overline{U}_0}{2} {\boldsymbol{v}_n}^2(0) \\
&\quad + \frac{\overline{U}_0}{2N^2} {\boldsymbol{\psi}_n}^2(L_1) - \frac{\overline{U}_0}{2N^2} {\boldsymbol{\psi}_n}^2(0) \\
&\quad - \frac{1}{\lambda_n} \boldsymbol{u}_n(L_1) \cdot \boldsymbol{\psi}_n(L_1) + \frac{1}{\lambda_n} \boldsymbol{u}_n(0) \cdot \boldsymbol{\psi}_n(0) \\
&\quad - \frac{1}{\lambda_n} \int_0^{L_1} (\Lambda \boldsymbol{\psi}_n \cdot \boldsymbol{v}_n + \Lambda \boldsymbol{v}_n \cdot \boldsymbol{\psi}_n) \, \mathrm{d}x \Bigg].
\end{aligned}$$

We now separate the supercritical and subcritical modes, and drop those terms that vanish according to the boundary conditions in Eqs. (3.17), (3.18). There

remains:

$$
\begin{aligned}
(AU, U) = \sum_{n>n_c} \Bigg[& \frac{\overline{U}_0}{2} \boldsymbol{u}_n{}^2(L_1) + \frac{\overline{U}_0}{2} \boldsymbol{v}_n{}^2(L_1) + \frac{\overline{U}_0}{2N^2} \boldsymbol{\psi}_n{}^2(L_1) \\
& - \frac{1}{\lambda_n} \boldsymbol{u}_n(L_1) \cdot \boldsymbol{\psi}_n(L_1) \Bigg] \\
& + \sum_{1 \le n \le n_c} \Bigg[\frac{\overline{U}_0}{2} \boldsymbol{u}_n{}^2(L_1) - \frac{\overline{U}_0}{2} \boldsymbol{u}_n{}^2(0) + \frac{\overline{U}_0}{2} \boldsymbol{v}_n{}^2(L_1) \\
& + \frac{\overline{U}_0}{2N^2} \boldsymbol{\psi}_n{}^2(L_1) - \frac{\overline{U}_0}{2N^2} \boldsymbol{\psi}_n{}^2(0) \\
& - \frac{1}{\lambda_n} \boldsymbol{u}_n(L_1) \cdot \boldsymbol{\psi}_n(L_1) + \frac{1}{\lambda_n} \boldsymbol{u}_n(0) \cdot \boldsymbol{\psi}_n(0) \Bigg] \\
& + \sum_{n \ge 1} \frac{1}{\lambda_n} \int_0^{L_1} (-\Lambda \boldsymbol{\psi}_n \cdot \boldsymbol{v}_n - \Lambda \boldsymbol{v}_n \cdot \boldsymbol{\psi}_n) \, \mathrm{d}x.
\end{aligned}
$$

We then write

$$
(AU,\ U) = \sum_{n>n_c} \mathrm{I}_n + \sum_{1 \le n \le n_c} \mathrm{II}_n + \mathrm{III},
$$

where

$$
\begin{aligned}
\mathrm{I}_n = & \frac{\overline{U}_0}{2} \boldsymbol{u}_n{}^2(L_1) + \frac{\overline{U}_0}{2} \boldsymbol{v}_n{}^2(L_1) + \frac{\overline{U}_0}{2N^2} \boldsymbol{\psi}_n{}^2(L_1) \\
& - \frac{1}{\lambda_n} \boldsymbol{u}_n(L_1) \cdot \boldsymbol{\psi}_n(L_1), \\
\mathrm{II}_n = & \frac{\overline{U}_0}{2} \boldsymbol{u}_n{}^2(L_1) - \frac{\overline{U}_0}{2} \boldsymbol{u}_n{}^2(0) + \frac{\overline{U}_0}{2} \boldsymbol{v}_n{}^2(L_1) + \frac{\overline{U}_0}{2N^2} \boldsymbol{\psi}_n{}^2(L_1) \\
& - \frac{\overline{U}_0}{2N^2} \boldsymbol{\psi}_n{}^2(0) - \frac{1}{\lambda_n} \boldsymbol{u}_n(L_1) \cdot \boldsymbol{\psi}_n(L_1) + \frac{1}{\lambda_n} \boldsymbol{u}_n(0) \cdot \boldsymbol{\psi}_n(0), \\
\mathrm{III} = & \sum_{n \ge 1} \frac{1}{\lambda_n} \int_0^{L_1} (-\Lambda \boldsymbol{\psi}_n \cdot \boldsymbol{v}_n - \Lambda \boldsymbol{v}_n \cdot \boldsymbol{\psi}_n) \, \mathrm{d}x.
\end{aligned}
$$

We see that I_n is the sum of a quadratic form and the positive term $\overline{U}_0 \boldsymbol{v}_n^2(L_1)/2$. We find the determinant for the quadratic form to be $\lambda_n^{-2} - \overline{U}_0^2/N^{-2}$, which is < 0, thanks to the fact that $\overline{U}_0 - N/\lambda_n > 0$ for each supercritical mode. Hence, we have

$$
\mathrm{I}_n \ge 0, \qquad \text{for each } n > n_c.
$$

Using the boundary conditions in Eq. (3.18) for the subcritical modes, we find that

$$
\begin{aligned}
\mathrm{II}_n = & \frac{\overline{U}_0}{2} \boldsymbol{v}_n^2(L_1) + \left(\overline{U}_0 + \frac{N}{\lambda_n} \right) \boldsymbol{u}_n^2(L_1) + \left(\frac{N}{\lambda_n} - \overline{U}_0 \right) \boldsymbol{u}_n^2(0) \\
& \ge 0, \qquad \text{for each } 1 \le n \le n_c.
\end{aligned}
$$

The last inequality is due to the fact that $\overline{U}_0 - 1/\lambda_n < 0$ for each subcritical mode. We also find an upper bound on the absolute value of III:

$$\begin{aligned}
|\mathrm{III}| &\le \sum_{n\ge 1} \frac{1}{N\lambda_n} \int_0^{L_1} |(\Lambda \boldsymbol{\psi}_n \cdot \boldsymbol{v}_n + \Lambda \boldsymbol{v}_n \cdot \boldsymbol{\psi}_n)| \, \mathrm{d}x \\
&\le \sum_{n\ge 1} \frac{c_1'}{\lambda_n} \left(\int_0^{L_1} |\frac{\boldsymbol{\psi}_n}{N}|^2 \, \mathrm{d}x \right)^{\frac{1}{2}} \left(\int_0^{L_1} |\boldsymbol{v}_n|^2 \, \mathrm{d}x \right)^{\frac{1}{2}} \\
&\le \frac{c_1'}{2\lambda_1} \sum_{n\ge 1} \left(\int_0^{L_1} |\frac{\boldsymbol{\psi}_n}{N}|^2 \, \mathrm{d}x + \int_0^{L_1} |\boldsymbol{v}_n|^2 \, \mathrm{d}x \right),
\end{aligned}$$

where c_1' is the norm of the matrix Λ. Hence, $\delta \ge c_1'/2\lambda_n$ for every $n \ge 1$. Then,

$$|\mathrm{III}| \le \frac{c_1'}{2\lambda_1} |U|_H^2,$$

and, for any $\delta > c_1'/2\lambda_1$, the operator $A + \delta I$ is positive.

That $A^* + \delta I \ge 0$ (possibly with a different δ) can be shown in a similar way. And, we can always choose a constant δ such that the operators $A + \delta I$ and $A^* + \delta I$ are both positive. This completes the proof of Lemma 3.1. □

LEMMA 3.2. *$D(A)$ is dense in H, and A is a closed operator.*

PROOF. To show that $D(A)$ is dense in H, consider $U = (\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{\psi}) \in H$. Since $(\boldsymbol{\mathcal{D}}(\mathcal{M}'))^3$, the set of $\mathcal{C}^\infty$ functions with compact support in $\mathcal{M}'$, is dense in $\mathbf{L}^2(\mathcal{M}')$, U can be approximated in $\mathbf{L}^2(\mathcal{M}')$ by elements of $(\boldsymbol{\mathcal{D}}(\mathcal{M}'))^3$, say $\Phi^j = (\Phi_{\boldsymbol{u}}^j, \Phi_{\boldsymbol{v}}^j, \Phi_{\boldsymbol{\psi}}^j)$, $1 \le j \le \infty$. Since P is continuous in $(\mathbf{L}^2(\mathcal{M}'))^3$, $(P\Phi_{\boldsymbol{u}}^j, P\Phi_{\boldsymbol{v}}^j, \Phi_{\boldsymbol{\psi}}^j)$ also converge to $PU = U$ in H as $j \longrightarrow \infty$. The function Φ^j are compactly supported in $\mathcal{M}'$, and, by the definition of P, the functions $P\Phi^j$ are also compactly supported in $\mathcal{M}'$. The necessary boundary conditions are satisfied, and it is then clear that $(P\Phi_{\boldsymbol{u}}^j, P\Phi_{\boldsymbol{v}}^j, \Phi_{\boldsymbol{\psi}}^j)$ belong to $D(A)$.

To show that A is closed, we need to show that for a sequence $\{U^j\}_{j\ge 1}$ in $D(A)$ such that

$$U^j \longrightarrow U \quad \text{in } H, \tag{3.32}$$

$$AU^j \longrightarrow F \quad \text{in } H, \tag{3.33}$$

with $U,\ F \in H$, then $U \in D(A)$ and $F = AU$. For each component of U^j, U, and F, Eqs. (3.32) and (3.33) mean

$$\begin{cases} \boldsymbol{u}^j \longrightarrow \boldsymbol{u} & \text{in } H_{\boldsymbol{u}}, \\ \boldsymbol{v}^j \longrightarrow \boldsymbol{v} & \text{in } H_{\boldsymbol{v}}, \\ \boldsymbol{\psi}^j \longrightarrow \boldsymbol{\psi} & \text{in } H_{\boldsymbol{\psi}}, \end{cases} \tag{3.34}$$

and

$$\begin{cases} \overline{U}_0 \boldsymbol{u}_x^j - f\boldsymbol{v}^j - P[\displaystyle\int_z^0 \boldsymbol{\psi}_x^j(x,z')\,\mathrm{d}z'] \longrightarrow F_{\boldsymbol{u}} \quad \text{in } H_{\boldsymbol{u}}, \\ \overline{U}_0 \boldsymbol{v}_x^j + f\boldsymbol{u}^j - P[\displaystyle\int_z^0 \Lambda\boldsymbol{\psi}^j(x,z')\,\mathrm{d}z'] \longrightarrow F_{\boldsymbol{v}} \quad \text{in } H_{\boldsymbol{v}}, \\ \overline{U}_0 \boldsymbol{\psi}_x^j + N^2 \displaystyle\int_z^0 (\boldsymbol{u}_x^j + \Lambda\boldsymbol{v}^j)\,\mathrm{d}z' \longrightarrow F_{\boldsymbol{\psi}} \quad \text{in } H_{\boldsymbol{\psi}}. \end{cases} \tag{3.35}$$

With regard to each mode, Eq. (3.34) implies that

$$\begin{cases} \boldsymbol{u}_n^j \longrightarrow \boldsymbol{u}_n \quad \text{in } \mathbf{L}^2(0,L_1), \\ \boldsymbol{v}_n^j \longrightarrow \boldsymbol{v}_n \quad \text{in } \mathbf{L}^2(0,L_1), \\ \boldsymbol{\psi}_n^j \longrightarrow \boldsymbol{\psi}_n \quad \text{in } \mathbf{L}^2(0,L_1); \end{cases} \tag{3.36}$$

Eq. (3.34) also implies that $\sum_{n\geq 1} |\boldsymbol{u}_n^j|^2_{L^2(0,L_1)}$, $\sum_{n\geq 1} |\boldsymbol{v}_n^j|^2_{L^2(0,L_1)}$, and $\sum_{n\geq 1} |\boldsymbol{\psi}_n^j|^2_{L^2(0,L_1)}$ are uniformly bounded in j, by a bound of the $|U^n|^2_H$. Similarly, Eq. (3.35) implies that

$$\begin{cases} \overline{U}_0 \boldsymbol{u}_{nx}^j - \dfrac{1}{\lambda_n}\boldsymbol{\psi}_{nx}^j - f\boldsymbol{v}_n^j \equiv F_{\boldsymbol{u},n}^j \longrightarrow F_{\boldsymbol{u},n} \quad \text{in } \mathbf{L}^2(0,L_1), \\ \overline{U}_0 \boldsymbol{v}_{nx}^j + f\boldsymbol{u}_n^j - \dfrac{1}{\lambda_n}\Lambda\boldsymbol{\psi}_n^j \equiv F_{\boldsymbol{v},n}^j \longrightarrow F_{\boldsymbol{v},n} \quad \text{in } \mathbf{L}^2(0,L_1), \\ -\dfrac{N^2}{\lambda_n}\boldsymbol{u}_{nx}^j + \overline{U}_0 \boldsymbol{\psi}_{nx}^j - \dfrac{N^2}{\lambda_n}\Lambda\boldsymbol{v}_n^j \equiv F_{\boldsymbol{\psi},n}^j \longrightarrow F_{\boldsymbol{\psi},n} \quad \text{in } \mathbf{L}^2(0,L_1), \end{cases} \tag{3.37}$$

and that $\sum_{n\geq 1} |F_{\boldsymbol{u},n}^j|^2_{L^2(0,L_1)}$, $\sum_{n\geq 1} |F_{\boldsymbol{v},n}^j|^2_{L^2(0,L_1)}$, and $\sum_{n\geq 1} |F_{\boldsymbol{\psi},n}^j|^2_{L^2(0,L_1)}$ are uniformly bounded in j, by a bound of the $|AU^n|^2_H$. By the second equation in Eq. (3.37), we have

$$\boldsymbol{v}_{nx}^j = \frac{1}{\overline{U}_0}\left(-f\boldsymbol{u}_n^j + \frac{1}{\lambda_n}\Lambda\boldsymbol{\psi}_n^j + F_{\boldsymbol{v},n}^j\right). \tag{3.38}$$

Each term on the right-hand side of Eq. (3.38) converges in $\mathbf{L}^2(0,L_1)$, and therefore $\boldsymbol{v}_{nx}^j$ also converges in $\mathbf{L}^2(0,L_1)$. In addition, since on the right-hand side of Eq. (3.38), $\sum_{n\geq 1} |\boldsymbol{u}_n^j|^2_{L^2(0,L_1)}$, $\sum_{n\geq 1} |\boldsymbol{\psi}_n^j|^2_{L^2(0,L_1)}$, and $\sum_{n\geq 1} |F_{\boldsymbol{v},n}^j|^2_{L^2(0,L_1)}$ are all uniformly bounded in j, $\sum_{n\geq 1} |\boldsymbol{v}_{nx}^j|^2_{L^2(0,L_1)}$ is also uniformly bounded in j. These two facts imply that $\boldsymbol{v}_x^j$ converges in $\mathbf{L}^2(\mathcal{M}')$. Combining this result with Eq. (3.34), we conclude that $\boldsymbol{v}_x$ belongs to $\mathbf{L}^2(\mathcal{M}')$, and

$$\boldsymbol{v}_x^j \longrightarrow \boldsymbol{v}_x \quad \text{in } \mathbf{L}^2(\mathcal{M}'). \tag{3.39}$$

By the first and third equations in Eq. (3.37), we obtain

$$\boldsymbol{u}_{nx} = \frac{1}{\overline{U}_0^2 - N^2/\lambda_n^2}\left[(f\overline{U}_0 + \frac{N^2}{\lambda_n^2}\Lambda)\boldsymbol{v}_n + \overline{U}_0 F_{\boldsymbol{u},n} + \frac{1}{\lambda_n}F_{\boldsymbol{\psi},n}\right], \tag{3.40}$$

$$\boldsymbol{\psi}_{nx} = \frac{1}{\overline{U}_0^2 - N^2/\lambda_n^2}\left[\frac{N^2}{\lambda_n}(f + \overline{U}_0\Lambda)\boldsymbol{v}_n + \frac{N^2}{\lambda_n}F_{\boldsymbol{u},n} + \overline{U}_0 F_{\boldsymbol{\psi},n}\right]. \tag{3.41}$$

Following the similar idea as for $\boldsymbol{v}_x$, we can show that $\boldsymbol{u}_x$ and $\boldsymbol{\psi}_x$ belong to $\mathbf{L}^2(\mathcal{M}')$, and

$$\boldsymbol{u}_x^j \longrightarrow \boldsymbol{u}_x \quad \text{in } \mathbf{L}^2(\mathcal{M}'), \tag{3.42}$$

$$\boldsymbol{\psi}_x^j \longrightarrow \boldsymbol{\psi}_x \quad \text{in } \mathbf{L}^2(\mathcal{M}'). \tag{3.43}$$

To finish the proof, it remains to check that $U \in D(A)$ and $AU = F$. It is implied in the argument above that for each mode,

$$\begin{cases} \boldsymbol{u}_n^j \longrightarrow \boldsymbol{u}_n & \text{in } \mathbf{H}^1(0, L_1), \\ \boldsymbol{v}_n^j \longrightarrow \boldsymbol{v}_n & \text{in } \mathbf{H}^1(0, L_1), \\ \boldsymbol{\psi}_n^j \longrightarrow \boldsymbol{\psi}_n & \text{in } \mathbf{H}^1(0, L_1). \end{cases} \tag{3.44}$$

By the Sobolev embedding theorem, the convergences also hold in the space $\boldsymbol{C}([0, L_1])$, and thus the boundary conditions pass to the limit. Hence $U \in D(A)$. We infer from Eqs. (3.34), (3.39), (3.42), and (3.43) that $AU^j \longrightarrow AU$ in H. By Eq. (3.33), we have $AU = F$. This completes the proof of Lemma 3.2. □

LEMMA 3.3. *A^* is a closed operator.*

PROOF. This is a consequence of Lemma 3.2 and of the following standard lemma, which can be found in RUDIN [1991] and other functional analysis books. □

LEMMA 3.4. *If T is a densely defined operator in H, then T^* is a closed operator.*

We now state a well-known result and give a direct proof for the convenience of the reader (see HILLE and PHILLIPS [1974] and, for Banach spaces, see BRÉZIS [1970]):

LEMMA 3.5. *Let A and A^* be linear unbounded operators in H with domains $D(A)$ and $D(A^*)$, respectively, and let A^* be the adjoint operator of A (as an unbounded operator). It is also assumed that both $D(A)$ and $D(A^*)$ are dense in H. If furthermore A and A^* are both positive and closed, then $A + \mu_0 I$ and $A^* + \mu_0 I$ are onto for every $\mu_0 > 0$.*

PROOF. Consider $\epsilon > 0$, which will eventually converge to zero. For each value of ϵ, we construct a bilinear form b_ϵ on $D(A)$:

$$b_\epsilon(U, \widetilde{U}) = \epsilon(AU, A\widetilde{U})_H + (AU, \widetilde{U})_H + \mu_0(U, \widetilde{U})_H. \tag{3.45}$$

It is easy to check that b_ϵ is bilinear, bounded, and coercive on $D(A)$. Then, by the Lax–Milgram theorem, for any given $F \in H$, there exists a unique $U_\epsilon \in D(A)$ such that

$$\epsilon(AU_\epsilon, A\widetilde{U})_H + (AU_\epsilon, \widetilde{U})_H + \mu_0(U_\epsilon, \widetilde{U})_H = (F, \widetilde{U})_H \tag{3.46}$$

holds for any $\widetilde{U} \in D(A)$. For each ϵ, we observe that

$$\widetilde{U} \longrightarrow (AU_\epsilon, A\widetilde{U})_H = \frac{1}{\epsilon}(F - \mu_0 U_\epsilon - AU_\epsilon, \widetilde{U})_H \tag{3.47}$$

is a linear functional on $D(A)$ continuous for the norm of H. By the definition of the domain $D(A^*)$ of A^* (see, e.g., RUDIN [1991]), this means that

$$AU_\epsilon \in D(A^*), \tag{3.48}$$

and

$$A^*AU_\epsilon = \frac{1}{\epsilon}(F - \mu_0 U_\epsilon - AU_\epsilon) \text{ in } H. \tag{3.49}$$

We then write Eq. (3.49) as

$$\epsilon A^*AU_\epsilon + AU_\epsilon + \mu_0 U_\epsilon = F. \tag{3.50}$$

Multiplying Eq. (3.50) by U_ϵ, we obtain

$$\epsilon(AU_\epsilon, AU_\epsilon)_H + (AU_\epsilon, U_\epsilon)_H + \mu_0(U_\epsilon, U_\epsilon)_H = (F, U_\epsilon)_H. \tag{3.51}$$

Since $(AU_\epsilon, AU_\epsilon)_H \geq 0$ and $(AU_\epsilon, U_\epsilon)_H \geq 0$ by the assumption that A is positive, we then have

$$|U_\epsilon|_H \leq c|F|_H, \tag{3.52}$$

where c is a constant independent of ϵ. Therefore, there exists a subsequence $\epsilon' \longrightarrow 0$ such that

$$U_{\epsilon'} \rightharpoonup U \quad \text{weakly in } H, \tag{3.53}$$

for some $U \in H$. Multiplying Eq. (3.50) by AU_ϵ, we also obtain

$$\epsilon(A^*AU_\epsilon, AU_\epsilon)_H + (AU_\epsilon, AU_\epsilon)_H + \mu_0(AU_\epsilon, U_\epsilon)_H = (F, AU_\epsilon)_H. \tag{3.54}$$

Since $(A^*AU_\epsilon, AU_\epsilon)_H \geq 0$ and $(AU_\epsilon, U_\epsilon)_H \geq 0$, we have

$$|AU_\epsilon|_H \leq |F|_H. \tag{3.55}$$

This implies that there exists a subsequence, still denoted ϵ', such that

$$AU_{\epsilon'} \rightharpoonup \chi \quad \text{weakly in } H, \tag{3.56}$$

for some $\chi \in H$. By the assumption that the operator A is closed, and by Eqs. (3.53) and (3.56), we see that

$$U \in D(A) \quad \text{and} \quad \chi = AU. \tag{3.57}$$

We find from Eq. (3.50) that

$$A^*(\epsilon A U_\epsilon) = F - AU_\epsilon - \mu_0 U_\epsilon. \tag{3.58}$$

Since both AU_ϵ and U_ϵ converge weakly in H,

$$A^*(\epsilon A U_\epsilon) \rightharpoonup \sigma = F - AU - \mu_0 U \quad \text{weakly in } H. \tag{3.59}$$

And, since $|AU_\epsilon|_H$ is bounded independently of ϵ, we find that $\sigma = 0$, i.e.,

$$(A + \mu_0 I)U = F. \tag{3.60}$$

Thus, the claim that $A + \mu_0 I$ is onto for any $\mu_0 > 0$ is proved. That $A^* + \mu_0 I$ is onto for any $\mu_0 > 0$ can be proved in a similar way. □

Proof of Theorem 3.1. Let

$$U = e^{\delta t} U^b, \tag{3.61}$$

where δ is the positive constant chosen in the proof of Lemma 3.1. Inserting Eq. (3.61) into Eq. (3.19), we obtain an initial value problem for U^b:

$$\begin{cases} \dfrac{\mathrm{d}U^b}{\mathrm{d}t} + (A + \delta)U^b = \widetilde{F}, \\ U^b(0) = U_0. \end{cases} \tag{3.62}$$

If we can show that the system in Eq. (3.62) is well-posed and U^b satisfies Eq. (3.26), then, by the relation in Eq. (3.61), U satisfies Eq. (3.26) too, and of course, Eq. (3.19) is well-posed. We have in fact verified the hypotheses of Theorem 2.1 in Lemmas 3.1 to 3.5 for the operator $A + \delta I$ of Theorem 3.1. Now we readily apply Theorem 2.1 and complete the proof of Theorem 3.1. □

3.3.4. Treatment of the constant mode ($n = 0$)

We now introduce (propose) the boundary conditions for the zero mode, which is important because it contains much energy. What follows is valid whether the boundary conditions are homogeneous or not for the modes $n \geq 1$.

We start with a technical point which has no mathematical relevance, especially for the linearized equations for which the modes are decoupled; it has, however, a computational

and physical importance, in particular in the nonlinear case when all the modes are coupled: the function $\boldsymbol{\phi}_0$ will be decomposed in the sum

$$\boldsymbol{\phi}_0 = \bar{\boldsymbol{\phi}}_0 + \boldsymbol{\phi}_0', \tag{3.63}$$

where $\bar{\boldsymbol{\phi}}_0$, which is not unique, is one of the constant solutions[13] of

$$\Lambda \bar{\boldsymbol{\phi}}_0 = -\boldsymbol{f}_0, \tag{3.64}$$

with Λ and $\boldsymbol{f}_0$ defined in Eqs. (3.8) and (3.13), respectively, while $\boldsymbol{\phi}_0'$ needs to be determined.[14]

Then, for $\boldsymbol{u}_0$, $\boldsymbol{v}_0$, and $\boldsymbol{\phi}_0'$, we propose the following boundary conditions:

$$\begin{cases} \boldsymbol{u}_0(0,t) = \boldsymbol{u}_l(t), \\ \boldsymbol{v}_0(0,t) = \boldsymbol{v}_l(t), \\ \boldsymbol{\phi}_0'(0,t) = \boldsymbol{\phi}_l'(t). \end{cases} \tag{3.65}$$

Of course, the third equation in Eq. (3.65) is the same as $\boldsymbol{\phi}_0(0,t) = \boldsymbol{\phi}_l(t) = \bar{\boldsymbol{\phi}}_0 + \boldsymbol{\phi}_l'(t)$. We can obtain $\boldsymbol{v}_{0x}(0,t)$ from second equation of Eq. (3.12), i.e.,

$$\boldsymbol{v}_{0x}(0,t) = -\frac{1}{\overline{U}_0}(\boldsymbol{v}_{lt} + f\boldsymbol{u}_l + \Lambda\boldsymbol{\phi}_l'). \tag{3.66}$$

We then multiply the first equation of Eq. (3.12) by Λ,

$$(\Lambda\boldsymbol{u}_0)_t + \overline{U}_0(\Lambda\boldsymbol{u}_0)_x + \Lambda\boldsymbol{\phi}_{0x}' - f\Lambda\boldsymbol{v}_0 = 0, \tag{3.67}$$

and differentiate the second equation of Eq. (3.12) with respect to x,

$$(\boldsymbol{v}_{0x})_t + \overline{U}_0\boldsymbol{v}_{0xx} + f\boldsymbol{u}_{0x} + \Lambda\boldsymbol{\phi}_{0x}' = 0. \tag{3.68}$$

By subtracting Eq. (3.68) from Eq. (3.67), we obtain, thanks to third equation of Eq. (3.12),

$$(\Lambda\boldsymbol{u}_0 - \boldsymbol{v}_{0x})_t + \overline{U}_0(\Lambda\boldsymbol{u}_0 - \boldsymbol{v}_{0x})_x = 0. \tag{3.69}$$

The value of $\Lambda\boldsymbol{u}_0 - \boldsymbol{v}_{0x}$ at $x = 0$ is known, and therefore we can solve the equation above for $\Lambda\boldsymbol{u}_0 - \boldsymbol{v}_{0x}$. Once we have found $\Lambda\boldsymbol{u}_0 - \boldsymbol{v}_{0x}$, say $\Lambda\boldsymbol{u}_0 - \boldsymbol{v}_{0x} = k(x,t)$, then with third equation of Eq. (3.12), we have

$$\begin{cases} \boldsymbol{u}_{0x} + \Lambda\boldsymbol{v}_0 = 0, \\ \boldsymbol{v}_{0x} - \Lambda\boldsymbol{u}_0 = -k(x,t). \end{cases} \tag{3.70}$$

We can solve this linear ODE system with the boundary conditions for $\boldsymbol{u}_0$ and $\boldsymbol{v}_0$ at $x = 0$, which are given.

[13] Det $\Lambda = 0$, and Λ is of rank 2.

[14] As we observed, Λ is essentially the mathematical representation of $\partial/\partial y$ in the finite-element Galerkin procedure, and $\bar{\boldsymbol{\phi}}_0$ is the part of the basic geostrophic flow alluded to in Eq. (3.1).

We now have $\boldsymbol{u}_0$ and $\boldsymbol{v}_0$; we can solve first equation of Eq. (3.12) for $\boldsymbol{\phi}_0 = \bar{\boldsymbol{\phi}}_0 + \boldsymbol{\phi}_0'$, since the boundary condition for $\boldsymbol{\phi}_0$ (or $\boldsymbol{\phi_0}'$) is also given at $x = 0$.

We leave as an exercise to the reader to find the suitable regularity assumptions for the data $\boldsymbol{u}_l$, $\boldsymbol{v}_l$, and $\boldsymbol{\phi}_l'$.

3.3.5. The case of nonhomogeneous boundary conditions

In practical simulations, we want to be able to solve the PEs Eq. (3.7) with *nonhomogeneous* boundary conditions at $x = 0$ and L_1. We write Eq. (3.7) in a form similar to Eq. (3.19) corresponding to the elimination of $\boldsymbol{w}$ and $\boldsymbol{\phi}$ and the exclusion of the zero mode:

$$\begin{cases} \dfrac{\partial U}{\partial t} + \mathcal{A}U = F, \\ U(t=0) = U_0. \end{cases} \tag{3.71}$$

Here, $U = (\boldsymbol{u},\ \boldsymbol{v},\ \boldsymbol{\psi})$ and $F = (F_{\boldsymbol{u}},\ F_{\boldsymbol{v}},\ F_{\boldsymbol{\psi}})$ are like before, and $\mathcal{A}$ is the differential operator represented by the right-hand side of Eq. (3.25) or (3.27). The proposed boundary conditions for U at $x = 0$ and L_1 will be derived from given functions $U^{g,l}(z, t)$ and $U^{g,r}(z, t)$. In this subsection, we will demonstrate how to derive from $U^{g,l}$ and $U^{g,r}$ the boundary conditions for U so that the initial boundary value problem corresponding to Eq. (3.71) is well-posed.

As pointed out in Section 3.2.3, the subcritical and the supercritical modes require different boundary conditions. From the physical and computational points of view, we can assume that all the components of $U^{g,l}$ and $U^{g,r}$ are available. The mathematical issue is then to determine which components are needed for each mode. The normal mode expansions for $U^{g,l}$ and $U^{g,r}$ are given:

$$\begin{cases} U^{g,l}(z,t) = \Big(\displaystyle\sum_{n\geq 0} \boldsymbol{u}_n^{g,l}(t)\mathcal{U}_n(z), \sum_{n\geq 0} \boldsymbol{v}_n^{g,l}(t)\mathcal{U}_n(z), \sum_{n\geq 1} \boldsymbol{\psi}_n^{g,l}(t)\mathcal{W}_n(z) \Big), \\ U^{g,r}(z,t) = \Big(\displaystyle\sum_{n\geq 0} \boldsymbol{u}_n^{g,r}(t)\mathcal{U}_n(z), \sum_{n\geq 0} \boldsymbol{v}_n^{g,r}(t)\mathcal{U}_n(z), \sum_{n\geq 1} \boldsymbol{\psi}_n^{g,r}(t)\mathcal{W}_n(z) \Big). \end{cases} \tag{3.72}$$

From $U^{g,l}$, and $U^{g,r}$, we construct $U^g = U^g(z, t)$,

$$U^g(z,t) = \Big(\sum_{n\geq 1} \boldsymbol{u}_n^g(t)\mathcal{U}_n(z), \sum_{n\geq 1} \boldsymbol{v}_n^g(t)\mathcal{U}_n(z), \sum_{n\geq 1} \boldsymbol{\psi}_n^g(t)\mathcal{W}_n(z) \Big), \tag{3.73}$$

where, for each $n \geq 1$, $\boldsymbol{u}_n^g$, $\boldsymbol{v}_n^g$, and $\boldsymbol{\psi}_n^g$ are determined by the following equations:

$$\begin{cases} \boldsymbol{u}_n^g(t) - \dfrac{1}{N}\boldsymbol{\psi}_n^g(t) = \boldsymbol{u}_n^{g,l}(t) - \dfrac{1}{N}\boldsymbol{\psi}_n^{g,l}(t), \\ \boldsymbol{v}_n^g(t) = \boldsymbol{v}_n^{g,l}(t), \qquad\qquad \text{if } 1 \leq n \leq n_c, \\ \boldsymbol{u}_n^g(t) + \dfrac{1}{N}\boldsymbol{\psi}_n^g(t) = \boldsymbol{u}_n^{g,r}(t) + \dfrac{1}{N}\boldsymbol{\psi}_n^{g,r}(t), \end{cases} \tag{3.74}$$

i.e.,

$$\begin{cases} \boldsymbol{u}_n^g(t) = \dfrac{1}{2}(\boldsymbol{u}_n^{g,l}(t) + \boldsymbol{u}_n^{g,r}(t)) + \dfrac{1}{2N}(\boldsymbol{\psi}_n^{g,r}(t) - \boldsymbol{\psi}_n^{g,l}(t)), \\ \boldsymbol{v}_n^g(t) = \boldsymbol{v}_n^{g,l}(t), \qquad\qquad \text{if } 1 \le n \le n_c, \\ \boldsymbol{\psi}_n^g(t) = \dfrac{N}{2}(\boldsymbol{u}_n^{g,r}(t) - \boldsymbol{u}_n^{g,l})(t) + \dfrac{1}{2}(\boldsymbol{\psi}_n^{g,l}(t) + \boldsymbol{\psi}_n^{g,r}(t)), \end{cases} \tag{3.75}$$

and

$$\begin{cases} \boldsymbol{u}_n^g(t) = \boldsymbol{u}_n^{g,l}(t), \\ \boldsymbol{v}_n^g(t) = \boldsymbol{v}_n^{g,l}(t), \qquad \text{if } n > n_c, \\ \boldsymbol{\psi}_n^g(t) = \boldsymbol{\psi}_n^{g,l}(t). \end{cases} \tag{3.76}$$

We observe here that U^g is independent of x, i.e., $\partial U^g/\partial x = 0$, and that, for a.e. $t \in (0,\ T)$, $U^g \in H$ provided that $U^{g,l}$ and $U^{g,r}$ are sufficiently smooth: indeed $U^g \in \mathbf{L}^2(\mathcal{M}')^3$, and the integral conditions appearing in Eq. (3.20) are automatically satisfied since the mode 0 is not present here ($n \ge 1$).

Then, we set

$$U = U^{\#} + U^g. \tag{3.77}$$

We observe that $U^{\#} \in D(A)$ (if $U^{\#}$ is sufficiently smooth). Then, setting $F^{\#} = F - \partial U^g/\partial t - \mathcal{A}U^g$ and $U_0^{\#} = U_0 - U^g|_{t=0}$, we see that $U^{\#}$ is the solution of the following problem:

$$\begin{cases} \dfrac{\mathrm{d}U^{\#}}{\mathrm{d}t} + AU^{\#} = F^{\#}, \\ U^{\#}(t=0) = U_0^{\#}. \end{cases} \tag{3.78}$$

Like Eq. (3.19), Eq. (3.78) corresponds to the case with homogeneous boundary conditions. In order to apply Theorem 3.1 to Eq. (3.78), we would need to have

$$U_0^{\#} = U_0 - U^g|_{t=0} \in D(A), \tag{3.79}$$

and

$$F^{\#},\ \frac{\mathrm{d}F^{\#}}{\mathrm{d}t} \in L^1(0,\ T;\ \mathbf{L}^2(\mathcal{M})^3). \tag{3.80}$$

It is easily seen that Eqs. (3.79) and (3.80) are satisfied if the following hypotheses are verified (up to Eq. (3.85), and see also Eq. (3.89)):

$$U_0 \in \mathbf{L}^2(\mathcal{M})^3, \quad \frac{\partial U_0}{\partial x} \in \mathbf{L}^2(\mathcal{M})^3, \tag{3.81}$$

$$F \in L^1(0,\ T;\ \mathbf{L}^2(\mathcal{M})^3), \quad \frac{\partial F}{\partial t} \in L^1(0,\ T;\ \mathbf{L}^2(\mathcal{M})^3), \tag{3.82}$$

$$\frac{\partial^k U^{g,l}}{\partial t^k},\ \frac{\partial^k U^{g,r}}{\partial t^k} \in L^1(0,\ T;\ \mathbf{L}^2(-L_3,\ 0)^3) \quad \text{for } k = 0,\ 1,\ 2. \tag{3.83}$$

In addition, we must assume that U_0, $U^{g,l}$, and $U^{g,r}$ satisfy the compatibility conditions which will guarantee that the boundary conditions included in Eq. (3.79) are satisfied. Denoting the function U_0 of initial values by $(\widetilde{\boldsymbol{u}}_0, \widetilde{\boldsymbol{v}}_0, \widetilde{\boldsymbol{\psi}}_0)$,[15] the compatibility conditions for U_0, $U^{g,l}$, and $U^{g,r}$ which guarantee that $U_0^{\#} \in D(A)$ are written:

$$\begin{cases} \displaystyle\int_{-L_3}^{0} \widetilde{\boldsymbol{u}}_0(x=0,z)\mathcal{U}_n(z) - \frac{1}{N}\int_{-L_3}^{0} \widetilde{\boldsymbol{\psi}}_0(x=0,z)\mathcal{W}_n(z)\,\mathrm{d}z \\ \qquad = \displaystyle\int_{-L_3}^{0} \boldsymbol{u}^{g,l}(z,t=0)\mathcal{U}_n(z) - \frac{1}{N}\int_{-L_3}^{0} \boldsymbol{\psi}^{g,l}(z,t=0)\mathcal{W}_n(z)\,\mathrm{d}z, \\ \displaystyle\int_{-L_3}^{0} \widetilde{\boldsymbol{v}}_0(x=0,z)\mathcal{U}_n(z)\,\mathrm{d}z = \int_{-L_3}^{0} \boldsymbol{v}^{g,l}(z,t=0)\mathcal{U}_n(z)\,\mathrm{d}z, \qquad \text{for } 1 \le n \le n_c, \\ \displaystyle\int_{-L_3}^{0} \widetilde{\boldsymbol{u}}_0(x=L_1,z)\mathcal{U}_n(z) + \frac{1}{N}\int_{-L_3}^{0} \widetilde{\boldsymbol{\psi}}_0(x=L_1,z)\mathcal{W}_n(z)\,\mathrm{d}z \\ \qquad = \displaystyle\int_{-L_3}^{0} \boldsymbol{u}^{g,r}(z,t=0)\mathcal{U}_n(z) + \frac{1}{N}\int_{-L_3}^{0} \boldsymbol{\psi}^{g,r}(z,t=0)\mathcal{W}_n(z)\,\mathrm{d}z, \end{cases} \tag{3.84}$$

and

$$\begin{cases} \displaystyle\int_{-L_3}^{0} \widetilde{\boldsymbol{u}}_0(x=0,z)\mathcal{U}_n(z)\,\mathrm{d}z = \int_{-L_3}^{0} \boldsymbol{u}^{g,l}(z,t=0)\mathcal{U}_n(z)\,\mathrm{d}z, \\ \displaystyle\int_{-L_3}^{0} \widetilde{\boldsymbol{v}}_0(x=0,z)\mathcal{U}_n(z)\,\mathrm{d}z = \int_{-L_3}^{0} \boldsymbol{v}^{g,l}(z,t=0)\mathcal{U}_n(z)\,\mathrm{d}z, \qquad \text{for } n > n_c, \\ \displaystyle\int_{-L_3}^{0} \widetilde{\boldsymbol{\psi}}_0(x=0,z)\mathcal{W}_n(z)\,\mathrm{d}z = \int_{-L_3}^{0} \boldsymbol{\psi}^{g,l}(z,t=0)\mathcal{W}_n(z)\,\mathrm{d}z. \end{cases} \tag{3.85}$$

It should be noted here that Eqs. (3.81)–(3.85) are sufficient conditions for Eqs. (3.79) and (3.80). They have been chosen for their relative simplicity; Eqs. (3.84) and (3.85) indeed guarantee that the boundary conditions required in Eq. (3.79) ($U_0^{\#} \in D(A)$) are satisfied.

Now we can apply Theorem 3.1 to the system in Eq. (3.78), and we obtain a unique solution $U^{\#}$ that satisfies the analog of Eq. (3.26). We then recover U via Eq. (3.77), and we easily see that

$$U \in \mathcal{C}([0,T];\mathbf{L}^2(\mathcal{M})^3), \tag{3.86}$$

$$\frac{\partial U}{\partial x} \in L^{\infty}(0,T;\mathbf{L}^2(\mathcal{M})^3). \tag{3.87}$$

[15] The tildes here on $\widetilde{\boldsymbol{u}}_0$, $\widetilde{\boldsymbol{v}}_0$, and $\widetilde{\boldsymbol{\psi}}_0$ are meant to distinguish these initial data from the zero modes of $U(t)$, which do not appear in fact in this subsection.

We will also have

$$\frac{\partial U}{\partial t} \in L^\infty(0, T; \mathbf{L}^2(\mathcal{M})^3), \tag{3.88}$$

provided we further require that

$$\frac{\partial U^{g,l}}{\partial t}, \ \frac{\partial U^{g,r}}{\partial t} \in L^\infty(0, T; \mathbf{L}^2(-L_3, 0)^3). \tag{3.89}$$

The boundary conditions that U satisfies, expressing the fact that $U^\#(t) = U(t) - U^g(t)$ belongs to $D(A)$ for a.e. t, are as follows:

For the subcritical modes $1 \le n \le n_c$:

$$\begin{cases} \displaystyle\int_{-L_3}^0 \boldsymbol{u}(0, z, t)\mathcal{U}_n(z) - \frac{1}{N}\int_{-L_3}^0 \boldsymbol{\psi}(0, z, t)\mathcal{W}_n(z)\,\mathrm{d}z \\ \qquad \displaystyle= \int_{-L_3}^0 \boldsymbol{u}^{g,l}(z, t)\mathcal{U}_n(z) - \frac{1}{N}\int_{-L_3}^0 \boldsymbol{\psi}^{g,l}(z, t)\mathcal{W}_n(z)\,\mathrm{d}z, \\ \displaystyle\int_{-L_3}^0 \boldsymbol{v}(0, z, t)\mathcal{U}_n(z)\,\mathrm{d}z = \int_{-L_3}^0 \boldsymbol{v}^{g,l}(z, t)\mathcal{U}_n(z)\,\mathrm{d}z, \\ \displaystyle\int_{-L_3}^0 \boldsymbol{u}(L_1, z, t)\mathcal{U}_n(z) + \frac{1}{N}\int_{-L_3}^0 \boldsymbol{\psi}(L_1, z, t)\mathcal{W}_n(z)\,\mathrm{d}z \\ \qquad \displaystyle= \int_{-L_3}^0 \boldsymbol{u}^{g,r}(z, t)\mathcal{U}_n(z) + \frac{1}{N}\int_{-L_3}^0 \boldsymbol{\psi}^{g,r}(z, t)\mathcal{W}_n(z)\,\mathrm{d}z, \end{cases} \tag{3.90}$$

and, *for the supercritical mode* $n > n_c$:

$$\begin{cases} \displaystyle\int_{-L_3}^0 \boldsymbol{u}(0, z, t)\mathcal{U}_n(z)\,\mathrm{d}z = \int_{-L_3}^0 \boldsymbol{u}^{g,l}(z, t)\mathcal{U}_n(z)\,\mathrm{d}z, \\ \displaystyle\int_{-L_3}^0 \boldsymbol{v}(0, z, t)\mathcal{U}_n(z)\,\mathrm{d}z = \int_{-L_3}^0 \boldsymbol{v}^{g,l}(z, t)\mathcal{U}_n(z)\,\mathrm{d}z, \\ \displaystyle\int_{-L_3}^0 \boldsymbol{\psi}(0, z, t)\mathcal{W}_n(z)\,\mathrm{d}z = \int_{-L_3}^0 \boldsymbol{\psi}^{g,l}(z, t)\mathcal{W}_n(z)\,\mathrm{d}z. \end{cases} \tag{3.91}$$

We summarize the result concerning the case of nonhomogeneous boundary conditions in a theorem.

THEOREM 3.2. *Let H be the Hilbert space defined in Eq. (3.20), A the linear operator defined in Eq. (3.25), $\mathcal{A}$ the corresponding differential operator, and let $D(A)$ be the domain of the operator A in H. We assume that the data U_0, F, $U^{g,l}$, and $U^{g,r}$ satisfy the regularity conditions in Eqs. (3.81)–(3.83), and in addition, U_0, $U^{g,l}$, and $U^{g,r}$ satisfy the compatibility conditions in Eqs. (3.84) and (3.85). Then, the initial boundary value problem corresponding to Eq. (3.71) supplemented with the boundary conditions in*

Eqs. (3.90) and (3.91) has a unique solution U, and U satisfies Eqs. (3.86) and (3.87); U will also satisfy Eq. (3.88) if we further assume Eq. (3.89) *for* $U^{g,l}$ *and* $U^{g,r}$.

3.4. Special case of impenetrable boundaries

We now consider another interesting model with only one degree of freedom for v (one unknown component for v). In this case, we require that v vanishes at $y = 0$ and 1, which in physics corresponds to impenetrable boundaries at the North and South. To impose this boundary condition, we use a single mode in the y-direction for v, namely $\widetilde{h}_2$ (see Fig. 3.1), and the other unknowns u, w, ϕ, and ψ are decomposed as they were in Section 3.2.1; hence

$$\begin{cases} u \simeq u_1(x,z,t)h_1(y) + u_2(x,z,t)h_2(y) + u_3(x,z,t)h_3(y), \\ v \simeq v_2(x,z,t)\widetilde{h}_2(y), \\ w \simeq w_1(x,z,t)h_1(y) + w_2(x,z,t)h_2(y) + w_3(x,z,t)h_3(y), \\ \phi \simeq \phi_1(x,z,t)h_1(y) + \phi_2(x,z,t)h_2(y) + \phi_3(x,z,t)h_3(y), \\ \psi \simeq \psi_1(x,z,t)h_1(y) + \psi_2(x,z,t)h_2(y) + \psi_3(x,z,t)h_3(y). \end{cases} \tag{3.92}$$

Then, we introduce Eq. (3.92) into Eq. (3.3). We perform the same operations as after Eq. (3.3), except that we multiply the equation for v, second equation of Eq. (3.3), by $\widetilde{h}_2$, integrate over $(0, 1)$, and divide it by $\int_0^1 \widetilde{h}_2^2 dy$. Thus, we arrive at the following approximation of the system in Eq. (3.3):

$$\begin{cases} \boldsymbol{u}_t + \overline{U}_0 \boldsymbol{u}_x + \boldsymbol{\phi}_x + f v_2 \boldsymbol{\sigma_1} = 0, \\ v_{2t} + \overline{U}_0 v_{2x} + f \boldsymbol{\sigma_3} \cdot \boldsymbol{u} + \boldsymbol{\sigma_4} \cdot \boldsymbol{\phi} + \dfrac{3}{2} f \overline{U}_0 = 0, \\ \boldsymbol{\psi}_t + \overline{U}_0 \boldsymbol{\psi}_x + N^2 \boldsymbol{w} = 0, \\ \boldsymbol{u}_x + v_2 \boldsymbol{\sigma_2} + \boldsymbol{w}_z = 0, \\ \boldsymbol{\psi} = \boldsymbol{\phi}_z. \end{cases} \tag{3.93}$$

The vector notation, i.e., $\boldsymbol{u} = (u_1,\ u_2,\ u_3)$, $\boldsymbol{w} = (w_1,\ w_2,\ w_3)$, etc., has been used, and

$$\boldsymbol{\sigma_1} = \frac{1}{2}\begin{pmatrix} -1 \\ 1 \\ -1 \end{pmatrix}, \quad \boldsymbol{\sigma_2} = 3\begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \quad \boldsymbol{\sigma_3} = \frac{1}{4}\begin{pmatrix} 1 \\ -6 \\ 1 \end{pmatrix}, \quad \boldsymbol{\sigma_4} = \frac{3}{2}\begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}. \tag{3.94}$$

The dot in Eq. (3.93) represents the dot product in the Euclidean space. We note here that all the equations in Eq. (3.93) are vector equations except the second one, which is a scalar equation for the scalar unknown v_2.

From here on, we proceed essentially as in Sections 3.2.2, 3.2.3, and 3.3, and below we only highlight the differences with the previous case.

The normal mode expansion for $\boldsymbol{u}$, $\boldsymbol{w}$, $\boldsymbol{\phi}$, and $\boldsymbol{\psi}$ are the same as before, and for v_2 it reads

$$v_2(x,z,t) = \sum_{n\geq 0} v_{2n}(x,t)\mathcal{U}_n(z). \tag{3.95}$$

The system for the zero mode is

$$\begin{cases} \boldsymbol{u}_{0t} + \overline{U}_0 \boldsymbol{u}_{0x} + \boldsymbol{\phi}_{0x} + f v_{20}\boldsymbol{\sigma_1} = 0, \\ v_{20t} + \overline{U}_0 v_{20x} + f\boldsymbol{\sigma_3}\cdot\boldsymbol{u}_0 + \boldsymbol{\sigma_4}\cdot\boldsymbol{\phi}_0 + \dfrac{3}{2}L_3^{\frac{1}{2}} f\overline{U}_0 = 0, \\ \boldsymbol{u}_{0x} + v_{20}\boldsymbol{\sigma_2} = 0. \end{cases} \tag{3.96}$$

For $n \geq 1$, the system is

$$\begin{cases} \boldsymbol{u}_{nt} + \overline{U}_0 \boldsymbol{u}_{nx} + \boldsymbol{\phi}_{nx} + f v_{2n}\boldsymbol{\sigma_1} = 0, \\ v_{2nt} + \overline{U}_0 v_{2nx} + f\boldsymbol{\sigma_3}\cdot\boldsymbol{u}_n + \boldsymbol{\sigma_4}\cdot\boldsymbol{\phi}_n = 0, \\ \boldsymbol{\psi}_{nt} + \overline{U}_0 \boldsymbol{\psi}_{nx} + N^2\boldsymbol{w}_n = 0, \\ \boldsymbol{u}_{nx} + v_{2n}\boldsymbol{\sigma_2} + \lambda_n \boldsymbol{w}_n = 0, \\ -\lambda_n \boldsymbol{\phi}_n = \boldsymbol{\psi}_n. \end{cases} \tag{3.97}$$

Eliminating $\boldsymbol{\phi}_n$ and $\boldsymbol{w}_n$ from Eq. (3.97), we obtain a system for $\boldsymbol{u}_n$, $\boldsymbol{v}_n$, and $\boldsymbol{\psi}_n$ ($n \geq 1$), namely:

$$\begin{cases} \boldsymbol{u}_{nt} + \overline{U}_0 \boldsymbol{u}_{nx} - \dfrac{1}{\lambda_n}\boldsymbol{\psi}_{nx} + f v_{2n}\boldsymbol{\sigma_1} = 0, \\ \boldsymbol{v}_{nt} + \overline{U}_0 \boldsymbol{v}_{nx} + f\boldsymbol{\sigma_3}\boldsymbol{u}_n - \dfrac{1}{\lambda_n}\boldsymbol{\sigma_4}\cdot\boldsymbol{\psi}_n = 0, \\ \boldsymbol{\psi}_{nt} - \dfrac{N^2}{\lambda_n}\boldsymbol{u}_{nx} + \overline{U}_0 \boldsymbol{\psi}_{nx} - \dfrac{N^2}{\lambda_n} v_{2n}\boldsymbol{\sigma_2} = 0. \end{cases} \tag{3.98}$$

The coefficient matrix associated with the first-order derivative terms (with respect to x) terms in the first and third equations in Eq. (3.98) is

$$\begin{pmatrix} \overline{U}_0 & -\dfrac{1}{\lambda_n} \\ -\dfrac{N^2}{\lambda_n} & \overline{U}_0 \end{pmatrix}.$$

This matrix has two eigenvalues: $U_0 + N/\lambda_n$ and $U_0 - N/\lambda_n$. The first eigenvalue is always positive, while the second one could be positive, which corresponds to

supercritical modes, or negative, which corresponds to subcritical modes. As before, let n_c denote the number of subcritical modes, and for each $n \geq 1$, we also introduce the variables $\boldsymbol{\xi}_n = \boldsymbol{u}_n - \boldsymbol{\psi}_n/N$, $\boldsymbol{\eta}_n = \boldsymbol{u}_n + \boldsymbol{\psi}_n/N$. By an analysis similar to that in Section 3.2.3, we are led to propose the following boundary conditions for the subcritical modes:

$$\begin{cases} \boldsymbol{\xi}_n(0, t) = 0, \\ v_{2n}(0, t) = 0, \qquad \text{for } 1 \leq n \leq n_c, \\ \boldsymbol{\eta}_n(L_1, t) = 0, \end{cases} \tag{3.99}$$

and for the supercritical modes:

$$\begin{cases} \boldsymbol{\xi}_n(0, t) = 0, \\ v_{2n}(0, t) = 0, \qquad \text{for } n > n_c. \\ \boldsymbol{\eta}_n(0, t) = 0. \end{cases} \tag{3.100}$$

Again, we want to transform Eq. (3.93) (except the zero mode, which needs a separate treatment) into an abstract initial value problem of the form

$$\begin{cases} \dfrac{\mathrm{d}U}{\mathrm{d}t} + AU = F, \\ U(0) = U_0. \end{cases} \tag{3.101}$$

For this purpose, we introduce the following function spaces:

$$\begin{aligned} &H = H_{\boldsymbol{u}} \times H_{v_2} \times H_{\boldsymbol{\psi}}, \\ &H_{\boldsymbol{u}} = \left\{ \boldsymbol{u} \in \mathbf{L}^2(\mathcal{M}') \;\middle|\; \int_{-L_3}^{0} \boldsymbol{u}(x, z)\, \mathrm{d}z = \mathbf{0}, \text{ for a.e. } x \in (0, L_1) \right\}, \\ &H_{v_2} = \left\{ v_2 \in L^2(\mathcal{M}') \;\middle|\; \int_{-L_3}^{0} v_2(x, z)\, \mathrm{d}z = 0, \text{ for a.e. } x \in (0, L_1) \right\}, \\ &H_{\boldsymbol{\psi}} = \mathbf{L}^2(\mathcal{M}'). \end{aligned} \tag{3.102}$$

We endow H with the inner product

$$(U, \widetilde{U})_H = \int_{\mathcal{M}'} (\boldsymbol{u} \cdot \widetilde{\boldsymbol{u}} + v_2 \widetilde{v_2} + \frac{1}{N^2} \boldsymbol{\psi} \cdot \widetilde{\boldsymbol{\psi}})\, \mathrm{d}\mathcal{M}' \quad \text{for } U, \widetilde{U} \in H. \tag{3.103}$$

With this inner product, H is a Hilbert space. We let P denote the orthogonal projector from $\mathbf{L}^2(\mathcal{M}')$ onto $H_{\boldsymbol{u}}$. For convenience, we also use P for the orthogonal projector from $L^2(\mathcal{M}')$ onto H_{v_2}.

The boundary conditions for $\boldsymbol{u}$, v_2, and $\boldsymbol{\psi}$ follow those we chose above, mode by mode (see Eqs. (3.99), (3.100)); hence:

For the subcritical modes ($1 \leq n \leq n_c$),

$$\begin{cases} \displaystyle\int_{-L_3}^{0} \boldsymbol{u}(0,z)\mathcal{U}_n(z) - \frac{1}{N}\int_{-L_3}^{0} \boldsymbol{\psi}(0,z)\mathcal{W}_n(z)\,\mathrm{d}z = \mathbf{0}, \\ \displaystyle\int_{-L_3}^{0} v_2(0,z)\mathcal{U}_n(z) = \mathbf{0}, \\ \displaystyle\int_{-L_3}^{0} \boldsymbol{u}(L_1,z)\mathcal{U}_n(z) + \frac{1}{N}\int_{-L_3}^{0} \boldsymbol{\psi}(L_1,z)\mathcal{W}_n(z)\,\mathrm{d}z = \mathbf{0}, \end{cases} \tag{3.104}$$

and *for the supercritical modes* ($n > n_c$),

$$\begin{cases} \displaystyle\int_{-L_3}^{0} \boldsymbol{u}(0,z)\mathcal{U}_n(z)\,\mathrm{d}z = \mathbf{0}, \\ \displaystyle\int_{-L_3}^{0} v_2(0,z)\mathcal{U}_n(z)\,\mathrm{d}z = \mathbf{0}, \\ \displaystyle\int_{-L_3}^{0} \boldsymbol{\psi}(0,z)\mathcal{W}_n(z)\,\mathrm{d}z = \mathbf{0}. \end{cases} \tag{3.105}$$

The domain of the operator A is defined as

$$\begin{aligned} D(A) = \{\, U \in H \mid U_x \in \mathbf{L}^2(\mathcal{M}') \times L^2(\mathcal{M}') \times \mathbf{L}^2(\mathcal{M}'), \\ \text{and } U \text{ verifies the BC's (3.104) and (3.105)} \,\}. \end{aligned} \tag{3.106}$$

For each $U \in D(A)$, AU is given by

$$AU = \begin{pmatrix} \overline{U}_0 \boldsymbol{u}_x + f v_2 \boldsymbol{\sigma_1} - P[\displaystyle\int_z^0 \boldsymbol{\psi}_x(x,z')\,\mathrm{d}z'] \\ \overline{U}_0 v_{2x} + f\boldsymbol{\sigma_3}\cdot\boldsymbol{u} - P[\displaystyle\int_z^0 \boldsymbol{\sigma_4}\cdot\boldsymbol{\psi}(x,z')\,\mathrm{d}z'] \\ \overline{U}_0 \boldsymbol{\psi}_x + N^2 \displaystyle\int_z^0 (\boldsymbol{u}_x + v_2\boldsymbol{\sigma_2})\,\mathrm{d}z' \end{pmatrix}. \tag{3.107}$$

In the process of establishing the well-posedness of the initial value problem associated with our new model, we need to determine the adjoint operator A^* of A (as an unbounded operator in H) and its domain $D(A^*)$. We now list, without details of the calculations, the definitions of the operator A^* and its domain $D(A^*)$. The functions $\widetilde{U} = (\widetilde{\boldsymbol{u}}, \widetilde{v}_2, \widetilde{\boldsymbol{\psi}})$ in $D(A^*)$ satisfy the following boundary conditions.

For the subcritical modes ($1 \le n \le n_c$):

$$\begin{cases} \displaystyle\int_{-L_3}^{0} \widetilde{\boldsymbol{u}}(L_1, z)\mathcal{U}_n(z) - \frac{1}{N}\int_{-L_3}^{0} \widetilde{\boldsymbol{\psi}}(L_1, z)\mathcal{W}_n(z)\,\mathrm{d}z = \mathbf{0}, \\ \displaystyle\int_{-L_3}^{0} \widetilde{v}_2(L_1, z)\mathcal{U}_n(z) = \mathbf{0}, \\ \displaystyle\int_{-L_3}^{0} \widetilde{\boldsymbol{u}}(0, z)\mathcal{U}_n(z) + \frac{1}{N}\int_{-L_3}^{0} \widetilde{\boldsymbol{\psi}}(0, z)\mathcal{W}_n(z)\,\mathrm{d}z = \mathbf{0}, \end{cases} \tag{3.108}$$

and *for the supercritical modes* ($n > n_c$):

$$\begin{cases} \displaystyle\int_{-L_3}^{0} \widetilde{\boldsymbol{u}}(L_1, z)\mathcal{U}_n(z)\,\mathrm{d}z = \mathbf{0}, \\ \displaystyle\int_{-L_3}^{0} \widetilde{v}_2(L_1, z)\mathcal{U}_n(z)\,\mathrm{d}z = \mathbf{0}, \\ \displaystyle\int_{-L_3}^{0} \widetilde{\boldsymbol{\psi}}(L_1, z)\mathcal{W}_n(z)\,\mathrm{d}z = \mathbf{0}. \end{cases} \tag{3.109}$$

The domain $D(A^*)$ is then defined as follows:

$$D(A^*) = \{\, \widetilde{U} \in H \mid \widetilde{U}_x \in (\mathbf{L}^2(\mathcal{M}') \times L^2(\mathcal{M}') \times \mathbf{L}^2(\mathcal{M}'), \text{ and } \widetilde{U} \text{ verifies the BC's (3.108) and (3.109)} \,\}. \tag{3.110}$$

For each $\widetilde{U} \in D(A^*)$, $A^*\widetilde{U}$ is given by

$$A^*\widetilde{U} = \begin{pmatrix} \displaystyle\sum_{n\ge1}(-\overline{U}_0\widetilde{\boldsymbol{u}}_{nx} + \frac{1}{N\lambda_n}\widetilde{\boldsymbol{\psi}}_{nx} + f\widetilde{v}_{2n}\boldsymbol{\sigma_3})\mathcal{U}_n \\ \displaystyle\sum_{n\ge1}(-\overline{U}_0\widetilde{v}_{2nx} + f\widetilde{\boldsymbol{u}}_n\cdot\boldsymbol{\sigma_1} - \frac{1}{N\lambda_n}\widetilde{\boldsymbol{\psi}}_n\cdot\boldsymbol{\sigma_2})\mathcal{U}_n \\ \displaystyle\sum_{n\ge1}(\frac{N}{\lambda_n}\widetilde{\boldsymbol{u}}_{nx} - \overline{U}_0\widetilde{\boldsymbol{\psi}}_{nx} - \frac{N}{\lambda_n}\widetilde{v}_{2n}\boldsymbol{\sigma_4})\mathcal{W}_n \end{pmatrix}, \tag{3.111}$$

where the $(\boldsymbol{u}_n, v_{2n}, \boldsymbol{\psi}_n)$, for $n \ge 1$, are the normal modes of $\widetilde{U}$.

The following theorem, which is a copy of Theorem 3.1 with minor modifications, gives the well-posedness result about the system in Eq. (3.101).

THEOREM 3.3. *Let H, A, and $D(A)$ be defined as above. Then, the initial value problem in Eq. (3.101) is well-posed. That is, for every $U_0 \in D(A)$ and $F \in L^1(0, T; H)$, with $F' \in L^1(0, T; H)$, Eq. (3.101) has a unique solution U such that*

$$U \in C([0, T]; H) \cap L^\infty(0, T; D(A)), \quad \frac{\mathrm{d}U}{\mathrm{d}t} \in L^\infty(0, T; H). \tag{3.112}$$

Theorem 3.3 is also a direct consequence of Theorem 2.1. The verification of the hypotheses of Theorem 2.1 can be done similarly as in Section 3.3.

For the system of the zero mode in Eq. (3.96), we can also decompose $\boldsymbol{\phi}_0$ into two parts:

$$\boldsymbol{\phi}_0 = \bar{\boldsymbol{\phi}}_0 + \boldsymbol{\phi}_0', \tag{3.113}$$

where $\bar{\boldsymbol{\phi}}_0$ is one of the stationary solutions of the equation

$$\boldsymbol{\sigma_4} \cdot \bar{\boldsymbol{\phi}}_0 = -\frac{3}{2} L_3^{\frac{1}{2}} f \overline{U}_0. \tag{3.114}$$

Then, we impose the boundary conditions on the left boundary:

$$\begin{cases} \boldsymbol{u}_0(0,t) = \boldsymbol{u}^l(t), \\ v_{20}(0,t) = v_2^l(t), \\ \boldsymbol{\phi}_0'(0,t) = \boldsymbol{\phi}'^l(t). \end{cases} \tag{3.115}$$

With the boundary conditions above, we can treat the zero mode in a way similar to that in Section 3.3.4. First, by combining the first and second equations (now with $\boldsymbol{\phi}_0'$) of Eq. (3.96), we find and then solve the resulting equation for $\boldsymbol{\sigma_4} \cdot \boldsymbol{u}_0 - v_{20x}$. Once $\boldsymbol{\sigma_4} \cdot \boldsymbol{u}_0 - v_{20x}$ is known, say $\boldsymbol{\sigma_4} \cdot \boldsymbol{u}_0 - v_{20x} = K(x,t)$, we can solve for $\boldsymbol{u}_0$ and v_{20} from this expression and the third equation of Eq. (3.96). Then, when $\boldsymbol{u}_0$ and v_{20} are known, the first equation in Eq. (3.96) gives $\boldsymbol{\phi}_0'$. We leave it as an exercise for the reader to check the details and to address the case of nonhomogeneous boundary conditions as in Section 3.3.5.

4. The full 3D linear case

4.1. Equations and preliminary results

In this section, we consider the PEs in space dimension 3. We focus on the linearized equations since the boundary condition difficulty is already fully present in the linear case (see Section 2 above).

This section is organized as follows: we first recall the PEs and their linearized form. We also recall the normal mode expansion of the unknowns and their decomposition into the subcritical and supercritical modes. These two sets of modes necessitate different treatments and, unlike in Sections 2 and 3 above, the study of the supercritical modes is not straightforward. This Section 4.1 also contains (Section 4.1.3) a study of the associated stationary operator $\mathcal{A}$, a trace theorem adapted to this stationary operator which shows that if $U = (u, v, \psi)$ and $\mathcal{A}U$ are square integrable, then the traces of v and ψ are defined on the whole boundary and the trace of u is defined on part of the boundary (Section 4.1.4); finally, Section 4.1 finishes with the study of the zero mode in the modal decomposition (Section 4.1.5). Section 4.2 is devoted to the study of the subcritical modes for which the stationary problem, partly elliptic and partly hyperbolic, possesses a regularity result.

Section 4.3 is devoted to the study of the supercritical modes handled in a different manner; the stationary problem is then fully hyperbolic, and it does not produce any regularity. Finally, in Section 4.4, we consider the full linearized PEs containing both the subcritical and the supercritical modes, and we prove our main existence and uniqueness results for homogeneous and nonhomogeneous boundary conditions.

Note that the boundary conditions proposed here for the subcritical modes are different than those studied in Sections 2 and 3 in dimensions 2 and 2.5; this change is of no importance in view of the computational objectives (see Section 2). The related open problem is the determination of *all* the sets of boundary conditions making the nonviscous PEs well-posed. The full nonlinear PEs with boundary conditions similar to those proposed here will be studied in a separate work.

4.1.1. Equations and normal modes expansion

We now recall the PEs and their normal mode expansion. The reader is referred to Section 1 for further details. The nonlinear 3D PEs read:

$$\begin{cases} \dfrac{\partial \widetilde{\mathbf{v}}}{\partial t} + (\widetilde{\mathbf{v}} \cdot \nabla)\widetilde{\mathbf{v}} + \widetilde{w}\,\dfrac{\partial \widetilde{\mathbf{v}}}{\partial z} + k \times \widetilde{\mathbf{v}} + \dfrac{1}{\rho_0}\nabla \widetilde{p} = 0, \\ \dfrac{\partial \widetilde{p}}{\partial z} = -\widetilde{\rho}\, g, \\ \nabla \cdot \widetilde{\mathbf{v}} + \dfrac{\partial \widetilde{w}}{\partial z} = 0, \\ \dfrac{\partial \widetilde{T}}{\partial t} + (\widetilde{\mathbf{v}} \cdot \nabla)\widetilde{T} + \widetilde{w}\dfrac{\partial \widetilde{T}}{\partial z} = 0, \ \ \widetilde{\rho} = \widetilde{\rho}(\widetilde{T}). \end{cases} \tag{4.1}$$

The notations are as follows: $\widetilde{\mathbf{u}} = (\widetilde{u}, \widetilde{v}, \widetilde{w})$ the velocity of the water, $\widetilde{\mathbf{v}}$ the horizontal velocity, $\widetilde{\rho}$ the density, $\widetilde{p}$ the pressure, $\widetilde{T}$ the temperature, and $\widetilde{\rho} = \widetilde{\rho}(\widetilde{T})$ the equation of state. In agreement with the Boussinesq approximation, the density ρ is constant everywhere, $\widetilde{\rho} = \rho_0$, except in the second equation of Eq. (4.1). The salinity equation is not present in Eq. (4.1), but this would raise little additional difficulty to take into account the salinity S. As indicated before, the viscosity is not present in Eq. (4.1); this is a crucial point in this study. Equation (4.1) correspond to the β-plane approximation of the PEs near the latitude $\theta = \theta_0$, and $f = f_0 + \beta y$, $f_0 = \Omega \sin\theta_0$, where Ω is the angular velocity of the earth, and $\beta = (df/dy)$ at $\theta = \theta_0$, i.e., $\beta = f_0/a$ at midlatitudes, ($\theta_0 = \pi/4$); k is the unit vector along the south to north poles; g is the gravitational constant. The domain occupied by the water is $\mathcal{M} = (0, L_1) \times (0, L_2) \times (-L_3, 0)$ in the $Oxyz$ system of coordinates.

We linearize Eq. (4.1) around the simple uniform stratified flow (Eq. (4.2)):

$$\overline{u} = \overline{U}_0, \ \overline{v} = 0, \ \overline{T} = \overline{T}(z), \ \overline{\rho} = \rho_0(1 - \alpha(\overline{T} - T_0)), \tag{4.2}$$

where $\overline{U}_0 > 0$, $\rho_0 > 0$, and $T_0 > 0$ are reference average values of the density and the temperature, $\alpha > 0$ is a constant and $\overline{T}$ and $\overline{\rho}$ are linear in z. We introduce the Brunt–Väisälä (buoyancy) frequency

$$N^2 = -\frac{g}{\rho_0}\frac{\mathrm{d}\overline{\rho}}{\mathrm{d}z},$$

and we assume that N does not depend on z. We write $\phi = p/\rho_0$, $\psi = gT/T_0$, and we set $\widetilde{u} = \overline{u} + u$, $\widetilde{v} = \overline{v} + v$, etc. We obtain as in Eq. (1.13):

$$\begin{cases} u_t + \overline{U}_0 u_x - fv + \phi_x = 0, \\ v_t + \overline{U}_0 v_x + fu + \phi_y = 0, \\ \psi_t + \overline{U}_0 \psi_x + N^2 w = 0, \\ u_x + v_y + w_z = 0, \\ \phi_z = \psi. \end{cases} \tag{4.3}$$

4.1.2. Normal modes expansion

As indicated in Section 1, the first step of the analysis of Eq. (4.3) consists, by separation of variables, in looking for solutions of the form

$$\begin{cases} u(x, y, z, t) = \mathcal{U}(z)\, \hat{u}(x, y, t), \;\; v(x, y, z, t) = \mathcal{V}(z)\, \hat{v}(x, y, t), \\ \psi(x, y, z, t) = \Psi(z)\, \hat{\psi}(x, y, t), \\ w(x, y, z, t) = \mathcal{W}(z)\, \hat{w}(x, y, t), \;\; \phi(x, y, z, t) = \Phi(z)\, \hat{\phi}(x, y, t). \end{cases} \tag{4.4}$$

Substituting these expressions in Eq. (4.3), we end up with the following systems (see Section 1 above), for $n \geq 1$,

$$\begin{cases} \dfrac{\partial u_n}{\partial t} + \overline{U}_0 \dfrac{\partial u_n}{\partial x} - fv_n + \dfrac{\partial \phi_n}{\partial x} = 0, \\ \dfrac{\partial v_n}{\partial t} + \overline{U}_0 \dfrac{\partial v_n}{\partial x} + fu_n + \dfrac{\partial \phi_n}{\partial y} = 0, \\ \dfrac{\partial \psi_n}{\partial t} + \overline{U}_0 \dfrac{\partial \psi_n}{\partial x} + N^2 w_n = 0, \\ \phi_n = -\dfrac{1}{\lambda_n} \psi_n, \quad w_n = -\dfrac{1}{\lambda_n} \left(\dfrac{\partial u_n}{\partial x} + \dfrac{\partial v_n}{\partial y} \right). \end{cases} \tag{4.5}$$

For $n = 0$, $w_0 = \psi_0 = 0$, and there remains

$$\begin{cases} \dfrac{\partial u_0}{\partial t} + \overline{U}_0 \dfrac{\partial u_0}{\partial x} - fv_0 + \dfrac{\partial \phi_0}{\partial x} = 0, \\ \dfrac{\partial v_0}{\partial t} + \overline{U}_0 \dfrac{\partial v_0}{\partial x} + fu_0 + \dfrac{\partial \phi_0}{\partial y} = 0, \\ \dfrac{\partial u_0}{\partial x} + \dfrac{\partial v_0}{\partial y} = 0. \end{cases} \tag{4.6}$$

Note that since the considered problem is linear, there is no coupling between the different modes; see, e.g., Section 2 above for the nonlinear case which introduces these couplings.

We will study the zero mode separately (see Section 4.1.5), and for $n \geq 1$, we use the last two equations of Eq. (4.5) and rewrite the first three in the form

$$\begin{cases} \dfrac{\partial u_n}{\partial t} + \overline{U}_0 \dfrac{\partial u_n}{\partial x} - f v_n - \dfrac{1}{\lambda_n} \dfrac{\partial \psi_n}{\partial x} = 0, \\ \dfrac{\partial v_n}{\partial t} + \overline{U}_0 \dfrac{\partial v_n}{\partial x} + f u_n - \dfrac{1}{\lambda_n} \dfrac{\partial \psi_n}{\partial y} = 0, \\ \dfrac{\partial \psi_n}{\partial t} + \overline{U}_0 \dfrac{\partial \psi_n}{\partial x} - \dfrac{N^2}{\lambda_n} \left(\dfrac{\partial u_n}{\partial x} + \dfrac{\partial v_n}{\partial y} \right) = 0. \end{cases} \tag{4.7}$$

As indicated before, our aim is to propose boundary conditions for Eqs. (4.5)–(4.7) which make these equations well–posed and consequently the Eq. (4.3) also. As we shall see (see also Sections 2 and 3 above), the boundary conditions are different depending on whether

$$1 \leq n \leq n_c \quad \text{or} \quad n > n_c,$$

where n_c, λ_{n_c} are such that

$$\frac{n_c \pi}{L_3} = \lambda_{n_c} < \frac{N}{\overline{U}_0} < \lambda_{n_c+1} = \frac{(n_c+1)\pi}{L_3}. \tag{4.8}$$

We will not study the nongeneric case where $L_3 N / \pi \overline{U}_0$ is an *integer*.

The modes $0 \leq n \leq n_c$ are called *subcritical*, and the modes $n > n_c$ are called *supercritical*. It is convenient to introduce the sub and supercritical components of the functions defined by

$$u^0 = P_0 u = \mathcal{U}_0 u_0, \, u^I = P_I u = \sum_{n=1}^{n_c} \mathcal{U}_n u_n, \quad u^{II} = P_{II} u = \sum_{n>n_c}^{\infty} \mathcal{U}_n u_n, \tag{4.9}$$

and similarly for all the other functions; of course, the zero mode u^0 is a subcritical mode, but as we will see, we need to treat it separately. With these notations, the Eqs. (4.3), (4.5), and (4.7) are equivalent to the following systems:

$$\begin{cases} u^0_t + \bar{U}_0 u^0_x - f v^0 + \phi^0_x = 0, \\ v^0_t + \bar{U}_0 v^0_x + f u^0 + \phi^0_y = 0, \\ u^0_x + v^0_y = 0, \end{cases} \tag{4.10}$$

$$\begin{cases} u^I_t + \overline{U}_0 u^I_x - f v^I + \phi^I_x = 0, \\ v^I_t + \overline{U}_0 v^I_x + f u^I + \phi^I_y = 0, \\ \psi^I_t + \overline{U}_0 \psi^I_x + N^2 w^I = 0, \end{cases} \tag{4.11}$$

$$\begin{cases} u_t^{II} + \overline{U}_0 u_x^{II} - f v^{II} + \phi_x^{II} = 0, \\ v_t^{II} + \overline{U}_0 u_x^{II} + f u^{II} + \phi_y^{II} = 0, \\ \psi_t^{II} + \overline{U}_0 \psi_x^{II} + N^2 w^{II} = 0, \end{cases} \tag{4.12}$$

with the additional relations $\phi = \phi(\psi)$, $w = w(u, v)$:

$$\begin{cases} \phi^I = -\sum_{n=1}^{n_c} \frac{1}{\lambda_n} \psi_n \mathcal{U}_n, \quad w^I = -\sum_{n=1}^{n_c} \frac{1}{\lambda_n} (u_{nx} + v_{ny}) \mathcal{W}_n, \\ \phi^{II} = -\sum_{n>n_c} \frac{1}{\lambda_n} \psi_n \mathcal{U}_n, \quad w^{II} = -\sum_{n>n_c} \frac{1}{\lambda_n} (u_{nx} + v_{ny}) \mathcal{W}_n. \end{cases} \tag{4.13}$$

We will also set $U = (u, v, \psi)$, $U^0 = P_0 U$, $U^I = P_I U$, $U^{II} = P_{II} U$.

Hereafter, our aim will be to study separately the subcritical and supercritical modes, proposing suitable boundary conditions for them, and to combine them and obtain existence, uniqueness, and regularity of the solution U. In each case, we will study one (subcritical/supercritical) mode separately and then combine them for the whole subcritical and supercritical components. We now conclude this section with some remarks concerning the stationary (time independent) equations associated with Eqs. (4.6), (4.7) and by a trace theorem which will be used repeatedly in the sequel.

4.1.3. The stationary equations associated with Eqs. (4.6)–(4.7).
The (physical) spatial domain under consideration will be $\mathcal{M} = \mathcal{M}' \times (-L_3, 0)$, where $\mathcal{M}'$ is the interface atmosphere/ocean, $\mathcal{M}' = (0, L_1) \times (0, L_2)$.

We introduce, componentwise, the differential operators $\mathcal{A}_n = (\mathcal{A}_{n1}, \mathcal{A}_{n2}, \mathcal{A}_{n3})$ operating on $U_n = (u_n, v_n, \psi_n)$,

$$\mathcal{A}_n U_n = \begin{cases} \overline{U}_0 u_{nx} - \frac{1}{\lambda_n} \psi_{nx}, \\ \overline{U}_0 v_{nx} - \frac{1}{\lambda_n} \psi_{ny}, \\ \overline{U}_0 \psi_{nx} - \frac{N^2}{\lambda_n} (u_{nx} + v_{ny}), \end{cases} \tag{4.14}$$

with $\overline{U}_0$, N and $\lambda_n > 0$ as above.

Our object here is to study (recall) the nature of the stationary (time independent) equations in $\mathcal{M}'$:

$$\mathcal{A}_n U_n = F_n = (F_{un}, F_{vn}, F_{\psi n}), \quad n \geq 1. \tag{4.15}$$

We momentarily drop the indices n for the sake of simplicity and although this is not of direct use in the sequel, it is useful to look for the characteristics of the differential system $\mathcal{A}U = F$. We write this system in the matrix form

$$\mathcal{E} U_x + \mathcal{G} U_y = F, \tag{4.16}$$

with

$$\mathcal{E} = \begin{pmatrix} \overline{U}_0 & 0 & -\frac{1}{\lambda} \\ 0 & \overline{U}_0 & 0 \\ -\frac{N^2}{\lambda} & 0 & \overline{U}_0 \end{pmatrix}, \quad \mathcal{G} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -\frac{1}{\lambda} \\ 0 & -\frac{N^2}{\lambda} & 0 \end{pmatrix},$$

and the equation of the characteristics (see, e.g., JOHN [1982]) is given by

$$\det(\mathcal{E}dx - \mathcal{G}dy) = 0,$$

i.e.,

$$\det \begin{vmatrix} \overline{U}_0\mu & 0 & -\frac{\mu}{\lambda} \\ 0 & \overline{U}_0\mu & \frac{1}{\lambda} \\ -\frac{N^2\mu}{\lambda} & \frac{N^2}{\lambda} & \overline{U}_0\mu \end{vmatrix} = 0,$$

with $\mu = \mathrm{d}x/\mathrm{d}y$. Hence, the equation for μ:

$$\overline{U}_0\mu \left[\left(\overline{U}_0^2 - \frac{N^2}{\lambda^2} \right) \mu^2 - \frac{N^2}{\lambda^2} \right] = 0. \tag{4.17}$$

The (real) solution $\mu_0 = 0$ exists in all cases, producing the characteristics $x =$ constant (parallel to the background flow $\overline{U}_0 e_x$). This corresponds to the first equation:

$$\frac{\partial}{\partial x} \left(\overline{U}_0 u - \frac{\psi}{\lambda} \right) = F_u + fv.$$

Then, in the supercritical case, $\overline{U}_0^2 - N^2\lambda^{-2} > 0$ and we have two more real characteristics

$$\frac{\mathrm{d}x_\pm}{\mathrm{d}y} = \mu_\pm = \pm \frac{N}{\lambda} \left(\overline{U}_0^2 - \frac{N^2}{\lambda^2} \right)^{-1/2}, \tag{4.18}$$

whereas in the subcritical case, these two characteristics are imaginary.

For the stationary zero mode, we obtain from Eq. (4.6) after dropping the Coriolis term:

$$\begin{aligned} \overline{U}_0 u_x + \phi_x &= F_u, \\ \overline{U}_0 v_x + \phi_y &= F_v, \\ u_x + v_y &= 0. \end{aligned} \tag{4.19}$$

By elimination of ϕ, we find

$$\overline{U}_0 \left(u_{xy} - v_{xx} \right) = F_{u,y} - F_{v,x},$$

and hence, we find the fully elliptic equation

$$v_{xx} + v_{yy} = \frac{1}{\overline{U}_0} \left(F_{v,x} - F_{u,y} \right). \tag{4.20}$$

We infer from this remark that the stationary system $\mathcal{A}_n U_n = F_n$ is fully elliptic for the zero mode, partly hyperbolic and partly elliptic for the other subcritical modes (one real characteristic) and fully hyperbolic in the supercritical case (three real characteristics). This remark will underlie the studies in Sections 4.2 and 4.3 although as we said, we do not use it directly.

4.1.4. A trace theorem

We consider the same differential operator $\mathcal{A} = (\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3)$, as in Eq. (4.14) operating on $U = (u, v, \psi)$, but the indices n are dropped for the sake of simplicity:

$$\mathcal{A}U = \begin{cases} \overline{U}_0 u_x - \dfrac{1}{\lambda}\psi_x, \\ \overline{U}_0 v_x - \dfrac{1}{\lambda}\psi_y, \\ \overline{U}_0 \psi_x - \dfrac{N^2}{\lambda}(u_x + v_y), \end{cases} \tag{4.21}$$

with $\overline{U}_0$, N, $\lambda = \lambda_n > 0$ as above, and we consider the space[16]

$$\mathcal{X} = \left\{ U \in L^2(\mathcal{M}')^3, \mathcal{A}U \in L^2(\mathcal{M}')^3 \right\}, \tag{4.22}$$

endowed with its natural Hilbert norm $(|U|^2_{L^2(\Gamma_i)^3} + |\mathcal{A}U|^2_{L^2(\Gamma_i)^3})^{\frac{1}{2}}$. We have

THEOREM 4.1. *If $U = (u, v, \psi) \in \mathcal{X}$, the traces of v and ψ are defined on all of $\partial\mathcal{M}'$, the trace of u is defined at $x = 0$ and L_1, and they belong to the respective spaces $H_x^{-1}(0, L_1)$ and $H_y^{-1}(0, L_2)$. Furthermore, the trace operators are linear continuous in the corresponding spaces, e.g., $U \in \mathcal{X} \to u|_{x=0}$ is continuous from $\mathcal{X}$ into $H_y^{-1}(0, L_2)$.*

PROOF. Let us write $\mathcal{A}U = F = (f_1, f_2, f_3)$. Since $U = (u, v, \psi) \in L^2(\mathcal{M}')^3 = L^2_x(0, L_1; L^2_y(0, L_2)^3)$, we see that $U_y = \partial U/\partial y$ belongs to $L^2_x(0, L_1; H_y^{-1}(0, L_2)^3)$. From $\mathcal{A}_2 U = \overline{U}_0 v_x - \lambda^{-1}\psi_y = f_2 \in L^2(\mathcal{M}')$, we conclude that $v_x \in L^2_x(0, L_1; H_y^{-1}(0, L_2))$ and $v \in \mathcal{C}([0, L_1]; H_y^{-1}(0, L_2))$, so that its traces at $x = 0$ and L_1 are defined and belong to $H_y^{-1}(0, L_2)$. We then have

$$\overline{U}_0 u_x - \lambda^{-1}\psi_x = f_1 \in L^2_x(0, L_1; L^2_y(0, L_2)),$$
$$\overline{U}_0 \psi_x - (N^2/\lambda)u_x = f_2 - (N^2/\lambda)v_y \in L^2_x(0, L_1; H_y^{-1}(0, L_2))$$

so that both u_x and ψ_x belong to the last space and

$$u, \psi \in \mathcal{C}([0, L_1]; H_y^{-1}(0, L_2)).$$

[16] We will write $\mathcal{A}_n$, $\mathcal{X}_n$ when it is necessary to emphasize the dependence on n through λ $(\lambda = \lambda_n)$.

Their traces are defined as well at $x = 0$ and L_1. Finally, we write

$$\overline{U}_0 v_x - \lambda^{-1}\psi_y = f_2,$$
$$(\overline{U}_0 - N^2/\lambda^2\overline{U}_0)\psi_x - (N^2/\lambda)v_y = f_3 + N^2 f_1/\lambda\overline{U}_0$$

from which we conclude that v_y and $\psi_y \in L^2_y(0, L_2; H_x^{-1}(0, L_1))$ and thus v and $\psi \in \mathcal{C}_y([0, L_2]; H_x^{-1}(0, L_1))$ and their traces are both defined at $y = 0$ and L_2. Finally, all the mappings above are continuous, and the theorem is proved. □

REMARK 4.1. Although the values of $\overline{U}_0$, N, $\lambda = \lambda_n$ are intended to be those above, Theorem 4.1 extends to operators $\mathcal{A}$ with the same structure and more general constant coefficients, and it will be used in this way at times.

4.1.5. The zero mode

The equations for this mode appear in Eq. (4.6), but for the convenience of the notations, the subscripts are now changed to superscripts. Due to the form of the third equation, we proceed by analogy with the incompressible Navier–Stokes equations, and we determine first $\mathbf{u}^0 = (u^0, v^0)$ and then ϕ^0 by solving a Neumann problem. The natural function space for $\mathbf{u}^0$ is

$$H^0 = \left\{\mathbf{u}^0 = (u^0, v^0) \in L^2(\mathcal{M}')^2, u^0_x + v^0_y = 0, \mathbf{u}^0 \cdot \mathbf{n} = 0 \text{ on } \partial\mathcal{M}'\right\}, \tag{4.23}$$

where $\mathbf{n} = (n_x, n_y)$ is the unit outward normal on $\partial\mathcal{M}'$. Recall (see, e.g., TEMAM [2001]) that the trace of $\mathbf{u}^0 \cdot \mathbf{n}$ on $\partial\mathcal{M}'$ makes sense for $\mathbf{u}^0 \in L^2(\mathcal{M}')^2$ with $\operatorname{div} \mathbf{u}^0 = u^0_x + v^0_y \in L^2(\mathcal{M}')$. If the test function $\widetilde{\mathbf{u}}^0 = (\widetilde{u}^0, \widetilde{v}^0) \in H^0$ is smooth, we classically see that Eq. (4.6) implies that

$$\frac{\mathrm{d}}{\mathrm{d}t}(\mathbf{u}^0, \widetilde{\mathbf{u}}^0)_{H^0} + \bar{U}_0(\mathbf{u}^0_x, \widetilde{\mathbf{u}}^0) + f(e_z \wedge \mathbf{u}^0, \widetilde{\mathbf{u}}^0) = 0, \tag{4.24}$$

where $e_z = (0, 0, 1)$. Conversely, if there exists u^0 such that Eq. (4.24) is satisfied for all such $\widetilde{\mathbf{u}}^0$, then there exists ϕ^0 such that Eq. (4.6) is satisfied.

We then introduce the linear unbounded operator A^0 in H,

$$A^0\mathbf{u}^0 = P_{H^0}\left(\bar{U}_0\frac{\partial \mathbf{u}^0}{\partial x} + f e_z \wedge \mathbf{u}^0\right), \tag{4.25}$$

with domain

$$D(A^0) = \left\{\mathbf{u}^0 \in H^0, \mathbf{u}^0_x \in L^2(\mathcal{M}')^2\right\}, \tag{4.26}$$

where P_{H^0} is the orthogonal projector in $L^2(\mathcal{M}')^2$ onto H^0. Equation (4.24) is then equivalent to the evolution equation

$$\frac{\mathrm{d}\mathbf{u}^0}{\mathrm{d}t} + A^0\mathbf{u}^0 = 0. \tag{4.27}$$

Using the Hille–Phillips–Yoshida theorem, it is easy to see that Eq. (4.27) with initial condition $\mathbf{u}^0(0)$ given in H^0 or $D(A^0)$ produces a well-posed initial value problem. For that purpose, it is sufficient to show that $-A^0$ is the infinitesimal generator of a contraction semi-group in H^0. Since the operator $\mathbf{u}^0 \longrightarrow P_{H^0}(fe_z \wedge \mathbf{u}^0)$ is continuous in H^0, it suffices to show that $\bar{A}^0\mathbf{u}^0 = P_{H^0}\bar{U^0}\mathbf{u}_{0x}$ with domain $D(\bar{A}^0)$ is dense in H^0 and $\bar{A}^0$ is closed which is easy; also $\bar{A}^0 \geq 0$ as

$$\begin{aligned}(\bar{A}^0\mathbf{u}^0, \mathbf{u}^0)_{H^0} &= \overline{U}_0 \int_{\mathcal{M}'} \mathbf{u}^0_x \cdot \mathbf{u}^0 \mathrm{d}\mathcal{M}' \\ &= \frac{\overline{U}_0}{2} \int_0^{L_2} \Big[|\mathbf{u}^0|^2(L_1, y) - |\mathbf{u}^0|^2(0, y)\Big] \mathrm{d}y \\ &= 0,\end{aligned} \tag{4.28}$$

the integration in x being justified for $\mathbf{u}^0 \in D(A^0)$. We need also to show that $\bar{A}^{0*}$ is positive, but this results from the fact that $\bar{A}^{0*} = -\bar{A}^0$, with the same domain.[17] We refrain from giving all the details of the proof for this partial result and refer the reader to Section 4.4 for the complete analysis.

4.2. *Subcritical modes*

We now proceed and study the subcritical modes $1 \leq n \leq n_c$.

4.2.1. *One subcritical mode* ($1 \leq n \leq n_c$)

We temporarily drop the indices n and first want to set and study an initial value problem for Eq. (4.7) when the mode is subcritical, i.e. (see Eq. (4.8)):

$$\lambda = \lambda_n < \frac{N}{\overline{U}_0} \quad (0 \leq n \leq n_c).$$

There are several possible choices of suitable boundary conditions see, e.g., different ones in Rousseau, Temam and Tribbia [2005b] for a related situation. Here, for a simple subcritical mode $1 \leq n \leq n_c$, we choose the following boundary conditions:

$$\begin{cases} \psi = 0 \text{ at } x = L_1, \quad \text{and } y = 0, L_2, \\ v = 0 \text{ and } u = \psi/\lambda\overline{U}_0 \text{ at } x = 0, \end{cases} \tag{4.29}$$

and we introduce the space

$$D(A) = \Big\{U \in L^2(\mathcal{M}')^3, \mathcal{A}U \in L^2(\mathcal{M}')^3, U \text{ satisfies (4.29)}\Big\}, \tag{4.30}$$

and the operator[18]

$$AU = \mathcal{A}U.$$

[17] Note that $A^{0*} = -A^0$ as well, and of course $D(A^{0*}) = D(A^0) = D(\bar{A}^0)$.

[18] When needed we will write also $\mathcal{A}_n$, A_n, $D(A_n)$ to emphasize the dependence on n ($\lambda = \lambda_n$).

In view of Theorem 4.1, the traces appearing in Eq. (4.29) are well defined when $U \in L^2(\mathcal{M}')^3$ and $\mathcal{A}U \in L^2(\mathcal{M}')^3$ so that the definition of $D(A_n)$ in (4.30) makes sense.

REMARK 4.2. As indicated above, the boundary conditions in Eq. (4.29) are different than those in Section 2 for dimension 2 (after neglecting the dependence on y).

We proceed with a regularity result for U in $D(A)$ (see the comments in Section 4.1.3).

THEOREM 4.2. *If $U = (u, v, \psi) \in D(A)$, then v and ψ belong to $H^1(\mathcal{M}')$ and u_x belongs to $L^2(\mathcal{M}')$*

PROOF. For $U \in D(A)$, we set $AU = F = (f_1, f_2, f_3)$. Then, in the distribution sense,

$$\begin{cases} \overline{U}_0 u_x - \dfrac{1}{\lambda}\psi_x = f_1, \\ \overline{U}_0 v_x - \dfrac{1}{\lambda}\psi_y = f_2, \\ \overline{U}_0 \psi_x - \dfrac{N^2}{\lambda}(u_x + v_y) = f_3. \end{cases} \tag{4.31}$$

Combining the first and third equations, we find

$$\overline{U}_0 \left(1 - \frac{N^2}{\lambda^2 \overline{U}_0^2}\right) \psi_x - \frac{N^2}{\lambda} v_y = \frac{N^2}{\lambda \overline{U}_0} f_1 + f_3. \tag{4.32}$$

Combining this equation with the second equation of Eq. (4.31), we obtain

$$(\overline{U}_0^2 - \frac{N^2}{\lambda^2})\psi_{xx} - \frac{N^2}{\lambda^2}\psi_{yy} = \frac{N^2}{\lambda} f_{1x} + \frac{N^2}{\lambda} f_{2y} + \overline{U}_0 f_{3x}. \tag{4.33}$$

Note that this equation is elliptic in the subcritical case; of course, a similar elliptic equation can be derived for v, but we will not use it. We associate to this equation the boundary condition $\psi = 0$ at $y = 0$, L_2 and $x = L_1$ contained in Eq. (4.29). Then, for the side $x = 0$ of $\mathcal{M}'$, a suitable boundary condition is given by Eq. (4.32) in which $v_y = 0$ since $v = 0$ at $x = 0$; hence,

$$-\frac{\partial \psi}{\partial n} = \psi_x = \overline{U}_0^{-1} \left(1 - \frac{N^2}{\lambda^2 \bar{U}_0^2}\right)^{-1} \left(\frac{N^2}{\lambda \overline{U}_0} f_1 + f_2\right), \text{ at } x = 0. \tag{4.34}$$

The right-hand side of Eq. (4.34) does not make sense on $x = 0$ for $F \in L^2(\mathcal{M}')^3$. So, we proceed as follows: we approximate F in $L^2(\mathcal{M}')^3$ by a sequence of smooth functions $F_m \in \mathcal{C}^\infty(\mathcal{M}')^3$. For each m, the right-hand side of Eq. (4.34) makes sense, and we find a unique solution ψ_m of Eqs. (4.32), (4.34), and $\psi_m = 0$ on the other

sides of $\mathcal{M}'$. Of course, ψ_m is $\mathcal{C}^\infty$ on $\mathcal{M}'$ away from the corners and $\psi_m \in H^1(\mathcal{M}')$ (at least), see GRISVARD [1985]. Then, from ψ_m, we determine the corresponding $v = v_m$ up to an additive constant: v_{mx} and v_{my} are given by Eq. (4.32) and the second equation of Eq. (4.31), and these equations are compatible (i.e., $v_{mxy} = v_{myx}$) because of Eq. (4.33). Note that v_m belongs to $H^1(\mathcal{M}')$ at least, its trace on the side $x = 0$ of $\mathcal{M}'$ is defined, $v_{my} = 0$ on this side because of Eq. (4.34). Hence, $v_m = 0$ on $x = 0$ by choosing properly the constant. Finally, u_m is determined by the first equation of Eq. (4.31) and the boundary condition $U_m = \psi_m/\lambda\overline{U}_0$ at $x = 0$. In conclusion, $U_m = (u_m, v_m, \psi_m)$ that we just constructed belongs to $D(A)$ and satisfies $AU_m = F_m$.

To pass to the limit $m \to \infty$, we obtain the suitable a priori estimates as follows: we multiply the second equation of Eq. (4.31) by $-(N^2/\lambda)\psi_{my}$, Eq. (4.32) by $-\overline{U}_0\psi_{mx}$, integrate over $\mathcal{M}'$ and add these equations. We find

$$\overline{U}_0^2\left(\frac{N^2}{\lambda^2\overline{U}_0^2} - 1\right)\int_{\Gamma_i}\psi_{mx}^2 \mathrm{d}\mathcal{M}' + \frac{N^2}{\lambda^2}\int_{\mathcal{M}'}\psi_{my}^2 \mathrm{d}\mathcal{M}'$$

$$+ \frac{\bar{U}_0 N^2}{\lambda}\int_{\Gamma i}\left(v_{my}\psi_{mx} - v_{mx}\psi_{my}\right)\mathrm{d}\mathcal{M}' = \tag{4.35}$$

$$= -\int_{\mathcal{M}'}\left(\frac{N^2}{\lambda}f_{m2}\psi_{my} + \frac{N^2}{\lambda}f_{m1}\psi_{mx} + \overline{U}_0 f_{m2}\psi_{mx}\right)\mathrm{d}\mathcal{M}'.$$

The integrals involving $v\psi$ cancel each other because it is legitimate to integrate by parts (enough regularity) and, by integration by parts, taking into account the boundary conditions in Eq. (4.29) for U_m, we find

$$\int_{\mathcal{M}'} v_{my}\psi_{mx}\mathrm{d}\mathcal{M}' = \int_{\Gamma_i} v_{mx}\psi_{my}\mathrm{d}\mathcal{M}'. \tag{4.36}$$

Since $N^2 > \lambda^2\overline{U}_0^2$, we then easily infer from Eq. (4.35) that

$$|\nabla\psi_m|_{L^2(\mathcal{M}')^2} \leq c|F_m|_{L^2(\Gamma_i)^2} \leq \text{ const.} \tag{4.37}$$

Thanks to the boundary conditions on ψ_m, we have a Poincaré inequality which guarantees that

$$|\psi_m|_{L^2(\Gamma_i)} \leq \text{ const,} \tag{4.38}$$

and ψ_m is bounded in $H^1(\mathcal{M}')$. As for the construction of v_m, the second equation of Eqs. (4.31), (4.32) and $v_m = 0$ on $x = 0$ then show that v_m is bounded in $H^1(\mathcal{M}')$. Finally, $\overline{U}_0 u_m - \psi_m/\lambda$ and its x derivative are bounded in $L^2(\mathcal{M}')$ so that u_m and u_{mx} are bounded in $L^2(\mathcal{M}')$ as well.

Passing to the limit $m \to \infty$, we obtain $U_m \to \overline{U}$, with $\overline{U} \in D(A)$ and $A\overline{U} = F, \overline{U}$ satisfying the desired regularity properties. To conclude, we need to show that $\overline{U} = U$, i.e., A is one-to-one.

We thus consider $U \in D(A)$ such that $AU = 0$. Then, U satisfies Eq. (4.31) with $f_1 = f_2 = f_3 = 0$ and the boundary conditions in Eq. (4.29): ψ, v also satisfy Eqs. (4.32), (4.33), and (4.34) with $F = 0$. The mixed Neumann–Dirichlet problem of which ψ is solution shows that $\psi = 0$; then $v = 0$ because of Eq. (4.32), the second equation of Eq. (4.31) and $v = 0$ at $x = 0$. Finally, $u = 0$ because of the first equation of Eq. (4.31) and the boundary condition $\overline{U}_0 u - \psi\lambda = 0$ at $x = 0$.

Theorem 4.2 is thus proved. □

*4.2.2. Positivity of A and A**

We endow the space $H = L^2(\mathcal{M}')^3$ with the Hilbert scalar product and norm

$$(U, \widetilde{U})_H = \int_{\Gamma_i} \left(u\widetilde{u} + v\widetilde{v} + \frac{1}{N^2}\psi\widetilde{\psi} \right) \mathrm{d}\mathcal{M}', \; |U|_H = \{(U, U)_H\}^{1/2}.$$

Our aim is now to prove that A and its adjoint A^* defined below are positive in the sense

$$\begin{cases} (AU, U)_H \geq 0, \quad \forall U \in D(A), \\ (A^*U, U)_H \geq 0, \quad \forall U \in D(A^*). \end{cases} \tag{4.39}$$

These properties are needed to apply the Hille–Phillips–Yoshida theorem (see Section 4.4). The result for U is now easy, thanks to Theorem 4.2. Indeed the following easy calculations are now legitimate, $\forall \; U \in D(A)$:

$$\begin{aligned} (AU, U)_H &= \int_{\Gamma_i} \left[\left(\overline{U}_0 u_x - \frac{1}{\lambda}\psi_x \right) u + \left(\overline{U}_0 v_x - \frac{1}{\lambda}\psi_y \right) v \right. \\ &\quad \left. + \frac{\overline{U}_0}{N^2}\psi_x\psi - \frac{1}{\lambda}(u_x + v_y)\psi \right] \mathrm{d}\mathcal{M}' \\ &= \frac{\overline{U}_0}{2} \int_0^{L_2} (u^2 + v^2)(L_1, y)\mathrm{d}y \\ &\quad - \int_0^{L_2} \left[\frac{\overline{U}_0}{2}(u^2 + \frac{1}{N^2}\psi^2)(0, y) - \frac{1}{\lambda}(u\psi)(0, y) \right] \mathrm{d}y \\ &\geq \frac{\overline{U}_0}{2} \int_0^{L_2} \left((\lambda\overline{U}_0)^{-2} - N^{-2} \right) \psi^2(0, y)\mathrm{d}y \geq 0. \end{aligned} \tag{4.40}$$

All the integrations by parts above are easy to justify for functions in $H^1(\mathcal{M}')$. We just want to emphasize those involving u. If u and $\widetilde{u}$ belong to $L^2_y(0, L_2; H^1_x(0, L_1))$, then $u, \widetilde{u} \in L^2_y(0, u_2; \mathcal{C}_x([0, L_1]))$ and for a.e. $y \in (0, L_2)$:

$$\int_0^{L_1} (u_x\widetilde{u} + u\widetilde{u}_x)(x, y)\mathrm{d}x = (u\widetilde{u})(1, y) - (u\widetilde{u})(0, y)$$

and integrating in y,

$$\int_{\Gamma_i} (u_x\widetilde{u} + u\widetilde{u}_x)\mathrm{d}x\mathrm{d}y = \int_0^{L_2} [(u, \widetilde{u})(1, y) - (u\widetilde{u})(0, y)]\mathrm{d}y. \tag{4.41}$$

To prove Eq. (4.40), we apply Eq. (4.41) with $\widetilde{u} = u$, ψ, and v.

We now turn to the definition of the formal adjoint A^* of A and its domain $D(A^*)$, in the sense of the adjoint of a linear unbounded operator (see RUDIN [1991]). For that purpose, we first compute $(AU, \widetilde{U})_H$ for U and $\widetilde{U}$ smooth. By integration by parts, using Stokes formula, we find:

$$\begin{aligned}(AU, \widetilde{U})_H &= \int_{\Gamma_i} \left[\left(\overline{U}_0 u_x - \frac{1}{\lambda}\psi_x \right) \widetilde{u} + \left(\overline{U}_0 v_x - \frac{1}{\lambda}\psi_y \right) \widetilde{v} \right. \\ &\quad \left. + \frac{\overline{U}_0}{N^2}\psi_x \widetilde{\psi} - \frac{1}{\lambda}(u_x + v_y)\widetilde{\psi} \right] \mathrm{d}\mathcal{M}' \\ &= I_0 + I_1,\end{aligned} \tag{4.42}$$

where I_0 stands for the integrals in $\mathcal{M}'$ and I_1 for the integrals on $\partial\mathcal{M}'$. For I_0, we have

$$I_0 = \int_{\Gamma_i} \left(\mathcal{A}_1^* \widetilde{U} u + \mathcal{A}_2^* \widetilde{U} v + N^{-2} \mathcal{A}_3^* \widetilde{U} \psi \right) \mathrm{d}\mathcal{M}', \tag{4.43}$$

with $\mathcal{A}^* \widetilde{U} = (\mathcal{A}_1^* \widetilde{U}, \mathcal{A}_2^*, \widetilde{U}, \mathcal{A}_3^* \widetilde{U})$, and

$$\mathcal{A}^* \widetilde{U} = \begin{cases} -\overline{U}_0 \widetilde{u}_x + \dfrac{1}{\lambda}\widetilde{\psi}_x, \\ -\overline{U}_0 \widetilde{v}_x + \dfrac{1}{\lambda}\widetilde{\psi}_y, \\ -\overline{U}_0 \widetilde{\psi}_x + \dfrac{N^2}{\lambda}(\widetilde{u}_x + \widetilde{v}_y). \end{cases} \tag{4.44}$$

For I_1, taking into account the boundary conditions in Eq. (4.29), there remains:

$$\begin{aligned} I_1 &= \int_0^{L_2} \left[\overline{U}_0 (u\widetilde{u})(L_1, y) + \overline{U}_0 (v\widetilde{v})(L_1, y) - \frac{1}{\lambda}(u\widetilde{\psi})(L_1, y) \right] \mathrm{d}y \\ &\quad + \int_0^{L_2} \left(-\frac{\overline{U}_0}{N^2} + \frac{1}{\lambda^2 \overline{U}_0} \right) (\psi\widetilde{\psi})(0, y) \mathrm{d}y \\ &\quad - \lambda^{-1} \int_0^{L_1} \left[(v\widetilde{\psi})(x, L_2) + (v\widetilde{\psi})(x, 0) \right] \mathrm{d}x. \end{aligned}$$

According to RUDIN [1991], $D(A^*)$ consists of the $\widetilde{U}$ in H such that $U \to (AU, \widetilde{U})_H$ is continuous on $D(A)$ for the topology (norm) of H. If U is restricted to the class of $\mathcal{C}^\infty$ functions with compact support in $\mathcal{M}'$ (endowed with the norm of H), then $I_1 = 0$, and $U \to I_0$ can only be continuous if $\mathcal{A}^* \widetilde{U}$ as defined in Eq. (4.44) belongs to $L^2(\mathcal{M}')^3$. We then observe that Theorem 4.1 applies to $\mathcal{A}^*$ as well and to more general constant coefficients operators. Hence, if $\widetilde{U} \in D(A^*)$, then $\widetilde{U} \in L^2(\mathcal{M}')^3$ with $\mathcal{A}^* \widetilde{U} \in L^2(\mathcal{M}')^3$, and the traces of $\widetilde{U}$ are defined as in Theorem 4.1. We now restrict U to the class of $\mathcal{C}^\infty$

functions on $\bar{\mathcal{M}}'$ which belong to $D(A)$. Then, the expressions above of I_0 and I_1 show that $U \to (AU, \widetilde{U})_H$ can only be continuous in U for the topology (norm) of H if the following boundary conditions are satisfied:

$$\begin{cases} \widetilde{\psi} = 0 \text{ at } y = 0, L_2 \text{ and } x = 0, \\ \widetilde{v} = 0 \text{ and } \widetilde{u} = \widetilde{\psi}/\lambda \overline{U}_0 \text{ at } x = L_1. \end{cases} \tag{4.45}$$

Hence, we conclude that[19]

$$D(A^*) = \left\{ \widetilde{U} \in L^2(\mathcal{M}')^3, \mathcal{A}^* \widetilde{U} \in L^2(\mathcal{M}')^3, \text{ and } \widetilde{U} \text{ satisfies (4.45)} \right\}. \tag{4.46}$$

We have shown indeed that $D(A^*)$ is included in the right-hand side of Eq. (4.46). Now, with exactly the same reasoning as in Theorem 4.2, we can show that if $\widetilde{U} = (\widetilde{u}, \widetilde{v}, \widetilde{\psi}) \in D(A^*)$, then $\widetilde{v}$ and $\widetilde{\psi}$ belong to $H^1(\mathcal{M}')$ and $\widetilde{u}_x$ belongs to $L^2(\mathcal{M}')$.

Thus, using again Eq. (4.41), we see that for every U in $D(A)$ and $\widetilde{U}$ in $D(A^*)$ (not necessarily $\mathcal{C}^\infty$), then $(AU, \widetilde{U})_H = I_0 + I_1$ as above, with $I_1 = 0$ and I_0 as in Eq. (4.43), so that $U \to (AU, \widetilde{U})_H$ is continuous on $D(A)$ for the norm of H. The opposite inclusion is proven and Eq. (4.46) is established. This reasoning also shows that, for every $\widetilde{U} \in D(A^*)$, $A^* \widetilde{U} = \mathcal{A}^* \widetilde{U}$, $\mathcal{A}^*$ as in Eq. (4.44).

It is now easy to prove the positivity of A^*, i.e., the second statement in Eq. (4.39). We proceed as in Eq. (4.40), using Eq. (4.41):

$$\begin{aligned} (A^*U, U)_H = \int_{\Gamma_i} \Bigg[& \left(-\overline{U}_0 u_x + \frac{1}{\lambda} \psi_x \right) u \\ & + \left(-\overline{U}_0 v_x + \frac{1}{\lambda} \psi_y \right) v + \left(-\frac{\bar{U}_0}{N^2} \psi_x + \frac{1}{\lambda}(u_x + v_y) \right) \psi \Bigg] \mathrm{d}\mathcal{M}' \\ & \text{(using the boundary conditions in Eq. (4.45))} \\ = & \frac{\bar{U}_0}{2} \int_0^{L_2} (u^2 + v^2)(0, y) \mathrm{d}y \\ & - \int_0^{L_2} \left[\frac{\bar{U}_0}{2} (u^2 + \frac{1}{N^2} \psi^2)(L_1, y) - \frac{1}{\lambda} (u\psi)(L_1, y) \right] \mathrm{d}y \\ \geq & \frac{\bar{U}_0}{2} \int_0^{L_2} \left((\lambda \overline{U}_0)^{-2} - N^{-2} \right) \psi^2(L_1, y) \mathrm{d}y \geq 0. \end{aligned}$$

Hence the positivity. Note that we cannot just write $(A^*U, U)_H = (AU, U)_H \geq 0$ because U in $D(A^*)$ may not belong to $D(A)$. In summary, we have proven the following theorem:

THEOREM 4.3. *For every $U \in D(A_n)$, as defined in Eq. (4.30), we have $(A_n U, U)_{L^2(\mathcal{M}')^3} \geq 0$. Similarly, for every $U \in D(A_n^*)$ defined in Eq. (4.46), we have $(A_n^* U, U)_{L^2(\mathcal{M}')^3} \geq 0$.*

[19] Similarly, we write $\mathcal{A}_n^*$, A_n^*, $D(A_n^*)$ when the dependence on n needs to be emphasized ($\lambda = \lambda_n$).

REMARK 4.3. Based on the previous results, we can show that, for each n, $1 \leq n \leq n_c$, $-A = -A_n$ is the infinitesimal generator of a contraction semi-group. Then, by application of the Hille–Yoshida theorem, we can solve the initial and boundary value problem associated with Eq. (4.7) for every such n. We refrain from developing this and will instead establish a well-posedness result for all modes together, see Section 4.4.

4.3. Supercritical modes

We now consider the initial and boundary value problem for one single supercritical mode, i.e., Eq. (4.5) or equivalently Eq. (4.7) when $n > n_c$. We temporarily drop the indices n and write, e.g.,

$$\lambda = \lambda_n > \frac{N}{\overline{U}_0}. \tag{4.47}$$

*4.3.1. The operator A and its adjoint A**

Here, for one supercritical mode, we choose the following boundary conditions:

$$\begin{cases} u, v, \text{ and } \psi = 0 \text{ at } x = 0, \\ \text{and } \psi = 0 \text{ at } y = 0 \text{ and } L_2. \end{cases} \tag{4.48}$$

In this case, the operator $A = A_n$ is defined by $AU = \mathcal{A}U$ as in Eq. (4.21), and

$$D(A) = \left\{ U \in H = L^2(\mathcal{M}')^2, \mathcal{A}U \in L^2(\mathcal{M}'), U \text{ satisfies Eq. (4.48)} \right\}. \tag{4.49}$$

Note that, according to Theorem 4.1, the traces of u, v, ψ appearing in Eqs. (4.48) and (4.49) are well defined when $U \in L^2(\mathcal{M}')^3$ and $\mathcal{A}U \in L^2(\mathcal{M}')^3$.

In view of proving that $-A = -A_n$ is the infinitesimal generator of a contraction semigroup, our main task is now to show that

$$\begin{cases} (AU, U)_H \geq 0, \forall U \in D(A), \quad \text{and} \\ (A^*U, U)_H \geq 0, \forall U \in D(A^*)^{20}, \end{cases} \tag{4.50}$$

where A^* is defined below. Our approach for Eq. (4.50) is, however, different from the subcritical case which was based on the regularity result Theorem 4.2. In the supercritical case, the equations are hyperbolic and there are no similar regularity results. Instead we are going to prove that $(AU, U)_H \geq 0$ when U is sufficiently regular; then we define A^* and prove that $(A^*U, U)_H \geq 0$ for every U, sufficiently regular, in the domain of A^*; and finally, by passage to the limit, we prove Eq. (4.50) for all functions in $D(A)$ and $D(A^*)$.

4.3.2. Positivity of A

We prove that $(AU, U)_H \geq 0$ when U belongs to $D(A)$ and is sufficiently regular (say in $\mathcal{C}^2(\bar{\mathcal{M}}')^3$):

$$\begin{aligned}
(AU, U)_H &= \int_{\mathcal{M}'} [(\overline{U}_0 u_x - \frac{1}{\lambda}\psi_x)u + (\overline{U}_0 v_x - \frac{1}{\lambda}\psi_y)v \\
&\quad + \frac{1}{N^2}(\overline{U}_0\psi_x - \frac{N^2}{\lambda}(u_x + v_y))\psi]\mathrm{d}\mathcal{M}' \\
&= \text{ (using Eqs. (4.48) and (4.47))} \\
&= \frac{\overline{U}_0}{2}\int_0^{L_2}(u^2 + v^2 + \frac{1}{N^2}\psi^2)(L_1, y)\mathrm{d}y \\
&\quad - \frac{1}{\lambda}\int_0^{L_2}(u\psi)(L_1, y)\mathrm{d}y \\
&= \frac{\overline{U}_0}{2}\int_0^{L_2} v^2(L_1, y)\mathrm{d}y + \\
&\quad + \frac{\overline{U}_0}{2}\int_0^{L_2}(u^2 + \frac{1}{N^2}\psi^2 - \frac{2}{\lambda\overline{U}_0}u\psi)(L_1, y)\mathrm{d}y \\
&\geq 0.
\end{aligned} \tag{4.51}$$

4.3.3. The adjoint A^*

Assume that $U \in D(A)$ and $\widetilde{U} \in H$ are smooth functions; then as in Eq. (4.42):

$$\begin{aligned}
(AU, \widetilde{U})_H &= \int_{\mathcal{M}'}\Big[(\overline{U}_0 u_x - \frac{1}{\lambda}\psi_x)\widetilde{u} + (\overline{U}_0 v_x - \frac{1}{\lambda}\psi_y)\widetilde{v} \\
&\quad + \frac{1}{N^2}(\overline{U}_0\psi_x - \frac{N^2}{\lambda}(u_x + v_y))\widetilde{\psi}\Big]\mathrm{d}\mathcal{M}' \\
&= I_0 + I_1,
\end{aligned} \tag{4.52}$$

where I_0 stands for the integrals on $\mathcal{M}'$ and I_1 for the integrals on $\partial\mathcal{M}'$. For I_0, we have

$$I_0 = \int_{\mathcal{M}'}\left(\mathcal{A}_1^*\widetilde{U}u + \mathcal{A}_2^*\widetilde{U}v + N^{-2}\mathcal{A}_3^*\widetilde{U}\psi\right)\mathrm{d}\mathcal{M}',$$

with $\mathcal{A}^*\widetilde{U} = (\mathcal{A}_1^*\widetilde{U}, \mathcal{A}_2^*\widetilde{U}, \mathcal{A}_3^*\widetilde{U})$ as in Eq. (4.44). For I_1, taking into account the boundary conditions in Eq. (4.48), there remains:

$$\begin{aligned}
I_1 &= \int_0^{L_2}\overline{U}_0[(u\widetilde{u}) + (v\widetilde{v}) + N^{-2}(\psi\widetilde{\psi})](L_1, y)\mathrm{d}y \\
&\quad - \int_0^{L_2}\lambda^{-1}(\psi\widetilde{u} + u\widetilde{\psi})(L_1, y)\mathrm{d}y \\
&\quad - \int_0^{L_1}\lambda^{-1}[(v\widetilde{\psi})(L_2, y) - (v\widetilde{\psi})(0, y)]\mathrm{d}y.
\end{aligned} \tag{4.53}$$

According to RUDIN [1991] , $D(A^*)$ consists of the $\widetilde{U}$ in H such that $U \longrightarrow (AU, \widetilde{U})_H$ is continuous on $D(A)$ for the topology (norm) of H. If U is restricted to the class of $\mathcal{C}^\infty$ functions with compact support in $\mathcal{M}'$ (endowed with the norm of H), then $I_1 = 0$, and $U \longrightarrow I_0$ can only be continuous if $\mathcal{A}^*\widetilde{U}$ as defined in Eq. (4.44) belongs to $L^2(\mathcal{M}')^3$. If $\widetilde{U}$ belongs to H and $\mathcal{A}^*\widetilde{U}$ belongs to $L^2(\mathcal{M}')^3$, then we already observed that Theorem 4.1 applies to $\mathcal{A}^*$ as well. Consequently, the traces of $\widetilde{U}$ are defined as in Theorem 4.1, and the calculations in Eq. (4.52) are now valid for any such $\widetilde{U}$ (and U in $D(A)$ not necessarily smooth). We now restrict U to the class of $\mathcal{C}^\infty$ function on $\bar{\mathcal{M}}'$ which belong to $D(A)$. Then, the expressions above of I_0 and I_1 show that $U \longrightarrow (AU, \widetilde{U})_H$ can only be continuous in U for the topology (norm) of H if the following boundary conditions are satisfied:

$$\begin{cases} \widetilde{u}, \widetilde{v} \text{ and } \widetilde{\psi} = 0 \text{ at } x = L_1, \\ \text{and } \widetilde{\psi} = 0 \text{ at } y = 0 \text{ and } L_2. \end{cases} \tag{4.54}$$

Conversely, if $\widetilde{U} \in H$, $\mathcal{A}^*\widetilde{U} \in L^2(\mathcal{M}')^3$ and the conditions in Eq. (4.54) are satisfied, then the calculations in Eq. (4.52) are valid, $I_1 = 0$, and $U \longrightarrow (AU, \widetilde{U})_H$ is continuous on $D(A)$ for the norm of H. Hence $\widetilde{U} \in D(A^*)$ and we conclude[21] that

$$D(A^*) = \left\{ \widetilde{U} \in L^2(\mathcal{M}')^3, \mathcal{A}^*\widetilde{U} \in L^2(\mathcal{M}')^3, \quad \text{and } \widetilde{U} \text{ satisfies (4.54)} \right\}; \tag{4.55}$$

and that $A^*\widetilde{U} = \mathcal{A}^*\widetilde{U}$ for $\widetilde{U}$ in $D(A^*)$, $\mathcal{A}^*$ as in Eq. (4.44).

*4.3.4. Positivity of A and A**

The proof of the positivity is not done as in the subcritical case, since the regularity result of Theorem 4.2 is not available in this case. Instead, for A, to prove that $(AU, U)_H \geq 0$, for U in $D(A)$, we will construct a sequence of smooth functions $U_n \in D(A)$ such that, as $n \longrightarrow \infty$,

$$U_n \longrightarrow U \text{ in } H \text{ strongly},$$
$$AU_n \rightharpoonup AU \text{ in } H \text{ weakly}.$$

Then, $(AU_n, U_n)_H \longrightarrow (AU, U)_H$ and since $(AU_n, U_n)_H \geq 0$ by Eq. (4.51), $(AU, U)_H \geq 0$ follows. The proof for A^* would be similar.

Given $U \in D(A)$, with $F = (f_1, f_2, f_3) = AU \in H$, we observe that the calculations in Eqs. (4.31)–(4.33) are still valid, but now since $\lambda > N/\widetilde{U}_0$, Eq. (4.33) is hyperbolic. In fact, we are now going to treat Eq. (4.33) as a second-order evolution equation in x (wave equation) in which x is the time-like variable and y is the spatial variable. For such a wave equation, we need to prescribe ψ and ψ_x at $x = 0$, and ψ at $y = 0$ and L_2. These values of ψ are given equal to 0, and we are missing ψ_x which we infer from the first and third equations of Eq. (4.31) when U is smooth, which we assume for the

[21] Remember that A, A^* depend on n through $\lambda = \lambda_n$; we write A_n, A_n^* when the dependance on n needs to be emphasized.

moment. Indeed since $v = 0$ at $x = 0$, $v_y = 0$ and these equations, restricted to $x = 0$, become a system

$$\overline{U}_0 u_x - \lambda^{-1}\psi_x = f_1,$$
$$\overline{U}_0 \psi_x - N^2\lambda^{-1}u_x = f_3,$$

which allows us to compute u_x and ψ_x at $x = 0$; hence for ψ_x:

$$\psi_x(0, y) = \frac{1}{\overline{U}_0^2 - N^2\lambda^{-2}} \left(\frac{N^2}{\lambda} f_1(L_1, y) + \overline{U}_0 f_3(L_1, y)\right), 0 < y < L_2. \tag{4.56}$$

We continue to assume that all functions $(f_1, f_2, f_3, u, v, \psi)$ are sufficiently regular, and we integrate Eq. (4.33) from 0 to x. Setting

$$\Psi(x, y) = \int_0^x \psi(x', y)\mathrm{d}y, \tag{4.57}$$

we obtain:

$$-(\overline{U}_0^2 - \frac{N^2}{\lambda^2})(\psi_x(x, y) - \psi_x(0, y))$$
$$-\frac{N^2}{\lambda^2}\Psi_{yy}(x, y) = \frac{N^2}{\lambda}(f_1(x, y) - f_1(0, y))$$
$$+\frac{N^2}{\lambda}F_{2y}(x, y) + \overline{U}_0(f_3(x, y) - f_3(0, y)),$$

where

$$F_i(x, y) = \int_0^x f_i(x', y)\mathrm{d}x'. \tag{4.58}$$

Taking Eq. (4.56) into account, there remains

$$(\overline{U}_0^2 - \frac{N^2}{\lambda^2})\Psi_{xx} - \frac{N^2}{\lambda^2}\Psi_{yy} = \frac{N^2}{\lambda}f_1 + \frac{N^2}{\lambda}F_{2y} + \overline{U}_0 f_3, \tag{4.59}$$

which we aim to consider for $x > 0$, with the initial and boundary conditions:

$$\begin{cases} \Psi = 0 \text{ and } \Psi_x = \psi = 0 \text{ at } x = 0, \\ \Psi = 0 \text{ at } y = 0 \text{ and } L_2. \end{cases} \tag{4.60}$$

We obtain a priori estimates for Ψ in a standard way by multiplying Eq. (4.59) by Ψ_x, integrating in y and integrating by parts. We find

$$\frac{1}{2}(\overline{U}_0^2 - \frac{N^2}{\lambda^2})\frac{\mathrm{d}}{\mathrm{d}x}\int_0^{L_2}\Psi_x^2(x, y)\mathrm{d}y + \frac{N^2}{2\lambda^2}\frac{\mathrm{d}}{\mathrm{d}x}\int_0^{L_2}\Psi_y^2(x, y)\mathrm{d}y$$

$$= -\int_0^{L_2}[(\frac{N^2}{\lambda}f_1 - \frac{N^2}{\lambda}F_{2y} + \overline{U}_0 f_3)\Psi_x](x, y)\mathrm{d}y.$$

We then integrate in x from 0 to x to obtain, using Eq. (4.60):

$$\begin{aligned}&\frac{1}{2}(\overline{U}_0^2 - \frac{N^2}{\lambda^2})\int_0^{L_2} \Psi_x^2(x, y)\mathrm{d}y + \frac{N^2}{2\lambda^2}\int_0^{L_2} \Psi_y^2(x, y)\mathrm{d}y \\ &= -\int_0^x \int_0^{L2} [(\frac{N^2}{\lambda} f_1 - \frac{N^2}{\lambda} F_{2y} + \overline{U}_0 f_3)\Psi_x](x', y)\mathrm{d}x'\mathrm{d}y.\end{aligned} \tag{4.61}$$

The term involving F_{2y} can be integrated by parts, using Eq. (4.60); we find, all functions being sufficiently regular:

$$\begin{aligned}&\frac{N^2}{\lambda}\int_0^x \int_0^{L_2} (F_{2y}\Psi_x)(x', y)\mathrm{d}x'\mathrm{d}y \\ &\quad = \frac{N^2}{\lambda}\int_0^x \int_0^{L_2} (F_2\Psi_{xy})(x', y)\mathrm{d}x'\mathrm{d}y \\ &\quad = \frac{N^2}{\lambda}\int_0^{L_2} (F_2\Psi_y)(x, y)\mathrm{d}y - \frac{N^2}{\lambda}\int_0^x \int_0^{L_2} (F_{2x}\Psi_y)(x', y)\mathrm{d}x'\mathrm{d}y \\ &\quad = \frac{N^2}{\lambda}\int_0^{L_2} \Psi_y(x, y)\int_0^x f_2(x', y)\mathrm{d}x'\mathrm{d}y - \frac{N^2}{\lambda}\int_0^x \int_0^{L_2} (f_2\Psi_y)(x', y)\mathrm{d}x'\mathrm{d}y.\end{aligned}$$

We insert this expression in Eq. (4.61) and integrate Eq. (4.61) in x from 0 to L_1, which leads to:

$$\begin{aligned}&(\overline{U}_0^2 - \frac{N^2}{\lambda^2})\int_{\mathcal{M}'} \Psi_x^2(x, y)\mathrm{d}x\mathrm{d}y + \frac{N^2}{\lambda^2}\int_{\mathcal{M}'} \Psi_y^2(x, y)\mathrm{d}x\mathrm{d}y \\ &\quad = \frac{2N^2}{\lambda}\int_{\mathcal{M}'} \Psi_y(x, y)(\int_0^x f_2(x', y)\mathrm{d}x')\mathrm{d}y\mathrm{d}x \\ &\qquad - 2\int_{\mathcal{M}'} [\frac{N^2}{\lambda} f_1\Psi_x + \frac{N^2}{\lambda}(\int_0^x f_2(x', y)\mathrm{d}x')\Psi_y + \overline{U}_0 f_3\Psi_x](x, y)\mathrm{d}x\mathrm{d}y.\end{aligned} \tag{4.62}$$

Since $\overline{U}_0 > N/\lambda$, we easily deduce from Eq. (4.62) an estimate

$$\begin{aligned}&\int_{\mathcal{M}'} (\Psi_x^2 + \Psi_y^2)(x, y)\mathrm{d}x\mathrm{d}y \\ &\leq \kappa_1(|f_1|^2_{L^2(\mathcal{M}')} + |f_2|^2_{L^2(\mathcal{M}')} + |f_3|^2_{L^2(\mathcal{M}')}),\end{aligned} \tag{4.63}$$

where κ_1 depends only on the data, namely, $L_1, L_2, \overline{U}_0, N$, and λ. Alternatively, Eq. (4.63) can be written as

$$\int_{\mathcal{M}'} (\psi^2 + \Psi_y^2)(x, y)\mathrm{d}x\mathrm{d}y \leq \kappa_1 |F|^2_{L^2(\mathcal{M}')^3}. \tag{4.64}$$

The calculations above have been made under the assumption that $U \in D(A)$ (and $AU = F$) are sufficiently regular. The lemma below extends Eq. (4.64) to all U in $D(A)$.

LEMMA 4.1. *In the supercritical case (i.e., assuming Eq. (4.47)), Eq. (4.64) is valid for every* $U = (u, v, \psi)$ *in* $D(A)$. *There also exists a constant* κ_2 *depending only on the data such that*

$$|U|_H \leq \kappa_2 |AU|_H, \quad \forall\ U \in D(A). \tag{4.65}$$

PROOF. Given U in $D(A)$, then $AU = F = (f_1, f_2, f_3)$ belongs to $H = L^2(\mathcal{M}')^3$, and it can be approximated in $L^2(\mathcal{M}')^3$ by a sequence of smooth functions $F_m = (f_{1m}, f_{2m}, f_{3m})$ which are $\mathcal{C}^\infty$ with compact support in $\mathcal{M}'$. With these F_m, we solve Eq. (4.59) with boundary and initial conditions in Eq. (4.60) so that we obtain the Ψ_m which satisfy Eq. (4.64).

As $n \to \infty$, the F_m converge to F in $L^2(\mathcal{M}')^3$ and the Ψ_m converge to $\bar{\Psi}$ weakly in $H^1(\mathcal{M}')$, where $\bar{\Psi}$ is the (unique) solution of Eqs. (4.59), (4.60) in $H^1(\mathcal{M}')$. We then define $\bar{\psi} = \partial\bar{\Psi}/\partial x$ which satisfies Eq. (4.33) in the distributional sense, and Eq. (4.64) is satisfied by $\bar{\psi}$, $\bar{\Psi}$ and F. By inspection of Eq. (4.59), we notice that $\bar{\Psi}_{yy}$, and F_{2y} belong to $L^2_x(0, L_1; H_y^{-1}(0, L_2))$ so that

$$\bar{\psi}_x = \bar{\Psi}_{xx} \in L^2_x(0, L_1; H_y^{-1}(0, L_2)),$$

and $\bar{\psi} \in \mathcal{C}_x([0, L_1]; H_y^{-1}(0, L_2))$. Hence $\bar{\psi}(0, \cdot)$ is defined and it vanishes according to Eq. (4.60).

Now, integrating in x the first and second equations of Eq. (4.31) and imposing $\bar{u} = \bar{v} = 0$ at $x = 0$, we define $\bar{u}$ and $\bar{v}$ by setting

$$\begin{aligned} \bar{U}_0\bar{u} - \frac{1}{\lambda}\bar{\psi} &= \int_0^x f_1 \mathrm{d}x', \\ \bar{U}_0\bar{v} - \frac{1}{\lambda}\bar{\psi}_y \mathrm{d}x' &= \int_0^x f_2 \mathrm{d}x'. \end{aligned} \tag{4.66}$$

We want to show that the third equation of Eq. (4.31) is satisfied as well; differentiating the first equation of Eq. (4.66) in x and the second equation (4.66) in y, we find:

$$\bar{U}_0\bar{u}_x - \frac{1}{\lambda}\bar{\psi}_x = f_1, \quad \bar{U}_0\bar{v}_y - \frac{1}{\lambda}\bar{\Psi}_{yy} = F_{2y},$$

and then

$$\begin{aligned} \bar{U}_0\bar{\psi}_x - \frac{N^2}{\lambda}(\bar{u}_x + \bar{v}_y) &= \\ &= \bar{U}_0\bar{\psi}_x - \frac{N^2}{\lambda\bar{U}_0}\left(\frac{1}{\lambda}\bar{\psi}_x + f_1 + \frac{1}{\lambda}\bar{\Psi}_{yy} + F_{2y}\right) \\ &= \left(\bar{U}_0 - \frac{N^2}{\lambda^2\bar{U}_0}\right)\bar{\psi}_x - \frac{N^2}{\lambda^2\bar{U}_0}\bar{\Psi}_{yy} - \frac{N^2}{\lambda\bar{U}_0}(f_1 + F_{2y}) \\ &= \text{(by Eq. (4.59))} \\ &= f_3 \end{aligned}$$

so that all three equations of Eq. (4.31) are satisfied by $\bar{U}$. Furthermore, $\bar{U}$ satisfies the boundary conditions in Eq. (4.60), and we conclude that $\bar{U} \in D(A)$ and $A\bar{U} = F$. Since $AU = F$ as well, we will conclude that $\bar{U} = U$ by showing that A is one-to-one.

To show that A is one-to-one, consider $\widetilde{U} \in D(A)$ such that $A\widetilde{U} = 0$. Then, $\widetilde{\Psi}$ defined by Eq. (4.57) satisfies Eqs. (4.59) and (4.60). At this point we do not know that $\widetilde{\Psi} \in H^1(\mathcal{M}')$, but, at least, we infer from Eq. (4.57) that $\widetilde{\Psi} \in L^2(\mathcal{M}')$ since $\widetilde{\psi} \in L^2(\mathcal{M}')$. We then infer from LIONS and MAGENES [1972] that Eqs. (4.59)–(4.60) have a unique solution in $L^2(\mathcal{M}')$ so that $\widetilde{\Psi} = 0$. From this, we conclude that $\widetilde{\psi} = 0$ and $\widetilde{u}$ and $\widetilde{v}$ also vanish since they satisfy Eq. (4.66) because of the boundary conditions at $x = 0$. Hence, $\widetilde{U} = 0$ and A is one-to-one.

Returning to U, we conclude at this point that ψ and Ψ satisfy Eq. (4.64), which was the first statement in this lemma.

There remains to prove Eq. (4.65); $|\psi|_{L^2(\mathcal{M}')} \leq \kappa |AU|_H$ follows from Eq. (4.64), and the similar results for u and v follow from Eq. (4.66) (and Eq. (4.64)).

The proof of the Lemma is complete. □

We can now prove (4.50).[22]

THEOREM 4.4. *In the supercritical case (i.e., assuming Eq. (4.47)), for every* $U \in D(A_n)$, A_n *defined in Eq. (4.49), we have* $(A_n U, U)_{L^2(\mathcal{M}')^3} \geq 0$. *Similarly, we have* $(A_n^* U, U)_{L^2(\mathcal{M}')^3} \geq 0$, *for every* U *in* $D(A_n^*)$, A_n^* *and* $D(A_n^*)$ *defined in Eq. (4.55).*

PROOF. We prove the result for A, the proof would be similar for A^*.

Considering $U \in D(A)$, we approximate $AU = F$ by a sequence of smooth functions F_m as in Lemma 4.1. To each function F_m, we associate $U_m \in D(A)$ such that $AU_m = F_m$: each U_m is constructed exactly as we constructed $\bar{U}$ in Lemma 4.1, and U_m is smooth. We easily check that as $m \to \infty$, U_m weakly converges in H to U, whereas $AU_m = F_m$ strongly converges in H to $AU = F$. Hence

$$(AU_m, U_m)_H \longrightarrow (AU, U)_H,$$

and since $(AU_m, U_m)_H \geq 0$ by Eq. (4.51), U_m being sufficiently regular, we conclude that $(AU, U)_H \geq 0$. □

REMARK 4.4. As indicated in Remark 4.3 and based on the previous results, we can show for each $n > n_c$ that $-A = -A_n$ is the infinitesimal generator of a contraction semi-group. Then, by application of the Hille–Yoshida theorem, we can solve the initial and boundary value problem associated with Eq. (4.7), for each such n. We refrain from developing this, and we will study all subcritical and supercritical modes at once (together) in the next section.

[22] We recall that A and A^* depend on n as $\lambda = \lambda_n$.

4.4. The initial and boundary value problem for the full system

In this section, we aim to combine the results of the previous sections and to investigate the well-posedness for Eq. (4.3) associated with the suitable initial and boundary conditions. We successively consider the case of homogeneous and nonhomogeneous boundary conditions.

4.4.1. The homogeneous boundary condition case

As explained in Eq. (4.9), the function U and its respective components are decomposed in the form $U = U^0 + U^I + U^{II}$. Accordingly, the basic function space H will be $L^2(\mathcal{M})^3$ or

$$H = H^0 \times \dot{L}^2(\mathcal{M})^3,$$

where H^0 is the same as H_0 in Eq. (4.23), and $\dot{L}^2(\mathcal{M})$ consists of the orthogonal, in $L^2(\mathcal{M})$, of the space of functions independent of z. Like in Section 4.1.5, the elements of H^0 will be the vectors $\mathbf{u}^0 = (u^0, v^0)$. The elements of $\dot{L}^2(\mathcal{M})^3$ will be the triplets $U = (u, v, \psi)$; each of these functions possesses an expansion of the form (Eq. (1.19)) from which we can accordingly identify the functions with the product of their components, and the space $L^2(\mathcal{M})$ with the product of an infinite sequence of spaces $L^2(\mathcal{M}')$. The space H is a subspace of $L^2(\mathcal{M})^3$, just remembering that $\psi_0 = 0$, and its natural scalar product and norms are essentially those of $L^2(\mathcal{M})^3$, more precisely,

$$\begin{aligned}(U, \widetilde{U})_H &= \big((u, v, \psi), (\widetilde{u}, \widetilde{v}, \widetilde{\psi})\big)_{L^2(\mathcal{M})^3} \\ &= (u, \widetilde{u})_{L^2(\mathcal{M})} + (v, \widetilde{v})_{L^2(\mathcal{M})} + \frac{1}{N^2}(\psi, \widetilde{\psi})_{L^2(\mathcal{M})}, \\ |U|_H &= [(U, U)_H]^{1/2}.\end{aligned}$$

Each U can be seen as the sum of its three components

$$U = U^0 + U^I + U^{II}, \tag{4.67}$$

or it can be identified with the infinite sequence of its components $\{U_n\}_{n\geq 0}$, in which case[23]

$$|U|_H^2 = |\mathbf{u}^0|^2_{L^2(\mathcal{M}')^2} + \Sigma_{n=1}^{\infty} |U_n|^2_{L^2(\mathcal{M})^2}.$$

4.4.2. The semigroup

We now introduce the operator A and its domain $D(A)$ in H. We have $D(A) = D(A^0) \times D(A^I) \times D(A^{II})$, where the space $D(A^0)$ is the same as in Eq. (4.26),

$$D(A^0) = \big\{\mathbf{u}^0 \in H^0, \mathbf{u}^0_x \in L^2(\mathcal{M}')^2\big\}. \tag{4.68}$$

[23] Remember that $\psi^0 = 0$ so that $U^0 = \mathbf{u}^0 = (u^0, v^0)$.

Then (compared to Eq. (4.30)),

$$D(A^I) = \{U^I = (U_1, \ldots, U_{n_c}),\ \ U_n \in L^2(\mathcal{M}')^3, \mathcal{A}_n U_n \in L^2(\mathcal{M}')^3, \\ n = 1, \ldots, n_c,\ \ U^I \text{ satisfies, (4.70)}\}, \tag{4.69}$$

$$\begin{cases} \psi^I = 0 \text{ at } x = L_1, \quad \text{and } y = 0, L_2, \\ v^I = 0 \text{ and } u^I + \phi^I/\overline{U}_0 = 0 \text{ at } x = 0. \end{cases} \tag{4.70}$$

Here we introduced for convenience the function $\phi = \psi^0 + \phi^I + \phi^{II} = \{\phi_n\}_{n\geq 0}$, with, according to Eq. (4.5),

$$\phi_n = -\frac{1}{\lambda_n}\psi_n, \quad n \geq 1. \tag{4.71}$$

Finally (compared to Eq. (4.49)),

$$D(A^{II}) = \Big\{U^{II} = \{U_n\}_{n>n_c}, U_n \in L^2(\mathcal{M}')^3, \mathcal{A}_n U_n \\ \in L^2(\mathcal{M}')^3, n = 1, \ldots, n_c,\ \ U^{II} \text{ satisfies Eq. (4.73)}\Big\}, \tag{4.72}$$

$$\begin{cases} u^{II} = v^{II} = \psi^{II} = 0 \quad \text{at } x = 0, \\ \psi^{II} = 0 \quad \text{at } y = 0 \text{ and } L_2. \end{cases} \tag{4.73}$$

For $U = (\mathbf{u}_0, U^I, U^{II})$ in $D(A)$, we set $AU = (A^0\mathbf{u}^0, A^I U^I, A^{II}U^{II})$, where

$$A^0\mathbf{u}^0 = P_{H^0}\left(\overline{U}_0 \frac{\partial \mathbf{u}^0}{\partial x} + f e_z \wedge \mathbf{u}^0\right)$$

as in Eq. (4.25) and we define $A^I U^I$ and $A^{II}U^{II}$ componentwise by setting

$$A_n U_n = \mathcal{A}_n U_n, \quad \text{for } 1 \leq n,$$

$\mathcal{A}_n$ as in Eq. (4.21) with $\lambda = \lambda_n$.

We now need to define the adjoint A^* of A and prove that A and A^* are positive which will follow promptly from the results in the previous sections.

For the adjoint, it is easy to see that

$$D(A^*) = D(A^{0*}) \times D(A^{I*}) \times D(A^{II**}), \tag{4.74}$$

with $D(A^{0*}) = D(A^0)$ as shown in Section 4.1.5, $D(A^{I*})$ defined in Eq. (4.46) and $D(A^{II*})$ defined in Eq. (4.55). Indeed, according to Rudin [1991], $\widetilde{U} \in D(A^*)$ if and only if,

$$U \to (AU, \widetilde{U})_H = (A^0\mathbf{u}^0, \widetilde{\mathbf{u}}^0) + (A^I U^I, \widetilde{U}^I) + (A^{II}U^{II}, \widetilde{U}^{II})$$

is continuous on $D(A)$ for the topology (norm) of H. Considering successively $U = (\mathbf{u}^0, 0, 0)$, $U = (0, U^I, 0)$, and $U = (0, 0, U^{II})$, we obtain that $D(A)$ is included in the space in the right-hand side of Eq. (4.74). Conversely, any $\widetilde{U}$ in the right-hand side of Eq. (4.74) belongs to $D(A)$ and hence Eq. (4.74) is proven.

We can prove the following

THEOREM 4.5. *The operator $-A$ is infinitesimal generator of a semigroup of contractions in H.*

PROOF. According to YOSIDA [1980] and HILLE and PHILLIPS [1974], it suffices to show that

i) A and A^* are closed operators, and their domains $D(A)$ and $D(A^*)$ are dense in H.
ii) A and A^* are positive:

$$\begin{aligned} &(AU, U)_H \geq 0, \quad \forall U \in D(A), \\ &(A^*U, U)_H \geq 0, \quad \forall U \in D(A^*). \end{aligned} \tag{4.75}$$

For i) we observe, as is well-known, that $D(A^*)$ (resp. $D(A)$) dense in H implies that A (resp. A^*) is closed. We proceed componentwise for, say, $D(A) : D(A^0)$ defined in Eq. (4.68) is dense in H^0 since the $\mathcal{C}^\infty$ functions $\mathbf{u}^0 = (u^0, v^0)$ with compact support in $\mathcal{M}'$ and such that div $\mathbf{u}^0 = u^0_x + v^0_y = 0$ are dense in H^0; see, e.g., TEMAM [2001]; and for $D(A^I)$ and $D(A^{II})$, we simply observe that the $\mathcal{C}^\infty$ functions with compact support in $\mathcal{M}'$ are dense in $L^2(\mathcal{M}')$.

Finally, for Eq. (4.75), we proceed componentwise and use the results of the previous sections, e.g., for A:

$$(AU, U)_H = (A^0\mathbf{u}^0, \mathbf{u}^0)_{H^0} + (A^I U^I, U^I)_{H^I} + (A^{II}U^{II}, U^{II})_{H^{II}}. \tag{4.76}$$

The first term in the right-hand side of Eq. (4.76) has been shown to be positive (= 0 in fact, see Eq. (4.28)). The second term is equal to

$$\sum_{n=1}^{nc} (A_n U_n, U_n)_{L^2(\mathcal{M}')^3},$$

and each of these terms is positive as shown in Eq. (4.39). Finally, the third term

$$(A^{II}U^{II}, U^{II})_{H^{II}} = \sum_{n>n_c} (A_n U_n, U_n)_{L^2(\Gamma_i)^3},$$

and each term of the series is positive according to Eq. (4.50). □

The initial and boundary value problems

We now consider the whole system of three-dimensional linearized PEs, namely Eq. (4.3), and introduce the initial and boundary conditions. We start with the homogeneous boundary conditions and treat subsequently the case of nonhomogeneous boundary conditions.

As implied by the previous sections, the boundary conditions will be different for the subcritical and supercritical components of $U = (u, v, \psi) = (U^0, U^I, U^{II})$. Hence, for $U^0 = \mathbf{u}^0(\psi^0 = 0)$, we set (see Eq. (4.26)):

$$\mathbf{u}^0 \cdot \mathbf{n} = 0 \text{ on } \partial\mathcal{M}'. \tag{4.77}$$

For U^I, according to Eq. (4.29), the boundary conditions read

$$\begin{cases} \psi^I = 0 \text{ at } x = L_1, \quad \text{and } y = 0, L_2, \\ v^I = 0 \text{ and } u_n = \psi_n/\lambda_n \overline{U}_0 \text{ at } x = 0, n = 1, \ldots, n_c. \end{cases} \tag{4.78}$$

For U^{II}, the boundary conditions are inferred from Eq. (4.49) and read

$$\begin{cases} u^{II} = v^{II} = \psi^{II} = 0 \text{ at } x = 0, \\ \text{and } \psi^{II} = 0 \text{ at } y = 0 \text{ and } L_2. \end{cases} \tag{4.79}$$

All these boundary conditions are taken into account in the domain $D(A)$ of A. Finally, if we add the initial conditions

$$U(0) = (u(0), v(0), \psi(0)) = U_0 = (u_0, v_0, \psi_0), \tag{4.80}$$

then the initial and boundary value problem consisting of Eqs. (4.3) and (4.77)–(4.80) is equivalent to the abstract initial value problem

$$\frac{\mathrm{d}U}{\mathrm{d}t} + AU = F, \tag{4.81}$$

$$U(0) = U_0. \tag{4.82}$$

Note that $F = (F_u, F_v, F_w)$ which does not appear in Eq. (4.3) is added here for mathematical generality and to study below the case of nonhomogeneous boundary conditions. By Theorem 4.5 this problem is now solved by the Hille–Yoshida theorem, and we have

THEOREM 4.6. *Let H, A, and $D(A)$ be defined as in Eqs. (4.67)–(4.73). Then, the initial value problem in Eqs. (4.81)–(4.82) is well-posed. That is, for every $U_0 \in D(A)$ and $F \in L^1(0, T, H)$, with $F' = \mathrm{d}F/\mathrm{d}t$ in $L^1(0, T; H)$, Eqs. (4.81)–(4.82) have a unique solution U such that*

$$U \in \mathcal{C}([0, T]; H) \cap L^\infty(0, T; D(A)), \quad \frac{\mathrm{d}U}{\mathrm{d}t} \in L^\infty(0, T; H). \tag{4.83}$$

4.4.3. The nonhomogeneous boundary conditions

We now turn to the case of nonhomogeneous boundary conditions for Eqs. (4.77)–(4.79), i.e., we want to solve Eq. (4.3) with Eqs. (4.77)–(4.79) in which the boundary conditions

are now nonhomogeneous, and with initial condition in Eq. (4.80). We assume that all boundary data are inferred from a function $U^g = (\mathbf{u}^{g0}, U^{gI}, U^{gII})$ which is defined in $\mathcal{M} \times [0, T]$. We also assume that U^g is given by its normal modes expansion:

$$U^g(x, y, z, t) = \left(\sum_{n\geq 0} u_n^g(x, y, t)\, \mathcal{U}_n(z), \sum_{n\geq 0} v_n^g(x, y, t)\, \mathcal{U}_n(z), \right.$$
$$\left. \sum_{n\geq 1} \psi_n^g(x, y, t)\, \mathcal{W}_n(z) \right). \tag{4.84}$$

We now set

$$U = U^{\#} + U^g,$$

and observe that $U^{\#} \in D(A)$ if $U^{\#}$ is smooth enough (homogeneous boundary conditions). Then, $U^{\#}$ will be sought as the solution of the linear evolution equation

$$\begin{aligned} &\frac{\mathrm{d}U^{\#}}{\mathrm{d}t} + AU^{\#} = F^{\#}, \\ &U^{\#}(0) = U_0^{\#}, \end{aligned} \tag{4.85}$$

where

$$U_0^{\#} = U_0 - U^g|_{t=0} \tag{4.86}$$

and

$$F^{\#} = F - \frac{\partial U^g}{\partial t} - \mathcal{A}U^g. \tag{4.87}$$

Here $\mathcal{A}U^g$ is defined by its normal mode expansion, where each $(\mathcal{A}U^g)_n$ is equal to $\mathcal{A}_n U_n^g$, $\mathcal{A}_n$ as in Eq. (4.21).

Theorem 4.6 will be applicable to Eq. (4.85), and we will obtain the desired existence and uniqueness result for U, provided we assume that $U_0^{\#}$ and $F^{\#}$ satisfy the hypotheses of Theorem 4.6. It is very easy to give sufficient (non necessarily optimal) conditions on U^g which guarantee that $U_0^{\#} \in D(A)$ and $F^{\#}$ and $\mathrm{d}F^{\#}/\mathrm{d}t$ are in $L^1(0, T; \mathbf{L}^2(\mathcal{M})^3)$. We assume, e.g., the following:

$$\begin{aligned} &U_0, \frac{\partial U_0}{\partial x}, \frac{\partial U_0}{\partial y} \in L^2(\mathcal{M})^3, \quad \text{and div } \mathbf{u}_0^0 = 0, \\ &F, \frac{\partial F}{\partial t} \in L^1(0, T; L^2(\mathcal{M})^3), \\ &U^g, \frac{\partial U^g}{\partial t}, \frac{\partial U^g}{\partial x}, \frac{\partial U^g}{\partial y}, \frac{\partial^2 U^g}{\partial t^2}, \frac{\partial^2 U^g}{\partial x \partial t}, \frac{\partial^2 U^g}{\partial y \partial t} \in \mathcal{C}([0, T]; L^2(\mathcal{M})^3). \end{aligned} \tag{4.88}$$

In addition, we require that U_0 and U^g satisfy certain compatibility conditions, for $t = 0$, and $(x, y) \in \partial\mathcal{M}'$, conditions which guarantee that $U_0^{\#} \in D(A)$. Setting

$U_0 = (\widetilde{u}_0, \widetilde{v}_0, \widetilde{\psi}_0) = (\widetilde{U}_0^0, \widetilde{U}_0^I, \widetilde{U}_0^{II})$,[24] we require

$$\begin{aligned}
\widetilde{\mathbf{u}}^0 \cdot \mathbf{n} &= \mathbf{u}^{0g} \cdot \mathbf{n}, \quad \text{on } \partial\mathcal{M}', \text{ at } t = 0, \\
\widetilde{\psi}_0^I &= \widetilde{\psi}^{gI} \text{ at } t = 0 \text{ and } x = L_1, \text{ or } y = 0 \text{ or } L_2, \\
\widetilde{v}_0^I &= \widetilde{v}^{gI} \text{ and } \widetilde{u}_{on} = \widetilde{\psi}_{on}/\lambda_n \overline{U}_0 = \widetilde{u}_n^g - \widetilde{\psi}_n^g/\lambda_n \overline{U}_0 \\
&\qquad \text{at } x = 0 \text{ and } t = 0, n = 1, \ldots, n_c, \\
\widetilde{u}_0^{II} - \widetilde{u}^{gII} &= \widetilde{v}_0^{II} - \widetilde{v}^{gII} = \widetilde{\psi}_0^{II} - \widetilde{\psi}^{gII}, \text{ at } x = 0 \text{ and } t = 0, \\
\widetilde{\psi}_0^{II} - \widetilde{\psi}^{gII} &= 0, \text{ at } t = 0 \text{ and } y = 0 \text{ or } L_2.
\end{aligned} \tag{4.89}$$

With the regularity hypotheses (Eq. (4.88)) and the compatibility hypotheses (Eq. (4.89)), we obtain U satisfying

$$\begin{aligned}
&U \in \mathcal{C}([0, T]; L^2(\mathcal{M})^3), \\
&\mathcal{A}U \in L^\infty(0, T; L^2(\mathcal{M})^3), \\
&\frac{\partial U}{\partial t} \in L^\infty(0, T; L^2(\mathcal{M})^3),
\end{aligned} \tag{4.90}$$

and the boundary conditions for $0 < t < T$:

$$\begin{aligned}
\mathbf{u}^0 \cdot \mathbf{n} &= \mathbf{u}^g \cdot \mathbf{n} \text{ on } \partial\mathcal{M}', -L_3 < z < 0, \\
\psi^I &= \psi^{gI} \text{ at } x = L_1 \text{ and } y = 0, L_2, \\
v^I &= v^{gI} \text{ at } x = 0, \\
u_n^I - \psi_n^I/\lambda_n \overline{U}_0 &= u_n^{gI} - \psi_n^{gI}/\lambda_n \overline{U}_0, \text{ at } x = 0, \quad n = 1, \ldots, n_c, \\
u^{II} &= u^{gII}, v^{II} = v^{gII}, \psi^{II} = \psi^{gII} \text{ at } x = 0, \\
\psi^{II} &= \psi^{gII} \text{ at } y = 0 \text{ and } L_2.
\end{aligned} \tag{4.91}$$

In summary, we have proven the following theorem:

THEOREM 4.7. *We assume that U_0, F, and U^g are given satisfying the hypotheses (Eqs. (4.88) and (4.89)). Then there exists a unique U solution of the PEs (Eq. (4.3)), satisfying the regularity properties (Eq. (4.90)), the boundary condition (Eq. (4.91)) and the initial condition (Eq. (4.82)).*

5. Conclusion

In this chapter we have analyzed the inviscid linearized primitive equations considering successively the dimensions 2, 2.5, and 3. In accordance with the previously known result

[24] The tildes here on u_0, v_0, ψ_0, etc. are intended to distinguish these initial data from the zero modes of $U(t)$.

that these equations cannot be well-posed for any set of local boundary conditions, we have proposed nonlocal boundary conditions and established their suitability.

In space dimension 2, numerical simulations have been performed for both the linearized and nonlinear nonviscous primitive equations. The suitability of the boundary conditions that we have proposed is numerically confirmed in the linear case. In the nonlinear case, the same boundary conditions have been used and have shown to be also numerically suitable. Furthermore, the nonoccurence of numerical blow-up is an indication that these boundary conditions are appropriate in the nonlinear case as well.

Future work in this domain will consist on the theoretical side in considering more complicated background flows in the linear case and the full nonlinear equations. On the computational side, the three-dimensional case and more involved equations (richer physics) should be investigated.

6. Acknowledgments

This work was partially supported by the National Science Foundation under the grant NSF-DMS-0604235, and by the Research Fund of Indiana University.

References

BENNETT, A.F., CHUA, B.S. (1999). Open boundary conditions for Lagrangian geophysical fluid dynamics. *J. Comput. Phys.* **153** (2), 418–436.

BENNETT, A.F., KLOEDEN, P.E. (1978). Boundary conditions for limited-area forecasts. *J. Atmos. Sci.* **35** (6), 990–996.

BERNARDI, C., MADAY, Y. (1997). Spectral methods. In: *Handbook of Numerical Analysis, vol. V* (North-Holland, Amsterdam, The Netherlands), pp. 209–485.

BLAYO, E., DEBREU, L. (2005). Revisiting open boundary conditions from the point of view of characteristic variables. *Ocean Model.* **9**, 231–252.

BRÉZIS, H. (1970). On some degenerate nonlinear parabolic equations. In: *Nonlinear Functional Analysis (Proc. Sympos. Pure Math., Vol. XVIII, Part 1, Chicago, Ill., 1968)* (American Mathematical Society, Providence, RI), pp. 28–38.

BRÉZIS, H. (1973). *Opérateurs Maximaux Monotones Et Semi-Groupes De Contractions Dans Les Espaces De Hilbert* (North Holland Publishing Co, Amsterdam, The Netherlands).

BURQ, N., GÉRARD, P. (2003). *Contrôle Optimal Des Équations Aux Dérivées Partielles* (Ecole Polytechnique, Palaiseau, France).

CAO, C., TITI, E. (2007). Global well-posedness of the threedimensional viscous primitive equations of large scale ocean and atmosphere dynamics. *Ann. Math.* **165** (2), To appear.

CHARNEY, J.G., FJÖRTOFT, R., VON NEUMANN, J. (1950). Numerical integration of the barotropic vorticity equation. *Tellus* **2**, 237–254.

DUMAS, F., LAZURE, P. (2007). An external-internal mode coupling for a 3d hydrodynamical model for applications at regional scale (mars). *Adv. Water Resour.*

GOTTLIEB, D., HESTHAVEN, J.S. (2001). Spectral methods for hyperbolic problems. *J. Comput. Appl. Math.* **128** (1–2), 83–131.

GRISVARD, P. (1985). *Elliptic Problems in Nonsmooth Domains* (Monographs and Studies in Mathematics. Pitman, Boston, MA).

HENRY, D. (1981). *Geometric Theory of Semilinear Parabolic Equations, vol. 840* (Lecture Notes in Mathematics. Springer Verlag, Berlin, Germany).

HILLE, E., PHILLIPS, R. (1974). *Functional Analysis and Semi-Groups* Third printing of the revised edition of 1957, American Mathematical Society Colloquium Publications, vol. XXXI. (American Mathematical Society, Providence, RI).

JOHN, F. (1982). *Partial Differential Equations*, fourth ed. (Springer Verlag, New York).

KOBELKOV, G. (2006). Existence of a solution 'in the large' for the 3D large-scale ocean dynamics equations. *C. R. Math. Acad. Sci. Paris* **343** (4), 283–286.

KUKAVICA, I., ZIANE, M. (2007). On the regularity of the primitive equations of the ocean. To appear.

LIONS, J. (1965). *Problèmes aux limites dans les équations aux dérivées partielles*. Les Presses de l'Université de Montréal, Montreal, Que. Reedited in Lions (2003).

LIONS, J. (2003). *Selected Work, vol. 1* (EDS Sciences, Paris, France).

LIONS, J., TEMAM, R., WANG, S. (1992a). New formulations of the primitive equations of atmosphere and applications. *Nonlinearity* **5** (2), 237–288.

LIONS, J., TEMAM, R., WANG, S. (1992b). On the equations of the large-scale ocean. *Nonlinearity* **5** (5), 1007–1053.

LIONS, J.-L., MAGENES, E. (1972). *Non-Homogeneous Boundary Value Problems and Applications, vols. I and II* (Springer-Verlag, New-York-Heidelberg), Translated from the French by P. Kenneth.

MADEC, G., DELECLUSE, P., IMBARD, M., LÉVY, C. (1998). *OPA 8.1. Ocean General Circulation Model Reference Manual* (Institut Pierre Simon Laplace).

OLIGER, J., SUNDSTRÖM, A. (1978). Theoretical and practical aspects of some initial boundary value problems in fluid dynamics. *SIAM J. Appl. Math.* **35** (3), 419–446.

PAZY, A. (1983). Semigroups of operators in Banach spaces. In: *Equadiff 82 (Würzburg, 1982)*. In: Lecture Notes in Mathematics, **1017** (Springer, Berlin, Germany), pp. 508–524.

PEDLOSKY, J. (1987). *Geophysical Fluid Dynamics*, second ed. (Springer).

PETCU, M., TEMAM, R., ZIANE, M. (2008). Mathematical problems for the primitive equations with viscosity. In: Ciarlet, P.G. (eds.), *Handbook of Numerical Analysis. Special Issue on Some Mathematical Problems in Geophysical Fluid Dynamics* (Elsevier, New York).

ROUSSEAU, A., TEMAM, R., TRIBBIA, J. (2004). Boundary layers in an ocean related system. *J. Sci. Comput.* **21** (3), 405–432.

ROUSSEAU, A., TEMAM, R., TRIBBIA, J. (2005a). Boundary conditions for an ocean related system with a small parameter. In: Chen, G-Q., Gasper, G., Jerome, J.J. (eds.), *Nonlinear PDEs and Related Analysis*. Contemporary Mathematics, vol. 371 (AMS, Providence, RI), pp. 231–263.

ROUSSEAU, A., TEMAM, R., TRIBBIA, J. (2005b). Boundary conditions for the 2D linearized PEs of the ocean in the absence of viscosity. *Discrete Contin. Dyn. Syst.* **13** (5), 1257–1276.

ROUSSEAU, A., TEMAM, R., TRIBBIA, J. (2007). Numerical simulations of the inviscid primitive equations in a limited domain. In: Calgaro, C., Coulombel, J-F., Goudon, T. (eds.), *Analysis and Simulation of Fluid Dynamics*, Advances in Mathematical Fluid Mechanics.

RUDIN, W. (1991). *Functional Analysis. International Series in Pure and Applied Mathematics*, second ed. (McGraw-Hill Inc., New York).

SALMON, R. (1998). *Lectures on Geophysical Fluid Dynamics* (Oxford University Press, New York).

TEMAM, R. (2001). *Navier-Stokes Equations: Theory and Numerical Analysis* (AMS Chelsea Publishing, Providence, RI) Reprint of the 1984 edition.

TEMAM, R., TRIBBIA, J. (2003). Open boundary conditions for the primitive and Boussinesq equations. *J. Atmos. Sci.* **60** (21), 2647–2660.

TEMAM, R., ZIANE, M. (2004). Some mathematical problems in geophysical fluid dynamics. In: Friedlander, S., Serre, D. (eds.), *Handbook of Mathematical Fluid Dynamics* (North Holland).

WARNER, T., PETERSON, R., TREADON, R. (1997). A tutorial on lateral boundary conditions as a basic and potentially serious limitation to regional numerical weather prediction. *Bull. Am. Meteorol. Soc.* **78** (11), 2599–2617.

WASHINGTON, W., PARKINSON, C. (1986). *An Introduction to Threedimensional Climate Modelling*. (Oxford Univ. Press).

YOSIDA, K. (1980). *Functional Analysis*, sixth ed. (Springer-Verlag, Berlin, Germany).

Some Mathematical Problems in Geophysical Fluid Dynamics

Madalina Petcu

Laboratoire de Mathématiques et Applications, UMR 6086, Universîté de Poitiers, France

Roger M. Temam

The Institute for Scientific Computing and Applied Mathematics, Indiana University, Bloomington, IN 47405, USA

Mohammed Ziane

University of Southern California, Mathematics Department, 1042 W. 36 Place, Los Angeles, CA 90049, USA

Abstract

This chapter reviews the recently developed mathematical setting of the primitive equations (PEs) of the atmosphere, the ocean, and the coupled atmosphere and ocean. The mathematical issues that are considered here are the existence, uniqueness, and regularity of solutions for the time-dependent problems in space dimensions 2 and 3, the PEs being supplemented by a variety of natural boundary conditions. The emphasis is on the case of the ocean that encompasses most of the mathematical difficulties. This chapter is devoted to the PEs in the presence of viscosity, while the PEs without viscosity are considered in the chapter by Rousseau, Temam, and Tribbia in the same volume.

Whereas the theory of PEs without viscosity is just starting, the theory of PEs with viscosity has developed since the early 1990s and has now reached a satisfactory level of completion. The theory of the PEs was initially developed by analogy with that of the incompressible Navier Stokes equations, but the most recent developments reported in this chapter have shown that unlike the incompressible Navier-Stokes equations and the celebrated Millenium Clay problem, the PEs with viscosity are well-posed in space dimensions 2 and 3, when supplemented with fairly

Computational Methods for the Atmosphere and the Oceans
Special Volume (Roger M. Temam and Joseph J. Tribbia, Guest Editors) of
HANDBOOK OF NUMERICAL ANALYSIS, VOL. XIV
P.G. Ciarlet (Editor)

Copyright © 2009 Elsevier B.V.
All rights reserved

ISSN 1570-8659
DOI 10.1016/S1570-8659(08)00212-3

general boundary conditions. This chapter is essentially self-contained, and all the mathematical issues related to these problems are developed.
A guide and summary of results for the physics-oriented reader is provided at the end of the Introduction (Section 1.4).

1. Introduction

The aim of this chapter is to address some mathematical aspects of the equations of geophysical fluid dynamics, namely existence, uniqueness, and regularity of solutions.

The equations of geophysical fluid dynamics are the equations governing the motion of the atmosphere and the ocean, and they are derived from the conservation equations from physics, namely conservation of mass, momentum, energy, and some other components such as salt for the ocean and humidity (or chemical pollutants) for the atmosphere. The basic equations of conservation of mass and momentum, i.e., the three-dimensional (3D) compressible Navier-Stokes equations, contain however too much information, and we cannot hope to numerically solve these equations with enough accuracy in a foreseeable future. Owing to the difference in sizes of the vertical and horizontal dimensions, both in the atmosphere and in the ocean (10–20 km versus several thousands of kilometers), the most natural simplification leads to the so-called primitive equations (PEs), which we study in this chapter.

We continue this introduction by briefly describing the physical and mathematical backgrounds of the PEs.

1.1. Physical background

The PEs are based on the so-called hydrostatic approximation, in which the conservation of momentum in the vertical direction is replaced by the simpler, hydrostatic equation (see Eq. (2.25)).

As far as we know, the PEs were essentially introduced by Richardson in 1922; when it appeared that they were still too complicated, they were abandoned and, instead, attention was focused on simpler models, such as the barotropic and the geostrophic and quasi-geostrophic models, considered in the late 1940s by von Neumann and his collaborators, in particular Charney. With the increase of computing power, interest eventually returned to the PEs, which are now the core of many Global Circulation Models (GCMs) or Ocean Global Circulation Models (OGCMs) available at the National Center for Atmospheric Research and elsewhere. GCMs and OGCMs are very complex models that contain many physical components (for the atmosphere, the chemistry (equations of concentration of pollutants), the physics of the cloud (radiation of solar energy, concentration of vapor), the vegetation, the topography, the albedo, for the oceans, phenomena such as the sea ice or the topography of the bottom of the oceans). Nevertheless, the PEs that describe the dynamics of the air or the water and the balance of energy are the central components for the dynamics of the air or the water. For some phenomena, there is need to give up the hydrostatic hypothesis and then nonhydrostatic models are considered, such as

TABLE 1.1
Level of physical complexity (richness)

Three-dimensional Navier-Stokes equations
⇓
Nonhydrostatic models
⇓
Primitive equations
(hydrostatic equations)
⇓
Shallow water equations
⇓
Quasi-geostrophic models
⇓
Two-dimensional barotropic equations

in LAPRISE [1992] or SMOLARKIEWICZ, MARGOLIN and WYSZOGRODZKI [2001]; these models stand at an intermediate level of physical complexity between the full Navier-Stokes equations and the PEs hydrostatic equations. Research on nonhydrostatic models is ongoing and, at this time, there is no agreement, in the physical community, for a specific model.

In this hierarchy of models for geophysical fluid dynamics, let us add also the shallow water equation corresponding essentially to a vertically integrated form of the Navier-Stokes equations; from the physical point of view, they stand as an intermediate model between the primitive and the quasi-geostrophic equations.

In summary, in terms of physical relevance and the level of complexity of the physical phenomena they can account for, the hierarchy of models in geophysical fluid dynamics is as in Table 1.1.

We remark here also that much study is needed for the boundary conditions from both the physical and the mathematical point of views. As we said, our aim in this chapter is the study of mathematical properties of the PEs.

In the above, and in all of this chapter, the PEs that we consider are the PEs with viscosity; the PEs without viscosity are studied in the chapter by ROUSSEAU, TEMAM, and TRIBBIA [2008] in this volume. The PEs without viscosity raise questions of a totally different nature. In particular, whereas the PEs with viscosity bear some similarity with the incompressible Navier-Stokes equations as we explain below, the PEs without viscosity are different from the Euler equations of incompressible inviscid flows in many respects (see the already quoted chapter of Rousseau, Temam, and Tribbia).

1.2. Mathematical background

The level of mathematical complexity of the equations in Table 1.1 is not the same as the level of physical complexity: at both ends, the quasi-geostrophic models and barotropic equations are mathematically well understood (at least in the presence of viscosity; see WANG [1992a,b], and despite its well-known limitations, the mathematical theory of the incompressible Navier-Stokes equations is also relatively well understood. On the other hand, nonhydrostatic models are mathematically out of reach, and there are much less

mathematical results available for the shallow water equations than for the Navier-Stokes equations, even in space dimension 2 (see, however Orenga [1995]).

The mathematical theory of the (viscous) PEs has developed in two stages. The first stage ranging from the article of Lions, Temam and Wang [1993a,b] to the review article by Temam and Ziane [2004] concentrated on the analogy of the PEs with the 3D incompressible Navier-Stokes equations. Indeed, and as we show below, the PEs although physically "poorer" than the Navier-Stokes equations, in some sense, they structurally more complicated than the incompressible Navier-Stokes equations.

Indeed, this is due to the fact that the nonlinear term in the Navier-Stokes equations, also called inertial term, is of the form

$$\text{velocity} \times \text{first-order derivatives of velocity},$$

whereas the nonlinear term for the PEs is of the form

$$\text{first-order derivatives of horizontal velocity} \times \text{first-order derivatives of horizontal velocity}.$$

The mathematical study of the PEs was initiated by Lions, Temam and Wang [1992a,b]. They produced a mathematical formulation of the PEs that resembles that of the Navier-Stokes due to Leray and obtained the existence for all time of weak solutions (see Section 2 and the original articles by Lions, Temam and Wang [1992a,b, 1995] in the list of references). Further works, conducted during the 1990s and especially during the past few years, have improved and supplemented the early results of these authors by a set of results that, essentially, brings the mathematical theory of the PEs to that of the 3D incompressible Navier-Stokes equations, despite the added complexity mentioned above; this added complexity is overcome by a nonisotropic treatment of the equations (of certain nonlinear terms), in which the horizontal and vertical directions are treated differently.

In summary, the following results have been obtained, which were presented in the review article by Temam and Ziane [2004] and appear herein in Sections 2 and 3:

(i) Existence of weak solutions for all time (dimensions 2 and 3) (see Sections 2).
(ii) In space dimension 3, existence of a strong solution for a limited time (local in time existence) (see Section 3.1).
(iii) In space dimension 2, existence and uniqueness for all time of a strong solution (see Section 3.3).
(iv) Uniqueness of a weak solution in space dimension 2 (see Section 3.4).

In the above, the terminology that is normally used for Navier-Stokes equations: the weak solutions are those with finite (fluid) kinematic energy ($L^\infty(L^2)$ and $L^2(H^1)$), and the strong solutions are those with finite (fluid) enstrophy ($L^\infty(H^1)$ and $L^2(H^2)$). Essential in the most recent developments (ii)–(iv) above is the H^2-regularity result for a Stokes-type problem appearing in the PEs, the analog of the H^2-regularity in the Cattabriga–Solonnikov results on the usual Stokes problem; the whole Section 4 is devoted to this problem.

The second stage of the mathematical theory of the (viscous) PEs is more recent. It is based on the observation that the pressure-like function (the surface pressure) is in fact a two-dimensional (2D) function (a function of the horizontal variables and time) and because of that the 3D PEs are also close to a 2D system. Technically, by suitable estimates of the surface pressure, the difficulties related to the pressure are overcome. This approach was developed in the two independent articles (with different proofs) by CAO and TITI [2007] and by KOBELKOV [2006], for the case of an ocean with a flat bottom and Neuman boundary condition. The case of a varying-bottom topography the Dirichlet boundary condition studied in the subsequent article of KUKAVICA and ZIANE [2007a,b]. These three articles combine the above mentioned results of local existence of a strong solution and the new a priori estimates to show that the strong solution is defined for all time. These newest results appear in Section 3.2.

1.3. Content of this chapter

Because of space limitation, it was not possible to consider all relevant cases here. Relevant cases include,

the ocean, the atmosphere, and the coupled ocean and atmosphere,

on the one hand, and, on the other hand, the study of global phenomena on the sphere (involving the writing of the equations in spherical coordinates), and the study of mid-latitude regional models in which the equations are projected on a space tangent to the sphere (the Earth), corresponding to the so-called β-plane approximation: here, $0x$ is the west–east axis, $0y$ is the south–north axis, and $0z$ is the ascending vertical.

In this chapter, we have chosen to concentrate on the cases mathematically most significant. Hence for each case, after a brief description of the equations on the sphere (in spherical coordinates), we concentrate our efforts on the corresponding β-plane Cartesian coordinates). Indeed, in general, going from the β-plane Cartesian coordinates to the spherical case necessitates only the proper handling of terms involving lower order derivatives; full details concerning the spherical case can be found also in the original articles by LIONS, TEMAM and WANG [1992a,b, 1995]).

In the Cartesian case of emphasis, generally, we first concentrate our attention on the ocean. Indeed, as we will see in Section 2, the domain occupied by the ocean contains corners (in dimension 2) or wedges (in dimension 3); some regularity issues occur in this case, which must be handled using the theory of regularity of elliptic problems in nonsmooth domains (GRISVARD [1985], KOZLOV, MAZYA and ROSSMANN [1997], MAZYA and ROSSMANN [1994]). For the atmosphere or the coupled atmosphere–ocean (CAO), the difficulties are similar or easier to handle—hence, most of the mathematical efforts will be devoted to the ocean in Cartesian coordinates.

In Section 2, we describe the governing equations and derive the result of existence of weak solutions with a method different from that of the original articles by LIONS, TEMAM and WANG [1992a,b, 1995], thus allowing more generality (for the ocean, the atmosphere, and the CAO).

In Section 3, we study the existence of strong solutions in space dimensions 3 and 2 and a wealth of other mathematical results, regularity in H^m–higher Sobolev spaces,

$\mathcal{C}^\infty$–regularity, Gevrey regularity, and *backward uniqueness*. We establish in dimension 3 the existence and uniqueness of strong solutions on a limited interval of time (Section 3.1) and then for all time (see Section 3.2). In dimension 2, we prove the existence and uniqueness, for all time, of such strong solutions (see Section 3.3). Section 3.4 contains a technical result. In Section 3.5, we consider the 2D space-periodic case and prove the existence of solutions for all time, in all H^m, $m \geq 2$. In Section 3.6, we prove the Gevrey regularity of the solutions and in Section 3.7.2 the backward uniqueness result.

Section 4 is technically very important, and many results of Sections 2 and 3 rely on it: this section contains the proof of the H^2-regularity of elliptic problems, which arise in the PEs. This proof relies, as we said, on the theory of regularity of solutions of elliptic problems in nonsmooth domains. It is shown there that the solutions to certain elliptic problems enjoy certain regularity properties (H^2-regularity, i.e., the function and their first and second derivatives are square integrable); the problems corresponding to the (horizontal) velocity, the temperature, and the salinity are successively considered. The study in Section 4 contains many specific aspects that are explained in detail in the introduction to that section.

More explanations and references will be given in the introduction of or within each section.

As mentioned earlier, the mathematical formulation of the equations of the atmosphere, ocean, and CAO was derived by Lions, Temam and Wang [1992a,b, 1995]. For each of these problems, these articles also contain the proof of existence of weak solutions for all time (in dimension 3 with a proof that easily extends to dimension 2). An alternative slightly more general proof of this result is given in Section 2. Concerning the strong solutions, the proof given here of the local existence in dimension 3 is based on the article by Hu, Temam and Ziane [2003]. An alternate proof of this result is due to Guillén-González, Masmoudi and Rodríguez-Bellido [2001]. In dimension 2, the proof of existence and uniqueness of strong solutions, for all time, for the considered system of equations and boundary conditions is new and based on an unpublished manuscript by Ziane [2000]. This result is also established, for a simpler system (without temperature and salinity), by Bresch, Kazhikhov and Lemoine [2004]. Most of the results of Sections 3.4–3.7.2 are due to M. Petcu, alone or in collaboration with D. Wirosoetisno.

1.4. Summary of results for the physics-oriented reader

The physics-oriented reader will recognize in Eqs. (2.1)–(2.5) the basic conservation laws: conservations of momentum, mass, energy and salt for the ocean, equation of state. In Eqs. (2.6) and (2.7) appears the simplification due to the Boussinesq approximation, and in Eqs. (2.11)–(2.16) the simplifications resulting from the hydrostatic balance assumption. Hence Eqs. (2.11)–(2.16) are the PEs of the ocean. The PEs of the atmosphere appear in Eqs. (2.116)–(2.121), and those of the CAO are described in Section 2.5. Concerning, to begin, the ocean, the first task is to write these equations,

supplemented by the initial and boundary conditions, as an initial value problem in a phase space H of the form

$$\frac{\mathrm{d}U}{\mathrm{d}t} + AU + B(U, U) + E(U) = \ell, \tag{1.1}$$

$$U(0) = U_0, \tag{1.2}$$

where U is the set of prognostic variables of the problem, i.e., the horizontal velocity $\mathbf{v} = (u, v)$, the temperature T, and the salinity S, $U = (\mathbf{v}, T, S)$; (see Eq. (2.66)). The phase space H consists, for its fluid mechanics part, of (horizontal) vector fields with finite kinetic energy. We then study the stationary solutions of Eq. (1.1) in Section 2.2.2, and in Theorem 2.2; we prove the existence for all times of weak solutions of Eqs. (1.1) and (1.2), which are solutions in $L^\infty(0, t_1; L^2)$ and $L^2(0, t_1; H^1)$ (bounded kinetic energy and square integrable enstrophy for the fluid mechanics part). A parallel study is conducted for the atmosphere and the CAO in Sections 2.4 and 2.5. Section 4 is mathematically very important although technical.

For the physics-oriented reader, the most important results are those of Sections 2 and 3. Section 2 contains the "weak" formulation of the PEs and shows the extensive use of the balance of energy principles to prove them. The tools of balance of energy are also those needed for the study of stability of numerical results, and they are therefore both physically and computationally revelant.

The main results of Section 3 are the existence and uniqueness of strong solutions for all time, now both available in space dimensions 2 and 3. Noteworthy also in this section are the results concerning the Gevrey regularity of the solutions, which implies in particular an exponential decay of the Fourier coefficients, results that have been used in the recent articles by Temam and Wirosoetisno [2007, 2008] to prove that the PEs can be approximated by a finite-dimensional model up to an exponentially small error. The results of existence and uniqueness for all time of strong solutions are also important for the study of the dynamical system generated by the PE (attractors, etc) (see the first developments of this theory in the article by Ju [2007], and quoted therein some previous partial results).

The study presented in this chapter is only a small part of the mathematical problems on geophysical flows, but we believe it is an important part. We did not try to produce here an exhaustive bibliography. Further mathematical references on geophysical flows will be given in the text, (see also the bibliography of the articles and books that we quote). There is also of course a very large literature in the physical context; we only mentioned some of the books that were very useful to us such as Haltiner and Williams [1980], Pedlovsky [1987], Trenberth [1992], Washington and Parkinson [1986], and Zeng [1979].

The mathematical theory presented in this chapter focuses on questions of existence, uniqueness, and regularity of solutions, the so-called issue of well-posedness. From the geophysics point of view, these issues relate, according to von Neumann [1963], to the short-term forecasting. The other issues as described in von Neumann [1963] relate to the long-term climate and intermediate climate dynamics. Pertaining to the long-term

climates are the questions of *attractors* for the PEs, which have been addressed in e.g., JU [2007], LIONS, TEMAM and WANG [1992a,b, 1993a, 1995] (see also the references therein). For intermediate climate dynamics, the mathematical issues relate to successive bifurcations, transition, and instabilities (see, e.g., MA and WANG [2005a,b], and the chapter by SIMONNET, DIJKSTRA and GHIL [2008] in this volume).

Besides the efforts of the authors, we mention in several places that this study is based on joint works with Lions, Wang, Hu, Wirosoetisno and others. Their help is gratefully acknowledged, and we pay tribute to the memory of Jacques-Louis Lions. The authors thank Denis Serre and Shouhong Wang for their careful reading of an earlier version of this manuscript and for their numerous comments that significantly improved the manuscript. They extend also their gratitude to Daniele Le Meur and Teresa Bunge who typed significant parts of the manuscript.

This chapter is an updated version of the article by TEMAM and ZIANE [2004]. It is included in this volume by invitation of PG Ciarlet, editor of the *Handbook of Numerical Analysis*. The authors thank PG Ciarlet for his invitation and the Elsevier Company for endorsing it.

2. The PEs: weak formulation, existence of weak solutions

As explained in the introduction to this chapter, our aim in this section is first to present the derivation of the PEs from the basic physical conservation laws. We then describe the natural boundary conditions. Then, on the mathematical side, we introduce the function spaces and derive the mathematical formulation of the PEs. Finally, we derive the existence for all time of weak solutions.

We successively consider the ocean, the atmosphere, and the CAO.

2.1. The PEs of the ocean

Our aim in this section is to describe the PEs of the ocean (see Section 2.1.1), and we then describe the corresponding boundary conditions and the associated initial and boundary value problems (Section 2.1.2).

2.1.1. The PEs

It is considered that the ocean is made up of a slightly compressible fluid with Coriolis force. The full set of equations of the large-scale ocean are the following: the conservation of momentum equation, the continuity equation (conservation of mass), the thermodynamics equation (that is the conservation of energy equation), the equation of state, and the equation of diffusion for the salinity S:

$$\rho \frac{\mathrm{d}\mathbf{V}_3}{\mathrm{d}t} + 2\rho\Omega \times \mathbf{V}_3 + \nabla_3 p + \rho \mathbf{g} = D, \tag{2.1}$$

$$\frac{\mathrm{d}\rho}{\mathrm{d}t} + \rho \operatorname{div}_3 \mathbf{V}_3 = 0, \tag{2.2}$$

$$\frac{\mathrm{d}T}{\mathrm{d}t} = Q_T, \tag{2.3}$$

$$\frac{\mathrm{d}S}{\mathrm{d}t} = Q_S, \tag{2.4}$$

$$\rho = f(T, S, p). \tag{2.5}$$

Here, $\mathbf{V}_3$ is the 3D velocity vector, $\mathbf{V}_3 = (u, v, w)$, ρ, p, T are the density, pressure, and temperature, S is the concentration of salinity, $\mathbf{g} = (0, 0, g)$ is the gravity vector, D is the molecular dissipation, and Q_T and Q_S are the heat and salinity diffusions; Ω is the rotation vector. The analytic expressions of D, Q_T, and Q_S will be given below. We denote by Δ_3, ∇_3, div_3, the 3D Laplacian, gradient, and divergence, leaving Δ, ∇, div to their 2D versions more frequently used.

The Boussinesq approximation. From both the theoretical and the computational points of views, the above systems of equations of the ocean seem to be too complicated to study. So it is necessary to simplify them according to some physical and mathematical considerations. The Mach number for the flow in the ocean is not large and therefore, as a starting point, we can make the so-called *Boussinesq approximation* in which the density is assumed constant, $\rho = \rho_0$, except in the buoyancy term and in the equation of state.

This amounts to replacing Eqs. (2.1) and (2.2) by

$$\rho_0 \frac{\mathrm{d}\mathbf{V}_3}{\mathrm{d}t} + 2\rho_0 \Omega \times \mathbf{V}_3 + \nabla_3 p + \rho \mathbf{g} = D, \tag{2.6}$$

$$\mathrm{div}_3\, \mathbf{V}_3 = 0. \tag{2.7}$$

Consider the spherical coordinate system (θ, ϕ, r), where θ $(-\pi/2 < \theta < \pi/2)$ stands for the latitude of the earth, ϕ $(0 \le \phi \le 2\pi)$ on the longitude of the earth, r for the radial distance, and $z = r - a$ for the vertical coordinate with respect to the sea level, and let $\mathbf{e}_\theta$, $\mathbf{e}_\phi$, and $\mathbf{e}_r$ be the unit vectors in the θ-, ϕ-, and z-directions, respectively. Then we write the velocity of the ocean in the form

$$\mathbf{V}_3 = v_\theta \mathbf{e}_\theta + v_\phi \mathbf{e}_\phi + v_r \mathbf{e}_r = \mathbf{v} + w, \tag{2.8}$$

where $\mathbf{v} = v_\theta \mathbf{e}_\theta + v_\phi \mathbf{e}_\phi$ is the horizontal velocity field and w is the vertical velocity.

Another common simplification is to replace, to first order, r by the radius a of the earth. This is based on the fact that the depth of the ocean is small compared with the radius of the earth. In particular,

$$\frac{\mathrm{d}}{\mathrm{d}t} = \frac{\partial}{\partial t} + \frac{v_\theta}{r}\frac{\partial}{\partial \theta} + \frac{v_\phi}{r\cos\theta}\frac{\partial}{\partial \phi} + v_r \frac{\partial}{\partial r} \tag{2.9}$$

becomes

$$\frac{\mathrm{d}}{\mathrm{d}t} = \frac{\partial}{\partial t} + \frac{v_\theta}{a}\frac{\partial}{\partial \theta} + \frac{v_\phi}{a\cos\theta}\frac{\partial}{\partial \phi} + v_z \frac{\partial}{\partial z}, \tag{2.10}$$

and taking the viscosity into consideration, we obtain the equations of the large-scale ocean with Boussinesq approximation, which are simply called *Boussinesq equations of the ocean*, i.e., Eqs. (2.11)–(2.16) hereafter (for the equation of state (Eq. (2.16)), see Remark 2.1):

$$\frac{\partial \mathbf{v}}{\partial t} + \nabla_{\mathbf{v}}\mathbf{v} + w\frac{\partial \mathbf{v}}{\partial z} + \frac{1}{\rho_0}\nabla p + 2\Omega \sin\theta \times \mathbf{v} - \mu_{\mathbf{v}}\Delta \mathbf{v} - \nu_{\mathbf{v}}\frac{\partial^2 \mathbf{v}}{\partial z^2} = 0, \tag{2.11}$$

$$\frac{\partial w}{\partial t} + \nabla_{\mathbf{v}} w + w\frac{\partial w}{\partial z} + \frac{1}{\rho_0}\frac{\partial p}{\partial z} + \frac{\rho}{\rho_0}g - \mu_{\mathbf{v}}\Delta w - \nu_{\mathbf{v}}\frac{\partial^2 w}{\partial z^2} = 0, \tag{2.12}$$

$$\operatorname{div}\mathbf{v} + \frac{\partial w}{\partial z} = 0, \tag{2.13}$$

$$\frac{\partial T}{\partial t} + \nabla_{\mathbf{v}} T + w\frac{\partial T}{\partial z} - \mu_T \Delta T - \nu_T \frac{\partial^2 T}{\partial z^2} = 0, \tag{2.14}$$

$$\frac{\partial S}{\partial t} + \nabla_{\mathbf{v}} S + w\frac{\partial S}{\partial z} - \mu_S \Delta S - \nu_S \frac{\partial^2 S}{\partial z^2} = 0, \tag{2.15}$$

$$\rho = \rho_0\big(1 - \beta_T(T - T_{\mathrm{r}}) + \beta_S(S - S_{\mathrm{r}})\big), \tag{2.16}$$

where $\mathbf{v}$ is the horizontal velocity of the water, w is the vertical velocity, and T_{r} and S_{r} are averaged (or reference) values of T and S. The diffusion coefficients $\mu_{\mathbf{v}}$, μ_{T}, μ_{S} and $\nu_{\mathbf{v}}$, ν_{T}, ν_{S} are different in the horizontal and vertical directions, accounting for some eddy diffusions in the sense of SMAGORINSKY [1963].

The differential operators are defined as follows. The (horizontal) gradient operator $\operatorname{grad} = \nabla$ is defined by

$$\operatorname{grad} p = \nabla p = \frac{1}{a}\frac{\partial p}{\partial \theta}\mathbf{e}_\theta + \frac{1}{a\cos\theta}\frac{\partial p}{\partial \phi}\mathbf{e}_\phi. \tag{2.17}$$

The (horizontal) divergence operator $\operatorname{div} = \nabla\cdot$ is defined by

$$\operatorname{div}(v_\theta \mathbf{e}_\theta + v_\phi \mathbf{e}_\phi) = \nabla \cdot \mathbf{v} = \frac{1}{a\cos\theta}\left(\frac{\partial(v_\theta \cos\theta)}{\partial\theta} + \frac{\partial v_\phi}{\partial \phi}\right). \tag{2.18}$$

The derivatives $\nabla_{\mathbf{v}}\tilde{\mathbf{v}}$ and $\nabla_{\mathbf{v}}\widetilde{T}$ of a vector function $\tilde{\mathbf{v}}$ and a scalar function $\widetilde{T}$ (covariant derivatives with respect to $\mathbf{v}$) are

$$\begin{aligned}\nabla_{\mathbf{v}}\tilde{\mathbf{v}} = {} & \left\{\frac{v_\theta}{a}\frac{\partial \tilde{v}_\theta}{\partial\theta} + \frac{v_\phi}{a\cos\theta}\frac{\partial \tilde{v}_\theta}{\partial\phi} - \frac{v_\phi \tilde{v}_\phi}{a}\cot\theta\right\}\mathbf{e}_\theta \\ & + \left\{\frac{v_\theta}{a}\frac{\partial \tilde{v}_\phi}{\partial\theta} + \frac{v_\phi}{a\cos\theta}\frac{\partial \tilde{v}_\phi}{\partial\phi} - \frac{\tilde{v}_\theta v_\phi}{a}\tan\theta\right\}\mathbf{e}_\phi,\end{aligned} \tag{2.19}$$

$$\nabla_{\mathbf{v}}\widetilde{T} = \frac{v_\theta}{a}\frac{\partial \widetilde{T}}{\partial\theta} + \frac{v_\phi}{a\cos\theta}\frac{\partial \widetilde{T}}{\partial\phi}. \tag{2.20}$$

Moreover, we have used the same notation Δ to denote the Laplace–Beltrami operators for both scalar functions and vector fields on S_a^2, the 2D sphere of radius a centered at 0. More precisely, we have

$$\Delta T = \frac{1}{a^2 \cos\theta}\left\{\frac{\partial}{\partial\theta}\left(\cos\theta\frac{\partial T}{\partial\theta}\right) + \frac{1}{\cos\theta}\frac{\partial^2 T}{\partial\phi^2}\right\}, \tag{2.21}$$

$$\begin{aligned}\Delta\mathbf{v} &= \Delta(v_\theta\mathbf{e}_\theta + v_\phi\mathbf{e}_\phi)\\ &= \left\{\Delta v_\theta - \frac{2\sin\theta}{a^2\cos^2\theta}\frac{\partial v_\phi}{\partial\phi} - \frac{v_\theta}{a^2\cos^2\theta}\right\}\mathbf{e}_\theta\\ &\quad + \left\{\Delta v_\phi - \frac{2\sin\theta}{a^2\cos^2\theta}\frac{\partial v_\theta}{\partial\phi} - \frac{v_\phi}{a^2\cos^2\theta}\right\}\mathbf{e}_\phi,\end{aligned} \tag{2.22}$$

where in Eq. (2.22), Δv_θ, and Δv_ϕ are defined by Eq. (2.21), and in Eq. (2.21), T is any given (smooth) function on S_a^2, the 2D sphere of radius a.

REMARK 2.1. The equation of state for the ocean is given by Eq. (2.5). Only empirical forms of the function $\rho = f(T, S, p)$ are known (see WASHINGTON and PARKINSON [1986, pp. 131–132]). This equation of state is generally derived on a phenomenological basis. It is natural to expect that ρ decreases if T increases and that ρ increases if S increases.

The simplest law is Eq. (2.16) corresponding to a linearization around average (or reference) values ρ_0, T_r, S_r of the density, the temperature, and the salinity, and β_T and β_S are positive constant expansion coefficients. Much of what follows extends to more general nonlinear equations of state.

REMARK 2.2. The replacement of r by Eq. (2.10) in the differential operators implies a change of metric in $\mathbb{R}^3$, where the usual metric is replaced by that of $\mathbb{S}_a^2 \times \mathbb{R}$, $\mathbb{S}_a^2$, the 2D sphere of radius a centered at O.

REMARK 2.3. In a classical manner, the Coriolis force $2\rho\Omega \times \mathbf{V}_3$ produces the term $2\Omega\sin\theta\mathbf{k} \times \mathbf{v}$ and a horizontal gradient term that is combined with the pressure so that p in Eq. (2.11) is the so-called *augmented pressure*.

The hydrostatic approximation. It is known that for large-scale ocean, the horizontal scale is much bigger than the vertical one (5–10 km versus a few thousands kilometers). Therefore, the scale analysis (see PEDLOVSKY [1987]) shows that $\partial p/\partial z$ and ρg are the dominant terms in Eq. (2.12), leading to the hydrostatic approximation

$$\frac{\partial p}{\partial z} = -\rho g, \tag{2.23}$$

which then replaces Eq. (2.12). The approximate relation is highly accurate for the large-scale ocean, and it is considered as a fundamental equation in oceanography. From the mathematical point of view, its justification relies on tools similar to those used in Section 4.1.

The rigorous mathematical justification of the hydrostatic approximation is given by AZÉRAD and GUILLÉN [2001]. In this chapter, the authors studied the asymptotic behavior of the incompressible Navier-Stokes equations when the depth goes to zero, and they proved that the solutions of the Navier-Stokes equations converge to a weak solution of the PEs. The mathematical details will not be discussed in this chapter, (see however AZÉRAD and GUILLÉN [2001], as well as Remark 4.1 in Section 4.1).

Using the hydrostatic approximation, we obtain the following equations called the *PEs of the large-scale ocean*:

$$\frac{\partial \mathbf{v}}{\partial t} + \nabla_{\mathbf{v}}\mathbf{v} + w\frac{\partial \mathbf{v}}{\partial z} + \frac{1}{\rho_0}\nabla p + 2\Omega \sin\theta \mathbf{k} \times \mathbf{v} - \mu_{\mathbf{v}}\Delta \mathbf{v} - \nu_{\mathbf{v}}\frac{\partial^2 \mathbf{v}}{\partial z^2} = F_{\mathbf{v}}, \tag{2.24}$$

$$\frac{\partial p}{\partial z} = -\rho g, \tag{2.25}$$

$$\operatorname{div} \mathbf{v} + \frac{\partial w}{\partial z} = 0, \tag{2.26}$$

$$\frac{\partial T}{\partial t} + \nabla_{\mathbf{v}} T + w\frac{\partial T}{\partial z} - \mu_{\mathrm{T}}\Delta T - \nu_T \frac{\partial^2 T}{\partial z^2} = F_{\mathrm{T}}, \tag{2.27}$$

$$\frac{\partial S}{\partial t} + \nabla_{\mathbf{v}} S + w\frac{\partial S}{\partial z} - \mu_{\mathrm{S}}\Delta S - \nu_S \frac{\partial^2 S}{\partial z^2} = F_{\mathrm{S}}, \tag{2.28}$$

$$\rho = \rho_0\big(1 - \beta_T(T - T_{\mathrm{r}}) + \beta_S(S - S_{\mathrm{r}})\big). \tag{2.29}$$

Note that $F_{\mathbf{v}}$, F_T, and F_S corresponding to volumic sources (of horizontal momentum, heat, and salt) vanish in reality; they are introduced here for mathematical generality. We also set $\Omega = \Omega\mathbf{k}$, where $\mathbf{k}$ is the unit vector in the direction of the poles (from south to north).

REMARK 2.4. At this stage, the unknown functions can be divided into two sets. The first one is called the *prognostic variables*, $\mathbf{v}$, T, S (four scalar functions); we aim to write the PEs as an initial-boundary value problem for these unknowns, and we set $U = (\mathbf{v}, T, S)$. The second set of variables comprises p, ρ, w; they are called the *diagnostic variables*. In Section 2.1.2, we will see how, using the boundary condition, one can, at each instant of time, express the diagnostic variables in terms of the prognostic variables (a fact that is already transparent for ρ in Eq. (2.29).

REMARK 2.5. We integrate Eq. (2.28) over the domain $\mathcal{M}$ occupied by the fluid that is described in Section 2.1.2. Using the Stokes formula and taking into account Eq. (2.26) and the boundary conditions (also described in Section 2.1.2), we arrive at

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathcal{M}} S\,\mathrm{d}\mathcal{M} = \int_{\mathcal{M}} F_S\,\mathrm{d}\mathcal{M}; \tag{2.30}$$

hence,

$$\int_{\mathcal{M}} S\,\mathrm{d}\mathcal{M}\Big|_t = \int_{\mathcal{M}} S\,\mathrm{d}\mathcal{M}\Big|_0 + \int_0^t \int_{\mathcal{M}} F_S\,\mathrm{d}\mathcal{M}\,\mathrm{d}t'.$$

In practical applications, $F_S = 0$ as we said, and the total amount of salt $\int_{\mathcal{M}} S \, \mathrm{d}\mathcal{M}$ is conserved. In all cases, we write

$$S' = S - \frac{1}{|\mathcal{M}|} \int_{\mathcal{M}} S \, \mathrm{d}\mathcal{M}, \quad F'_S = F_S - \frac{1}{|\mathcal{M}|} \int_{\mathcal{M}} F_S \, \mathrm{d}\mathcal{M}, \tag{2.31}$$

where $|\mathcal{M}|$ is the volume of $\mathcal{M}$, and we see that S' satisfies the same Eq. (2.15), with F_S replaced by F'_S. From now on, dropping the primes, we consider Eq. (2.15) as the equation for S' and we thus have

$$\int_{\mathcal{M}} S \, \mathrm{d}\mathcal{M} = 0, \quad \int_{\mathcal{M}} F_S \, \mathrm{d}\mathcal{M} = 0. \tag{2.32}$$

2.1.2. The initial and boundary value problems

We assume that the ocean fills a domain $\mathcal{M}$ of $\mathbb{R}^3$, which we describe as follows (see Figure 2.1):

The top of the ocean is a domain Γ_{i} included in the surface of the earth S_a (sphere centered at 0 of radius a). The bottom Γ_{b} of the ocean is defined by ($z = x_3 = r - a$)

$$z = -h(\theta, \varphi),$$

where h is a function of class $\mathcal{C}^2$ at least on $\overline{\Gamma}_{\mathrm{i}}$; it is assumed also that h is bounded from below,

$$0 < \underline{h} \leq h(\theta, \varphi) \leq \bar{h}, \, (\theta, \varphi) \in \Gamma_{\mathrm{i}}. \tag{2.33}$$

The lateral surface Γ_ℓ consists of the part of cylinder

$$(\theta, \varphi) \in \partial\Gamma_{\mathrm{i}}, \, -h(\theta, \varphi) \leq z \leq 0. \tag{2.34}$$

REMARK 2.6. Let us make two remarks concerning the geometry of the ocean; the first one is that for mathematical reasons, the depth is not allowed to be 0 ($h \geq \underline{h} > 0$) and thus "beaches" are excluded. The second one is that the top of the ocean is flat (spherical), not allowing waves; this corresponds to the so-called *rigid lid assumption* in oceanography. The assumption $h > 0$ can be relaxed for some of the following results, but this will not be discussed here. The rigid lid assumption can be also relaxed by the introduction of an additional equation for the free surface, but this also will not be considered.

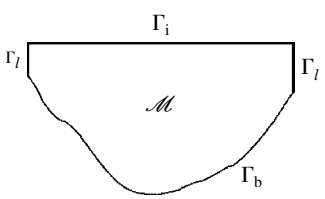


FIG. 2.1 The ocean $\mathcal{M}$.

Boundary conditions. There are several sets of natural boundary conditions can be associated with the PEs, for instance, the following:
On the top of the ocean Γ_{i} ($z = 0$):

$$\begin{aligned} &\nu_{\mathbf{v}} \frac{\partial \mathbf{v}}{\partial z} + \alpha_{\mathbf{v}} \big(\mathbf{v} - \mathbf{v}^{\mathrm{a}}\big) = \tau_{\mathbf{v}}, \quad w = 0, \\ &\nu_T \frac{\partial T}{\partial z} + \alpha_T \big(T - T^{\mathrm{a}}\big) = 0, \\ &\frac{\partial S}{\partial z} = 0. \end{aligned} \tag{2.35}$$

At the bottom of the ocean Γ_{b} ($z = -h(\theta, \varphi)$):

$$\begin{aligned} &\mathbf{v} = 0, \qquad w = 0, \\ &\frac{\partial T}{\partial \mathbf{n}_T} = 0, \qquad \frac{\partial S}{\partial \mathbf{n}_S} = 0. \end{aligned} \tag{2.36}$$

On the lateral boundary Γ_ℓ ($-h(\theta, \varphi) < z < 0$, $(\theta, \varphi) \in \partial\Gamma_{\mathrm{i}}$):

$$\mathbf{v} = 0, \quad w = 0, \quad \frac{\partial T}{\partial \mathbf{n}_T} = 0, \quad \frac{\partial S}{\partial \mathbf{n}_S} = 0. \tag{2.37}$$

Here, $\mathbf{n} = (n_H, n_z)$ is the unit outward normal on $\partial\mathcal{M}$ decomposed into its horizontal and vertical components; the co-normal derivatives $\partial/\partial\mathbf{n}_T$ and $\partial/\partial\mathbf{n}_S$ are those associated with the linear (temperature and salinity) operators, i.e.,

$$\begin{aligned} \frac{\partial}{\partial \mathbf{n}_T} &= \frac{\mu_T}{1 + |\nabla h|^2} \nabla_{\nabla h} + \left(\frac{\mu_T}{1 + |\nabla h|^2} + (\nu_T - \mu_T) \right) \frac{\partial}{\partial z}, \\ \frac{\partial}{\partial \mathbf{n}_S} &= \frac{\mu_S}{1 + |\nabla h|^2} \nabla_{\nabla h} + \left(\frac{\mu_S}{1 + |\nabla h|^2} + (\nu_S - \mu_S) \right) \frac{\partial}{\partial z}, \end{aligned} \tag{2.38}$$

where $\nabla_{\nabla h}$ is the (2D) covariant derivative in the direction of ∇h (see, e.g., in Lions, Temam and Wang [1993a] after Eq. (1.21) and after Eq. (3.27)).

Remark 2.7.

(i) The boundary conditions (which are the same) on Γ_{b} and Γ_ℓ express the no-slip boundary conditions for the water and the absence of fluxes of heat or salt. For Γ_{i}, $w = 0$ is the geometrical (kinematical) boundary condition required by the rigid lid assumption; the Neumann boundary condition on S expresses the absence of salt flux.
(ii) In general, the boundary conditions on $\mathbf{v}$ and T on Γ_{i} are not fully settled from the physical point of view. These correspond to some resolution of the viscous boundary layers on the top of the ocean. Here, $\alpha_{\mathbf{v}}$ and α_T are given ≥ 0, $\mathbf{v}^{\mathrm{a}}$ and T^{a} correspond to the values in the atmosphere, and $\tau_{\mathbf{v}}$ corresponds to the shear of the wind.
(iii) The first boundary condition Eq. (2.35) could be replaced by $\mathbf{v} = \mathbf{v}^{\mathrm{a}}$ expressing a no-slip condition between air and sea. However, such a boundary condition

necessitating an exact resolution of the boundary layer would not be practically (computationally) realistic, and as indicated in (ii), we use instead some classical resolution of the boundary layer (SCHLICHTING [1979]).

(iv) As we said the boundary condition of Γ_i is standard unless more involved interactions are taken into consideration. However, for Γ_b and Γ_ℓ, different combinations of the Dirichlet and Neumann boundary conditions can be (have been) considered. (LIONS, TEMAM and WANG [1992b]).

Beta-plane approximation. For midlatitude regional studies, it is usual to consider the beta-plane approximation of the equations in which $\mathcal{M}$ is a domain in the space $\mathbb{R}^3$ with Cartesian coordinates denoted as x, y, z or x_1, x_2, x_3. In the beta-plane approximation, $\Omega = 2f\mathbf{k}$, $f = f_0 + \beta y$, $\mathbf{k}$ the unit vector along the south to north poles, $\Delta = \partial^2/\partial x^2 + \partial^2/\partial y^2$, ∇ is the usual nabla vector $(\partial/\partial x, \partial/\partial y)$, and $\nabla_{\mathbf{v}} = u\partial/\partial x + v\partial/\partial y$ $(\mathbf{v} = (u, v))$. With these notations, the Eqs. (2.24)–(2.29) and the boundary conditions (Eqs. (2.35)–(2.38)) keep the same form; here the depth $h = h(x, y)$ satisfies, like Eq. (2.33),

$$0 < \underline{h} \leq h(x, y) \leq \bar{h}, \tag{2.39}$$

and the boundary of $\mathcal{M}$ consists of Γ_i, Γ_b, Γ_ℓ defined as before.

As indicated in the Introduction, we will emphasize in this chapter the regional model, which is slightly simpler, in particular because of the use of Cartesian coordinates. Usually, the general model in spherical coordinates simply requires the treatment of lower order terms.

From now on, we consider the regional (Cartesian coordinate) case.

The diagnostic variables. The first step in the mathematical formulation of the PEs consists in showing how to express the diagnostic variables in terms of the prognostic variables, thanks to the equations and boundary conditions.

Since $w = 0$ on Γ_i and Γ_b, integration of Eq. (2.26) in z gives

$$w = w(\mathbf{v}) = \int_z^0 \operatorname{div} \mathbf{v}\, dz' \tag{2.40}$$

and

$$\int_{-h}^0 \operatorname{div} \mathbf{v}\, dz = 0. \tag{2.41}$$

Note that

$$\operatorname{div} \int_{-h}^0 \mathbf{v}\, dz = \int_{-h}^0 \operatorname{div} \mathbf{v}\, dz + \nabla h \cdot \mathbf{v}\Big|_{z=-h},$$

and since $\mathbf{v}$ vanishes on Γ_b, condition (Eq. (2.41)), is the same as

$$\operatorname{div} \int_{-h}^0 \mathbf{v}\, dz = 0. \tag{2.42}$$

Similarly, integration of Eq. (2.25) in z gives

$$p = p_s + P, \qquad P = P(T, S) = g \int_z^0 \rho \, dz'. \tag{2.43}$$

Here, ρ is expressed in terms of T and S through Eq. (2.29), and $p_s = p_s(x, y, t) = p(x, y, 0, t)$ is the pressure at the surface of the ocean.

Hence, Eqs. (2.40) and (2.43) provide an expression of the diagnostic variables in terms of the prognostic variables (and the surface pressure), and Eq. (2.42) is an additional equation that, we will see, is mathematically related to the surface pressure.

REMARK 2.8. The introduction of the nonlocal constraint (Eq. (2.41)) and of the surface pressure p_s was first carried out by LIONS, TEMAM and WANG [1992a,b]. This new formulation has played a crucial role in much of the mathematical analysis of the PEs in various cases.

2.2. *Weak formulation of the PEs of the ocean: the stationary PEs*

We denote by U the triplet $(\mathbf{v}, T, S)$ (four scalar functions). In summary, the equations that we consider for the subsequent mathematical theory (the PEs) are Eqs. (2.24), (2.27), and (2.28), with $w = w(\mathbf{v})$ given by Eq. (2.40) and p given by Eq. (2.43) (ρ given by Eq. (2.29)); furthermore, $\mathbf{v}$ satisfies Eq. (2.41); hence,

$$\frac{\partial \mathbf{v}}{\partial t} + \nabla_{\mathbf{v}} \mathbf{v} + w \frac{\partial \mathbf{v}}{\partial z} + \frac{1}{\rho_0} \nabla p + 2f \mathbf{k} \times \mathbf{v} - \mu_{\mathbf{v}} \Delta \mathbf{v} - \nu_{\mathbf{v}} \frac{\partial^2 \mathbf{v}}{\partial z^2} = F_{\mathbf{v}}, \tag{2.44}$$

$$\frac{\partial T}{\partial t} + \nabla_{\mathbf{v}} T + w \frac{\partial T}{\partial z} - \mu_T \Delta T - \nu_T \frac{\partial^2 T}{\partial z^2} = F_T, \tag{2.45}$$

$$\frac{\partial S}{\partial t} + \nabla_{\mathbf{v}} S + w \frac{\partial S}{\partial z} - \mu_S \Delta S - \nu_S \frac{\partial^2 S}{\partial z^2} = F_S, \tag{2.46}$$

$$w = w(\mathbf{v}) = \int_z^0 \operatorname{div} \mathbf{v} \, dz', \tag{2.47}$$

$$\operatorname{div} \int_{-h}^0 \mathbf{v} \, dz = 0, \tag{2.48}$$

$$p = p_s + P, \quad P = P(T, S) = g \int_z^0 \rho \, dz', \tag{2.49}$$

$$\rho = \rho_0 \big(1 - \beta_T (T - T_r) + \beta_S (S - S_r)\big), \tag{2.50}$$

$$\int_{\mathcal{M}} S \, d\mathcal{M} = 0. \tag{2.51}$$

The boundary conditions are Eqs. (2.35)–(2.37).

2.2.1. Weak formulation and functional setting

For the weak formulation of this problem, we introduce the following function spaces V and H:

$$V = V_1 \times V_2 \times V_3, \qquad H = H_1 \times H_2 \times H_3,$$

$$V_1 = \left\{ \mathbf{v} \in H^1(\mathcal{M})^2, \operatorname{div} \int_{-h}^{0} \mathbf{v}\, \mathrm{d}z = 0, \mathbf{v} = 0 \text{ on } \Gamma_\mathrm{b} \cup \Gamma_\ell \right\},$$

$$V_2 = H^1(\mathcal{M}),$$

$$V_3 = \dot{H}^1(\mathcal{M}) = \left\{ S \in H^1(\mathcal{M}), \int_{\mathcal{M}} S\, \mathrm{d}\mathcal{M} = 0 \right\},$$

$$H_1 = \left\{ \mathbf{v} \in L^2(\mathcal{M})^2, \operatorname{div} \int_{-h}^{0} \mathbf{v}\, \mathrm{d}z = 0, n_H \cdot \int_{-h}^{0} \mathbf{v}\, \mathrm{d}z = 0 \text{ on } \partial\Gamma_\mathrm{i} \text{ (i.e., on } \Gamma_\ell) \right\},$$

$$H_2 = L^2(\mathcal{M}),$$

$$H_3 = \dot{L}^2(\mathcal{M}) = \left\{ S \in L^2(\mathcal{M}), \int_{\mathcal{M}} S\, \mathrm{d}\mathcal{M} = 0 \right\}.$$

These spaces are endowed with the following scalar products and norms:

$$\big((U, \widetilde{U})\big) = \big((\mathbf{v}, \tilde{\mathbf{v}})\big)_1 + K_T\big((T, \widetilde{T})\big)_2 + K_S\big((S, \widetilde{S})\big)_3,$$

$$\big((\mathbf{v}, \tilde{\mathbf{v}})\big)_1 = \int_{\mathcal{M}} \left(\mu_\mathbf{v} \nabla \mathbf{v} \cdot \nabla \tilde{\mathbf{v}} + \nu_\mathbf{v} \frac{\partial \mathbf{v}}{\partial z} \frac{\partial \tilde{\mathbf{v}}}{\partial z} \right) \mathrm{d}\mathcal{M},$$

$$\big((T, \widetilde{T})\big)_2 = \int_{\mathcal{M}} \left(\mu_T \nabla T \cdot \nabla \widetilde{T} + \nu_T \frac{\partial T}{\partial z} \frac{\partial \widetilde{T}}{\partial z} \right) \mathrm{d}\mathcal{M} + \int_{\Gamma_\mathrm{i}} \alpha_T T \widetilde{T}\, \mathrm{d}\Gamma_\mathrm{i},$$

$$\big((S, \widetilde{S})\big)_3 = \int_{\mathcal{M}} \left(\mu_S \nabla S \cdot \nabla \widetilde{S} + \nu_S \frac{\partial S}{\partial z} \frac{\partial \widetilde{S}}{\partial z} \right) \mathrm{d}\mathcal{M},$$

$$\big(U, \widetilde{U}\big)_H = \int_{\mathcal{M}} \big(\mathbf{v} \cdot \tilde{\mathbf{v}} + K_T T \widetilde{T} + K_S S \widetilde{S} \big)\, \mathrm{d}\mathcal{M},$$

$$\|U\| = \big((U, U)\big)^{1/2}, \qquad |U|_H = (U, U)_H^{1/2}.$$

Here, K_T and K_S are suitable positive constants chosen below. The norm on H is of course equivalent to the L^2-norm, and because of the Poincaré inequality, $\mathbf{v}$ vanishes on $\Gamma_\mathrm{b} \cup \Gamma_\ell$, and Eq. (2.51), $\|\cdot\|_i = ((\cdot,\cdot))_i^{1/2}$ is a Hilbert norm on V_i, and $\|\cdot\|$ is a Hilbert norm on V; more precisely we have, with $c_0 > 0$ a suitable constant depending on $\mathcal{M}$:

$$|U|_H \leq c_0 \|U\| \quad \forall U \in V. \tag{2.52}$$

Let $\mathcal{V}_1$ be the space of $\mathcal{C}^\infty$ (2D) vector functions $\mathbf{v}$, which vanish in a neighborhood of $\Gamma_\mathrm{b} \cup \Gamma_\ell$ and such that

$$\operatorname{div} \int_{-h}^{0} \mathbf{v}\, \mathrm{d}z = 0.$$

Then, $\mathcal{V}_1 \subset V_1$, and it has been shown by Lions, Temam and Wang [1992a] that

$$\mathcal{V}_1 \text{ is dense in } V_1. \tag{2.53}$$

We also denote by $\mathcal{V}_2 \subset V_2$ the set of $\mathcal{C}^\infty$ functions on $\overline{\mathcal{M}}$ and by $\mathcal{V}_3 \subset V_3$ the set of $\mathcal{C}^\infty$ functions on $\overline{\mathcal{M}}$ with zero average; $\mathcal{V} = \mathcal{V}_1 \times \mathcal{V}_2 \times \mathcal{V}_3$ is dense in V.

To derive the weak formulation of this problem, we consider a sufficiently regular test function $\widetilde{U} = (\tilde{\mathbf{v}}, \widetilde{T}, \widetilde{S})$ in V. We multiply Eq. (2.44) by $\tilde{\mathbf{v}}$, Eq. (2.45) by $K_T\widetilde{T}$, Eq. (2.46) by $K_S\widetilde{S}$, integrate over $\mathcal{M}$ and add the resulting equations; K_T, $K_S > 0$ are two constants to be chosen later on.

The term involving grad p_S vanishes; indeed, by the Stokes formula,

$$\int_{\mathcal{M}} \nabla p_S \cdot \tilde{\mathbf{v}} \,\mathrm{d}\mathcal{M} = \int_{\partial\mathcal{M}} p_S n_H \cdot \tilde{\mathbf{v}} \,\mathrm{d}(\partial\mathcal{M}) - \int_{\mathcal{M}} p_S \nabla \cdot \tilde{\mathbf{v}} \,\mathrm{d}\mathcal{M},$$

where $\mathbf{n} = (n_H, n_z)$ is the unit outward normal on $\partial\mathcal{M}$, and n_H its horizontal component. The integral on $\partial\mathcal{M}$ vanishes because $n_H \cdot \tilde{\mathbf{v}}$ vanishes on $\partial\mathcal{M}$; the remaining integral on $\mathcal{M}$ vanishes too since by Fubini's theorem, Eq. (2.48), and $\tilde{\mathbf{v}} = 0$ on Γ_{b},

$$\int_{\mathcal{M}} p_S \nabla \cdot \tilde{\mathbf{v}} \,\mathrm{d}\mathcal{M} = \int_{\Gamma_{\mathrm{i}}} p_S \int_{-h}^{0} \nabla \cdot \tilde{\mathbf{v}} \,\mathrm{d}z \,\mathrm{d}\Gamma_{\mathrm{i}} = \int_{\Gamma_{\mathrm{i}}} p_S \left(\nabla \cdot \int_{-h}^{0} \tilde{\mathbf{v}} \,\mathrm{d}z \right) \mathrm{d}\Gamma_{\mathrm{i}} = 0.$$

Using Stokes' formula and the boundary conditions (Eqs. (2.35)–(2.38)) we arrive after some easy calculations at

$$\left(\frac{\mathrm{d}}{\mathrm{d}t} U, \widetilde{U} \right)_H + a\big(U, \widetilde{U}\big) + b\big(U, U, \widetilde{U}\big) + e\big(U, \widetilde{U}\big) = \ell\big(\widetilde{U}\big). \tag{2.54}$$

The notations are as follows:

$$\begin{aligned}
&\big(U, \widetilde{U}\big)_H = \int_{\mathcal{M}} \big(\mathbf{v} \cdot \tilde{\mathbf{v}} + K_T T\widetilde{T} + K_S S\widetilde{S} \big) \,\mathrm{d}\mathcal{M}, \\
&a\big(U, \widetilde{U}\big) = a_1\big(U, \widetilde{U}\big) + K_T a_2\big(U, \widetilde{U}\big) + K_S a_3\big(U, \widetilde{U}\big), \\
&a_1\big(U, \widetilde{U}\big) = \int_{\mathcal{M}} \left(\mu_{\mathbf{v}} \nabla \mathbf{v} \cdot \nabla \tilde{\mathbf{v}} + \nu_{\mathbf{v}} \frac{\partial \mathbf{v}}{\partial z} \frac{\partial \tilde{\mathbf{v}}}{\partial z} \right) \mathrm{d}\mathcal{M} \\
&\qquad\qquad - \int_{\mathcal{M}} \overline{P}(T, S) \nabla \cdot \tilde{\mathbf{v}} \,\mathrm{d}\mathcal{M} + \int_{\Gamma_{\mathrm{i}}} \alpha_{\mathbf{v}} \mathbf{v} \tilde{\mathbf{v}} \,\mathrm{d}\Gamma_{\mathrm{i}}, \\
&\overline{P}(T, S) = g \int_z^0 (-\beta_T T + \beta_S S) dz' \quad \big(\text{see Eqs. (2.49) and (2.50)}\big), \\
&a_2\big(U, \widetilde{U}\big) = \int_{\mathcal{M}} \left(\mu_T \nabla T \cdot \nabla \widetilde{T} + \nu_T \frac{\partial T}{\partial z} \frac{\partial \widetilde{T}}{\partial z} \right) \mathrm{d}\mathcal{M} + \int_{\Gamma_{\mathrm{i}}} \alpha_T T \widetilde{T} \,\mathrm{d}\Gamma_{\mathrm{i}},
\end{aligned}$$

$$a_3\big(U,\widetilde{U}\big)=\int_{\mathcal{M}}\left(\mu_S\nabla S\cdot\nabla\widetilde{S}+\nu_S\frac{\partial S}{\partial z}\frac{\partial\widetilde{S}}{\partial z}\right)\mathrm{d}\mathcal{M},$$

$$b=b_1+K_Tb_2+K_Sb_3,$$

$$b_1\big(U,\widetilde{U},U^\sharp\big)=\int_{\mathcal{M}}\left(\mathbf{v}\cdot\nabla\tilde{\mathbf{v}}+w(\mathbf{v})\frac{\partial\tilde{\mathbf{v}}}{\partial z}\right)\mathbf{v}^\sharp\,\mathrm{d}\mathcal{M},$$

$$b_2\big(U,\widetilde{U},U^\sharp\big)=\int_{\mathcal{M}}\left(\mathbf{v}\cdot\nabla\widetilde{T}+w(\mathbf{v})\frac{\partial\widetilde{T}}{\partial z}\right)T^\sharp\,\mathrm{d}\mathcal{M},$$

$$b_3\big(U,\widetilde{U},U^\sharp\big)=\int_{\mathcal{M}}\left(\mathbf{v}\cdot\nabla\widetilde{S}+w(\mathbf{v})\frac{\partial\widetilde{S}}{\partial z}\right)S^\sharp\,\mathrm{d}\mathcal{M},$$

$$e\big(U,\widetilde{U}\big)=2\int_{\mathcal{M}}(f\mathbf{k}\times\mathbf{v})\cdot\tilde{\mathbf{v}}\,\mathrm{d}\mathcal{M}$$

and

$$\begin{aligned}\ell\big(\widetilde{U}\big)=&\int_{\mathcal{M}}\big(F_{\mathbf{v}}\tilde{\mathbf{v}}+K_TF_T\widetilde{T}+K_SF_S\widetilde{S}\big)\,\mathrm{d}\mathcal{M}\\&+\int_{\mathcal{M}}\left[g\int_z^0(1+\beta_TT_r-\beta_SS_{\mathrm{r}})\mathrm{d}z'\right]\nabla\cdot\tilde{\mathbf{v}}\,\mathrm{d}\mathcal{M}\\&+\int_{\Gamma_{\mathrm{i}}}\big[(g_{\mathbf{v}})\cdot\tilde{\mathbf{v}}+g_T\widetilde{T}\big]\,\mathrm{d}\Gamma_{\mathrm{i}},\end{aligned}\tag{2.55}$$

where (see Eq. (2.35))

$$g_{\mathbf{v}}=\tau_{\mathbf{v}}+\alpha_{\mathbf{v}}\mathbf{v}^{\mathrm{a}},\qquad g_T=\alpha_TT^{\mathrm{a}}.$$

For ℓ, we observe that if T_{r} and S_{r} are constant, then

$$\begin{aligned}&\int_{\mathcal{M}}\left[g\int_z^0(1+\beta_TT_{\mathrm{r}}-\beta_SS_{\mathrm{r}})\mathrm{d}z'\right]\nabla\cdot\mathbf{v}\,\mathrm{d}\mathcal{M}\\&\quad=\int_{\partial\mathcal{M}}\left[g\int_z^0(1+\beta_TT_{\mathrm{r}}-\beta_SS_{\mathrm{r}})\mathrm{d}z'\right]n_H\cdot\mathbf{v}\,\mathrm{d}(\partial\mathcal{M})\\&\quad=0.\end{aligned}\tag{2.56}$$

It is clear that each a_i, and thus a, is a bilinear continuous form on V; furthermore, if K_T and K_S are sufficiently large, a is coercive (a_2, a_3 are automatically coercive on V_2, V_3):

$$a(U,U)\geq c_1\|U\|^2\quad\forall U\in V\ (c_1>0).\tag{2.57}$$

Similarly, e is bilinear continuous on V_1 and even H_1, and

$$e(U,U)=0\quad\forall U\in H.\tag{2.58}$$

Before studying the properties of the form b, we introduce the space $V_{(2)}$:

$$V_{(2)} \text{ is the closure of } \mathcal{V} \text{ in } \big(H^2(\mathcal{M})\big)^4. \tag{2.59}$$

Then, we have the following

LEMMA 2.1. *The form b is trilinear continuous on $V \times V \times V_{(2)}$ and $V \times V_{(2)} \times V$,*[1]

$$\begin{aligned} &\big|b\big(U, \widetilde{U}, U^\sharp\big)\big| \\ &\quad \leq \begin{cases} c_2\|U\|\big\|\widetilde{U}\big\|\big\|U^\sharp\big\|_{V_{(2)}}, & \forall U, \widetilde{U} \in V, U^\sharp \in V_{(2)}, \\ c_2\|U\|\big\|\widetilde{U}\big\|_{V_{(2)}}\big\|U^\sharp\big\|, & \forall U, U^\sharp \in V, \widetilde{U} \in V_{(2)}, \end{cases} \end{aligned} \tag{2.60}$$

or

$$\begin{aligned} &\big|b\big(U, \widetilde{U}, U^\sharp\big)\big| \\ &\quad \leq c_2\|U\|\big|\widetilde{U}\big|_H^{1/2}\big\|\widetilde{U}\big\|^{1/2}\big\|U^\sharp\big\|_{V_{(2)}}, \quad \forall U, \widetilde{U} \in V, U^\sharp \in V_{(2)}. \end{aligned} \tag{2.61}$$

Furthermore,

$$b\big(U, \widetilde{U}, \widetilde{U}\big) = 0 \quad \textit{for } U \in V, \widetilde{U} \in V_{(2)}, \tag{2.62}$$

and

$$b\big(U, \widetilde{U}, U^\sharp\big) = -b\big(U, U^\sharp, \widetilde{U}\big) \tag{2.63}$$

for $U, \widetilde{U}, U^\# \in V$, and $\widetilde{U}$ or $U^\sharp$ in $V_{(2)}$.

PROOF. To show first that b is defined on $V \times V \times V_{(2)}$, let us consider the typical and most problematic term

$$\int_{\mathcal{M}} w(\mathbf{v})\frac{\partial \widetilde{T}}{\partial z}T^\sharp \,\mathrm{d}\mathcal{M}. \tag{2.64}$$

We have

$$\int_{\mathcal{M}} \left|w(\mathbf{v})\frac{\partial \widetilde{T}}{\partial z}T^\sharp\right| \mathrm{d}\mathcal{M} \leq \big|w(\mathbf{v})\big|_{L^2(\mathcal{M})}\left|\frac{\partial \widetilde{T}}{\partial z}\right|_{L^2(\mathcal{M})}\big|T^\sharp\big|_{L^\infty(\mathcal{M})}.$$

The first two terms in the right-hand side of this inequality are bounded by *const* $\cdot \|\mathbf{v}\|_1$ (using Eq. (2.40)) and $\|\widetilde{T}\|_2$. In dimension 3, $H^2(\mathcal{M}) \subset L^\infty(\mathcal{M})$ so that the third term

[1] For Eqs. (2.60) and (2.61), the specific form of V and $V_{(2)}$ is not important: b is as well trilinear continuous on $H^1(\mathcal{M})^4 \times H^2(\mathcal{M})^4 \times H^1(\mathcal{M})^4$ and $H^1(\mathcal{M})^4 \times H^1(\mathcal{M})^4 \times H^2(\mathcal{M})$, and the estimates are similar, the H^1 and H^2 norms replacing the V and $V_{(2)}$ norms.

is bounded by $const \cdot \|T^\sharp\|_{V_{(2)}}$, and hence the right-hand side of the last inequality is bounded by

$$c\|U\|\|\widetilde{U}\|\|U^\sharp\|_{V_{(2)}}.$$

With similar (and easier) inequalities for the other integral, we conclude that b is defined and trilinear continuous on $V \times V \times V_{(2)}$.

For the continuity on $V \times V_{(2)} \times V$, the typical term above is bounded by

$$\big|w(\mathbf{v})\big|_{L^2(\mathcal{M})}\left|\frac{\partial\widetilde{T}}{\partial z}\right|_{L^4(\mathcal{M})}\big|T^\sharp\big|_{L^4(\mathcal{M})},$$

which is bounded by

$$c\|\mathbf{v}\|_1\|\widetilde{T}\|_{H^2}\|T^\sharp\|_{H^1} \le c\|U\|\|\widetilde{U}\|_{V_{(2)}}\|U^\sharp\|$$

since $H^1(\mathcal{M}) \subset L^6(\mathcal{M})$, hence the second bound (Eq. (2.60)).

We easily prove Eqs. (2.62) and (2.63) by integration by parts for $U, \widetilde{U}, U^\sharp \in \mathcal{V}$; the relations are then extended by continuity to the other cases, using Eq. (2.60).

To establish the improvement (Eq. (2.61)) of the first inequality (Eq. (2.60)), we observe that $b(U, \widetilde{U}, U^\sharp) = -b(U, U^\sharp, \widetilde{U})$ and consider again the most typical term $\int_{\mathcal{M}} w(\mathbf{v})(\partial U^\sharp/\partial z) \times \widetilde{U}\,\mathrm{d}\mathcal{M}$ that we bound by

$$\big|w(\mathbf{v})\big|_{L^2}\left|\frac{\partial U^\sharp}{\partial z}\right|_{L^6}\big|\widetilde{U}\big|_{L^3}.$$

Remembering that $H^1 \subset L^6$ and $H^{1/2} \subset L^3$, in space dimension 3, we bound this term by

$$c\|\mathbf{v}\|\|U^\sharp\|_{V_{(2)}}\big|\widetilde{U}\big|^{1/2}\|\widetilde{U}\|^{1/2},$$

and Eq. (2.61) follows.

The operator form of the equation. We can write Eq. (2.54) in the form of an evolution equation in the Hilbert space $V'_{(2)}$. For that purpose, we observe that we can associate to the forms a, b, e above, the following operators:

- A linear, continuous from V into V' defined by
$$\langle AU, \widetilde{U}\rangle = a\big(U, \widetilde{U}\big) \quad \forall U, \widetilde{U} \in V,$$
- B bilinear, continuous from $V \times V$ into $V'_{(2)}$ defined by
$$\big\langle B\big(U, \widetilde{U}\big), U^\sharp\big\rangle = b\big(U, \widetilde{U}, U^\sharp\big) \quad \forall U, \widetilde{U} \in V,\ \forall U^\sharp \in V_{(2)},$$
- E linear, continuous from H into itself defined by
$$\langle E(U), \widetilde{U}\rangle = e\big(U, \widetilde{U}\big) \quad \forall U, \widetilde{U} \in H.$$

Since $V_{(2)} \subset V \subset H$, with continuous injections, each space being dense in the next one, we also have the Gelfand–Lions inclusions

$$V_{(2)} \subset V \subset H \subset V' \subset V'_{(2)}. \tag{2.65}$$

With this, we see that Eq. (2.54) is equivalent to the following operator evolution equation

$$\frac{\mathrm{d}U}{\mathrm{d}t} + AU + B(U, U) + E(U) = \ell, \tag{2.66}$$

understood in $V'_{(2)}$ and with ℓ defined in Eq. (2.55). To this equation, we will naturally add an initial condition:

$$U(0) = U_0. \tag{2.67}$$

□

2.2.2. The stationary PEs

We now establish the existence of solutions of the stationary PEs. Besides its intrinsic interest, this result will be needed in the next section for the study of the time-dependent case.

The equations to be considered are the same as Eq. (2.54), with the only difference that the derivatives $\partial \mathbf{v}/\partial t$, $\partial T/\partial t$, and $\partial S/\partial t$ are removed, and that the source terms $F_{\mathbf{v}}$, F_T, F_S are given independent of time t.

The weak formulation proceeds as before:

$$\begin{aligned} &\text{Given } F = (F_{\mathbf{v}}, F_T, F_S) \text{ in } H \left(\text{or } L^2(\mathcal{M})^4\right), \text{ and} \\ &g = (g_{\mathbf{v}}, g_T) \quad \text{in } L^2(\Gamma_{\mathrm{i}})^3, \text{ find } U = (\mathbf{v}, T, S) \in V, \text{ such that} \\ &a\big(U, \widetilde{U}\big) + b\big(U, U, \widetilde{U}\big) + e\big(U, \widetilde{U}\big) = \ell\big(\widetilde{U}\big), \text{ for every } \widetilde{U} \in V_{(2)}; \end{aligned} \tag{2.68}$$

a, b, e, and ℓ are the same as above.

We have the following result.

THEOREM 2.1. *We are given $F = (F_{\mathbf{v}}, F_T, F_S)$ in $L^2(\mathcal{M})^4$ (or in H) and $g = (g_{\mathbf{v}}, g_T)$ in $L^2(\Gamma_{\mathrm{i}})^3$; then problem (Eq. (2.68)) possesses at least one solution $U \in V$ such that*

$$\|U\| \leq \frac{1}{c_1} \|\ell\|_{V'}. \tag{2.69}$$

PROOF. The proof of existence is done by Galerkin method, a priori estimates and passage to the limit. The proof is essentially standard, but we give the details because of some specificities in this case.

We consider a family of elements $\{\Phi_j\}_j$ of $V_{(2)}$, which is free and total in V ($V_{(2)}$ is dense in V); for each $m \in \mathbb{N}$, we look for an approximate solution of Eq. (2.68), $U_m = \sum_{j=1}^{m} \xi_{jm} \Phi_j$, such that

$$a(U_m, \Phi_k) + b(U_m, U_m, \Phi_k) + e(U_m, \Phi_k) = \ell(\Phi_k), \quad k = 1, \ldots, m. \tag{2.70}$$

The existence of U_m is shown below. An a priori estimate on U_m is obtained by multiplying each equation (Eq. (2.70)) by ξ_{km} and summing for $k = 1, \ldots, m$. This amounts to replacing Φ_k by U_m in Eq. (2.70); since $b(U_m, U_m, U_m) = 0$ by Lemma 2.1, we obtain

$$a(U_m, U_m) = \ell(U_m)$$

and, with Eq. (2.57),

$$c_1 \|U_m\|^2 \leq \|\ell\|_{V'} \|U_m\|,$$
$$\|U_m\| \leq \frac{1}{c_1} \|\ell\|_{V'}. \tag{2.71}$$

From Eq. (2.71), we see that there exists $U \in V$ and a subsequence $U_{m'}$ such that $U_{m'}$ converges weakly to U as $m' \to \infty$. Since we cannot replace $\widetilde{U}$ by U in Eq. (2.68), it is useful to notice that

$$\|U\| \leq \liminf_{m' \to \infty} \|U_{m'}\| \leq \frac{1}{c_1} \|\ell\|_{V'},$$

so that Eq. (2.69) is satisfied. Then, we pass to the limit in Eq. (2.70) written with m', and k fixed less than or equal to m'. We observe below that

$$b(U_{m'}, U_{m'}, \Phi_k) \to b(U, U, \Phi_k) \tag{2.72}$$

so that at the limit, U satisfies Eq. (2.68) for $\widetilde{U} = \Phi_k$, k fixed arbitrary; hence Eq. (2.68) is valid for any $\widetilde{U}$ linear combination of Φ_k and, by continuity (Lemma 2.1), for $\widetilde{U} \in V_{(2)}$.

The proof is complete after we prove the results used above. □

Convergence of the b term. To prove Eq. (2.72), we first observe, with Eq. (2.63), that $b(U_{m'}, U_{m'}, \Phi_k) = -b(U_{m'}, \Phi_k, U_{m'})$. We also observe that each component of $U_{m'}$ converges weakly in $H^1(\mathcal{M})$ to the corresponding component of Φ_k. Therefore, by compactness, the convergence takes place in $H^{3/4}(\mathcal{M})$ strongly; by Sobolev embedding, $H^{3/4}(\mathcal{M}) \subset L^4(\mathcal{M})$ in dimension 3, and the convergence holds in $L^4(\mathcal{M})$ strongly. Writing $\Phi_k = \Phi = (\mathbf{v}_\Phi, T_\Phi, S_\Phi)$, the typical most problematic term is

$$\int_{\mathcal{M}} w(\mathbf{v}_{m'}) \frac{\partial \mathbf{v}_\Phi}{\partial z} \mathbf{v}_{m'} \, \mathrm{d}\mathcal{M}. \tag{2.73}$$

Since $\operatorname{div} \mathbf{v}_{m'}$ converges weakly to $\operatorname{div} \mathbf{v}$ in $L^2(\mathcal{M})$, $w(\mathbf{v}_{m'})$ converges weakly in $L^2(\mathcal{M})$ to $w(\mathbf{v})$; $\mathbf{v}_{m'}$ converges strongly to $\mathbf{v}$ in $L^4(\mathcal{M})$ as observed before, and since $\partial \mathbf{v}_\Phi / \partial z$ belongs to $L^4(\mathcal{M})$, the term above converges to the corresponding term where $\mathbf{v}_{m'}$ is replaced by $\mathbf{v}(U = (\mathbf{v}, T, S))$, hence Eq. (2.69).

Existence of U_m. Equation (2.70) amounts to a system of m nonlinear equations for the components of the vector $\xi = (\xi_1, \ldots, \xi_m)$, where we have written $\xi_{jm} = \xi_j$ for simplicity. Existence follows from the following consequence of the Brouwer fixed point theorem. (see LIONS [1969]).

LEMMA 2.2. *Let $\mathcal{F}$ be a continuous mapping of $\mathbb{R}^m$ into itself such that*

$$\big[\mathcal{F}(\xi), \xi\big] > 0 \quad \textit{for } [\xi] = k, \textit{ for some } k > 0, \tag{2.74}$$

where $[\cdot, \cdot]$ and $[\cdot]$ are the scalar product and norm in $\mathbb{R}^m$.

Then there exists $\xi \in \mathbb{R}^m$ with $[\xi] < k$ such that $\mathcal{F}(\xi) = 0$.

PROOF. If $\mathcal{F}$ never vanishes, then $\mathcal{G} = -k\mathcal{F}(\xi)/[\mathcal{F}(\xi)]$ is continuous on $\mathbb{R}^m$, and we can apply the Brouwer fixed point theorem to $\mathcal{G}$, which maps the ball $\mathcal{C}$ centered at 0 of radius k into itself. Then $\mathcal{G}$ has a fixed point ξ_0 in $\mathcal{C}$ and we have

$$\big[\mathcal{G}(\xi_0)\big] = [\xi_0] = k,$$
$$\big[\mathcal{G}(\xi_0), \xi_0\big] = -k\frac{[\mathcal{F}(\xi_0), \xi_0]}{[\mathcal{F}(\xi_0)]} = [\xi_0]^2.$$

This contradicts the hypothesis (Eq. (2.74)) on $\mathcal{F}$; the lemma is proven. □

We apply this lemma to Eq. (2.70) as follows: $\mathcal{F} = (\mathcal{F}_1, \ldots, \mathcal{F}_m)$, with

$$\mathcal{F}_k(\xi) = [\mathcal{F}(U_m), \Phi_k] = a(U_m, \Phi_k) + b(U_m, U_m, \Phi_k) + e(U_m, \Phi_k) - l(\Phi_k). \tag{2.75}$$

The space $\mathbb{R}^m$ is equipped with the usual Euclidean scalar product so that

$$\begin{aligned}\big[\mathcal{F}(\xi), \xi\big] &= \sum_{k=1}^{m} \mathcal{F}_k(\xi)\xi_k \\ &= a(U_m, U_m) + b(U_m, U_m, U_m) - l(U_m) \\ &\geq \big(\text{with Eqs. (2.57), (2.58), (2.62) and Schwarz' inequality}\big) \\ &\geq c_1 \|U_m\|^2 - \|\ell\|_{V'} \|U_m\|. \end{aligned} \tag{2.76}$$

Since the last expression converges to $+\infty$ as $\|U_m\| \sim [\xi]$ converges to $+\infty$, there exists $k > 0$ such that Eq. (2.74) holds. The existence of U_m follows.

REMARK 2.9. A perusal of the proof of Theorem 2.1 shows that we proved the following more general result.

LEMMA 2.3. *Let V and W be two Hilbert spaces with $W \subset V$, the injection being continuous. Assume that $\bar{a}$ is bilinear continuous coercive on V, and that $\bar{b}$ is trilinear*

continuous on $V \times W \times V$, $V \times V \times W$, *and continuous on* $V_w \times V_w \times W$, *where* V_w *is* V *equipped with the weak topology. Furthermore,*

$$b\big(U, \widetilde{U}, U^\sharp\big) = -b\big(U, U^\sharp, \widetilde{U}\big) \quad \text{if } U, \widetilde{U}, U^\sharp \in V \text{ and } \widetilde{U} \text{ or } U^\sharp \in W.$$

Then, for $\bar{l}$ *given in* V', *there exists at least one solution* U *of*

$$\bar{a}\big(U, \widetilde{U}\big) + \bar{b}\big(U, U, \widetilde{U}\big) = \bar{l}\big(\widetilde{U}\big) \quad \forall \widetilde{U} \in W, \tag{2.77}$$

which satisfies

$$\bar{a}(U, U) \leq \bar{\ell}(U). \tag{2.78}$$

Lemma 2.3 will be useful in the next section.

2.3. Existence of weak solutions for the PEs of the ocean

In this section, we establish the existence, for all time, of weak solutions for the equations of the ocean. The main result is Theorem 2.2 given at the end of the section.

We consider the PEs in their formulation (Eq. (2.54)), i.e., with the notations of Section 2.2:

$$\begin{aligned} &\text{Given } t_1 > 0,\ U_0 \text{ in } H,\ F = (F_{\mathbf{v}}, F_T, F_S) \text{ in } L^2(0, t_1; H), \text{ and} \\ &g = (g_{\mathbf{v}}, g_T) \text{ in } L^2\big(0, t_1; L^2(\Gamma_i)\big)^3, \text{ to find} \\ &U \in L^\infty(0, t_1; H) \cap L^2(0, t_1; V) \text{ such that} \\ &\Big(\frac{\mathrm{d}}{\mathrm{d}t}U, \widetilde{U}\Big) + a\big(U, \widetilde{U}\big) + b\big(U, U, \widetilde{U}\big) + e\big(U, \widetilde{U}\big) = \ell\big(\widetilde{U}\big) \quad \forall \widetilde{U} \in V_{(2)}, \end{aligned} \tag{2.79}$$

$$U(0) = U_0. \tag{2.80}$$

Alternatively, and as explained in the previous section (see Eqs. (2.66) and (2.67)), we can write Eqs. (2.79) and (2.80) in the form of an operator evolution equation

$$\frac{\mathrm{d}U}{\mathrm{d}t} + AU + B(U, U) + E(U) = \ell, \tag{2.81}$$

$$U(0) = U_0. \tag{2.82}$$

To establish the existence of weak solutions of this problem, we proceed by finite differences in time.[2]

[2]At this level of generality, it has not been possible to prove the existence of weak solutions to the PEs by any other classical method for parabolic equations. In particular, the proofs in the articles by Lions, Temam and Wang [1992a,b, 1995] based on the Galerkin method assume the H^2–regularity of the solutions of the GFD–Stokes problem, and this result is not available at this level of generality. We recall that the whole Section 4 is devoted to this regularity question.

Finite differences in time. Given $t_1 > 0$, which is arbitrary, we consider N an arbitrary integer and introduce the time-step $k = \Delta t = t_1/N$. By time discretization of Eqs. (2.79) and (2.80), we are naturally led to define a sequence of elements of V, U^n, $0 \le n \le N$, defined by

$$U^0 = U_0, \tag{2.83}$$

and then, recursively for $n = 1, \dots, N$ by

$$\begin{aligned}&\frac{1}{\Delta t}\big(U^n - U^{n-1}, \widetilde{U}\big)_H + a\big(U^n, \widetilde{U}\big) + b\big(U^n, U^n, \widetilde{U}\big) + e\big(U^n, \widetilde{U}\big)\\ &\quad = \ell^n\big(\widetilde{U}\big) \quad \forall \widetilde{U} \in V_{(2)}.\end{aligned} \tag{2.84}$$

Here, $\ell^n \in V'$ is given by

$$\ell^n\big(\widetilde{U}\big) = \frac{1}{\Delta t}\int_{(n-1)\Delta t}^{n\Delta t} \ell\big(t; \widetilde{U}\big)\,\mathrm{d}t, \tag{2.85}$$

where $\ell(t; \widetilde{U})$ is defined exactly as in Eq. (2.54), the dependence of ℓ on t reflecting now the dependence on t of F, $g_{\mathbf{v}}$, and g_T.

The existence, for all n, of $U^n \in V$ solution of Eq. (2.84) follows from Lemma 2.3, Eq. (2.84) being the same as Eq. (2.77); the notations are obvious, and the verification of the hypotheses of the lemma is easy; furthermore, by Eq. (2.78), and after multiplication by $2\Delta t$:

$$\begin{aligned}&\big|U^n\big|_H^2 - \big|U^{n-1}\big|_H^2 + \big|U^n - U^{n-1}\big|_H^2 + 2\Delta t a\big(U^n, U^n\big)\\ &\quad \le 2\Delta t \ell^n\big(U^n\big).\end{aligned} \tag{2.86}$$

For Eq. (2.86), we also used Eq. (2.58) and the elementary relation

$$2\big(\widetilde{U} - U^\sharp, \widetilde{U}\big)_H = \big|\widetilde{U}\big|_H^2 - \big|U^\sharp\big|_H^2 + \big|\widetilde{U} - U^\sharp\big|_H^2 \quad \forall \widetilde{U}, U^\sharp \in H. \tag{2.87}$$

A priori estimates. We now proceed and derive a priori estimates for the U^n and then for some associated approximate functions.

Using Eqs. (2.85), (2.54), and Schwarz' inequality, we bound $\Delta t\ \ell^n(U^n)$ by $\Delta t^{1/2}$ $(\xi^n)^{1/2} \times \|U^n\|$ with

$$\begin{aligned}\xi^n &= \int_{(n-1)\Delta t}^{n\Delta t} \xi(t)\,\mathrm{d}t,\\ \xi(t) &= c_1'\bigg[\big|F(t)\big|_H^2 + \int_{\mathcal{M}} \big|1 + \beta_T T_{\mathrm{r}}(t) - \beta_S S_{\mathrm{r}}(t)\big|^2\,\mathrm{d}\mathcal{M}\\ &\qquad + \int_{\Gamma_{\mathrm{i}}} \big(\big|g_{\mathbf{v}}(t)\big|^2 + \big|g_T(t)\big|^2\big)\,\mathrm{d}\Gamma_i\bigg],\end{aligned} \tag{2.88}$$

where c_1' is an absolute constant related to c_0 (see Eq. (2.52)). Hence, using also Eq. (2.57), we infer from Eq. (2.86) that

$$\begin{aligned}&\left|U^n\right|_H^2 - \left|U^{n-1}\right|_H^2 + \left|U^n - U^{n-1}\right|_H^2 + 2\Delta t c_1 \left\|U^n\right\|^2 \\ &\quad \le 2\Delta t^{1/2}\left(\xi^n\right)^{1/2}\left\|U^n\right\| \\ &\quad \le \Delta t c_1 \left\|U^n\right\|^2 + c_1^{-1}\xi^n.\end{aligned}$$

Hence,

$$\begin{aligned}&\left|U^n\right|_H^2 - \left|U^{n-1}\right|_H^2 + \left|U^n - U^{n-1}\right|_H^2 + \Delta t c_1 \left\|U^n\right\|^2 \\ &\quad \le c_1^{-1}\xi^n \quad \text{for } n = 1, \ldots, N.\end{aligned} \tag{2.89}$$

Summing all these relations for $n = 1, \ldots, N$, we find

$$\left|U^N\right|_H^2 + \sum_{n=1}^{N}\left[\left|U^n - U^{n-1}\right|_H^2 + \Delta t c_1 \left\|U^n\right\|^2\right] \le \kappa_1, \tag{2.90}$$

with

$$\kappa_1 = |U_0|^2 + \frac{1}{c_1}\int_0^{t_1} \xi(t)\,\mathrm{d}t.$$

Summing the relations (Eq. (2.89)) for $n = 1, \ldots, m$, with m fixed, $1 \le m \le N$, we obtain as well

$$\left|U^m\right|_H^2 \le \kappa_1 \quad \forall m = 0, \ldots, N. \tag{2.91}$$

Approximate functions. The subsequent steps follow closely the proof in Temam [1977], Chapter 3, Section 4, for the Navier-Stokes equations, and we will skip many details.[3]

We first introduce the approximate functions defined as follows on $(0, t_1)$ $(k = \Delta t)$; for the sake of simplicity, we assume that $U_0 \in V$ instead of H, but this is not necessary:

$$\begin{aligned}&U_k : (0, t_1) \mapsto V, \quad U_k(t) = U^n, \quad t \in \big((n-1)k, nk\big), \\ &\ell_k : (0, t_1) \mapsto V', \quad \ell_k(t) = \ell^n, \quad t \in \big((n-1)k, nk\big), \\ &\widetilde{U}_k : (0, t_1) \mapsto V, \\ &\widetilde{U}_k \text{ is continuous, linear on each interval } \big((n-1)k, nk\big) \text{ and} \\ &\widetilde{U}_k(nk) = U^n, \quad n = 0, \ldots, N.\end{aligned}$$

[3] The proof given here would apply to the Navier-Stokes equations in space dimension $d \ge 4$; it extends the proof given in Temam [1977], which is only valid for the Navier-Stokes equations in dimension $d = 2$ or 3.

An easy computation (see TEMAM [1977]) shows that

$$\left|U_k - \widetilde{U}_k\right|_{L^2(0,t_1;H)} \leq \left(\frac{k}{3}\right)^{1/2} \left(\sum_{n=1}^{N} \left|U^n - U^{n-1}\right|_H^2\right)^{1/2}, \tag{2.92}$$

and we infer from Eqs. (2.90) and Eq. (2.91) that

$$\begin{aligned} &U_k \text{ and } \widetilde{U}_k \text{ are bounded independently} \\ &\text{of } \Delta t \text{ in } L^\infty(0, t_1; H) \text{ and } L^2(0, t_1; V). \end{aligned} \tag{2.93}$$

We infer from Eqs. (2.92) and (2.93) that there exists $U \in L^\infty(0, t_1; H) \cap L^2(0, t_1; V)$, and a subsequence $k' \to 0$, such that, as $k' \to 0$,

$$\begin{aligned} &U_{k'} \text{ and } \widetilde{U}_{k'} \rightharpoonup U \text{ in } L^\infty(0, t_1; H) \text{ weak star} \\ &\text{and in } L^2(0, t_1; V) \text{ weakly}, \end{aligned} \tag{2.94}$$

$$U_{k'} - \widetilde{U}_{k'} \to 0 \text{ in } L^2(0, t_1; H) \text{ strongly}. \tag{2.95}$$

Further a priori estimates and compactness. With the notations above and those used for Eqs. (2.66) and (2.67) (or Eqs. (2.81) and (2.82)), we see that the scheme (Eq. (2.84)) can be rewritten as

$$\frac{\mathrm{d}\widetilde{U}_k}{\mathrm{d}t} + AU_k + B(U_k, U_k) + E(U_k) = \ell_k, \quad 0 < t < t_1, \tag{2.96}$$

$$\widetilde{U}_k(0) = U_0. \tag{2.97}$$

From Eqs. (2.93) and (2.61), we see that $B(U_k, U_k)$ is bounded in $L^{4/3}(0, t_1; V'_{(2)})$; since the other terms in Eq. (2.96) are bounded in $L^2(0, t_1; V')$ independently of k, we conclude that

$$\frac{\mathrm{d}\widetilde{U}_k}{\mathrm{d}t} \text{ is bounded in } L^{4/3}\big(0, t_1; V'_{(2)}\big). \tag{2.98}$$

We then infer from Eq. (2.93) and the Aubin compactness theorem (see, e.g., TEMAM [1977]) that as $k' \to 0$,

$$\widetilde{U}_{k'} \to U \text{ in } L^2(0, t_1; H) \text{ strongly}, \tag{2.99}$$

and the same is true for U_k because of Eq. (2.95).

Passage to the limit. The passage to the limit $k' \to 0$ ($k = \Delta t$) follows now closely that of Theorem 4.1, Chapter 2 in TEMAM [1977], we skip the details.

We consider $\widetilde{U} \in \mathcal{V}$ (which is dense in $V_{(2)}$, see Eq. (2.59)) and a scalar function ψ in $\mathcal{C}^1([0, t_1])$ such that $\psi(t_1) = 0$. We take the scalar product in H of Eq. (2.96) with $\widetilde{U}\psi$, integrate from 0 to t_1 and integrate by parts the first term, and we arrive at

$$
\begin{aligned}
&-\int_0^{t_1} \big(\widetilde{U}_k, \widetilde{U}\big)_H \psi' \,\mathrm{d}t + \int_0^{t_1} \big[a\big(U_k, \widetilde{U}\big) + b\big(U_k, U_k, \widetilde{U}\big) + e\big(U_k, \widetilde{U}\big)\big]\psi \,\mathrm{d}t \\
&\quad = \big(U_0, \widetilde{U}\big)_H \psi(0) + \int_0^{t_1} \ell_k\big(\widetilde{U}\big)\psi \,\mathrm{d}t. \qquad (2.100)
\end{aligned}
$$

We can pass to the limit in Eq. (2.100) for the sequence k'; for the b term, we proceed somehow as for Eq. (2.72). For the nonlinear term, we write

$$
\int_0^{t_1} b\big(U_k, U_k, \widetilde{U}\big)\psi \,\mathrm{d}t = -\int_0^{t_1} b\big(U_k, \widetilde{U}, U_k\big)\psi \,\mathrm{d}t,
$$

and considering the typical most problematic term, we show that as $k' \to 0$,

$$
\int_0^{t_1}\int_{\mathcal{M}} w(\mathbf{v}_k)\frac{\partial \widetilde{T}}{\partial z} T_{k'}\psi \,\mathrm{d}\mathcal{M}\,\mathrm{d}t \to \int_0^{t_1}\int_{\mathcal{M}} w(\mathbf{v})\frac{\partial \widetilde{T}}{\partial z} T\psi \,\mathrm{d}\mathcal{M}\,\mathrm{d}t; \qquad (2.101)
$$

this follows from the fact that $\operatorname{div} \mathbf{v}_{k'}$ converges to $\operatorname{div} \mathbf{v}$ weakly in $L^2(\mathcal{M} \times (0, t_1))$ that $T_{k'}$ converges to T strongly in $L^2(\mathcal{M} \times (0, t_1))$, and $\psi\, \partial\widetilde{T}/\partial z$ belongs to $L^\infty(\mathcal{M} \times (0, t_1))$.

From this, we conclude that U satisfies

$$
\begin{aligned}
&-\int_0^{t_1} \big(U, \widetilde{U}\big)_H \psi' \,\mathrm{d}t + \int_0^{t_1} \big[a\big(U, \widetilde{U}\big) + b\big(U, U, \widetilde{U}\big) + e\big(U, \widetilde{U}\big)\big]\psi \,\mathrm{d}t \\
&\quad = \big(U_0, \widetilde{U}\big)\psi(0) + \int_0^{t_1} \ell\big(\widetilde{U}\big)\psi \,\mathrm{d}t \qquad (2.102)
\end{aligned}
$$

for all $\widetilde{U}$ in $\mathcal{V}$ and all ψ of the indicated type. Also, by continuity (Lemma 2.1), Eq. (2.101) is valid as well for all $\widetilde{U}$ in $V_{(2)}$ since $\mathcal{V}$ is dense in $V_{(2)}$ by Eq. (2.59).

It is then standard to infer from Eq. (2.101) that U is solution of Eqs. (2.79) and (2.80); this leads us to the main result of this section.

THEOREM 2.2. *The domain $\mathcal{M}$ is as before. We are given $t_1 > 0$, U_0 in H and $F = (F_{\mathbf{v}}, F_T, F_S)$ in $L^2(0, t_1; H)$ or $L^2(0, t_1; L^2(\mathcal{M})^4)$; $g = (g_{\mathbf{v}}, g_T)$ is given in $L^2(0, t_1; L^2(\Gamma_i)^3)$. Then there exists*

$$
U \in L^\infty(0, t_1; H) \cap L^2(0, t_1; V), \qquad (2.103)
$$

which is solution of Eqs. (2.79) and (2.80) (or Eqs. (2.81) and (2.82)); furthermore, U is weakly continuous from $[0, t_1]$ into H.

2.4. The PEs of the atmosphere

In this section, we briefly describe the PEs of the atmosphere, introduce their mathematical (weak) formulation, and state without proof the existence of weak solutions; the proof is essentially the same as for the ocean.

We start from the conservation equations similar to Eqs. (2.1)–(2.5). In fact, Eqs. (2.1) and (2.2) are the same; the equation of energy conservation is slightly different from Eq. (2.3) because of the compressibility of the air; the state equation is that of perfect gas instead of Eq. (2.5); finally, instead of the concentration of salt in the water, we consider the amount q of water in air. Hence, we have

$$\rho\frac{\mathrm{d}\mathbf{V}_3}{\mathrm{d}t} + 2\rho\boldsymbol{\Omega}\times\mathbf{V}_3 + \nabla_3 p + \rho\mathbf{g} = D, \tag{2.104}$$

$$\frac{\mathrm{d}\rho}{\mathrm{d}t} + \rho\,\mathrm{div}_3\,\mathbf{V}_3 = 0, \tag{2.105}$$

$$c_p\frac{\mathrm{d}T}{\mathrm{d}t} - \frac{RT}{p}\frac{\mathrm{d}p}{\mathrm{d}t} = Q_T, \tag{2.106}$$

$$\frac{\mathrm{d}q}{\mathrm{d}t} = Q_q, \tag{2.107}$$

$$p = R\rho T. \tag{2.108}$$

Here, D, Q_T, and Q_q contain the dissipation terms. As we said, the difference between Eqs. (2.106) and (2.3) is due to the compressibility of the air; in Eq. (2.106), $c_p > 0$ is the specific heat of the air at constant pressure and R is the specific gas constant for the air; Eq. (2.108) is the equation of state for the air.

The hydrostatic approximation. We decompose $\mathbf{V}_3$ into its horizontal and vertical components as in (Eq. (2.8)), $\mathbf{V}_3 = (\mathbf{v}, w)$, and we use the approximation (Eq. (2.10)) of $\mathrm{d}/\mathrm{d}t$. Also, as for the ocean, we use the hydrostatic approximation, replacing the equation of conservation of vertical momentum (third equation (2.104)) by the hydrostatic equation (Eq. (2.23)). We find

$$\frac{\partial\mathbf{v}}{\partial t} + \nabla_{\mathbf{v}}\mathbf{v} + w\frac{\partial\mathbf{v}}{\partial z} + \frac{1}{\rho}\nabla p + 2\Omega\sin\theta\,\mathbf{k}\times\mathbf{v} - \mu_{\mathbf{v}}\Delta\mathbf{v} - \nu_{\mathbf{v}}\frac{\partial^2\mathbf{v}}{\partial z^2} = 0, \tag{2.109}$$

$$\frac{\partial p}{\partial z} = -\rho g, \tag{2.110}$$

$$\frac{\partial\rho}{\partial t} + \rho\left(\nabla\mathbf{v} + \frac{\partial w}{\partial z}\right) + \mathbf{v}\nabla\rho + w\frac{\partial\rho}{\partial z} = 0, \tag{2.111}$$

$$\frac{\partial T}{\partial t} + \nabla_{\mathbf{v}}T + w\frac{\partial T}{\partial z} - \mu_T\Delta T - \nu_T\frac{\partial^2 T}{\partial z^2} - \frac{RT}{p}\frac{\mathrm{d}p}{\mathrm{d}t} = Q_T, \tag{2.112}$$

$$\frac{\partial q}{\partial t} + \nabla_{\mathbf{v}}q + w\frac{\partial q}{\partial z} - \mu_q\Delta q - \nu_q\frac{\partial^2 q}{\partial z^2} = 0, \tag{2.113}$$

$$p = R\rho T. \tag{2.114}$$

The right-hand side of Eq. (2.112), which is different from Q_T in Eq. (2.106) now represents the solar heating.

Change of vertical coordinate. Since ρ does not vanish, the hydrostatic equation (Eq. (2.110)) implies that p is a strictly decreasing function of z, and we are thus allowed to use p as the vertical coordinate; hence, in spherical geometry, the independent variables are now φ, θ, p, and t. By an abuse of notation, we still denote by $\mathbf{v}$, T, q, and ρ these functions expressed in the φ, θ, p, and t variables. We also denote by ω the vertical component of the wind, and one can show (see, e.g., HALTINER and WILLIAMS [1980]) that

$$\omega = \frac{\mathrm{d}p}{\mathrm{d}t} = \frac{\partial p}{\partial t} + \nabla_{\mathbf{v}} p + w\frac{\partial p}{\partial z}; \tag{2.115}$$

in Eq. (2.115), p is a dependent variable expressed as a function of φ, θ, z, and t.

In this context, the PEs of the atmosphere become

$$\frac{\partial \mathbf{v}}{\partial t} + \nabla_{\mathbf{v}}\mathbf{v} + \omega\frac{\partial \mathbf{v}}{\partial p} + 2\Omega \sin\theta\, \mathbf{k} \times \mathbf{v} + \nabla\Phi - L_{\mathbf{v}}\mathbf{v} = F_{\mathbf{v}}, \tag{2.116}$$

$$\frac{\partial \Phi}{\partial p} + \frac{RT}{p} = 0, \tag{2.117}$$

$$\operatorname{div}\mathbf{v} + \frac{\partial \omega}{\partial p} = 0, \tag{2.118}$$

$$\frac{\partial T}{\partial t} + \nabla_{\mathbf{v}} T + \omega\frac{\partial T}{\partial p} - \frac{R\overline{T}}{c_p p}\omega - L_T T = F_T, \tag{2.119}$$

$$\frac{\partial q}{\partial t} + \nabla_{\mathbf{v}} q + \omega\frac{\partial q}{\partial p} - L_q q = F_q, \tag{2.120}$$

$$p = R\rho T. \tag{2.121}$$

We have denoted by $\Phi = gz$ the geopotential (z is now a function of φ, θ, p, and t); $L_{\mathbf{v}}$, L_T, and L_q are the Laplace operators, with suitable eddy viscosity coefficients, expressed in the φ, θ, and p variables. Hence, for example,

$$L_{\mathbf{v}}\mathbf{v} = \mu_{\mathbf{v}}\Delta\mathbf{v} + \nu_{\mathbf{v}}\frac{\partial}{\partial p}\left[\left(\frac{gp}{R\overline{T}}\right)^2 \frac{\partial \mathbf{v}}{\partial p}\right], \tag{2.122}$$

with similar expressions for L_T and L_q. Note that F_T corresponds to the heating of the sun, whereas $F_{\mathbf{v}}$ and F_q, which vanish in reality, are added here for mathematical generality. In Eq.(2.119), T has been replaced by $\overline{T}$ in the term $RT\omega/c_p p$ (see in LIONS, TEMAM and WANG [1995] a better approximation of $RT\omega/c_p p$ involving an additional term). With additional precautions, and using the maximum principle for the temperature as in EWALD and TEMAM [2001], we could keep the exact term $RT\omega/c_p p$.

The change of variable gives for $\partial^2\mathbf{v}/\partial z^2$ a term different from the coefficient of $\nu_{\mathbf{v}}$. The expression above is a simplified form of this coefficient; the simplification is legitimate because $\nu_{\mathbf{v}}$ is a very small coefficient; in particular, T has been replaced by $\overline{T}$ (known), which is an average value of the temperature.

Pseudogeometrical domain. For physical and mathematical reasons, we do not allow the pressure to go to zero and assume that $p \geq p_0$, with $p_0 > 0$ "small." Physically, in the very high atmosphere (p being very small), the air is ionized and the equations above are not valid anymore; mathematically, with $p > p_0$, we avoid the appearance of singular terms as, for example, in the expressions of $L_{\mathbf{v}}$, L_T, and L_q. The pressure is then restricted to an interval $p_0 < p < p_1$, where p_1 is a value of the pressure smaller in average than the pressure on earth so that the isobar $p = p_1$ is slightly above the earth and the isobar $p = p_0$ is an isobar high in the sky. We study the motion of the air between these two isobars; as we said, for $p < p_0$, we would need a different set of equations, and for the "thin" portion of air between the earth and the isobar $p = p_1$, another specific simplified model would be necessary.

For the whole atmosphere, the domain is

$$\mathcal{M} = \big\{(\varphi, \theta, p),\, p_0 < p < p_1\big\},$$

and its boundary consists first of an upper part Γ_{u}, $p = p_0$ and a lower part $p = p_1$ which is divided into two parts: Γ_{i} the part of $p = p_1$ at the interface with the ocean, and Γ_{e} the part of $p = p_1$ above the Earth.

Boundary conditions. Typically the boundary conditions are as follows:

On the top of the atmosphere Γ_{u} ($p = p_0$):

$$\frac{\partial \mathbf{v}}{\partial p} = 0, \quad \omega = 0, \quad \frac{\partial T}{\partial p} = 0, \quad \frac{\partial q}{\partial p} = 0. \tag{2.123}$$

Above the Earth on Γ_{e}:

$$\begin{aligned} &\mathbf{v} = 0, \quad \omega = 0, \\ &\nu_T \frac{\partial T}{\partial p} + \alpha_T (T - T_{\mathrm{e}}) = 0, \\ &\frac{\partial q}{\partial p} = g_q. \end{aligned} \tag{2.124}$$

Above the ocean on Γ_{i}:

$$\begin{aligned} &\nu_{\mathbf{v}} \left(\frac{gp}{R\overline{T}} \right)^2 \frac{\partial \mathbf{v}}{\partial p} + \alpha_{\mathbf{v}} \big(\mathbf{v} - \mathbf{v}^{\mathrm{s}}\big) = \tau_{\mathbf{v}}, \quad \omega = 0, \\ &\nu_T \left(\frac{gp}{R\overline{T}} \right)^2 \frac{\partial T}{\partial p} + \alpha_T \big(T - T^{\mathrm{s}}\big) = 0, \\ &\frac{\partial q}{\partial p} = g_q. \end{aligned} \tag{2.125}$$

REMARK 2.10. The equations on Γ_{i} (Eq. (2.125)) are similar to those on Γ_{i} for the ocean (Eq. (2.35)), with different values of the coefficients $\nu_{\mathbf{v}}$, ν_T, ...; comparison between these two sets of boundary conditions is made in Section 2.5 devoted to the CAO system.

In Eq. (2.124), T_e is the (given) temperature on the earth, and $\mathbf{v}^s$ and T^s are the (given) velocity and temperature of the sea. The boundary conditions (Eq. (2.123)) on Γ_u are physically reasonable boundary conditions; they can be replaced by other boundary conditions (e.g., $\mathbf{v} = 0$), which can be treated mathematically in a similar manner.

Regional problems and beta-plane approximation. It is reasonable to study regional problems, in particular at midlatitudes, and in this case, we use the beta-plane approximation. In this case, as for the ocean, we use the Cartesian coordinates denoted by x, y, z or x_1, x_2, x_3, and $\Omega = (f_0 + \beta y)\mathbf{k}$. The equations are exactly the same as Eqs. (2.115)–(2.122), but now $\Delta = \partial^2/\partial x^2 + \partial^2/\partial y^2$, ∇ is the usual nabla vector $(\partial/\partial x, \partial/\partial y)$ and $\nabla_{\mathbf{v}} = u\partial/\partial x + v\partial/\partial y$, $\mathbf{v} = (u, v)$. The domain $\mathcal{M}$ is now some portion of the whole atmosphere:

$$\mathcal{M} = \big\{(\varphi, \theta, p), (\theta, \varphi) \in \Gamma_i \cup \Gamma_e, p_0 < p < p_1\big\},$$

where $\Gamma_i \cup \Gamma_e$ is only part of the isobar $p = p_1$. The boundary of $\mathcal{M}$ consists of Γ_u, Γ_i, and Γ_e defined as before and of a lateral boundary

$$\Gamma_\ell = \big\{(\varphi, \theta, p), p_0 < p < p_1, (\varphi, \theta) \in \partial\Gamma_u\big\}.$$

The boundary conditions are the same as before on Γ_u, Γ_i, and Γ_e, and on Γ_ℓ, the conditions would be as follows:

Boundary conditions on Γ_ℓ

$$\mathbf{v} = 0, \quad \omega = 0, \quad \frac{\partial T}{\partial \mathbf{n}_T} = 0, \quad \frac{\partial q}{\partial \mathbf{n}_q} = 0. \tag{2.126}$$

Here, $\partial/\partial\mathbf{n}_T$ and $\partial/\partial\mathbf{n}_q$ are defined as in Eq. (2.37). Compared with Eq. (2.37), $\mathbf{v} = 0$, $\omega = 0$, is not a physically satisfactory boundary condition; we would rather assume that $(\mathbf{v}, \omega)$ has a nonzero prescribed value; however, in the mathematical treatment of this boundary condition, we would then recover $(\mathbf{v}, \omega) = 0$ after removing a background flow; the necessary modifications are minor.

Below we only discuss the regional case.

Prognostic and diagnostic variables. The unknown functions are regrouped in two sets: the prognostic variables $U = (\mathbf{v}, T, q)$ for which and initial value problem will be defined, and the diagnostic variables ω, ρ, Φ $(= gz)$, which can be defined at each instant of time as functions (functionals) of the prognostic variables, using the equations and boundary conditions. In fact, ω is determined in terms of $\mathbf{v}$ very much as in the case of the ocean:

$$\omega = \omega(\mathbf{v}) = -\int_{p_0}^{p} \operatorname{div} \mathbf{v}\, \mathrm{d}p', \tag{2.127}$$

with

$$\int_{p_0}^{p_1} \operatorname{div} \mathbf{v}\, \mathrm{d}p = 0. \tag{2.128}$$

Then ρ is determined by the equation of state (Eq. (2.121)), and Φ is a function of p and T determined by integration of Eq. (2.117)

$$\Phi = \Phi_s + \int_p^{p_1} \frac{RT(p')}{p'}\,\mathrm{d}p'; \tag{2.129}$$

in Eq. (2.129), $\Phi_s = \Phi|_{p=p_1}$ is the geopotential at $p = p_1$, i.e., g times the height of the isobar $p = p_1$. This is an auxiliary unknown, and its introduction has been a crucial step on the basis of the new mathematical formulation of the PEs of the atmosphere in Lions, Temam and Wang [1992a].

Weak formulation of the PEs. For the weak formulation of the PEs, we introduce function spaces similar to those considered for the ocean, namely,

$$\begin{aligned}
&V = V_1 \times V_2 \times V_3, \qquad H = H_1 \times H_2 \times H_3,\\
&V_1 = \left\{\mathbf{v} \in H^1(\mathcal{M})^2, \operatorname{div} \int_{p_0}^{p_1} \mathbf{v}\,\mathrm{d}p = 0, \mathbf{v} = 0 \text{ on } \Gamma_e \cup \Gamma_\ell\right\},\\
&V_2 = V_3 = H^1(\mathcal{M}),\\
&H_1 = \left\{\mathbf{v} \in L^2(\mathcal{M})^2, \operatorname{div} \int_{p_0}^{p_1} \mathbf{v}\,\mathrm{d}p = 0,\right.\\
&\qquad \left.\mathbf{n}_H \cdot \int_{p_0}^{p_1} \mathbf{v}\,\mathrm{d}p = 0 \text{ on } \partial\Gamma_u \text{ (i.e., on } \Gamma_\ell)\right\},\\
&H_2 = H_3 = L^2(\mathcal{M}).
\end{aligned}$$

These spaces are endowed with scalar products similar to those for the ocean:

$$\begin{aligned}
&\big((U, \widetilde{U})\big) = \big((\mathbf{v}, \tilde{\mathbf{v}})\big)_1 + K_T\big((T, \widetilde{T})\big)_2 + K_q\big((q, \tilde{q})\big)_3,\\
&\big((\tilde{\mathbf{v}}, \tilde{\mathbf{v}})\big)_1 = \int_{\mathcal{M}} \left(\nabla\mathbf{v} \cdot \nabla\tilde{\mathbf{v}} + \left(\frac{gp}{R\overline{T}}\right)^2 \frac{\partial\mathbf{v}}{\partial p}\frac{\partial\tilde{\mathbf{v}}}{\partial p}\right)\mathrm{d}\mathcal{M},\\
&\big((T, \widetilde{T})\big)_2 = \int_{\mathcal{M}} \left(\nabla T \cdot \nabla\widetilde{T} + \left(\frac{gp}{R\overline{T}}\right)^2 \frac{\partial T}{\partial p}\frac{\partial\widetilde{T}}{\partial p}\right)\mathrm{d}\mathcal{M} + \int_{\Gamma_i} \alpha_T T\widetilde{T}\,\mathrm{d}\Gamma_i,\\
&\big((q, \tilde{q})\big)_3 = \int_{\mathcal{M}} \left(\nabla q \cdot \nabla\tilde{q} + \left(\frac{gp}{R\overline{T}}\right)^2 \frac{\partial q}{\partial p}\frac{\partial\tilde{q}}{\partial p}\right)\mathrm{d}\mathcal{M} + K'_q \int_{\mathcal{M}} q\tilde{q}\,\mathrm{d}\mathcal{M},\\
&\big(U, \widetilde{U}\big)_H = \int_{\mathcal{M}} \big(\mathbf{v} \cdot \tilde{\mathbf{v}} + K_T T\widetilde{T} + K_q q\tilde{q}\big)\,\mathrm{d}\mathcal{M},\\
&\qquad \|U\| = \big((U, U)\big)^{1/2}, \quad |U|_H = (U, U)_H^{1/2}.
\end{aligned}$$

Here, K_T and K_q are suitable positive constants chosen below, and $K'_q > 0$ is a constant of suitable (physical) dimension. The norm on H is of course equivalent to the L^2-norm

and, thanks to the Poincaré inequality, $\|\cdot\|$ is a Hilbert norm on V; more precisely, there exists a suitable constant $c_0 > 0$ (different from that in Eq. (2.52)) such that

$$|U|_H \le c_0\|U\| \quad \forall U \in V. \tag{2.130}$$

We denote by $\mathcal{V}_1$ the space of $\mathcal{C}^\infty$ ($\mathbb{R}^2$ valued) vector functions $\mathbf{v}$, which vanish in a neighborhood of $\Gamma_e \cup \Gamma_\ell$ and such that

$$\operatorname{div} \int_{p_0}^{p_1} \mathbf{v}\,\mathrm{d}p = 0.$$

Let $\mathcal{V}_2 = \mathcal{V}_3$ be the space of $\mathcal{C}^\infty$ functions on $\mathcal{M}$, and $\mathcal{V} = \mathcal{V}_1 \times \mathcal{V}_2 \times \mathcal{V}_3$; then, as in Eq. (2.53):

$$\mathcal{V}_1 \text{ is dense in } V_1, \qquad \mathcal{V} \text{ is dense in } V. \tag{2.131}$$

We also introduce $V_{(2)}$ the closure of $\mathcal{V}$ in $H^2(\mathcal{M})^4$.

The weak formulation of the PEs of the atmosphere takes the form:

$$\begin{aligned}&\left(\frac{\mathrm{d}U}{\mathrm{d}t}, \widetilde{U}\right)_H + a\big(U, \widetilde{U}\big) + b\big(U, U, \widetilde{U}\big) + e\big(U, \widetilde{U}\big)\\ &\quad = \ell\big(\widetilde{U}\big) \quad \forall \widetilde{U} \in V_{(2)}.\end{aligned} \tag{2.132}$$

Here, $b = b_1 + K_T b_2 + K_q b_3$ and e are essentially as for the ocean, replacing, for b, $\partial/\partial z$ by $\partial/\partial p$ and $w(\mathbf{v})$ by $\omega(\mathbf{v})$. Then $a = a_1 + K_T a_2 + K_q a_3$, with

$$\begin{aligned}a_1\big(U, \widetilde{U}\big) &= \int_{\mathcal{M}} \left(\mu_{\mathbf{v}} \nabla\mathbf{v}\cdot\nabla\tilde{\mathbf{v}} + \nu_{\mathbf{v}}\left(\frac{gp}{R\overline{T}}\right)^2 \frac{\partial\mathbf{v}}{\partial p}\frac{\partial\tilde{\mathbf{v}}}{\partial p}\right)\mathrm{d}\mathcal{M}\\ &\quad - \int_{\mathcal{M}} \left(\int_p^{p_1} \frac{RT}{p'}\,\mathrm{d}p'\,\nabla\tilde{v}\right)\mathrm{d}\mathcal{M} + \int_{\Gamma_i} \alpha_{\mathbf{v}}\mathbf{v}\tilde{\mathbf{v}}\,\mathrm{d}\Gamma_i,\\ a_2\big(U, \widetilde{U}\big) &= \int_{\mathcal{M}} \left(\mu_T \nabla T\cdot\nabla\widetilde{T} + \nu_T\left(\frac{gp}{R\overline{T}}\right)^2 \frac{\partial T}{\partial p}\frac{\partial\widetilde{T}}{\partial p}\right)\mathrm{d}\mathcal{M}\\ &\quad - \int_{\mathcal{M}} \frac{R\overline{T}(p)}{c_p p}\omega(\mathbf{v})\widetilde{T}\,\mathrm{d}\mathcal{M} + \int_{\Gamma_i} \alpha_T T\widetilde{T}\,\mathrm{d}\Gamma_i,\\ a_3\big(U, \widetilde{U}\big) &= \int_{\mathcal{M}} \left(\mu_S \nabla q\cdot\nabla\tilde{q} + \nu_S\left(\frac{gp}{R\overline{T}}\right)^2 \frac{\partial q}{\partial p}\frac{\partial\tilde{q}}{\partial p}\right)\mathrm{d}\mathcal{M}.\end{aligned}$$

Finally,

$$\begin{aligned}\ell\big(\widetilde{U}\big) &= \int_{\mathcal{M}} \big(F_{\mathbf{v}}\tilde{v} + K_T F_T\widetilde{T} + K_q F_q\tilde{q}\big)\,\mathrm{d}\mathcal{M}\\ &\quad + \int_{\Gamma_i} g_{\mathbf{v}}\tilde{v} + g_T\widetilde{T}\,\mathrm{d}\Gamma_i + \int_{\Gamma_e} g_T\widetilde{T}\,\mathrm{d}\Gamma_e,\\ g_{\mathbf{v}} &= \tau_{\mathbf{v}} + \alpha_{\mathbf{v}}\mathbf{v}^s, \qquad g_T = \alpha_T T^s \quad \text{on } \Gamma_i, \qquad g_T = \alpha_T T_e \quad \text{on } \Gamma_e.\end{aligned}$$

We find that there exists $c_1, c_2 > 0$ such that

$$a(U, U) + c_2 \int_{\mathcal{M}} q^2 \, \mathrm{d}\mathcal{M} \geq c_1 \|U\|^2 \quad \forall U \in V,$$

and that

$$e(U, U) = 0 \quad \forall U \in V.$$

Properties of b are similar to those in Lemma 2.1, with $V_{(2)}$ defined as the closure of $\mathcal{V}$ in $H^2(\mathcal{M})^4$.

The boundary and initial value problem. As for the ocean, the weak formulation reads as follows:

We are given $t_1 > 0$, U_0 in H, $F = (F_{\mathbf{v}}, F_T, F_q)$ in $L^2(0, t_1; H)$ (or $L^2(0, t_1; L^2(\mathcal{M})^4)$, $g_{\mathbf{v}}$ in $L^2(0, t_1; L^2(\Gamma_{\mathrm{i}})^2)$, g_T in $L^2(0, t_1; \Gamma_{\mathrm{i}} \cup \Gamma_{\mathrm{e}})$). We look for U:

$$U \in L^\infty(0, t_1; H) \cap L^2(0, t_1; V) \tag{2.133}$$

such that

$$\left(\frac{\mathrm{d}U}{\mathrm{d}t}, \widetilde{U}\right)_H + a\big(U, \widetilde{U}\big) + b\big(U, U, \widetilde{U}\big) + e\big(U, \widetilde{U}\big) = \ell\big(\widetilde{U}\big) \quad \forall \widetilde{U} \in V_{(2)}, \tag{2.134}$$

$$U(0) = U_0. \tag{2.135}$$

REMARK 2.11. We can introduce the operators A, B, and E and write Eq. (2.134) in an operator form, as Eq. (2.66).

The analog of Theorem 2.2 can be proved in exactly the same way:

THEOREM 2.3. *The domain $\mathcal{M}$ is a before. We are given $t_1 > 0$, U_0 in H, $F = (F_{\mathbf{v}}, F_T, F_q)$ in $L^2(0, t_1; H)$ (or $L^2(0, t_1; L^2(\mathcal{M})^4)$, $g_{\mathbf{v}}$ in $L^2(0, t_1; L^2(\Gamma_{\mathrm{i}})^2)$, g_T in $L^2(0, t_1; \Gamma_{\mathrm{i}} \cup \Gamma_{\mathrm{e}})$). Then there exists U that satisfies Eqs. (2.134) and (2.135). Furthermore, U is weakly continuous from $[0, t_1]$ in H.*

2.5. The CAO

After considering the ocean and the atmosphere separately, we consider in this section the CAO. The model presented here was first introduced LIONS, TEMAM and WANG [1995]; it is amenable to the mathematical and numerical analysis and is physically sound. The model was derived by carefully examining the boundary layer near the interface Γ_{i} between the ocean and the atmosphere. Although some processes are still not fully understood from the physical point of view, the derivation of the boundary condition is based on the work of GILL [1982] and HANEY [1971].

We will present the equations and boundary conditions, the variational formulation and arrive to a point where the mathematical treatment is the same as for the ocean and the atmosphere.

The pseudogeometrical domain. Let us first introduce the pseudogeometrical domain. Let h_0 be a typical length (height) for the atmosphere; for harmonization with the ocean, we introduce the vertical variable $\eta = z$, for $z < 0$ (in the ocean) and

$$\eta = h_0\left(\frac{p_1 - p}{p_1 - p_0}\right), \quad 0 < \eta < h_0, \tag{2.136}$$

for the atmosphere. The pseudogeometrical domain is

$$\mathcal{M} = \mathcal{M}^{\mathrm{a}} \cup \mathcal{M}^{\mathrm{s}} \cup \Gamma_{\mathrm{i}},$$

where $\mathcal{M}^{\mathrm{s}}$ is the ocean defined as in Section 2.1.2, $\mathcal{M}^{\mathrm{a}}$ is the atmosphere, $0 < \eta < h_0$, and Γ_{i} is, as before, the interface between the ocean and the atmosphere.

All quantities will now be defined as for the atmosphere alone, adding when needed, a superscript "a" or as for the ocean alone, adding a superscript "s."[4] Hence, with obvious notations, the boundary of $\mathcal{M}$ consists of

$$\Gamma_\ell^{\mathrm{s}} \cup \Gamma_{\mathrm{b}} \cup \Gamma_{\mathrm{e}} \cup \Gamma_{\mathrm{u}}. \tag{2.137}$$

The governing equations. In $\mathcal{M}^{\mathrm{a}}$, the variable is $U^{\mathrm{a}} = (v^{\mathrm{a}}, T^{\mathrm{a}}, q)$, and in $\mathcal{M}^{\mathrm{s}}$, the vari- able is $U^{\mathrm{s}} = (v^{\mathrm{s}}, T^{\mathrm{s}}, S)$; we set also $U = \{U^{\mathrm{a}}, U^{\mathrm{s}}\}$, or alternatively $v = \{v^{\mathrm{a}}, v^{\mathrm{s}}\}$, $T = \{T^{\mathrm{a}}, T^{\mathrm{s}}\}$.

The equations for U^{s} are exactly as in Eqs. (2.24)–(2.29), introducing only a superscript "s" for w, p, ρ_0, $F_{\mathbf{v}}$, F_T, F_S, T_{r}, and S_{r}, as well as the eddy viscosity coefficients $\mu_{\mathbf{v}}$, $\nu_{\mathbf{v}}$, etc.

The equations for U^{a} are exactly as in Eqs. (2.116)–(2.122), introducing again a superscript "a" for ω, p, ρ, $F_{\mathbf{v}}$, F_T, and F_q, as well as the various coefficients.[5] Of course, the variable p is replaced by η following Eq. (2.136), and the differential operators are changed accordingly.

Boundary conditions. Except for Γ_{i}, the boundary conditions are the same as for the ocean and the atmosphere taken separately. Hence, we recover the conditions (Eq. (2.123)) on Γ_{u} and (Eq. (2.124)) on Γ_{e}, adding the superscripts "a."

Let us now describe the boundary conditions on Γ_{i}. These conditions were introduced by LIONS, TEMAM and WANG [1995]; we refer the reader to this monograph for justification and a detailed discussion.

We first have the geometrical (kinematical) condition:

$$w^{\mathrm{s}} = \omega^{\mathrm{a}} = 0 \quad \text{on } \Gamma_{\mathrm{i}}, \tag{2.138}$$

which expresses that $\eta = 0$ ($z = 0$) is indeed the upper limit of the ocean (under the rigid lid hypothesis) and $\eta = 0$ is the lower limit of the atmosphere (the isobar $p = p_1$).

[4] "s" for sea rather than "o" for ocean, which could be confused with zero.

[5] An additional term linear in ω^{a} with a coefficient depending on p^{a} appears in the equation for T^{a} in LIONS, TEMAM and WANG [1995], (see equation (1.11), p. 4, and footnote 2, p.15 of LIONS, TEMAM and WANG [1995]. This term does not affect the discussion hereafter.

Then for the velocity, we express the fact that the tangential shear stresses exerted by the atmosphere on the ocean have opposite values and vice versa, and this value is expressed as a function of the differences of velocities $\mathbf{v}^{\mathrm{a}} - \mathbf{v}^{\mathrm{s}}$ using a classical empirical model of resolution of boundary layers (see, e.g., GILL [1982], HANEY [1971], LIONS, TEMAM and WANG [1995]; the boundary layer model is used to model the boundary layer of the atmosphere that is most significant). These conditions read[6]

$$\rho_0^{\mathrm{s}}\nu_{\mathbf{v}}^{\mathrm{s}}\frac{\partial \mathbf{v}^{\mathrm{s}}}{\partial z} = -\bar{\rho}^{\mathrm{a}}\nu_{\mathbf{v}}^{\mathrm{a}}\frac{\partial \mathbf{v}^{\mathrm{a}}}{\partial z} = \bar{\rho}^{\mathrm{a}}C_D^{\mathrm{a}}(\alpha)\big(\mathbf{v}^{\mathrm{a}} - \mathbf{v}^{\mathrm{s}}\big)\big|\mathbf{v}^{\mathrm{a}} - \mathbf{v}^{\mathrm{s}}\big|^{\alpha}. \tag{2.139}$$

Here, $\alpha \geq 0$ and $C_D^{\mathrm{a}}(\alpha) \geq 0$ are coefficients from boundary layer theory, and $\bar{\rho}^{\mathrm{a}} > 0$ is an averaged value of the atmosphere density. Similar conditions hold for the temperatures, the salinity S in the ocean, and the humidity q in the atmosphere.

For the sake of simplicity, to keep the boundary condition linear, we take $\alpha = 0$; we also need to replace z by $\eta(p)$ in the atmosphere, (see LIONS, TEMAM and WANG [1995] for the details). In the end, we arrive at the following conditions on Γ_{i}

$$\begin{aligned}
&w^{\mathrm{a}} = \omega^{\mathrm{s}} = 0,\\
&\rho_0^{\mathrm{s}}\nu_{\mathbf{v}}^{\mathrm{s}}\frac{\partial \mathbf{v}^{\mathrm{s}}}{\partial z} = -\bar{\rho}^{\mathrm{a}}\nu_{\mathbf{v}}^{\mathrm{a}}\left(\frac{gp^{\mathrm{a}}}{R\overline{T}}\right)^2\frac{\partial v^{\mathrm{a}}}{\partial p^{\mathrm{a}}} = \alpha_{\mathbf{v}}\big(\mathbf{v}^{\mathrm{a}} - \mathbf{v}^{\mathrm{s}}\big),\\
&c_p^{\mathrm{s}}\rho_0^{\mathrm{s}}\nu_T^{\mathrm{s}}\frac{\partial T^{\mathrm{s}}}{\partial z} = -c_p^{a}\bar{\rho}^{\mathrm{a}}\nu_T^{\mathrm{a}}\left(\frac{gp^{\mathrm{a}}}{R\overline{T}}\right)^2\frac{\partial T^{\mathrm{a}}}{\partial z} = \alpha_T\big(T^{\mathrm{a}} - T^{\mathrm{s}}\big),\\
&\frac{\partial q}{\partial p^{\mathrm{a}}} = \frac{\partial S}{\partial z} = 0.
\end{aligned} \tag{2.140}$$

Weak formulation of the PEs. For the sake of simplicity, we restrict ourselves to a regional problem using the beta-plane approximation.

The function spaces that we introduce are similar to those used for the ocean and the atmosphere, hence

$$V = V_1 \times V_2 \times V_3, \qquad H = H_1 \times H_2 \times H_3,$$

where $V_i = V_i^{\mathrm{a}} \times V_i^{\mathrm{s}}$ and $H_i = H_i^{\mathrm{a}} \times H_i^{\mathrm{s}}$, the spaces V_i^{a}, H_i^{a}, V_i^{s}, and H_i^{s} being exactly like those of the atmosphere and the ocean, respectively. Alternatively we can write, with obvious notations, $V = V^{\mathrm{a}} \times V^{\mathrm{s}}$ and $H = H^{\mathrm{a}} \times H^{\mathrm{s}}$.

These spaces are endowed with the following scalar products:

$$\begin{aligned}
\big((U, \widetilde{U})\big) &= \big((U, \widetilde{U})\big)_{\mathrm{a}} + \big((U, \widetilde{U})\big)_{\mathrm{s}},\\
\big((U, \widetilde{U})\big)_{\mathrm{a}} &= \big((\mathbf{v}^{\mathrm{a}}, \tilde{\mathbf{v}}^{\mathrm{a}})\big)_{\mathrm{a},1} + K_T\big((T^{\mathrm{a}}, \widetilde{T}^{\mathrm{a}})\big)_{\mathrm{a},2} + K_q\big((q, \tilde{q})\big)_{\mathrm{a},3},\\
\big((U, \widetilde{U})\big)_{\mathrm{s}} &= \big((\mathbf{v}^{\mathrm{s}}, \tilde{\mathbf{v}}^{\mathrm{s}})\big)_{\mathrm{s},1} + K_T\big((T^{\mathrm{s}}, \widetilde{T}^{\mathrm{s}})\big)_{\mathrm{s},2} + K_S\big((S, \widetilde{S})\big)_{\mathrm{s},3},
\end{aligned}$$

[6]The same equation appears in LIONS, TEMAM and WANG [1995], with $\bar{\rho}^{\mathrm{a}}$ replaced by ρ^{a}. Replacing ρ^{a} by $\bar{\rho}^{\mathrm{a}}$ is a necessary simplification for the developments below.

$$
\begin{aligned}
\big(U,\widetilde{U}\big)_H &= \big(U,\widetilde{U}\big)_{\mathrm{a}} + \big(U,\widetilde{U}\big)_{\mathrm{s}},\\
\big(U,\widetilde{U}\big)_{\mathrm{a}} &= \int_{\mathcal{M}^{\mathrm{a}}} \big(\mathbf{v}^{\mathrm{a}}\cdot\tilde{\mathbf{v}}^{\mathrm{a}} + K_T T^{\mathrm{a}}\widetilde{T}^{\mathrm{a}} + K_q q\tilde{q}\big)\,\mathrm{d}\mathcal{M},\\
\big(U,\widetilde{U}\big)_{\mathrm{s}} &= \int_{\mathcal{M}^{\mathrm{s}}} \big(\mathbf{v}^{\mathrm{s}}\cdot\tilde{\mathbf{v}}^{\mathrm{s}} + K_T T^{\mathrm{s}}\widetilde{T}^{\mathrm{s}} + K_S S\widetilde{S}\big)\,\mathrm{d}\mathcal{M}.
\end{aligned}
$$

Note that K_T is chosen the same in the atmosphere and the ocean. By the Poincaré inequality, there exists a constant $c_0 > 0$ (different from those for the ocean and the atmosphere) such that

$$
|U|_H \le c_0\|U\| \quad \forall U \in V. \tag{2.141}
$$

From this we conclude that $\|\cdot\|$ is a Hilbert norm on V. We also introduce in a very similar way the spaces $\mathcal{V}$ and $V_{(2)}$.

With this, the weak formulation of the PEs of the CAO takes the form:

$$
\begin{aligned}
&\left(\frac{\mathrm{d}U}{\mathrm{d}t},\widetilde{U}\right)_H + a\big(U,\widetilde{U}\big) + b\big(U,U,\widetilde{U}\big) + e\big(U,\widetilde{U}\big)\\
&\qquad = \ell\big(\widetilde{U}\big) \quad \forall\widetilde{U}\in V_{(2)}, &(2.142)\\
&U(0) = U_0. &(2.143)
\end{aligned}
$$

Here,

$$
a = a_1 + a_2 + a_3, \quad b = b_1 + b_2 + b_3, \quad e = e^{\mathrm{a}} + e^{\mathrm{s}},
$$

where

$$
\begin{aligned}
a_1 &= \bar{\rho}^{\mathrm{a}} a_1^{\mathrm{a}} + \rho_0^{\mathrm{s}} a_1^{\mathrm{s}} + \alpha_{\mathbf{v}} \int_{\Gamma_{\mathrm{i}}} \big|\mathbf{v}^{\mathrm{a}} - \mathbf{v}^{\mathrm{s}}\big|^2\,\mathrm{d}\Gamma_{\mathrm{i}},\\
a_2 &= K_T c_p^{\mathrm{a}} \bar{\rho}^{\mathrm{a}} a_2^{\mathrm{a}} + K_T c_p^{\mathrm{s}} \rho_0^{\mathrm{s}} a_2^{\mathrm{s}} + \alpha_T \int_{\Gamma_{\mathrm{i}}} \big|T^{\mathrm{a}} - T^{\mathrm{s}}\big|^2\,\mathrm{d}\Gamma_{\mathrm{i}},\\
a_3 &= K_q a_3^{\mathrm{a}} + K_s a_3^{\mathrm{s}},\\
b_1 &= \bar{\rho}^{\mathrm{a}} b_1^{\mathrm{a}} + \rho_0^{\mathrm{s}} b_1^{\mathrm{s}},\\
b_2 &= K_T c_p^{\mathrm{a}} \bar{\rho}^{\mathrm{a}} b_2^{\mathrm{a}} + K_T c_p^{\mathrm{s}} \rho_0^{\mathrm{s}} b_2^{\mathrm{s}},\\
b_3 &= K_q b_3^{\mathrm{a}} + K_S b_3^{\mathrm{s}}.
\end{aligned}
$$

Here, of course, the forms a_i^{a}, b_i^{a} and e^{a} are those of the atmosphere, and a_i^{s}, b_i^{s} and e^{s} are those of the ocean.

The form ℓ is defined as for the ocean and the atmosphere, the terms concerning Γ_i being omitted. Hence,

$$\ell = \ell^a + \ell^s,$$

$$\ell^a(\widetilde{U}) = \int_{\mathcal{M}^a} \left(\bar{\rho}^a F_{\mathbf{v}}^a \tilde{\mathbf{v}}^a + K_T c_p^a \bar{\rho}^a F_T^a \widetilde{T}^a + K_q F_q \tilde{q}\right) \mathrm{d}\mathcal{M}^a + \int_{\Gamma_e} g_T^a \widetilde{T}^a \,\mathrm{d}\Gamma_e,$$

$$\ell^s(\widetilde{U}) = \int_{\mathcal{M}^s} \left(\rho_0^s F_{\mathbf{v}}^s \tilde{\mathbf{v}}^s + K_T c_p^s \rho_0^s \widetilde{T}^s + K_S F_S \widetilde{S}\right) \mathrm{d}\mathcal{M}^s + \int_{\mathcal{M}^s} (\beta_T T_r - \beta_S S_r) \nabla \cdot \tilde{\mathbf{v}}^s \,\mathrm{d}\mathcal{M}^s.$$

With these definitions, the properties of $a, b, e,$ and ℓ are exactly the same as for the ocean and atmosphere (separately), and we prove, exactly as before, the existence, for all time, of weak solutions.

THEOREM 2.4. *The domain $\mathcal{M}$ is as before. We are given $t_1 > 0$, U_0 in H and F in $L^2(0, t_1; H)$ (or $L^2(0, t_1; L^2(\mathcal{M})^8)$, and g_T^a in $L^2(0, t_1; \Gamma_e)$.*

Then there exists U that satisfies Eqs. (2.142) and (2.143), and

$$U \in L^\infty(0, t_1; H) \cap L^2(0, t_1; V).$$

Furthermore, U is weakly continuous from $[0, t_1]$ into H.

3. Strong solutions of the PEs in dimensions 2 and 3

In this section, we first show, in Section 3.1, the existence, local in time, of strong solutions to the PEs in space dimension 3, i.e., solutions whose norm in H^1 remains bounded for a limited time. Then, in Section 3.2, we show the existence and uniqueness, global in time, of strong solutions to the PEs in space dimension 3.

In Section 3.3, we consider the PEs in space dimension 2 in view of adapting to this case the results of Sections 2 and 3.1. The 2D PEs are presented in Section 3.3.1, as well as their weak formulation (Section 3.3.2). Strong solutions are considered in Section 3.3.3, and we show directly (unlike in dimension 3) that the strong solutions are now defined for all time $t > 0$.

Essential in all this section is the anisotropic treatment of the equations, the vertical direction $0z$ playing a different role compared with the horizontal ones ($0x$ in two dimensions, $0x$ and $0y$ in three dimensions).

3.1. Strong solutions in space dimension 3

In this section, we establish the local, in time, existence of strong solutions to the PEs of the ocean. The result that we obtain is similar to that for the 3D Navier-Stokes equations. The analysis given in this section also applies to the PEs of the atmosphere and the CAO equations using the notations and equations given in Section 2.

We first state the main result of Section 3.1, namely Theorem 3.1. We then prepare its proof in several steps. In Step 1, we consider the linearized PEs and establish the global existence of strong solutions. In Step 2, we use the solution of the linearized equations in order to reduce the PEs to a nonlinear evolution equation with zero initial data and homogeneous boundary conditions. We also provide the necessary a priori estimates for this new problem with zero initial data and homogeneous boundary conditions. In the last step, we actually prove Theorem 3.1; in particular, we show how one can establish the existence of solutions to this problem using the Galerkin approximation with basis consisting of the eigenvectors of A (which are in H^2, thanks to the regularity results of Section 4). We use the previous estimates and then pass to the limit.

In rectangular coordinates x, y, z or x_1, x_2, x_3, the domain filled by the ocean is as in Eqs. (2.33) and (2.39):

$$\mathcal{M} = \{(x, y, z), (x, y) \in \Gamma_i, \quad -h(x, y) < z < 0\}, \tag{3.1}$$

where Γ_i is the surface of the ocean, Γ_b its bottom, Γ_ℓ the lateral surface, and $0 < \underline{h} \leq h(x, y) < \overline{h}$.

The main result of this section is as follows.

THEOREM 3.1. *We assume that* Γ_{i} *is of class* $\mathcal{C}^3$ *and that* $h : \overline{\Gamma_{\mathrm{i}}} \to \mathbb{R}_+$ *is of class* $\mathcal{C}^3$; *we also assume that*

$$\nabla h \cdot n_{\Gamma_{\mathrm{i}}} = 0 \quad \text{on } \partial\Gamma_{\mathrm{i}}, \tag{3.2}$$

where $n_{\Gamma_{\mathrm{i}}}$ *is the unit outward normal on* $\partial\Gamma_{\mathrm{i}}$ (*in the plane* $0xy$). *Furthermore, we are given* U_0 *in* V, $F = (F_{\mathbf{v}}, F_T, F_S)$ *in* $L^2(0, t_1; H)$ *with* $\partial F/\partial t$ *in* $L^2(0, t_1; L^2(\mathcal{M})^4)$, *and* $g = (g_{\mathbf{v}}, g_T)$ *in* $L^2(0, t_1; H_0^1(\Gamma_{\mathrm{i}})^3)$ *with* $\partial g/\partial t$ *in* $L^2(0, t_1; H_0^1(\Gamma_{\mathrm{i}})^3)$.[7] *Then there exists* $t_* > 0, t_* = t_*(\|U_0\|)$, *and there exists a unique solution* $U = U(t)$ *of the PEs (Eq. (2.79)) such that*

$$U \in C\big([0, t_*]; V\big) \cap L^2\big(0, t_*; H^2(\mathcal{M})^4\big). \tag{3.3}$$

STEP 1. The first step in the proof of Theorem 3.1 is the study of the linear PEs of the ocean.

Hence, we consider the equations (to Eqs. (2.44)–(2.51)):

$$\frac{\partial \mathbf{v}^*}{\partial t} + \nabla p^* + 2f\mathbf{k} \times \mathbf{v}^* - \mu_{\mathbf{v}} \Delta \mathbf{v}^* - \nu_{\mathbf{v}} \frac{\partial^2 \mathbf{v}^*}{\partial z^2} = F_{\mathbf{v}}, \tag{3.4}$$

$$\frac{\partial p^*}{\partial z} = -\rho g, \tag{3.5}$$

$$\frac{\partial T^*}{\partial t} - \mu_T \Delta T^* - \nu_T \frac{\partial^2 T^*}{\partial z^2} = F_T, \tag{3.6}$$

[7] The hypotheses on $\partial F/\partial t$ and $\partial g/\partial t$ can be weakened.

$$\frac{\partial S^*}{\partial t} - \mu_S \Delta S^* - \nu_S \frac{\partial^2 S^*}{\partial z^2} = F_S, \tag{3.7}$$

$$\operatorname{div} \int_{-h}^{0} \mathbf{v}^* \, \mathrm{d}z' = 0, \tag{3.8}$$

$$\int_{\mathcal{M}} S^* \, \mathrm{d}\mathcal{M} = 0, \tag{3.9}$$

$$p^* = p_{\mathrm{s}}^* + g \int_{z}^{0} \rho^* \, \mathrm{d}z', \tag{3.10}$$

$$\rho^* = \rho_0\big(1 - \beta_T\big(T^* - T_{\mathrm{r}}\big) + \beta_S\big(S^* - S_{\mathrm{r}}\big)\big), \tag{3.11}$$

with the same initial and boundary conditions as for the full nonlinear problem, i.e., (see Eqs. (2.35)–(2.37)):

$$\begin{aligned} &\nu_{\mathbf{v}} \frac{\partial \mathbf{v}^*}{\partial z} + \alpha_{\mathbf{v}}\big(\mathbf{v}^* - \mathbf{v}^{\mathrm{a}}\big) = 0, \\ &\nu_T \frac{\partial T^*}{\partial z} + \alpha_T\big(T^* - T^{\mathrm{a}}\big) = 0, \qquad \frac{\partial S^*}{\partial z} = 0 \quad \text{on } \Gamma_{\mathrm{i}}, \end{aligned} \tag{3.12}$$

$$\begin{aligned} &\mathbf{v}^* = 0, \\ &\frac{\partial T^*}{\partial \mathbf{n}_T} = 0, \qquad \frac{\partial S^*}{\partial \mathbf{n}_S} = 0 \quad \text{on } \Gamma_{\mathrm{b}} \cup \Gamma_{\ell}, \end{aligned} \tag{3.13}$$

$$\mathbf{v}^* = \mathbf{v}_0, \qquad T^* = T_0, \quad \text{and} \quad S^* = S_0 \quad \text{at } t = 0. \tag{3.14}$$

Compared with the nonlinear problems (Eqs. (2.44)–(2.51)), and using the same notations as in Eqs. (2.66) and (2.67), we see that $U^* = (\mathbf{v}^*, T^*, S^*)$ is solution of the following equation written in functional form:

$$\frac{\mathrm{d}U^*}{\mathrm{d}t} + AU^* + E\big(U^*\big) = \ell, \tag{3.15}$$

$$U^*(0) = U_0. \tag{3.16}$$

The right-hand side ℓ of Eq. (3.15) is exactly the same as in Eq. (2.66) (see Eqs. (2.44)–(2.51) and the expression of ℓ in (Eq. (2.55)).

We also consider the solution $\overline{U} = (\bar{v}, \overline{T}, \overline{S})$ of the linear stationary problem, namely

$$-\mu_{\mathbf{v}} \Delta \bar{\mathbf{v}} - \nu_{\mathbf{v}} \frac{\partial^2 \bar{\mathbf{v}}}{\partial z^2} + \nabla \bar{p} = F_{\mathbf{v}} - 2f\mathbf{k} \times \bar{\mathbf{v}},$$

$$\frac{\partial \bar{p}}{\partial z} = -\bar{\rho} g,$$

$$-\mu_T \Delta \overline{T} - \nu_T \frac{\partial \overline{T}}{\partial z^2} = F_T,$$

$$
\begin{aligned}
&-\mu_T S\Delta\overline{S} - \nu_S \frac{\partial^2 \overline{S}}{\partial z^2} = F_S,\\
&\operatorname{div}\int_{-h}^{0} \bar{\mathbf{v}}\,\mathrm{d}z' = 0,\\
&\int_{\mathcal{M}} \overline{S}\,\mathrm{d}\mathcal{M} = 0,\\
&\bar{p} = \bar{p}_{\mathrm{s}} + g\int_{z}^{0} \bar{\rho}\,\mathrm{d}z',\\
&\bar{\rho} = \rho_0\big(1 - \beta_T\big(\overline{T} - T_{\mathrm{r}}\big) + \beta_S\big(\overline{S} - S_{\mathrm{r}}\big)\big),
\end{aligned}
\tag{3.17}
$$

with the boundary conditions

$$
\begin{aligned}
&\nu_{\mathbf{v}}\frac{\partial \bar{\mathbf{v}}}{\partial z} + \alpha_{\mathbf{v}}\big(\bar{\mathbf{v}} - \mathbf{v}^{\mathrm{a}}\big) = \tau_{\mathbf{v}},\\
&\nu_T\frac{\partial \overline{T}}{\partial z} + \alpha_T\big(\overline{T} - T^{\mathrm{a}}\big) = 0, \qquad \frac{\partial \overline{S}}{\partial z} = 0 \quad \text{on } \Gamma_{\mathrm{i}},\\
&\bar{\mathbf{v}} = 0, \qquad \frac{\partial \overline{T}}{\partial \mathbf{n}_T} = 0, \qquad \frac{\partial \overline{S}}{\partial \mathbf{n}_S} = 0 \quad \text{on } \Gamma_{\mathrm{b}} \cup \Gamma_{\ell}.
\end{aligned}
\tag{3.18}
$$

Note that (as in Eqs. (3.4)–(3.13)) the left- and right-hand sides in Eqs. (3.17) and (3.18) depend on the time t. The existence and uniqueness for (almost) every time t for Eqs. (3.17) and (3.18) follows from the Lax–Milgram theorem as explained, e.g., for the velocity $\bar{\mathbf{v}}$, in Section 4.4.1 and Proposition 4.1. Furthermore, the regularity results of Section 4 (in particular, Theorems 2.1–4.7) show that the solution belongs to $H^2(\mathcal{M})$, and that

$$
\begin{aligned}
&\|\bar{\mathbf{v}}\|^2_{\mathbb{H}^2(\mathcal{M})} + \big\|\overline{T}\big\|^2_{\mathbb{H}^2(\mathcal{M})} + \|\bar{\mathbf{v}}\|^2_{\mathbb{H}^2(\mathcal{M})} \le C_0\kappa_1,\\
&\kappa_1 = \kappa_1(F, \tau_{\mathbf{v}}, \mathbf{v}_{\mathrm{a}}, T_{\mathrm{a}}) = |F|^2_H + \|\tau_{\mathbf{v}}\|^2_{\mathbb{H}^1(\Gamma_{\mathrm{i}})} + \|\mathbf{v}_{\mathrm{a}}\|^2_{\mathbb{H}^1(\Gamma_{\mathrm{i}})} + \|T_{\mathrm{a}}\|^2_{\mathbb{H}^1(\Gamma_{\mathrm{i}})}.
\end{aligned}
\tag{3.19}
$$

Note that, again, each side of Eq. (3.19) depends on t, and Eq. (3.19) is valid for (almost) every t. The constant C_0 depends only on $\mathcal{M}$ according to the results of Section 4; in particular, C_0 is independent of t. Hypothesis (Eq. (3.2)) is precisely what is needed for the utilization of Theorem 4.5 used for Eq. (3.19). It is noteworthy that we have the same estimates as Eq. (3.19) for the derivatives $\partial\bar{\mathbf{v}}/\partial t$, $\partial\overline{T}/\partial t$, and $\partial\overline{S}/\partial t$, and κ_1 being replaced by κ_1', which is defined similarly in terms of the time derivatives $\partial F/\partial t$.

Now let $\tilde{\mathbf{v}} = \mathbf{v}^* - \bar{\mathbf{v}}$, $\tilde{p} = p^* - \bar{p}$, $\tilde{\rho} = \rho^* - \bar{\rho}$, $\widetilde{T} = T^* - \overline{T}$, and $\widetilde{S} = S^* - \overline{S}$. The equations satisfied by $(\tilde{v}, \widetilde{T}, \widetilde{S})$ are the same as Eqs. (3.4)–(3.14) but with $(F_{\mathbf{v}}, F_T, F_S) = -\mathrm{d}\overline{U}/\mathrm{d}t$, $\mathbf{v}_{\mathrm{a}} = \tau_{\mathbf{v}} = T_{\mathrm{a}} = T_{\mathrm{r}} = T_{\mathrm{s}} = 0$, and with initial data

$$
\tilde{\mathbf{v}}\big|_{t=0} = \mathbf{v}_0 - \bar{\mathbf{v}}(0), \quad \widetilde{T}\big|_{t=0} = T_0 - \overline{T}(0), \quad \text{and} \quad \widetilde{S}\big|_{t=0} = S_0 - \overline{S}(0). \tag{3.20}
$$

Compared with the nonlinear problem (Eqs. (2.44)–(2.51)), and using the same notation as in Eqs. (2.66) and (2.67), we see that $\widetilde{U} = (\tilde{\mathbf{v}}, \widetilde{T}, \widetilde{S})$ is solution of the following equation written in functional form:

$$\frac{\mathrm{d}\widetilde{U}}{\mathrm{d}t} + A\widetilde{U} + E\big(\widetilde{U}\big) = -\frac{\mathrm{d}\overline{U}}{\mathrm{d}t}, \tag{3.21}$$

$$\widetilde{U}(0) = \widetilde{U}_0 = U_0 - \overline{U}(0). \tag{3.22}$$

Note that the contribution from ℓ vanishes (see the expression in the equation preceding (Eq. (2.56)).

The existence for all time of a strong solution $\widetilde{U}$ to Eqs. (3.21) and (3.22) is classical, and we recall the estimate obtained by multiplying Eq. (3.21) by $A\widetilde{U}$ and integrating in time:

$$\sup_{0\le t\le t_1} \big\|\widetilde{U}(t)\big\|^2 + \int_0^{t_1} \big|A\widetilde{U}(s)\big|_H^2 \,\mathrm{d}s \le c\big\|\widetilde{U}(0)\big\|^2 + c\kappa_1'.$$

Here, we have used the analog of Eq. (3.19) for $\mathrm{d}\overline{U}/\mathrm{d}t$, (see the comments after Eq. (3.19)). From this, we obtain

$$\begin{aligned}&\sup_{0\le t\le t_1} \big\|U^*(t)\big\|^2 + \int_0^{t_1} \big|U^*(s)\big|_{H^2(\mathcal{M})^4}^2 \,\mathrm{d}s \\ &\quad\le c\big(\big\|U(0)\big\|^2 + \big\|\overline{U}(0)\big\|^2\big) + c\int_0^{t_1} \big|\overline{U}(s)\big|_{H^2(\mathcal{M})^4}^2 \,\mathrm{d}s + c\int_0^{t_1} \kappa_1'(s)\,\mathrm{d}s,\end{aligned} \tag{3.23}$$

and, with Eq. (3.19), we bound the right-hand side of Eq. (3.23) by an expression κ_2 of the form:

$$\kappa_2 = c\|U_0\|^2 + c_0\bigg(\kappa_1(0) + \int_0^{t_1} \big(\kappa_1 + \kappa_1'\big)(s)\,\mathrm{d}s\bigg). \tag{3.24}$$

Step 2. We will now use $U^* = (\mathbf{v}^*, T^*, S^*)$ and write the PEs of the ocean using the decomposition $U = U^* + U'$; we note that $U'(0) = 0$ and that $U' = (\mathbf{v}', T', S')$ satisfies homogeneous boundary conditions of the same type as U. More precisely, starting from the functional forms (Eqs. (2.66) and (2.67)) of the equation for U and using Eqs. (3.15) and (3.16), we see that U' satisfies

$$\begin{aligned}&\frac{\mathrm{d}U'}{\mathrm{d}t} + AU' + B\big(U', U^*\big) + B\big(U^*, U'\big) + B\big(U', U'\big) + E\big(U'\big) \\ &\quad= -AU^* - B\big(U^*, U^*\big),\end{aligned} \tag{3.25}$$

$$U'(0) = 0. \tag{3.26}$$

The existence of solution in Theorem 3.1 is proved by the existence of solution for this system (on some interval of time $(0, t_*)$). As usual, this proof of existence is based on

a priori estimates for the solutions U' of Eqs. (3.25) and (3.26). Some a priori estimates can be obtained by proceeding exactly as for Theorem 2.2, but additional estimates are needed here. Essential for these new estimates is another estimate on the bilinear operator B (or the trilinear form b), which is obtained by an *anisotropic treatment of certain integrals*. We have the following result (cf., to Lemma 2.1):

LEMMA 3.1. *In space dimension 3, the form b is trilinear continuous on* $H^2(\mathcal{M})^4 \times H^2(\mathcal{M})^4 \times L^2(\mathcal{M})^4$*, and we have*

$$\begin{aligned}\left|b\left(U, U^\flat, U^\sharp\right)\right| \leq c_3\big(&\|U\|_{H^1}\left\|U^\flat\right\|_{H^1}^{1/2}\left\|U^\flat\right\|_{H^2}^{1/2} \\ &+ \|U\|_{H^1}^{1/2}\|U\|_{H^2}^{1/2}\left\|U^\flat\right\|_{H^1}^{1/2}\left\|U^\flat\right\|_{H^2}^{1/2}\big)\left|U^\sharp\right|_{L^2}\end{aligned} \tag{3.27}$$

for every $(U, U^\flat, U^\sharp)$ *in this space.*

The proof of this lemma is given below. Using Lemma 3.1, we obtain a priori estimates on the solution U' of Eqs. (3.25) and (3.26). We denote by $A^{1/2}$ the square root of A so that

$$\left(A^{1/2}U, A^{1/2}\widetilde{U}\right)_H = a\left(U, \widetilde{U}\right) \quad \forall U, \widetilde{U} \in V.$$

Taking the scalar product of Eq. (3.25) with AU' in H, we obtain

$$\begin{aligned}&\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left|A^{1/2}U'\right|_H^2 + \left|AU'\right|_H^2 \\ &\quad = -b\left(U', U^*, AU'\right) - b\left(U^*, U', AU'\right) - b\left(U', U', AU'\right) \\ &\qquad - b\left(U^*, U^*, AU'\right) - \left(AU^*, AU'\right)_H - \left(E\left(U'\right), AU'\right)_H.\end{aligned} \tag{3.28}$$

We bound each term in the right-hand side of Eq. (3.28) as follows, using Lemma 3.1 for the b-terms:

$$\left|\left(AU^*, AU'\right)_H\right| \leq \frac{1}{12}\left|AU'\right|_H^2 + c\left|U^*\right|_{H^2}^2,\ ^8$$

$$\left|\left(E\left(U'\right), AU'\right)_H\right| \leq \frac{1}{12}\left|AU'\right|_H^2 + c\left|U'\right|_H^2,$$

$$\begin{aligned}\left|b\left(U', U^*, AU'\right)\right| &\leq c\left\|U'\right\|\left\|U^*\right\|_{H^2}\left|AU'\right|_H \\ &\quad + c\left\|U'\right\|^{1/2}\left\|U^*\right\|_{H^1}^{1/2}\left\|U^*\right\|_{H^2}^{1/2}\left|AU'\right|_H^{3/2} \\ &\leq \frac{1}{12}\left|AU'\right|_H^2 + c\left\|U'\right\|^2\left\|U^*\right\|_{H^2}^2\left(1 + \left\|U^*\right\|_{H^1}^2\right),\end{aligned}$$

$$\begin{aligned}\left|b\left(U^*, U', AU'\right)\right| &\leq c\left\|U^*\right\|_{H^1}^{1/2}\left\|U^*\right\|_{H^2}^{1/2}\left\|U'\right\|^{1/2}\left|AU'\right|_H^{3/2} \\ &\leq \frac{1}{12}\left|AU'\right|_H^2 + c\left\|U'\right\|^2\left\|U^*\right\|_{H^1}^2\left\|U^*\right\|_{H^2}^2,\end{aligned}$$

[8] U^* is not in $D(A)$ and $AU' \in V'$ because U^* does not satisfy the homogeneous boundary conditions; however this bound is valid (see the details in HU, TEMAM and ZIANE [2003]).

$$\begin{aligned}
&\left|b\left(U', U', AU'\right)\right| \leq c_4\left\|U'\right\|\left|AU'\right|_H^2,\\
&\left|b\left(U^*, U^*, AU'\right)\right| \leq c\left\|U^*\right\|_{H^1}\left\|U^*\right\|_{H^2}\left|AU'\right|_H\\
&\qquad\qquad \leq \frac{1}{12}\left|AU'\right|_H^2 + c\left\|U^*\right\|_{H^1}^2\left\|U^*\right\|_{H^2}^2.
\end{aligned}$$

Here, we used the fact that the norm $|AU'|_H$ is equivalent to the norm $|U'|_{H^2}$, thanks to the results of Section 4, and (this is easy) the fact that the norm $|A^{1/2}U'|_H$ is equivalent to the norm $\|U'\| = \|U'\|_{H^1}$.

Taking all these bounds into account, we infer from Eq. (3.28) that

$$\frac{\mathrm{d}}{\mathrm{d}t}\left|A^{1/2}U'\right|_H^2 + \left(1 - c_4\left|A^{1/2}U'\right|_H\right)\left|AU'\right|_H^2 \leq \lambda(t)\left|A^{1/2}U'\right|_H^2 + \mu(t), \tag{3.29}$$

with

$$\begin{aligned}
&\lambda(t) = c\left\|U^*\right\|_{H^2}^2 + \mu(t),\\
&\mu(t) = c\left\|U^*\right\|_{H^1}^2\left\|U^*\right\|_{H^2}^2.
\end{aligned}$$

We infer from Eqs. (3.23) and (3.24) (and from the precise expression of κ_2 in Eq. (3.24)) that λ and μ are integrable on $(0, t_1)$ and we set

$$\kappa_3 = \int_0^{t_1} \lambda(t)\,\mathrm{d}t.$$

By Gronwall's lemma and since $U'(0) = 0$, we have on some interval of time $(0, t_*)$ and as long as $1 - c_4|A^{1/2}U'|_H \geq 0$,

$$\begin{aligned}
&\left|A^{1/2}U'(t)\right|_H^2 \leq \left(\int_0^t \mu(s)\,\mathrm{d}s\right)\exp\left(\int_0^t \lambda(s)\,\mathrm{d}s\right),\\
&\left|A^{1/2}U'(t)\right|_H^2 \leq \exp\left(\int_0^t \lambda(s)\,\mathrm{d}s\right).
\end{aligned} \tag{3.30}$$

In fact, Eq. (3.29) is valid as long as $0 < t < t_*$ where t_* is smaller than t_1 and t^*, where t^* is either $+\infty$ or the time at which

$$\int_0^{t^*} \lambda(s)\,\mathrm{d}s = \log\left(\frac{1}{4c_4^2\kappa_3}\right). \tag{3.31}$$

We then have

$$\left|A^{1/2}U'(t)\right|_H^2 \leq \frac{1}{4c_4^2} \quad \text{for } 0 < t < t_*, \tag{3.32}$$

and returning to (3.30) we find also a bound

$$\int_0^{t_*} \left|AU'(t)\right|_H^2\,\mathrm{d}t \leq const. \tag{3.33}$$

STEP 3 (Proof of the existence in Theorem 3.1) As we said, the existence for Theorem 3.1 is shown by proving the existence of a solution U' of Eqs. (3.25) and (3.26) in $\mathcal{C}([0,t_*];V)\cap L^2(0,t_*;D(A))$. For that purpose, we implement a Galerkin method using the eigenvectors e_j of A:

$$Ae_j=\lambda_j e_j,\quad j\geq 1,\ 0<\lambda_1\leq\lambda_2\leq\cdots.$$

The results of Section 4.1 show that the e_j belong to $H^2(\mathcal{M})^4$ (since $D(A)\subset H^2(\mathcal{M})^4$). We look, for each $m>0$ fixed, for an approximate solution

$$U'_m=\sum_{j=1}^{m}\xi_{jm}(t)e_j,$$

satisfying, (cf., Eqs. (3.25) and (3.26))

$$\begin{aligned}&\left(\frac{\mathrm{d}U'_m}{\mathrm{d}t},e_k\right)_H+a\big(U'_m,e_k\big)+b\big(U'_m,U^*,e_k\big)\\&\quad+b\big(U^*,U'_m,e_k\big)+b\big(U'_m,U'_m,e_k\big)+e\big(U'_m,e_k\big)\\&=-a\big(U^*,e_k\big)-b\big(U^*,U^*,e_k\big),\quad k=1,\dots,m,\end{aligned}\tag{3.34}$$

and

$$U'_m(0)=0.$$

Multiplying Eq. (3.34) by $\xi_{km}(t)\lambda_k$, and adding these equations for $k=1,\dots,m$, we obtain the analog of Eq. (3.28) for U'_m. The same calculations as above show that U'_m satisfies the same estimates independent of m as Eqs. (3.32) and (3.33), with the same time t_* (also independent of m).

It is then straightforward to pass to the limit $m\to\infty$, and we obtain the existence.

STEP 4 (Proof of uniqueness in Theorem 3.1) The proof of uniqueness is easy. Consider two solutions U_1, and U_2 of the PEs; let $U=U_1-U_2$ and consider as above the associated functions $U'_i=U_i-U^*$ and $U'=U'_1-U'_2$. Then, U' satisfies

$$\frac{\mathrm{d}U'}{\mathrm{d}t}+AU'+B\big(U',U_2\big)+B\big(U_2,U'\big)+B\big(U',U'\big)+E\big(U'\big)=0,$$

$$U'(0)=0.$$

Treating this equation exactly as Eq. (3.25) we obtain an equation similar to Eq. (3.29) but with $\mu=0$ and a different λ:

$$\frac{\mathrm{d}}{\mathrm{d}t}\big|A^{1/2}U'\big|_H^2+\big(1-c_4\big|A'U\big|_H\big)\big|AU'\big|_H^2\leq\tilde{\lambda}(t)\big|A^{1/2}U'\big|_H^2.$$

The uniqueness follows using Gronwall's lemma.

To conclude this section and the proof of Theorem 3.1, we now prove Lemma 3.1.

PROOF OF LEMMA 3.1. We need only to show how the different integrals in b are bounded by the expressions appearing in the right-hand side of Eq. (3.27) and, in fact, we restrict ourselves to two typical terms, the other terms being treated in the same way.

The first one is bounded as follows:

$$\left| \int_{\mathcal{M}} \left[(\mathbf{v} \cdot \nabla) \mathbf{v}^\flat \right] \mathbf{v}^\sharp \, \mathrm{d}\mathcal{M} \right| \leq |\mathbf{v}|_{L^6} \left| \nabla \mathbf{v}^\flat \right|_{L^3} \left| \mathbf{v}^\sharp \right|_{L^2}. \tag{3.35}$$

By Sobolev embeddings and interpolation, we bound the right-hand side by

$$c \|\mathbf{v}\|_{H^1} \left\| \mathbf{v}^\flat \right\|_{H^2}^{1/2} \left| \mathbf{v}^\sharp \right|_{L^2},$$

which corresponds to the first term on the right-hand side of Eq. (3.27).

The second typical term that necessitates anisotropic estimates is bounded as follows:

$$\begin{aligned}
\left| \int_{\mathcal{M}} w(\mathbf{v}) \frac{\partial \mathbf{v}^\flat}{\partial z} \mathbf{v}^\sharp \, \mathrm{d}\mathcal{M} \right| &= \left| \int_{\Gamma_\mathrm{i}} \int_{-h}^{0} w(\mathbf{v}) \frac{\partial \mathbf{v}^\flat}{\partial z} \mathbf{v}^\sharp \, \mathrm{d}z \, \mathrm{d}\Gamma_\mathrm{i} \right| \\
&\leq \int_{\Gamma_\mathrm{i}} \left| w(\mathbf{v}) \right|_{L_z^\infty} \left| \frac{\partial \mathbf{v}^\flat}{\partial z} \right|_{L_z^2} \left| \mathbf{v}^\sharp \right|_{L_z^2} \mathrm{d}\Gamma_\mathrm{i} \\
&\leq \bar{h}^{1/2} \int_{\Gamma_\mathrm{i}} |\operatorname{div} \mathbf{v}|_{L_z^2} \left| \frac{\partial \mathbf{v}^\flat}{\partial z} \right|_{L_z^2} \left| \mathbf{v}^\sharp \right|_{L_z^2} \mathrm{d}\Gamma_\mathrm{i} \\
&\leq \text{(with Holder's inequality)} \\
&\leq \bar{h}^{1/2} |\operatorname{div} \mathbf{v}|_{L^4_{\Gamma_\mathrm{i}} L_z^2} \left| \frac{\partial \mathbf{v}^\flat}{\partial z} \right|_{L^4_{\Gamma_\mathrm{i}} L_z^2} \left| \mathbf{v}^\sharp \right|_{L^2_{\Gamma_\mathrm{i}} L_z^2}.
\end{aligned} \tag{3.36}$$

For a (scalar or vector) function ξ defined on $\mathcal{M}$, we have written ($1 \leq \alpha, \beta \leq \infty$):

$$\begin{aligned}
|\xi|_{L_z^\beta} &= |\xi|_{L_z^\beta}(x, y) = \left(\int_{-h(x,y)}^{0} \left| \xi(x, y, z) \right|^\beta \mathrm{d}z \right)^{1/\beta}, \\
|\xi|_{L^\alpha_{\Gamma_\mathrm{i}} L_z^\beta} &= \left| |\xi|_{L_z^\beta} \right|_{L^\alpha(\Gamma_\mathrm{i})} = \left(\int_{\Gamma_\mathrm{i}} |\xi|^\alpha_{L_z^\beta}(x, y) \, \mathrm{d}\Gamma_\mathrm{i} \right)^{1/\alpha}.
\end{aligned}$$

Notice also that $|\xi|_{L^2_{\Gamma_\mathrm{i}} L_z^2} = |\xi|_{L^2(\mathcal{M})}$ and that from the expression (Eq. (2.40)) of $w(\mathbf{v})$, and Eq. (2.39),

$$\left| w(\mathbf{v}) \right|_{L_z^\infty} = \left| \int_z^0 \operatorname{div} \mathbf{v} \, \mathrm{d}z' \right|_{L_z^\infty} \leq \bar{h}^{1/2} |\operatorname{div} \mathbf{v}|_{L_z^2}.$$

Now we remember that (in space dimension 2) there exists a constant $c = c(\Gamma_i)$ such that for every function ζ in $H^1(\Gamma_i)$,

$$|\zeta|_{L^4(\Gamma_i)} \leq c|\zeta|^{1/2}_{L^2(\Gamma_i)}|\zeta|^{1/2}_{H^1(\Gamma_i)}, \tag{3.37}$$

from which we infer, for a function ξ as above (setting $\zeta = |\xi|_{L^2_z}$),

$$|\xi|_{L^4_{\Gamma_i}L^2_z} \leq c|\xi|^{1/2}_{L^2_{\Gamma_i}L^2_z}\big||\xi|_{L^2_z}\big|^2_{H^1(\Gamma_i)}. \tag{3.38}$$

As before $|\xi|_{L^2_{\Gamma_i}L^2_z} = |\xi|_{L^2(\mathcal{M})}$, whereas

$$\big||\xi|_{L^2_z}\big|^2_{H^1(\Gamma_i)} = |\xi|^2_{L^2_{\Gamma_i}L^2_z} + \big|\nabla|\xi|_{L^2_z}\big|^2_{L^2(\Gamma_i)} = |\xi|^2_{L^2(\mathcal{M})} + |\nabla\theta|^2_{L^2(\Gamma_i)}, \tag{3.39}$$

with

$$\theta = \theta(x, y) = |\xi|_{L^2_z}(x, y) = \left(\int_{-h}^{0} \big|\xi(x, y, z)\big|^2 \,\mathrm{d}z\right)^{1/2}.$$

We intend to show that for ξ in $H^1(\mathcal{M})$ (scalar or vector ξ),

$$|\xi|_{L^4_{\Gamma_i}L^2_z} \leq c|\xi|^{1/2}_{L^2(\mathcal{M})}|\xi|^{1/2}_{H^1(\mathcal{M})} \tag{3.40}$$

for some suitable constant $c = c(\mathcal{M})$. For that purpose, we note that

$$\nabla\theta = \left(\int_{-h}^{0} |\xi|^2 \,\mathrm{d}z\right)^{1/2}\left\{\int_{-h}^{0} \xi\nabla\xi \,\mathrm{d}z + \big|\xi(-h)\big|^2\nabla h\right\},$$

where $\xi(-h)$ and below $\xi(z)$ are simplified notations for $\xi(x, y, -h(x, y))$ and $\xi(x, y, z)$, respectively. With the Schwarz inequality, we find that pointwise a.e. (for a.e. $(x, y) \in \Gamma_i$).

$$|\nabla\theta| \leq |\nabla\xi|_{L^2_z} + c|\xi|^{-1}_{L^2_z}\big|\xi(-h)\big|^2. \tag{3.41}$$

We have classically

$$\big|\xi(-h)\big|^2 = \big|\xi(z)\big|^2 - 2\int_{-h}^{z} \xi\frac{\partial\xi}{\partial z}\,\mathrm{d}z \leq \big|\xi(z)\big|^2 + 2|\xi|_{L^2_z}\left|\frac{\partial\xi}{\partial z}\right|_{L^2_z},$$

and by integration in z from $-h(x, y)$ to 0

$$h\big|\xi(-h)\big|^2 \leq |\xi|^2_{L^2_z} + 2\bar{h}|\xi|_{L^2_z}\left|\frac{\partial\xi}{\partial z}\right|_{L^2_z}.$$

Hence, (3.41) yields pointwise a.e.

$$|\nabla\theta| \leq |\nabla\xi|_{L^2_z} + c|\xi|_{L^2_z} + c\left|\frac{\partial\xi}{\partial z}\right|_{L^2_z}.$$

By integration on Γ_i, we find

$$|\nabla\theta|_{L^2(\Gamma_i)} \leq c|\xi|_{H^1(\mathcal{M})},$$

and then Eqs. (3.38) and (3.39) yield (3.40).

Having established Eq. (3.40), we return to Eq. (3.36): applying Eq. (3.40) with $\xi = \operatorname{div}\mathbf{v}$ and $\xi = \partial\tilde{\mathbf{v}}^\flat/\partial z$, we can bound the left-hand side of Eq. (3.36) by the second term on the right-hand side of Eq. (3.27), thus concluding the proof of Lemma 3.1. □

3.2. Strong solutions in dimension 3 (global existence)

The aim of this section is to present very recent results on the global existence of a strong solution for the PEs in space dimension 3. As pointed out in the general introduction of this chapter, this result was initially thought to be as difficult to prove as the global existence of strong solutions for the incompressible 3D Navier-Stokes equations. Recently, a series of three articles (Cao and Titi [2007], Kobelkov [2006], Kukavica and Ziane [2007a,b]) prove, by different methods, that the PEs have a 2D intrinsic structure and consequently prove the existence of global strong solutions.

The first articles are those, written independently, by Cao and Titi [2007] and Kobelkov [2006]. These articles, using different methods, address the case of a cylindrical domain with constant depth and Neumann boundary conditions. The case of nonconstant depth and the boundary conditions (2.35)–(2.37) is studied in the more recent articles by Kukavica and Ziane [2007a,b]. In this section, we present the proof of Kobelkov [2006] and state without proof the result of Kukavica and Ziane [2007a,b].

In Section 3.6, the method of Cao and Titi [2007] is described and used to prove the existence, globally in time, of even more regular solutions in the space periodic case.

In what follows, we do not work with both the salinity and the temperature since the equation for the salinity does not bring any additional difficulty. And instead of the temperature T, we use the density ρ, which is proportional to T.

As mentioned before, the domain we are working with is a cylinder until the end of Section 3.2:

$$\mathcal{M} = \left\{x = (x_1, x_2, x_3);\, (x_1, x_2) \in \mathcal{M}',\, x_3 \in [0, 1]\right\}, \tag{3.42}$$

where $\mathcal{M}' = \Gamma_i$ is a 2D domain with a boundary consisting of a finite number of smooth arcs intersecting at nonzero angles; Γ_i, Γ_ℓ, and Γ_b are defined as before, but the depth here is constant, $h = 1$. We also use the following notation:

$$\mathcal{M}_t = \mathcal{M} \times [0, t].$$

The PEs then read (with $z = x_3$ and $\nabla_{\mathbf{v}}$ as in Eq. (2.19)):

$$\begin{aligned}
&\frac{\partial \mathbf{v}}{\partial t} + \nabla_{\mathbf{v}}\mathbf{v} + w\frac{\partial \mathbf{v}}{\partial z} + \frac{1}{\rho_0}\nabla p + 2f\mathbf{k}\times\mathbf{v} - \nu_{\mathbf{v}}\Delta_3\mathbf{v} = F_{\mathbf{v}},\\
&\frac{\partial p}{\partial z} = -\rho g,\\
&\operatorname{div}\mathbf{v} + \frac{\partial w}{\partial z} = 0,\\
&\frac{\partial \rho}{\partial t} + \nabla_{\mathbf{v}}\rho + w\frac{\partial \rho}{\partial z} - \nu_\rho\Delta_3\rho = F_\rho,
\end{aligned} \tag{3.43}$$

The boundary conditions for Eq. (3.43) are:

$$\mathbf{v}\cdot\mathbf{n} = \frac{\partial \mathbf{v}}{\partial \mathbf{n}}\times\mathbf{n} = 0 \text{ on } \Gamma_l,\ w = 0 \text{ on } \Gamma_i\cup\Gamma_b, \tag{3.44}$$

$$\frac{\partial \mathbf{v}}{\partial \mathbf{n}} = 0 \text{ on } \Gamma_i\cup\Gamma_b,$$

$$\frac{\partial \rho}{\partial \mathbf{n}} = 0 \text{ on } \partial\mathcal{M}. \tag{3.45}$$

We also provide the system with the following initial condition:

$$\mathbf{v}(0,x) = \mathbf{v}_0(x), \quad \rho(0,x) = \rho_0(x), \tag{3.46}$$

where $\mathbf{v}_0$ has to satisfy the following compatibility condition:

$$\int_0^1 \operatorname{div}\mathbf{v}_0\,\mathrm{d}z = 0. \tag{3.47}$$

For simplicity, we consider the right-hand-side $F_{\mathbf{v}}$ of the first equation (Eq. (3.43)) to be zero, but the result stays valid for a nonzero forcing term (which is not of physical relevance of course).

In order to prove the existence of the strong solution for Eq. (3.43), we need to prove the appropriate a priori estimates. We start by obtaining a priori estimates on the pressure and on the density; we proceed in a formal way, i.e., assuming enough regularity.

A priori estimates on $\mathbf{v}$, ρ, *and* p

LEMMA 3.2. *The density* ρ *solution of the PEs satisfies the following estimate.*

$$\sup_{0\le t\le T}|\rho(t)|_{L^4}^4 + 12\nu_\rho\int_0^T |\rho\nabla_3\rho|_{L^2}^2\,\mathrm{d}t \le c|\rho_0|_{L^4}^4. \tag{3.48}$$

PROOF. The proof of the lemma is immediate; we take the L^2-scalar product of Eq. $(3.43)_4$ with ρ^3 and we find

$$\frac{1}{4}\frac{\mathrm{d}}{\mathrm{d}t}|\rho(t)|_{L^4}^4 + 3\nu_\rho|\rho\nabla_3\rho|_{L^2}^2 = 0. \tag{3.49}$$

Integrating Eq. (3.49) in t, we find Eq. (3.48). □

From the hydrostatic equation Eq. $(3.43)_2$, the following inequality is then deduced:

$$\sup_{0\le t\le T}|p_z(t)|_{L^4} \le c|\rho_0|_{L^4}. \tag{3.50}$$

We now estimate the L^2-norm of the velocity field.

LEMMA 3.3. *For a solution of the PEs, the following estimate holds*

$$\sup_{0\le t\le T}|\mathbf{v}(t)|_{L^2}^2 + \nu_{\mathbf{v}}\int_{L^2}^{T}|\nabla_3\mathbf{v}(t)|_{L^2}^2\mathrm{d}t \le c(|\mathbf{v_0}|_{L^2}^2 + |\rho_0|_{L^4}^2). \tag{3.51}$$

PROOF. We take the L^2-scalar product of Eq. $(3.43)_1$ by $\mathbf{v}$ and using the conservation of mass (Eq. $(3.43))_3$, and the boundary conditions, we find

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|\mathbf{v}(t)|_{L^2}^2 + \nu_{\mathbf{v}}|\nabla_3\mathbf{v}(t)|_{L^2}^2 - \frac{1}{\rho_0}(p,\ \mathrm{div}\ \mathbf{v})_{L^2} = 0. \tag{3.52}$$

We need to estimate the pressure term from Eq. (3.52):

$$\begin{aligned}|(p,\ \mathrm{div}\ \mathbf{v})_{L^2}| &= |(p, \frac{\partial w}{\partial z})_{L^2}| = |(p_z, w)_{L^2}|\\ &\le |p_z|_{L^2}|w|_{L^2} \le c|p_z|_{L^2}|\nabla\mathbf{v}|_{L^2}\\ &\le \frac{\nu_{\mathbf{v}}}{2}|\nabla\mathbf{v}|_{L^2}^2 + \frac{c}{\nu_{\mathbf{v}}}|p_z|_{L^2}.\end{aligned} \tag{3.53}$$

Using the estimates obtained above into Eq. (3.52), we find

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|\mathbf{v}(t)|_{L^2}^2 + \frac{\nu_{\mathbf{v}}}{2}|\nabla_3\mathbf{v}(t)|_{L^2}^2 \le \frac{c}{\nu_{\mathbf{v}}}|p_z(t)|_{L^2}^2 \le \frac{c}{\nu_{\mathbf{v}}}|\rho_0|_{L^4}^2; \tag{3.54}$$

we also took advantage of Eq. (3.50).

The proof of Lemma 3.3 is completed by applying the Gronwall lemma to Eq. (3.51). □

An immediate consequence of this result is as follows:

COROLLARY 3.1. *For the vertical velocity w, the following estimate holds:*

$$\int_0^T (|w(t)|_{L^2}^2 + |w_z(t)|_{L^2}^2)\mathrm{d}t \leq c. \tag{3.55}$$

We can now estimate the pressure in terms of the velocity.

LEMMA 3.4. *The pressure p from the PEs can be estimated as follows:*

$$|p|_{L^4} \leq c\left[(|\nabla_3|\mathbf{v}|^2|_{L^2}^{1/2} + |\mathbf{v}|_{L^4} + 1)|\mathbf{v}|_{L^4} + 1\right]. \tag{3.56}$$

PROOF. We write the pressure p as $p = p_1 + p_2$, where p_1 is the average of the pressure in the vertical direction:

$$p_1(x, y) = \int_0^1 p(x, y, z)\mathrm{d}z, \tag{3.57}$$

and p_2 is an antiderivative of p_z in z, of zero average in the vertical direction. Since the pressure p is determined up to a constant, we can assume that $(p_1^3, 1)_{L^2} = 0$. Indeed, p and p_1 being defined up to an additive constant c, and $((p_1 + c)^3, 1)_{L^2} = 0$ is an equation of the third degree in c, which always has a real solution.

Using Eq. (3.50), the pressure p_2 is estimated as follows:

$$\sup_{0\leq t\leq T} |p_2(t)|_{L^4} \leq c|\partial_z p_2(t)|_{L^4} = c|\partial_z p(t)|_{L^4} \leq c. \tag{3.58}$$

In order to estimate p_1, we introduce the following boundary value problem in $\mathcal{M}'$:

$$\begin{aligned} &\Delta q = p_1^3 \text{ in } \mathcal{M}', \\ &\frac{\partial q}{\partial n} = 0 \text{ on } \partial\mathcal{M}', \end{aligned} \tag{3.59}$$

with q of zero average. As mentioned before, Δ stands for the 2D Laplacian.

We note that since we imposed $(p_1^3, 1)_{L^2} = 0$, the solvability of Eq. (3.59) is ensured.

Let $\overline{\mathbf{v}}(x, y, t) = \int_0^1 \mathbf{v}(x, y, z, t)\mathrm{d}z$ be the vertical average of the velocity field. We can write then $\mathbf{v} = \overline{\mathbf{v}} + \tilde{\mathbf{v}}$, and we note that $\int_0^1 \tilde{\mathbf{v}}\mathrm{d}z = 0$. Due to the incompressibility equation (Eq. $(3.43)_3$), $\overline{\mathbf{v}}$ satisfies

$$\begin{aligned} &\mathrm{div}\,\overline{\mathbf{v}} = \int_0^1 \mathrm{div}\,\mathbf{v}\mathrm{d}z = -\int_0^1 \frac{\partial w}{\partial z}\mathrm{d}z = 0, \\ &\overline{\mathbf{v}}\cdot\mathbf{n} = \frac{\partial\overline{\mathbf{v}}}{\partial\mathbf{n}} \times \mathbf{n} = 0 \text{ on } \Gamma_l. \end{aligned} \tag{3.60}$$

From Eq. (3.60), we deduce then that $\overline{\mathbf{v}}$ is of the form $\overline{\mathbf{v}} = \mathrm{curl}\psi$.

We take the L^2-scalar product of Eq. $(3.43)_1$ with ∇q and integrating by parts we find

$$(\mathbf{v}_t, \nabla q)_{L^2} = -(\operatorname{div} \mathbf{v}_t, q)_{L^2} = (w_{tz}, q)_{L^2}$$
$$= \int_{\mathcal{M}'} q \left(\int_0^1 w_{tz} \mathrm{d}z \right) \mathrm{d}\mathcal{M}' = 0. \tag{3.61}$$

Using the boundary condition $\partial \mathbf{v}/\partial \mathbf{n} = 0$ on $\Gamma_i \cup \Gamma_b$, we also deduce

$$(\partial_z^2 \mathbf{v}, \nabla q)_{L^2} = 0. \tag{3.62}$$

By integration by parts, we also find

$$I = -(\Delta \mathbf{v}, \nabla q)_{L^2} = -(\Delta(\operatorname{curl} \psi + \tilde{\mathbf{v}}), \nabla q)_{L^2} \quad (\text{since } \int_0^1 \tilde{\mathbf{v}} \mathrm{d}z = 0)$$
$$= -(\operatorname{curl} \Delta\psi, \nabla q)_{L^2(\mathcal{M}')} = \int_{\partial \mathcal{M}'} \Delta\psi \left(\frac{\partial q}{\partial y} n_1 - \frac{\partial q}{\partial x} n_2 \right) \mathrm{d}(\partial \mathcal{M}'). \tag{3.63}$$

We need to estimate I. Since $\Delta\psi = \operatorname{curl} \overline{\mathbf{v}}$, we need to estimate $\operatorname{curl} \overline{\mathbf{v}}$ on $\partial\mathcal{M}'$. For each point $M_0 \in \partial\mathcal{M}'$, we consider the basis $(\tau, \mathbf{n})$, where $\mathbf{n}$ is the outward unit normal vector to $\partial\mathcal{M}'$, satisfying $\tau \times \mathbf{n} = e_z$ (where e_z is the unitary vertical vector oriented upward). Let $(x_\tau, x_\mathbf{n})$ be the coordinates of a point M_0 and $(u_\tau, u_\mathbf{n})$ the components of a vector u in the coordinate system $(\tau, \mathbf{n})$.

We know $\overline{\mathbf{v}} \cdot \mathbf{n} = 0$ on $\partial\mathcal{M}'$, so this means that the normal component of $\overline{\mathbf{v}}$ is zero: $\overline{\mathbf{v}}_\mathbf{n} = 0$. This implies $\partial \overline{\mathbf{v}}_\mathbf{n}/\partial\tau = 0$, and so $\operatorname{curl} \overline{\mathbf{v}} = \partial \overline{\mathbf{v}}_\mathbf{n}/\partial\tau - \partial \overline{\mathbf{v}}_\tau/\partial\mathbf{n} = -\partial \overline{\mathbf{v}}_\tau/\partial\mathbf{n}$.

We also know that $(\partial\overline{\mathbf{v}}/\partial\mathbf{n}) \times \mathbf{n} = 0$ on $\partial\mathcal{M}'$, and this reduces to $\partial \overline{\mathbf{v}}_\tau/\partial\mathbf{n} = 0$. We can now conclude that $\operatorname{curl} \overline{\mathbf{v}} = 0$, which implies $I = 0$.

Now, let us continue the estimation of the norm of p. We have

$$(\nabla_\mathbf{v} \mathbf{v}, \nabla q)_{L^2} + \left(w \frac{\partial \mathbf{v}}{\partial z}, \nabla q \right)_{L^2} + \frac{1}{\rho_0} (\nabla p, \nabla q)_{L^2} + 2f(\mathbf{k} \times \mathbf{v}, \nabla q)_{L^2} = 0. \tag{3.64}$$

We use the fact that $q = \Delta^{-1} p_1^3$, where Δ^{-1} is an operator inverse to Δ with the Neumann boundary condition.

The pressure term from Eq. (3.64) becomes

$$\frac{1}{\rho_0} (\nabla p, \nabla \Delta^{-1} p_1^3)_{L^2} = -\frac{1}{\rho_0} (p, p_1^3)_{L^2} = -\frac{1}{\rho_0} (p_1, p_1^3)_{L^2(\mathcal{M}')} = -\frac{1}{\rho_0} |p_1|^4_{L^4(\mathcal{M}')}. \tag{3.65}$$

For the Coriolis term, we find

$$2f|(\mathbf{k} \times \mathbf{v}, \nabla q)_{L^2}| \le c|\mathbf{v}|_{L^4} |\nabla \Delta^{-1} p_1^3|_{L^{4/3}}$$
$$\le c|\mathbf{v}|_{L^4} \left(\int_\mathcal{M} |p_1|^4 \mathrm{d}\mathcal{M} \right)^{3/4} = c|\mathbf{v}|_{L^4} |p_1|^3_{L^4}. \tag{3.66}$$

The scalar product containing the nonlinear terms is estimated as follows, using integration by parts and Sobolev embeddings:

$$
\begin{aligned}
&|(\nabla_{\mathbf{v}}\mathbf{v}, \nabla q)_{L^2} + \left(w\frac{\partial \mathbf{v}}{\partial z}, \nabla q\right)_{L^2}| = \\
&\quad = |(u\mathbf{v}, \partial_x \nabla q)_{L^2} + (v\mathbf{v}, \partial_y \nabla q)_{L^2} + (w\mathbf{v}, \partial_z \nabla q)_{L^2}| \\
&\quad = |(u\mathbf{v}, \partial_x \nabla \Delta^{-1} p_1^3)_{L^2} + (v\mathbf{v}, \partial_y \nabla \Delta^{-1} p_1^3)_{L^2}| \\
&\quad \le c \int_0^1 ||\mathbf{v}|^2|_{L^4(\mathcal{M}')} |p_1|^3_{L^4(\mathcal{M}')} \mathrm{d}z \\
&\quad \le c|p_1|^3_{L^4(\mathcal{M}')} \int_0^1 \left(|\nabla|\mathbf{v}|^2|^{1/2}_{L^2(\mathcal{M}')} + ||\mathbf{v}|^2|^{1/2}_{L^2(\mathcal{M}')}\right) \|\mathbf{v}|^2|^{1/2}_{L^2(\mathcal{M}')} \mathrm{d}z \\
&\quad \le c|p_1|^3_{L^4} \left(|\nabla|\mathbf{v}|^2|^{1/2}_{L^2} + |\mathbf{v}|_{L^4}\right) |\mathbf{v}|_{L^4}. \qquad (3.67)
\end{aligned}
$$

Gathering all the above estimates, we find

$$
|p_1|_{L^4} \le c\left[|\nabla_3 |\mathbf{v}|^2|^{1/2}_{L^2} + |\mathbf{v}|_{L^4} + 1\right] |\mathbf{v}|_{L^4}, \qquad (3.68)
$$

and now Lemma 3.4 follows immediately. □

In the proof of Theorem 3.2, we use a generalization of the Sobolev embeddings, for the case of a Lipschitz domain.

LEMMA 3.5. *If $f \in H^1(\mathcal{M})$, with $\mathcal{M}$ a 3D Lipschitz domain, then*

$$
|f|_{L^4} \le c(|f|_{L^2} + |\nabla_3 f|^{3/4}_{L^2} |f|^{1/4}_{L^2}). \qquad (3.69)
$$

PROOF. Let us recall the proof of this well-known result. For a function $g \in H^1_0(\mathcal{M})$ (the space of functions from $H^1(\mathcal{M})$ vanishing on the boundary), the following inequality is valid:

$$
|g|_{L^4} \le c|\nabla_3 g|^{3/4} |g|^{1/4}. \qquad (3.70)
$$

Let $\mathcal{M} \subset G$; f can be extended from $\mathcal{M}$ to G preserving the class and norm in such way that the extended function f vanishes on the boundary of G. We also know that

$$
|\tilde{f}|_{H^1(G)} \le c|f|_{H^1(\mathcal{M})}, \quad |\tilde{f}|_{L^4(G)} \le c|f|_{L^4(\mathcal{M})}. \qquad (3.71)
$$

Then,

$$
\begin{aligned}
|f|_{L^4(\mathcal{M})} &\le c|\tilde{f}|_{L^4(G)} \le c|\nabla_3 \tilde{f}|^{3/4}_{L^2(G)}|\tilde{f}|^{1/4}_{L^2(G)} \\
&\le c|f|^{1/4}_{L^2(\mathcal{M})}\left(|f|^2_{L^2(\mathcal{M})} + |\nabla_3 f|^2_{L^2(\mathcal{M})}\right)^{3/8} \\
&\le c\left(|f|_{L^2(\mathcal{M})} + |\nabla_3 f|^{3/4}_{L^2(\mathcal{M})}|f|^{1/4}_{L^2(\mathcal{M})}\right). \qquad (3.72)
\end{aligned}
$$

The lemma is proved. □

A similar result works for the 2D case, which will be used later on.

LEMMA 3.6. *If $f \in H^1(\mathcal{M})$, and $\mathcal{M}$ is a 2D Lipschitz domain, then*

$$
|f|_{L^4(\mathcal{M})} \le c(|f|_{L^2(\mathcal{M})} + |\nabla f|^{1/2}_{L^2(\mathcal{M})}|f|^{1/2}_{L^2(\mathcal{M})}). \qquad (3.73)
$$

LEMMA 3.7. *For* $\mathbf{v}$ *a solution of the PEs, the following estimate holds*

$$
\sup_{0\le t\le T} |\mathbf{v}|^4_{L^4} + \nu_{\mathbf{v}} \int_0^T |\mathbf{v}|\nabla_3 \mathbf{v}||^2_{L^2} \le c, \qquad (3.74)
$$

where c is a constant depending on the initial data.

PROOF. We take the L^2-scalar product of Eq. $(3.43)_1$ with $\mathbf{v}|\mathbf{v}|^2$:

$$
\begin{aligned}
&\frac{1}{4}\frac{\mathrm{d}}{\mathrm{d}t}|\mathbf{v}|^4_{L^4} + (\nabla_{\mathbf{v}}\mathbf{v}, \mathbf{v}|\mathbf{v}|^2)_{L^2} + (w\frac{\partial \mathbf{v}}{\partial z}, \mathbf{v}|\mathbf{v}|^2)_{L^2} \\
&\qquad + \frac{1}{\rho_0}(\nabla p, \mathbf{v}|\mathbf{v}|^2)_{L^2} - \nu_{\mathbf{v}}(\Delta_3 \mathbf{v}, \mathbf{v}|\mathbf{v}|^2)_{L^2} = 0. \qquad (3.75)
\end{aligned}
$$

The scalar products from Eq. (3.75) containing the nonlinear terms are estimated as follows:

$$
\begin{aligned}
(\nabla_{\mathbf{v}}\mathbf{v}, \mathbf{v}|\mathbf{v}|^2)_{L^2} + (w\frac{\partial \mathbf{v}}{\partial z}, \mathbf{v}|\mathbf{v}|^2)_{L^2} &= \frac{1}{4}\int_{\mathcal{M}} \nabla_{\mathbf{v}}|\mathbf{v}|^4 \mathrm{d}\mathcal{M} + \frac{1}{4}\int_{\mathcal{M}} w\frac{\partial}{\partial z}|\mathbf{v}|^4 \mathrm{d}\mathcal{M} \\
&= \frac{1}{4}\int_{\partial\mathcal{M}} \mathbf{v}\cdot\mathbf{n}\,|\mathbf{v}|^4 \mathrm{d}(\partial\mathcal{M}) + \frac{1}{4}\int_{\partial\mathcal{M}} w n_3 |\mathbf{v}|^4 \mathrm{d}(\partial\mathcal{M}) \\
&\quad - \frac{1}{4}\int_{\mathcal{M}} (\,\mathrm{div}\,\mathbf{v} + \frac{\partial w}{\partial z})|\mathbf{v}|^4 \mathrm{d}\mathcal{M} \\
&= 0. \qquad (3.76)
\end{aligned}
$$

We also have

$$
\begin{aligned}
-(\Delta_3 \mathbf{v}, \mathbf{v}|\mathbf{v}|^2)_{L^2} &= -\int_{\partial\mathcal{M}} \frac{\partial \mathbf{v}}{\partial \mathbf{n}}\cdot\mathbf{v}|\mathbf{v}|^2 \mathrm{d}(\partial\mathcal{M}) + (\nabla_3 \mathbf{v}, \nabla_3(\mathbf{v}|\mathbf{v}|^2))_{L^2} \\
&= \int_{\mathcal{M}} |\nabla_3 \mathbf{v}|^2 |\mathbf{v}|^2 \mathrm{d}\mathcal{M} + \int_{\mathcal{M}} |\nabla_3 |\mathbf{v}|^2|^2 \mathrm{d}\mathcal{M}. \qquad (3.77)
\end{aligned}
$$

It now remains to estimate the pressure term from Eq. (3.75):

$$\begin{aligned}|(\nabla p, \mathbf{v}|\mathbf{v}|^2)_{L^2}| &= |(p,\ \text{div}\ (\mathbf{v}|\mathbf{v}|^2))_{L^2}| \\ &\le |(p\mathbf{v}, \nabla|\mathbf{v}|^2)_{L^2}| + |(p|\mathbf{v}|^2,\ \text{div}\ \mathbf{v})_{L^2}| \\ &\le c|(p|\mathbf{v}|^2, |\nabla\mathbf{v}|)_{L^2}| \\ &\le c|(p_1|\mathbf{v}|^2, |\nabla\mathbf{v}|)_{L^2}| + c|(p_2|\mathbf{v}|^2, |\nabla\mathbf{v}|)_{L^2}|. \end{aligned} \tag{3.78}$$

For the term containing the pressure p_1, we have

$$\begin{aligned}|(p_1|\mathbf{v}|^2, |\nabla\mathbf{v}|)_{L^2}| &\le \int_0^1 |p_1|_{L^4(\mathcal{M}')}||\mathbf{v}|^2|_{L^4(\mathcal{M}')}|\nabla\mathbf{v}|_{L^2(\mathcal{M}')}\mathrm{d}z \\ &\le c|p_1|_{L^4(\mathcal{M}')}\int_0^1 (|\nabla|\mathbf{v}|^2|^{1/2}_{L^2(\mathcal{M}')} + ||\mathbf{v}|^2|^{1/2}_{L^2(\mathcal{M}')})||\mathbf{v}|^2|^{1/2}_{L^2(\mathcal{M}')}|\nabla\mathbf{v}|_{L^2(\mathcal{M}')}\mathrm{d}z \\ &\le c|p_1|_{L^4(\mathcal{M}')}(|\nabla|\mathbf{v}|^2|^{1/2}_{L^2(\mathcal{M})} + |\mathbf{v}|_{L^4(\mathcal{M})})|\mathbf{v}|_{L^4(\mathcal{M})}|\nabla\mathbf{v}|_{L^2(\mathcal{M})} \\ &\le (\text{ by Eq. 3.68}) \quad c(|\nabla|\mathbf{v}|^2|_{L^2(\mathcal{M})} + |\mathbf{v}|^2_{L^4(\mathcal{M})} + 1)|\mathbf{v}|^2_{L^4(\mathcal{M})}|\nabla\mathbf{v}|_{L^2(\mathcal{M})} \\ &\le \frac{\nu_\mathbf{v}}{4}|\nabla|\mathbf{v}|^2|^2_{L^2} + \frac{c}{\nu_\mathbf{v}}|\mathbf{v}|^4_{L^4}|\nabla\mathbf{v}|^2_{L^2} + c(|\mathbf{v}|^4_{L^4} + 1)|\nabla\mathbf{v}|_{L^2}. \end{aligned} \tag{3.79}$$

Using the same reasoning as above, we also have

$$\begin{aligned}|(p_2|\mathbf{v}|^2, |\nabla\mathbf{v}|)_{L^2}| &\le |p_2|_{L^4}||\mathbf{v}|^2|_{L^4}|\nabla\mathbf{v}|_{L^2} \\ &\le c\left(|\nabla_3|\mathbf{v}|^2|^{3/4}_{L^2} + |\mathbf{v}|^{3/2}_{L^4}\right)|\mathbf{v}|^{1/2}_{L^4}|\nabla\mathbf{v}|_{L^2} \\ &\le \frac{\nu_\mathbf{v}}{4}|\nabla_3|\mathbf{v}|^2|^2_{L^2} + c|\mathbf{v}|^4_{L^4} + c|\nabla\mathbf{v}|^2_{L^2}. \end{aligned} \tag{3.80}$$

Gathering all the estimates, we find

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}|\mathbf{v}|^4_{L^4} + \nu_\mathbf{v}\int_\mathcal{M}|\nabla_3\mathbf{v}|^2|\mathbf{v}|^2\mathrm{d}\mathcal{M} + \nu_\mathbf{v}\int_\mathcal{M}|\nabla_3|\mathbf{v}|^2|^2\mathrm{d}\mathcal{M} \\ &\qquad \le c(|\mathbf{v}|^4_{L^4} + 1)(|\nabla_3\mathbf{v}|^2_{L^2} + 1). \end{aligned} \tag{3.81}$$

Using the Gronwall lemma, Eq. (3.81) leads to

$$\sup_{0\le t\le T}|\mathbf{v}|^4_{L^4} + \nu_\mathbf{v}\int_0^T|\nabla_3|\mathbf{v}|^2|^2_{L^2(\mathcal{M})}\mathrm{d}t + \nu_\mathbf{v}\int_0^T||\nabla_3\mathbf{v}||\mathbf{v}||^2_{L^2(\mathcal{M})}\mathrm{d}t \le c. \tag{3.82}$$

□

COROLLARY 3.2. *For the pressure function, the following estimate holds*

$$\int_0^T |p|^4_{L^4}\mathrm{d}t \le c. \tag{3.83}$$

PROOF. Using Eqs. (3.56) and (3.74), the result follows immediately. □

Using all the estimates obtained till now, we can conclude that

$$\sup_{0\le t\le T}(|\mathbf{v}|_{L^4}^4+|\rho|_{L^4}^4)+c_0\int_0^T(|\rho\nabla_3\rho|_{L^2}^2 + ||\nabla_3\mathbf{v}||\mathbf{v}||_{L^2}^2+|\nabla_3|\mathbf{v}|^2|_{L^2}^2+|\rho|_{L^4}^4)\mathrm{d}t\le c. \tag{3.84}$$

A priori estimates on $\mathbf{v}_z$ *and* ρ_z In order to estimate $\mathbf{v}_z$ and ρ_z, we differentiate Eqs. $(3.43)_1$ and $(3.43)_4$ in z. Denoting $\mathbf{u}=\mathbf{v}_z$, we find

$$\begin{aligned}&\mathbf{u}_t-\nu_{\mathbf{v}}\Delta_3\mathbf{u}+2f\mathbf{k}\times\mathbf{u}+\frac{1}{\rho_0}\nabla p_z+\nabla_{\mathbf{v}}\mathbf{u}+w\frac{\partial\mathbf{u}}{\partial z}\\&+\nabla_{\mathbf{u}}\mathbf{v}+w_z\frac{\partial\mathbf{v}}{\partial z}=0,\\&\operatorname{div}\mathbf{u}=0,\\&\rho_{zt}+\nabla_{\mathbf{v}}\rho_z+w\frac{\partial\rho_z}{\partial z}+\nabla_{\mathbf{u}}\rho+w_z\frac{\partial\rho}{\partial z}-\nu_\rho\Delta_3\rho=0.\end{aligned} \tag{3.85}$$

LEMMA 3.8. *For* $\mathbf{u}=\mathbf{v}_z$*, the following estimate holds*

$$\sup_{0\le t\le T}|\mathbf{u}(t)|_{L^2}^2+\nu_{\mathbf{v}}\int_0^T|\nabla_3\mathbf{u}|_{L^2}^2\mathrm{d}t\le c. \tag{3.86}$$

PROOF. We take the L^2-scalar product of Eq. $(3.85)_1$ with $\mathbf{u}$:

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|\mathbf{u}(t)|_{L^2}^2+\nu_{\mathbf{v}}|\nabla_3\mathbf{u}|_{L^2}^2+\frac{1}{\rho_0}(\nabla p_z,\mathbf{u})_{L^2}+(\nabla_{\mathbf{u}}\mathbf{v}+w_z\frac{\partial v}{\partial z},\mathbf{u})_{L^2}=0. \tag{3.87}$$

We note here that $(\nabla_{\mathbf{v}},\mathbf{u},\mathbf{u})_{L^2}+(w\partial\mathbf{v}/\partial z,\mathbf{u})_{L^2}=0$ because of the incompressibility condition.

The pressure term from Eq. (3.87) is estimated as follows:

$$\begin{aligned}|(\nabla p_z,\mathbf{u})_{L^2}|&=|(p_z,\ \operatorname{div}\mathbf{u})|\le c|p_z|_{L^2}|\nabla_3\mathbf{u}|_{L^2}\\&\le c|\rho|_{L^2}|\nabla_3\mathbf{u}|_{L^2}\le c|\nabla_3\mathbf{u}|_{L^2}\le\frac{\nu_{\mathbf{v}}}{4}|\nabla_3\mathbf{u}|_{L^2}^2+c.\end{aligned} \tag{3.88}$$

The remaining nonlinear terms from Eq. (3.87) give us:

$$
\begin{aligned}
&|(\nabla_{\mathbf{u}}\mathbf{v}, \mathbf{u})_{L^2} - \left((\operatorname{div} \mathbf{v})\frac{\partial \mathbf{v}}{\partial z}, \mathbf{u}\right)_{L^2}| \\
&\quad \leq c\int_{\mathcal{M}} |\mathbf{v}||\nabla \mathbf{u}||\mathbf{u}|\,\mathrm{d}\mathcal{M} + |\int_{\mathcal{M}} (\operatorname{div} \mathbf{v}_z)\mathbf{v}\cdot\mathbf{u}\mathrm{d}\mathcal{M}| \\
&\qquad + |\int_{\mathcal{M}} (\operatorname{div} \mathbf{v})\mathbf{v}\cdot\mathbf{u}_z\mathrm{d}\mathcal{M}| \\
&\quad \leq c|\mathbf{v}|_{L^4}|\mathbf{u}|_{L^4}|\nabla\mathbf{u}|_{L^2} + |(\operatorname{div}\mathbf{v})\mathbf{v}|_{L^2}|\mathbf{u}_z|_{L^2} \\
&\quad \leq c|\mathbf{v}|_{L^4}|\mathbf{u}|_{L^2}^{1/4}|\nabla_3\mathbf{u}|_{L^2}^{7/4} + |(\operatorname{div}\mathbf{v})\mathbf{v}|_{L^2}|\mathbf{u}_z|_{L}^2 \\
&\quad \leq \frac{\nu_{\mathbf{v}}}{4}|\nabla_3\mathbf{u}|_{L^2}^2 + c|\mathbf{v}|_{L^4}^8|\mathbf{u}|_{L^2}^2 + c|(\operatorname{div}\mathbf{v})\mathbf{v}|_{L^2}^2. \qquad (3.89)
\end{aligned}
$$

Gathering all these estimates, we find:

$$
\frac{\mathrm{d}}{\mathrm{d}t}|\mathbf{u}|_{L^2}^2 + \nu_{\mathbf{v}}|\nabla_3\mathbf{u}|_{L^2}^2 \leq c|\mathbf{u}|_{L^2}^2 + c|(\operatorname{div}\mathbf{v})\mathbf{v}|_{L^2}^2. \qquad (3.90)
$$

From Eq. (3.84) and using the Gronwall lemma, the required result is proved. □

We now estimate the L^4-norm of $\mathbf{u}$.

LEMMA 3.9. *For* $\mathbf{u} = \mathbf{v}_z$, *the following estimate holds*

$$
\sup_{0\leq t\leq T}|\mathbf{u}|_{L^4}^4 + \nu_{\mathbf{v}}\int_0^T \big||\nabla_3\mathbf{u}||\mathbf{u}|\big|_{L^2}^2\mathrm{d}t + \nu_{\mathbf{v}}\int_0^T |\nabla_3|\mathbf{u}|^2|_{L^2}^2\mathrm{d}t \leq c. \qquad (3.91)
$$

PROOF. We take the L^2-scalar product of Eq. $(3.85)_1$ with $|\mathbf{u}|^2\mathbf{u}$. After integration by parts, we find

$$
\begin{aligned}
&\frac{1}{4}\frac{\mathrm{d}}{\mathrm{d}t}|\mathbf{u}|_{L^4}^4 + \nu_{\mathbf{v}}\int_{\mathcal{M}}|\nabla_3\mathbf{u}|^2|\mathbf{u}|^2\mathrm{d}\mathcal{M} + \frac{\nu_{\mathbf{v}}}{2}\int_{\mathcal{M}}|\nabla_3|\mathbf{u}|^2|^2 d\mathcal{M} \\
&\qquad + \frac{1}{\rho_0}(\nabla p_z, \mathbf{u}|\mathbf{u}|^2)_{L^2} + (\nabla_{\mathbf{u}}\mathbf{v} + w_z\frac{\partial\mathbf{v}}{\partial z}, |\mathbf{u}|^2\mathbf{u})_{L^2} = 0. \qquad (3.92)
\end{aligned}
$$

The pressure term is estimated in the following way[9]:

$$
\begin{aligned}
|(\nabla p_z, \mathbf{u}|\mathbf{u}|^2)_{L^2}| &\leq |(p_z, (\operatorname{div}\mathbf{u})|\mathbf{u}|^2)_{L^2}| + |(p_z, \mathbf{u}\cdot\nabla(|\mathbf{u}|^2))_{L^2}| \\
&\leq c|p_z|_{L^4}\big||\nabla\mathbf{u}||\mathbf{u}|\big|_{L^2}|\mathbf{u}|_{L^4} \\
&\leq c|\rho|_{L^4}\big||\nabla\mathbf{u}||\mathbf{u}|\big|_{L^2}|\nabla_3\mathbf{u}|_{L^2}^{3/4}|\mathbf{u}|_{L^2}^{1/4} \\
&\leq \frac{\nu_{\mathbf{v}}}{4}\big||\nabla_3\mathbf{u}||\mathbf{u}|\big|_{L^2}^2 + c|\nabla_3\mathbf{u}|_{L^2}^2 + c|\mathbf{u}|_{L^2}^2. \qquad (3.93)
\end{aligned}
$$

[9] In the following equations, $|\mathbf{u}|^2$ stands for the pointwise value of the norm of the vector $\mathbf{u}$ and similarly for the other quantities.

Here, we applied (for $\mathbf{u}$) the fact that in a 3D domain the estimate $|w|_{L^4} \leq c|\nabla_3 w|_{L^2}^{3/4}|w|_{L^2}^{1/4}$ is valid for a function w vanishing on part of the boundary.

We also need to estimate the last term from the left-hand side of Eq. (3.92):

$$\begin{aligned}
I_1 &= |(\nabla_{\mathbf{u}}\mathbf{v}, |\mathbf{u}|^2\mathbf{u})_{L^2}| \\
&= |\int_{\mathcal{M}} \mathbf{v}(\operatorname{div}\mathbf{u})|\mathbf{u}|^2\mathbf{u}\,\mathrm{d}\mathcal{M} + \int_{\mathcal{M}} \mathbf{v}\nabla_{\mathbf{u}}(|\mathbf{u}|^2\mathbf{u})\,\mathrm{d}\mathcal{M}| \\
&\leq \text{ (by integration by parts)} \\
&\leq \int_{\mathcal{M}} |\operatorname{div}\mathbf{u}||\mathbf{v}||\mathbf{u}|^3\mathrm{d}\mathcal{M} + \int_{\mathcal{M}} |\nabla\mathbf{u}||\mathbf{v}||\mathbf{u}|^3\mathrm{d}\mathcal{M} \\
&\leq c\big||\nabla_3\mathbf{u}||\mathbf{u}|\big|_{L^2}|\mathbf{v}|_{L^4}\left(\int_{\mathcal{M}}|\mathbf{u}|^8 d\mathcal{M}\right)^{1/4} \\
&\leq c\big||\nabla_3\mathbf{u}||\mathbf{u}|\big|_{L^2}|\mathbf{v}|_{L^4}||\mathbf{u}|^2|_{L^2}^{1/4}|\nabla_3|\mathbf{u}|^2|_{L^2}^{3/4} \\
&\leq \frac{\nu_{\mathbf{v}}}{8}\left(\|\nabla_3\mathbf{u}\|\mathbf{u}\|_{L^2}^2 + |\nabla_3|\mathbf{u}|^2|_{L^2}^2\right) + c|\mathbf{v}|_{L^4}^8|\mathbf{u}|_{L^4}^4. \qquad (3.94)
\end{aligned}$$

Integrating by parts, we also have

$$\begin{aligned}
I_2 &= \left|\left(w_z\frac{\partial\mathbf{v}}{\partial z}, |\mathbf{u}|^2\mathbf{u}\right)_{L^2}\right| = |(\operatorname{div}(\mathbf{v})\,\mathbf{u}, |\mathbf{u}|^2\mathbf{u})_{L^2}| \\
&\leq \int_{\mathcal{M}} |\mathbf{v}||\nabla|\mathbf{u}|^4|\mathrm{d}\mathcal{M} \\
&\leq c\int_{\mathcal{M}} |\mathbf{v}||\mathbf{u}|^2|\nabla|\mathbf{u}|^2|\mathrm{d}\mathcal{M} \\
&\leq c|\mathbf{v}|_{L^4}||\mathbf{u}|^2|_{L^4}|\nabla|\mathbf{u}|^2|_{L^2} \\
&\leq c|\mathbf{v}|_{L^4}|\nabla_3|\mathbf{u}|^2|_{L^2}^{3/4}|\mathbf{u}|_{L^4}^{1/2}|\nabla|\mathbf{u}|^2|_{L^2} \\
&\leq \frac{\nu_{\mathbf{v}}}{8}|\nabla_3|\mathbf{u}|^2|_{L^2}^2 + c|\mathbf{v}|_{L^4}^8|\mathbf{u}|_{L^4}^4. \qquad (3.95)
\end{aligned}$$

Therefore, from all the above inequalities, we find

$$\frac{\mathrm{d}}{\mathrm{d}t}|\mathbf{u}|_{L^4}^4 + \nu_{\mathbf{v}}\int_0^T \left[\|\mathbf{u}\|\nabla_3\mathbf{u}\|_{L^2}^2 + |\nabla_3|\mathbf{u}|^2|_{L^2}^2\right]\mathrm{d}t \leq c\left(|\mathbf{u}|_{L^4}^4 + |\nabla_3\mathbf{u}|_{L^2}^2 + |\mathbf{u}|_{L^2}^2\right). \qquad (3.96)$$

Applying the Gronwall lemma to Eq. (3.96) and using Eq. (3.86), Eq. (3.91) follows immediately. □

We now need to deduce the estimates for ρ_z.

LEMMA 3.10. *For ρ_z, the following estimate holds*

$$\sup_{0\le t\le T}|\rho_z|_{L^2}^2+\nu_\rho\int_0^T|\nabla_3\rho_z|_{L^2}^2\mathrm{d}t\le c. \tag{3.97}$$

PROOF. We take the L^2-scalar product of Eq. $(3.85)_3$ with ρ_z, and we find

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|\rho_z|_{L^2}^2+\nu_\rho|\nabla_3\rho_z|_{L^2}^2+\left(\nabla_{\mathbf{u}}\rho+w_z\frac{\partial\rho}{\partial z},\rho_z\right)_{L^2}=0. \tag{3.98}$$

We have

$$\begin{aligned}I_1=|\nabla_{\mathbf{u}}\rho,\rho_z)_{L^2}|&=|(\operatorname{div}\mathbf{u},\rho\rho_z)_{L^2}+(\rho\mathbf{u},\nabla\rho_z)_{L^2}|\\&\le|\operatorname{div}\mathbf{u}|_{L^2}|\rho|_{L^4}|\rho_z|_{L^4}+|\rho|_{L^4}|\mathbf{u}|_{L^4}|\nabla\rho_z|_{L^2}\\&\le c|\operatorname{div}\mathbf{u}|_{L^2}|\nabla_3\rho_z|_{L^2}^{3/4}|\rho_z|_{L^2}^{1/4}+c|\nabla\rho_z|_{L^2}\\&\le\frac{\nu_\rho}{4}|\nabla_3\rho_z|_{L^2}^2+c(|\operatorname{div}\mathbf{u}|_{L^2}^2+|\rho_z|_{L^2}^2)+c,\end{aligned} \tag{3.99}$$

and also

$$\begin{aligned}I_2=|\left(w_z\frac{\partial\rho}{\partial z},\rho_z\right)_{L^2}|&=|(\operatorname{div}\mathbf{v},\rho_z^2)_{L^2}|=2|(\mathbf{v},\rho_z\nabla\rho_z)_{L^2}|\\&\le c|\mathbf{v}|_{L^4}|\rho_z|_{L^4}|\nabla\rho_z|_{L^2}\\&\le c|\mathbf{v}|_{L^4}|\rho_z|_{L^2}^{1/4}|\nabla_3\rho_z|_{L^2}^{7/4}\\&\le\frac{\nu_\rho}{4}|\nabla_3\rho_z|_{L^2}^2+c|\mathbf{v}|_{L^4}^8|\rho_z|_{L^2}^2.\end{aligned} \tag{3.100}$$

Using these estimates, Eq. (3.98) leads to

$$\frac{\mathrm{d}}{\mathrm{d}t}|\rho_z|_{L^2}^2+\nu_\rho|\nabla_3\rho_z|_{L^2}^2\le c|\operatorname{div}\mathbf{u}|_{L^2}^2+c|\rho_z|_{L^2}^2+c, \tag{3.101}$$

and using the Gronwall lemma, Eq. (3.97) follows. □

A priori estimates on $\mathbf{v}_t$ and ρ_t For obtaining a priori estimates on $\mathbf{v}_t$ and ρ_t, we differentiate Eqs. $(3.43)_1$ and $(3.43)_4$ in t, and we obtain

$$\begin{aligned}&\mathbf{v}_{tt}-\nu_{\mathbf{v}}\Delta_3\mathbf{v}_t+2f\mathbf{k}\times\mathbf{v}_t+\nabla_{\mathbf{v}}\mathbf{v}_t+w\frac{\partial\mathbf{v}_t}{\partial z}+\nabla_{\mathbf{v}_t}\mathbf{v}+w_t\frac{\partial\mathbf{v}}{\partial z}+\frac{1}{\rho_0}\nabla p_t=0,\\&\rho_{tt}-\nu_\rho\Delta_3\rho_t+\nabla_{\mathbf{v}}\rho_t+w\frac{\partial\rho_t}{\partial z}+\nabla_{\mathbf{v}_t}\rho+w_t\frac{\partial\rho}{\partial z}=0.\end{aligned} \tag{3.102}$$

We can then find the following.

LEMMA 3.11. *For $\mathbf{v}_t$ and ρ_t, the following estimate holds:*

$$\sup_{0\le t\le T}\left(|\mathbf{v}_t(t)|_{L^2}^2+|\rho_t(t)|_{L^2}^2\right)+c_1\int_0^T\left(|\nabla_3\mathbf{v}_t|_{L^2}^2+|\nabla_3\rho_t|_{L^2}^2\right)\mathrm{d}t \le c_1\left(|\mathbf{v}_t(0)|_{L^2}^2+|\rho_t(0)|_{L^2}^2\right). \tag{3.103}$$

PROOF. We take the L^2-scalar product of Eq. $(3.102)_1$ with $\mathbf{v}_t$ and of Eq. $(3.102)_2$ with ρ_t, and we obtain

$$\begin{aligned}&\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|\mathbf{v}_t(t)|_{L^2}^2+\nu_{\mathbf{v}}|\nabla_3\mathbf{v}_t(t)|_{L^2}^2+\frac{1}{\rho_0}(\nabla p_t,\mathbf{v}_t(t))_{L^2}+(\nabla_{\mathbf{v}_t}\mathbf{v}+w_t\frac{\partial\mathbf{v}}{\partial z},\mathbf{v}_t)_{L^2}=0,\\&\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|\rho_t(t)|_{L^2}^2+\nu_\rho|\nabla_3\rho_t(t)|_{L^2}^2+(\nabla_{\mathbf{v}_t}\rho+w_t\frac{\partial\rho}{\partial z},\rho_t)_{L^2}=0.\end{aligned} \tag{3.104}$$

We need to estimate the scalar products in Eq. (3.104):

$$\begin{aligned}|(\nabla p_t,\mathbf{v}_t)_{L^2}|&=|(p_t,\operatorname{div}\mathbf{v}_t)_{L^2}|=|(p_t,w_{zt})_{L^2}|\\&=|(p_{tz},w_t)_{L^2}|=g|(\rho_t,w_t)_{L^2}|\\&\le c|\rho_t|_{L^2}|\nabla\mathbf{v}_t|_{L^2}\le\frac{\nu_{\mathbf{v}}}{8}|\nabla_3\mathbf{v}_t|^2+c|\rho_t|_{L^2}^2.\end{aligned} \tag{3.105}$$

For the last term of Eq. $(3.104)_1$, we have

$$\begin{aligned}|(\nabla_{\mathbf{v}_t}\mathbf{v},\mathbf{v}_t)_{L^2}|&\le|\nabla\mathbf{v}_t|_{L^2}|\mathbf{v}|_{L^4}|\mathbf{v}_t|_{L^4}\\&\le c|\mathbf{v}|_{L^4}|\mathbf{v}_t|_{L^2}^{1/4}|\nabla_3\mathbf{v}_t|_{L^2}^{7/4}\\&\le\frac{\nu_{\mathbf{v}}}{8}|\nabla_3\mathbf{v}_t|_{L^2}^2+c|\mathbf{v}|_{L^4}^8|\mathbf{v}_t|_{L^2}^2,\end{aligned} \tag{3.106}$$

and

$$\begin{aligned}|\left(w_t\frac{\partial\mathbf{v}}{\partial z},\mathbf{v}_t\right)_{L^2}|&\le|w_t|_{L^2}|\mathbf{v}_z|_{L^4}|\mathbf{v}_t|_{L^4}\\&\le c|\nabla\mathbf{v}_t|_{L^2}|\mathbf{v}_z|_{L^4}|\nabla_3\mathbf{v}_t|_{L^2}^{3/4}|\mathbf{v}_t|_{L^2}^{1/4}\\&\le\frac{\nu_{\mathbf{v}}}{8}|\nabla_3\mathbf{v}_t|_{L^2}^2+c|\mathbf{v}_z|_{L^4}^8|\mathbf{v}_t|_{L^2}^2.\end{aligned} \tag{3.107}$$

Gathering these estimates, the following inequality on $\mathbf{v}_t$ is obtained:

$$\frac{\mathrm{d}}{\mathrm{d}t}|\mathbf{v}_t(t)|_{L^2}^2+\frac{5}{4}\nu_{\mathbf{v}}|\nabla_3\mathbf{v}_t(t)|_{L^2}^2\le c(|\rho_t|_{L^2}^2+|\mathbf{v}_t|_{L^2}^2). \tag{3.108}$$

For the density, the last term from Eq. $(3.104)_2$ is estimated as follows:

$$\begin{aligned}|(\nabla_{\mathbf{v}_t}\rho,\rho_t)_{L^2}|&\le|\nabla\mathbf{v}_t|_{L^2}|\rho|_{L^4}|\rho_t|_{L^4}\\&\le c|\nabla\mathbf{v}_t|_{L^2}|\rho|_{L^4}\left(|\nabla_3\rho_t|_{L^2}^{3/4}|\rho_t|_{L^2}^{1/4}+|\rho_t|_{L^2}\right)\\&\le\frac{\nu_{\mathbf{v}}}{8}|\nabla_3\mathbf{v}_t|_{L^2}^2+\frac{\nu_\rho}{4}|\nabla_3\rho_t|_{L^2}^2\\&\quad+c|\rho_t|_{L^2}^2|\rho|_{L^2}^2|\rho|_{L^4}^2+c|\rho|_{L^4}^4|\rho_t|_{L^2}^2,\end{aligned} \tag{3.109}$$

and

$$\begin{aligned}|\Big(w_t\frac{\partial\rho}{\partial z},\,\rho_t\Big)_{L^2}| &\le |w_t|_{L^2}|\rho_z|_{L^4}|\rho_t|_{L^4}\\ &\le c|\nabla\mathbf{v}_t|_{L^2}|\rho_z|_{L^4}\big(|\nabla_3\rho_t|_{L^2}^{3/4}|\rho_t|_{L^2}^{1/4}+|\rho_t|_{L^2}\big)\\ &\le \frac{\nu_{\mathbf{v}}}{8}|\nabla_3\mathbf{v}_t|_{L^2}^2+\frac{\nu_\rho}{4}|\nabla_3\rho_t|_{L^2}^2+c|\rho_t|_{L^2}^2(|\rho_z|_{L^4}^2+|\rho_z|_{L^4}^4). \quad (3.110)\end{aligned}$$

The estimates on ρ lead to

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|\rho_t|_{L^2}^2+\frac{\nu_\rho}{2}|\nabla_3\rho_t|_{L^2}^2\le\frac{\nu_{\mathbf{v}}}{4}|\nabla_3\mathbf{v}_t|_{L^2}^2+c|\rho_t|_{L^2}^2. \quad (3.111)$$

By summing Eqs. (3.108) and (3.111), we find

$$\frac{\mathrm{d}}{\mathrm{d}t}(|\mathbf{v}_t(t)|_{L^2}^2+|\rho_t(t)|_{L^2}^2)+\nu_{\mathbf{v}}|\nabla_3\mathbf{v}_t|_{L^2}^2+\nu_\rho|\nabla_3\rho_t|_{L^2}^2\le c(|\mathbf{v}_t(t)|_{L^2}^2+|\rho_t(t)|_{L^2}^2), \quad (3.112)$$

and from the Gronwall lemma, Eq. (3.103) follows. □

Note that the right-hand side of Eq. (3.103) depends on values that are not input data of the problem. So we need to estimate the right-hand side of Eq. (3.103) in terms of the initial data.

From Eq. $(3.43)_4$, we can deduce

$$|\rho_t(0)|_{L^2}\le c(|\rho_0|_{H^2}+|\mathbf{v}_0|_{H^2}). \quad (3.113)$$

In order to bound $\mathbf{v}_t(0)$, we take the scalar product of Eq. $(3.43)_1$ with $\mathbf{v}_t$:

$$\begin{aligned}|\mathbf{v}_t|_{L^2}^2 =& -\nu_{\mathbf{v}}(\Delta_3\mathbf{v},\mathbf{v}_t)_{L^2}-2f(\mathbf{k}\times\mathbf{v},\mathbf{v}_t)_{L^2}-(\nabla p,\mathbf{v}_t)_{L^2}\\ &-(\nabla_{\mathbf{v}}\mathbf{v},\mathbf{v}_t)_{L^2}-\Big(w\frac{\partial\mathbf{v}}{\partial z},\mathbf{v}_t\Big)_{L^2}. \quad (3.114)\end{aligned}$$

Then,

$$\begin{aligned}&|(\Delta_3\mathbf{v},\mathbf{v}_t)_{L^2}|\le|\Delta_3\mathbf{v}|_{L^2}|\mathbf{v}_t|_{L^2}\le c|\mathbf{v}|_{H^2}|\mathbf{v}_t|_{L^2},\\ &|2f(\mathbf{k}\times\mathbf{v},\mathbf{v}_t)_{L^2}|\le c|\mathbf{v}|_{L^2}|\mathbf{v}_t|_{L^2}. \quad (3.115)\end{aligned}$$

For the pressure term, we write

$$|(\nabla p,\mathbf{v}_t)_{L^2}|=|(\nabla p_1,\mathbf{v}_t)_{L^2}+(\nabla p_2,\mathbf{v}_t)_{L^2}|. \quad (3.116)$$

Since p_1 is independent of t, one can easily see that

$$|(\nabla p_1,\mathbf{v}_t)_{L^2}|=|(p_1,\ \mathrm{div}\,\mathbf{v}_t)_{L^2}|=|(p_1,w_{tz})_{L^2}|=0. \quad (3.117)$$

For p_2, we use the fact that $\int_0^1 p_2 \mathrm{d}z = 0$ and so the following inequality holds:

$$|\nabla p_2|_{L^2} \leq c|\nabla p_{2,z}|_{L^2}, \tag{3.118}$$

and one can deduce the following:

$$|(\nabla p_2, \mathbf{v}_t)_{L^2}| \leq |\nabla p_2|_{L^2}|\mathbf{v}_t|_{L^2} \leq c|\nabla p_{2,z}|_{L^2}|\mathbf{v}_t|_{L^2} \leq c|\nabla \rho|_{L^2}|\mathbf{v}_t|_{L^2}. \tag{3.119}$$

The scalar products containing the nonlinear terms are bounded as follows:

$$|(\nabla_{\mathbf{v}}\mathbf{v}, \mathbf{v}_t)_{L^2}| \leq |\nabla_{\mathbf{v}}\mathbf{v}|_{L^2}|\mathbf{v}_t|_{L^2} \leq c|\mathbf{v}|^2_{H^2}|\mathbf{v}_t|_{L^2}, \tag{3.120}$$

and

$$|(w\frac{\partial \mathbf{v}}{\partial z}, \mathbf{v}_t)_{L^2}| \leq |w|_{L^6}|\frac{\partial \mathbf{v}}{\partial z}|_{L^3}|\mathbf{v}_t|_{L^2} \leq c|\mathbf{v}|^2_{H^2}|\mathbf{v}_t|_{L^2}. \tag{3.121}$$

Taking into account these estimates, Eq. (3.114) implies

$$|\mathbf{v}_t|_{L^2} \leq c(|\mathbf{v}|^2_{H^2} + |\rho|^2_{H^1}), \tag{3.122}$$

so we find

$$|\mathbf{v}_t(0)|_{L^2} + |\rho_t(0)|_{L^2} \leq c(|\mathbf{v}_0|_{H^2} + |\rho_0|_{H^2}). \tag{3.123}$$

Using Eq. (3.103), we conclude that

$$\sup_{0\leq t\leq T}(|\mathbf{v}_t(t)|^2_{L^2} + |\rho_t(t)|^2_{L^2}) + c_1\int_0^T (|\nabla_3 \mathbf{v}_t|^2_{L^2} + |\nabla_3 \rho_t|^2_{L^2})\mathrm{d}t \leq c_T(|\mathbf{v}_0|^2_{H^2} + |\rho_0|^2_{H^2}). \tag{3.124}$$

We can obtain a stronger estimate for $|\nabla_3 \rho|_{L^2}$: we take the scalar product of Eq. $(3.43)_4$ with ρ, and we find

$$\nu_\rho|\nabla_3 \rho|^2_{L^2} = -(\rho_t, \rho)_{L^2} \leq |\rho_t|_{L^2}|\rho|_{L^2} \leq c, \tag{3.125}$$

with c depending on the H^2 norms of the initial data.

Similarly, we find $|\nabla_3 \mathbf{v}|^2_{L^2} \leq c$.

We are now able to prove the existence and uniqueness of the strong solutions.

3.2.1. Existence and uniqueness of a strong solution

Our final result for this problem is the proof of existence and uniqueness of strong solutions of Eqs. (3.43)–(3.47), which belong to the following spaces (in space dimension 3):

$$\mathbf{V} = \{\mathbf{v} = (u, v) \in H^1(\mathcal{M}_T);\ \mathbf{v} \text{ satisfies Eq. (3.44)}, \mathbf{v}_z \in H^1(\mathcal{M}_T), \int_0^1 \mathrm{div}\ \mathbf{v}(t, x, y, z)\,\mathrm{d}z = 0\}, \tag{3.126}$$

$$\mathbf{R} = \{\rho \in H^1(\mathcal{M}_T);\ \rho_z \in H^1(\mathcal{M}_T)\}.$$

The uniqueness of such solutions can be proved classically, by estimating the L^2-norm of the difference of two solutions (see Section 3.1). The existence of such a solution follows from the existence of a local strong solution as in Section 3.1, which satisfies all the a priori estimates proved before. We note that the a priori estimates are independent of the length of the time interval, hence the solution cannot blow up before the end of the considered time interval.

Thus, we have proved the main result of this Section 3.2.

THEOREM 3.2. *We assume that $\mathcal{M}$ is cylindrical with constant depth as in Eq. (3.42). Let $\mathbf{v}_0 \in H^2(\mathcal{M})$, and $\rho_0 \in H^2(\mathcal{M})$ satisfying the boundary conditions (Eqs. (3.44), and (3.45)) and the compatibility condition (Eq. (3.47)). Then, for any $T > 0$, the problem (Eqs. (3.43)–(3.46)) has a unique (strong) solution $\mathbf{v} \in \mathbf{V}$, $\rho \in \mathbf{R}$ on $\mathcal{M}_T$ such that $\mathbf{v}$, $\mathbf{v}_z$, $\nabla_3\mathbf{v}$, $\nabla_3\mathbf{v}_z$, $\mathbf{v}_t$, and $\mathbf{v}_{tz}$ belong to $L^2(\mathcal{M}_T)$, and the norm $|\nabla_3\mathbf{v}|_{L^2}$ is continuous on t.*

REMARK 3.1. We note here that Theorem 3.2 is not valid for the case of the Navier-Stokes equations. Indeed the fact that, here, the pressure p_1 is only a function of the horizontal variables x and y is essential.

We now need to estimate the L^4-norm of $\mathbf{v}$.
A new result, extending Theorem 3.2, was obtained in the very recent articles by KUKAVICA and ZIANE [2007a,b]. They proved the existence of global strong solutions of the PEs of the ocean in the case of a nonflat bottom. The boundary conditions considered by them are the Dirichlet conditions on the side and the bottom boundaries. These conditions applied on a varying-bottom topography are physically interesting, and they were not covered by the previous works. The global existence proof is given for the actual primitive equations with the actual boundary conditions (KUKAVICA and ZIANE [2007b]). They then prove that if $F_{\mathbf{v}}$ is in $L^2(0, t_1; L^2(\mathcal{M})^2)$ for all $t_1 > 0$ and $\mathbf{v}_0 \in H^1(\mathcal{M})^2$, $\mathbf{v}_0 = 0$ on Γ_l and Γ_b, then $\mathbf{v}$, solution of Eqs. $(3.43)_1$ and $(3.43)_2$ and $\mathbf{v}(0, x, y, z) = \mathbf{v}_0(x, y, z)$, exists and is unique in $L^\infty(0, t_1; H^2(\mathcal{M})^2)$ and $L^2(0, t_1; H^2(\mathcal{M})^2)$, for all $t_1 > 0$.

3.3. Strong solutions of the 2D PEs: physical boundary conditions

In this section, we are concerned with the global existence and the uniqueness of strong solutions of the 2D PEs of the ocean. We will first derive the equations formally from the 3D PEs under the assumption of invariance with respect to the y-variable, i.e., we will assume that the initial data, the forcing terms, and depth function h are independent of the variable y. The uniqueness of weak solutions implies that the solution will be independent of y.

In Sections 3.3.1 and 3.3.2, we introduce the 2D PEs and present their weak formulation. In Sections 3.3.2 and 3.3.3, we show that the strong solutions provided by

an analog of Theorem 3.1 are in fact defined for all $t > 0$; this result is based on appropriate a priori estimates, which are described subsequently; the approach is different and somehow simpler compared with dimension 3 in Section 3.2.

3.3.1. The 2D PEs

We assume that the domain occupied by the ocean is represented by

$$\big\{(x, y, z) \in \mathbb{R}^3, x \in (0, L), y \in \mathbb{R}, -h(x) < z < 0\big\},$$

and we denote by $\mathcal{M}$ its cross section:

$$\mathcal{M} = \big\{(x, z), x \in (0, L), -h(x) < z < 0\big\}. \tag{3.127}$$

Here, L is a positive number, and $h : [0, L] \to \mathbb{R}$ satisfies $h \in \mathcal{C}^3([0, L])$,

$$h(x) \geq \underline{h} > 0 \quad \text{for } x \in (0, L), \qquad h'(0) = h'(L) = 0. \tag{3.128}$$

By dropping all the terms containing a derivative with respect to y in the 3D PEs (Eqs. (2.44)–(2.50)), we obtain the following system:

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + w\frac{\partial u}{\partial z} - \mu_{\mathbf{v}}\frac{\partial^2 u}{\partial x^2} - \nu_{\mathbf{v}}\frac{\partial^2 u}{\partial z^2} - fv + \frac{\partial p_{\mathrm{s}}}{\partial x} = g\int_z^0 \frac{\partial \rho}{\partial x}\,\mathrm{d}z' + F_u, \tag{3.129}$$

$$\frac{\partial v}{\partial t} + u\frac{\partial v}{\partial x} + w\frac{\partial v}{\partial z} - \mu_{\mathbf{v}}\frac{\partial^2 v}{\partial x^2} - \nu_{\mathbf{v}}\frac{\partial^2 v}{\partial z^2} + fu = F_v, \tag{3.130}$$

$$\frac{\partial T}{\partial t} + u\frac{\partial T}{\partial x} + w\frac{\partial T}{\partial z} - \mu_T\frac{\partial^2 T}{\partial x^2} - \nu_T\frac{\partial^2 T}{\partial z^2} = F_T, \tag{3.131}$$

$$\frac{\partial S}{\partial t} + u\frac{\partial S}{\partial x} + w\frac{\partial S}{\partial z} - \mu_S\frac{\partial^2 S}{\partial x^2} - \nu_S\frac{\partial^2 S}{\partial z^2} = F_S, \tag{3.132}$$

$$\frac{\partial u}{\partial x} + \frac{\partial w}{\partial z} = 0, \tag{3.133}$$

where

$$\rho = 1 - \beta_T(T - T_{\mathrm{r}}) + \beta_S(S - S_{\mathrm{r}}), \tag{3.134}$$

$$\int_{\mathcal{M}} S\,\mathrm{d}\mathcal{M} = 0. \tag{3.135}$$

Here, u and v are the two components of the horizontal velocity $\mathbf{v}$. Note that despite y-invariance, v does not vanish in the problem of physical relevance (unlike the 2D Navier-Stokes equations). The quantity p_{s} above is the same as p in Eq. (2.49), whereas the expression P in Eq. (2.49) has been replaced by $g\int_z^0 \rho\,\mathrm{d}z'$, with ρ being a function of T and S through Eq. (3.10). Finally, as in the 3D case, $F = (F_u, F_v, F_T, F_S)$ vanishes in the physical problem and it is added here for mathematical generality.

Boundary conditions. These equations are supplemented with the same set of boundary conditions and initial data as in Section 2. On the top boundary of $\mathcal{M}$, denoted $\Gamma_{\rm i}$, $\Gamma_{\rm i} = \{(x, y);\, x \in (0, L);\, z = 0\}$, we have (see also after Eq. (2.55))

$$\begin{aligned} &\nu_{\mathbf{v}} \frac{\partial u}{\partial z} + \alpha_{\mathbf{v}} u = g_u, \qquad \nu_{\mathbf{v}} \frac{\partial v}{\partial z} + \alpha_{\mathbf{v}} v = g_v, \\ &\nu_T \frac{\partial T}{\partial z} + \alpha_T T = g_T, \qquad \frac{\partial S}{\partial z} = 0. \end{aligned} \tag{3.136}$$

On the remaining part of the boundary, we assume the Dirichlet boundary condition for the velocity and the Neumann condition for the temperature and the salinity. That is

$$\begin{aligned} &(u, v, w) = (0, 0, 0) \quad \text{on } \Gamma_\ell \cup \Gamma_{\rm b}, \\ &\frac{\partial T}{\partial n_T} = \frac{\partial S}{\partial n_S} = 0 \quad \text{on } \Gamma_\ell \cup \Gamma_{\rm b}, \end{aligned} \tag{3.137}$$

where

$$\begin{aligned} &\Gamma_\ell = \big\{(x, y);\, x = 0 \text{ or } L,\, -h(x) < z < 0\big\}, \\ &\Gamma_{\rm b} = \big\{(x, z);\, x \in (0, L),\, z = -h(x)\big\}. \end{aligned} \tag{3.138}$$

We also have the initial data given by

$$u|_{t=0} = u_0, \qquad v|_{t=0} = v_0, \qquad T|_{t=0} = T_0, \quad \text{and} \quad S|_{t=0} = S_0. \tag{3.139}$$

3.3.2. Weak formulation and the main result

We now proceed, as in Section 2, toward the weak and functional formulations of this problem (Eq. (3.129)–(3.135)) with some simplifications due to the invariance with respect to y, and some other aspects that are specific to dimension 2.

We introduce, as in Section 2.2.1, the following spaces:

$$\begin{aligned} &V = V_1 \times V_2 \times V_3, \qquad H = H_1 \times H_2 \times H_3, \\ &V_1 = \left\{\mathbf{v} = (u, v) \in H^1(\mathcal{M})^2,\, \frac{\partial}{\partial x} \int_{-h(x)}^{0} u(x, z)\, \mathrm{d}z = 0,\, \mathbf{v} = 0 \text{ on } \Gamma_\ell \cup \Gamma_{\rm b}\right\}, \\ &V_2 = H^1(\mathcal{M}), \\ &V_3 = \dot{H}^1(\mathcal{M}) = \left\{S \in H^1(\mathcal{M}),\, \int_{\mathcal{M}} S\, \mathrm{d}\mathcal{M} = 0\right\}, \\ &H_1 = \left\{\mathbf{v} = (u, v) \in L^2(\mathcal{M})^2,\, \int_{-h(x)}^{0} u(x, z)\, \mathrm{d}z = 0,\, u = 0 \text{ on } \Gamma_\ell\right\}, \\ &H_2 = L^2(\mathcal{M}), \\ &H_3 = \dot{L}^2(\mathcal{M}) = \left\{S \in L^2(\mathcal{M}),\, \int_{\mathcal{M}} S\, \mathrm{d}\mathcal{M} = 0\right\}. \end{aligned} \tag{3.140}$$

The scalar products are defined exactly as in Section 2.2.1, ∇ being replaced by $\partial/\partial x$. The condition $\int_{-h(x)}^{0} u(x, z)\,\mathrm{d}z = 0$ comes from the fact that the derivative in x of this quantity vanishes (the 2D analog of Eq. (2.48)) and that this quantity vanishes at $x = 0$ and L (see Section 2.2.1 for the 3D analog).

We also introduce the spaces $\mathcal{V}_1$, $\mathcal{V}_2$, and $\mathcal{V}_3$, and $\mathcal{V} = \mathcal{V}_1 \times \mathcal{V}_2 \times \mathcal{V}_3$ with a similar definition.

Similarly, we consider the forms a, b, c, and e, defined exactly as in dimension 3, just deleting all quantities involving a y-derivative; the associated operators A, B, and E are defined in the same way.

With these notations, the weak formulation is exactly as in dimension 3 (see Eqs. (2.79) and (2.80) or, in operational form, Eqs. (2.81) and (2.82)). There is no new difficulty in proving the analog of Theorem 2.4 giving the existence, for all time, of weak solutions.

Similarly we can prove, exactly as in Section (3.1), an analog of Theorem 3.1. Our aim in this section is to show that $t_* = t_1$ in space dimension 2, for the t_* appearing in the statement of Theorem 3.1.

More precisely we will prove the following (cf., Theorems 3.1 and 2.2):

THEOREM 3.3. *We assume that $\mathcal{M}$ is as in Eq. (3.127) and that Eq. (3.128) is satisfied.*

We are given $t_1 > 0$, $U_0 \in V$, $F = (F_{\mathbf{v}}, F_T, F_S)$ and $g = (g_{\mathbf{v}}, g_T)$ such that F and $\mathrm{d}F/\mathrm{d}t$ are in $L^2(0, t_1; H)$ (or $L^2(0, t_1; L^2(\mathcal{M})^4)$) and g and $\mathrm{d}g/\mathrm{d}t$ are in $L^2(0, t_1; H_0^1(\Gamma_i)^3)$.

Then there exists a unique solution U of the PEs (Eqs. (2.79) *and (2.80)) such that*

$$U \in \mathcal{C}\big([0, t_1]; V\big) \cap L^2\big(0, t_1; H^2(\mathcal{M})^4\big). \tag{3.141}$$

PROOF. The proof of uniqueness is easy and is done as in Theorem 3.1 for dimension 3. To prove the existence of solutions, we start from the strong solution given by the 2D analog of Theorem 3.1 and prove by contradiction that $t_* = t_1$. Indeed, let us denote by $[0, t_0]$ the maximal interval of existence of a strong solution, i.e.,[10]

$$U \in L^\infty\big(0, t'; V\big). \tag{3.142}$$

for every $t' < t_0$, and Eq. (3.142) does not occur for $t' = t_0$ which means, in particular, that

$$\limsup_{t \to t_0 - 0} \big\| U(t) \big\| = +\infty. \tag{3.143}$$

We will show that Eq. (3.143) cannot occur: we will derive a finite bound for $\|U(t')\|$ on $[0, t_0]$, thus contradicting Eq. (3.143).

The bounds for $\|U(t)\|$ will be derived *sequentially*: we will show successively that u_z, and u_x are in $L^\infty(0, t_0; L^2(\mathcal{M}))$ and $L^2(0, t_0; H^1(\mathcal{M}))$, where $\varphi_x = \partial\varphi/\partial x$ and $\varphi_z = \partial\varphi/\partial z$; then, we will prove at once that $\mathbf{v}$, T, and S are in $L^\infty(0, t_0; H^1)$ and $L^2(0, t_0; H^2)$. In fact, we will give the proofs for u_z, u_x, and T; the other quantities being estimated in

[10] It is easy to see that if $U \in L^\infty(0, t'; V)$, then $U \in L^2(0, t'; H^2(\mathcal{M})^4)$ as well.

exactly the same way. For the sake of simplicity, we assume hereafter that $g = (g_{\mathbf{v}}, g_T) = 0$. When $g \neq 0$, we need to "homogenize" the boundary conditions by considering $U' = U - U^*$, with U^* defined exactly as in Section 3.1, then perform the following calculations for U'.[11]

Before we proceed, let us recall that we have already available the a priori estimates for U in $L^\infty(0, t_1; L^2)$ and $L^2(0, t_1; H^1)$ used to prove the analog of Theorem 2.2 (i.e., Eq. (2.93) in the discrete case). □

3.3.3. Vertical averaging

To derive the new a priori estimates, we need some operators related to vertical averaging that we now define.

For any function φ defined and integrable on $\mathcal{M}$, we set

$$\begin{aligned} &\overline{P}\varphi(x) = \int_{-h(x)}^{0} \varphi(x, z)\, \mathrm{d}z, \\ &P\varphi = \frac{1}{h}\overline{P}\varphi, \quad Q\varphi = \varphi - P\varphi. \end{aligned} \tag{3.144}$$

We now establish some useful properties of these operators, some simple, some more involved.

We first note that $PQ = 0$ so that

$$\int_{-h(x)}^{0} Q\varphi(x, z)\, \mathrm{d}z = 0, \quad \forall \varphi \in L^1(\mathcal{M}), \tag{3.145}$$

and

$$\int_{\mathcal{M}} (P\varphi)(x)(Q\psi)(x, z)\, \mathrm{d}x\, \mathrm{d}z = 0, \quad \forall \varphi, \psi \in L^2(\mathcal{M}). \tag{3.146}$$

Also, for all φ sufficiently regular,

$$\begin{aligned} \overline{P}\,\frac{\partial \varphi}{\partial x} &= \frac{\partial}{\partial x}\overline{P}\varphi - h'(x)\varphi\big(x, -h(x)\big), \\ \overline{P}\,\frac{\partial^2 \varphi}{\partial x^2} &= \frac{\partial^2}{\partial x^2}\overline{P}\varphi - 2h'(x)\frac{\partial \varphi}{\partial x}\big(x, h(x)\big) \\ &\quad + h'(x)^2 \frac{\partial \varphi}{\partial z}\big(x, -h(x)\big) - h''(x)\varphi\big(x, -h(x)\big). \end{aligned} \tag{3.147}$$

Now, if φ vanishes on Γ_b, $\varphi(x, -h(x)) = 0$, $0 < x < L$, then

$$\frac{\partial \varphi}{\partial x}\big(x, -h(x)\big) = h'(x)\frac{\partial \varphi}{\partial z}\big(x, -h(x)\big), \tag{3.148}$$

[11] Note that in Eq. (3.1), U^* was chosen so that the initial and boundary conditions for U' vanish. Here, we do not need to homogenize the initial condition, but we can use the same U^*.

and hence

$$\begin{aligned}
&\overline{P}\,\frac{\partial\varphi}{\partial x} = \frac{\partial}{\partial x}\overline{P}\varphi,\\
&\overline{P}\,\frac{\partial^2\varphi}{\partial x^2} = \frac{\partial^2}{\partial x^2}\overline{P}\varphi - h'(x)\frac{\partial\varphi}{\partial x}\big(x,-h(x)\big),\\
&P\frac{\partial^2\varphi}{\partial x^2}(x) = -\frac{h'}{h}\frac{\partial\varphi}{\partial x}\big(x,-h(x)\big) = -\frac{h'^2}{h}\frac{\partial\varphi}{\partial z}\big(x,-h(x)\big).
\end{aligned} \tag{3.149}$$

Finally, the following lemma will be needed.

LEMMA 3.12. *For any $v \in H^2(\mathcal{M})$ such that*

$$\nu\frac{\partial v}{\partial z} + \alpha v = 0 \quad \textit{on } \Gamma_{\mathrm{i}}, \qquad v = 0 \quad \textit{on } \Gamma_\ell \cup \Gamma_{\mathrm{b}},$$

we have

$$\begin{aligned}
\int_{\mathcal{M}} \frac{\partial^2 v}{\partial x^2}\frac{\partial^2 v}{\partial z^2}\,\mathrm{d}z\,\mathrm{d}x &= \int_{\mathcal{M}} \left|\frac{\partial^2 v}{\partial x\partial z}\right|^2 \mathrm{d}z\,\mathrm{d}x + \alpha\int_0^L \left|\frac{\partial v}{\partial x}(x,0)\right|^2 \mathrm{d}x\\
&\quad - \frac{1}{2}\int_0^L h''(x)\left|\frac{\partial v}{\partial z}\big(x,-h(x)\big)\right|^2 \mathrm{d}x.
\end{aligned} \tag{3.150}$$

PROOF. We give the proof for v smooth, say $v \in \mathcal{C}^3(\mathcal{M})$; the result extends then to $v \in H^2(\mathcal{M})$ using a density argument (that we skip).

We write

$$\begin{aligned}
\frac{\partial^2 v}{\partial x^2}\frac{\partial^2 v}{\partial z^2} &= \frac{\partial}{\partial x}\left(\frac{\partial v}{\partial x}\frac{\partial^2 v}{\partial z^2}\right) - \frac{\partial v}{\partial x}\frac{\partial^3 v}{\partial x\,\partial z^2}\\
&= \frac{\partial}{\partial x}\left(\frac{\partial v}{\partial x}\frac{\partial^2 v}{\partial z^2}\right) - \frac{\partial}{\partial z}\left(\frac{\partial v}{\partial x}\frac{\partial^2 v}{\partial x\,\partial z}\right) + \left|\frac{\partial^2 v}{\partial x\,\partial z}\right|^2.
\end{aligned} \tag{3.151}$$

We first integrate in z, and taking into account Eq. (3.147), we obtain

$$\begin{aligned}
&\int_{-h}^0 \frac{\partial^2 v}{\partial x^2}\frac{\partial^2 v}{\partial z^2}\,\mathrm{d}z\\
&\quad = I - h'(x)\frac{\partial v}{\partial x}\big(x,-h(x)\big)\frac{\partial^2 v}{\partial z^2}\big(x,-h(x)\big) - \frac{\partial v}{\partial x}(x,0)\frac{\partial^2 v}{\partial x\,\partial z}(x,0)\\
&\qquad + \frac{\partial v}{\partial x}\big(x,-h(x)\big)\frac{\partial^2 v}{\partial x\,\partial z}\big(x,-h(x)\big) + \int_{-h}^0 \left|\frac{\partial^2 v}{\partial x\,\partial z}\right|^2 \mathrm{d}z,\\
&I = \frac{\partial}{\partial x}\int_{-h}^0 \frac{\partial v}{\partial x}\frac{\partial^2 v}{\partial z^2}\,\mathrm{d}z.
\end{aligned} \tag{3.152}$$

We now integrate in x. The integral of $I = I(x)$ vanishes because $v(0, z) = v(L, z) = 0$ for all z so that

$$\left(\frac{\partial^2 v}{\partial z^2}\right)(0, z) = \left(\frac{\partial^2 v}{\partial z^2}\right)(L, z) = 0 \quad \forall z.$$

The third term on the right-hand side of Eq. (3.152) is equal to $\alpha \int_0^L |(\partial v/\partial x)(x, 0)|^2 \, \mathrm{d}x$. The sum of the second and fourth terms is equal to

$$\int_0^L \frac{\partial v}{\partial x}\big(x, -h(x)\big)\left(\frac{\partial^2 v}{\partial x\, \partial z} - h'\frac{\partial^2 v}{\partial z^2}\right)\big(x, -h(x)\big)\, \mathrm{d}x. \tag{3.153}$$

Setting $\varphi(x) = (\partial v/\partial z)(x, -h(x))$, we see that

$$\varphi'(x) = \frac{\partial^2 v}{\partial x\, \partial z}\big(x, -h(x)\big) - h'(x)\frac{\partial^2 v}{\partial z^2}\big(x, -h(x)\big),$$

and since $v(x, -h(x)) = 0$, we have

$$\frac{\partial v}{\partial x}\big(x, -h(x)\big) - h'(x)\frac{\partial v}{\partial z}\big(x, -h(x)\big) = 0,$$

and the integral in Eq. (3.153) is equal to

$$\int_0^L h'(x)\varphi(x)\varphi'(x)\, \mathrm{d}x = -\frac{1}{2}\int_0^L h''(x)\varphi^2(x)\, \mathrm{d}x;$$

for the last relation we have used $\varphi(0) = \varphi(L) = 0$. The lemma is proved. □

3.3.4. Estimates for u_z

To show that $u_z \in L^\infty(0, t_0; L^2(\mathcal{M})) \cap L^2(0, t_0; H^1(\mathcal{M}))$, we multiply Eq. (3.129) by Qu_{zz}, integrate over $\mathcal{M}$, and integrate by parts, and remember that $Qu = u$. Thus, for each term successively, omitting the variable t, we find

$$\begin{aligned}
\int_{\mathcal{M}} u_t Q(-u_{zz})\, \mathrm{d}\mathcal{M} &= -\int_{\mathcal{M}} u_t u_{zz}\, \mathrm{d}\mathcal{M} \\
&= -\int_0^L u_t u_z\big]_{-h}^0\, \mathrm{d}x + \int_{\mathcal{M}} u_{tz}u_z\, \mathrm{d}\mathcal{M} \\
&= \frac{\alpha_{\mathbf{v}}}{\nu_{\mathbf{v}}}\int_0^L u_t(x, 0)u(x, 0)\, \mathrm{d}x + \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|u_z|_{L^2}^2 \\
&= \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left(|u_z|_{L^2}^2 + \frac{\alpha_{\mathbf{v}}}{\nu_{\mathbf{v}}}\big|u(x, 0)\big|_{L^2(\Gamma_i)}^2\right),
\end{aligned}$$

$$\begin{aligned}\int_{\mathcal{M}} uu_x Q(-u_{zz})\,\mathrm{d}\mathcal{M} &= \int_{\mathcal{M}} uu_x u_{zz}\,\mathrm{d}\mathcal{M} + \int_{\mathcal{M}} uu_x Pu_{zz}\,\mathrm{d}\mathcal{M}\\ &= -\int_L^0 uu_x u_z\big]_{-h}^0\,\mathrm{d}x + \int_{\mathcal{M}} uu_{xz}u_z\,\mathrm{d}\mathcal{M}\\ &\quad + \int_{\mathcal{M}} u_x u_z^2\,\mathrm{d}\mathcal{M} + \int_{\mathcal{M}} uu_x Pu_{zz}\,\mathrm{d}\mathcal{M}\\ &= \frac{\alpha_{\mathbf{v}}}{\nu_{\mathbf{v}}}\int_0^L u^2(x,0)u_x(x,0)\,\mathrm{d}x + \frac{1}{2}\int_{\partial\mathcal{M}} un_x u_z^2\,\mathrm{d}(\partial\mathcal{M})\\ &\quad + \frac{1}{2}\int_{\mathcal{M}} u_x u_z^2\,\mathrm{d}\mathcal{M} + \int_{\mathcal{M}} uu_z Pu_{zz}\,\mathrm{d}\mathcal{M}\\ &= \frac{\alpha_{\mathbf{v}}}{3\nu_{\mathbf{v}}}u^3(x,0)\big]_0^L + \frac{1}{2}\int_{\mathcal{M}} u_x u_z^2\,\mathrm{d}\mathcal{M} + \int_{\mathcal{M}} uu_x Pu_{zz}\,\mathrm{d}\mathcal{M}\\ &= \frac{1}{2}\int_{\mathcal{M}} u_x u_z^2\,\mathrm{d}\mathcal{M} + \int_{\mathcal{M}} uu_x Pu_{zz}\,\mathrm{d}\mathcal{M}.\end{aligned}$$

In the relations above, $n = (n_x, n_z)$ is the unit outward normal on $\partial\mathcal{M}$, and we used the fact that $un_x = 0$ on $\partial\mathcal{M}$, and that $u(0,0) = u(L,0) = 0$ because $u = 0$ on Γ_ℓ.

$$\begin{aligned}\int_{\mathcal{M}} wu_z Q(-u_{zz})\,\mathrm{d}\mathcal{M} &= -\int_{\mathcal{M}} wu_z u_{zz}\,\mathrm{d}\mathcal{M} + \int_{\mathcal{M}} wu_z Pu_{zz}\,\mathrm{d}\mathcal{M}\\ &= -\frac{1}{2}\int_0^L wu_z^2\big]_{-h}^0\,\mathrm{d}x + \frac{1}{2}\int_{\mathcal{M}} w_z u_z^2\,\mathrm{d}\mathcal{M}\\ &\quad + \int_{\mathcal{M}} wu_z Pu_{zz}\,\mathrm{d}\mathcal{M}\\ &= (\text{since } w = 0 \text{ on } \Gamma_{\mathrm{i}} \text{ and } \Gamma_{\mathrm{b}} \text{ and } w_z = -u_z)\\ &= -\frac{1}{2}\int_{\mathcal{M}} u_x u_z^2\,\mathrm{d}\mathcal{M} + \int_{\mathcal{M}} wu_z Pu_{zz}\,\mathrm{d}\mathcal{M}.\end{aligned}$$

Since Pu_{zz} is independent of z and w vanishes on Γ_{i} and Γ_{b}, we have

$$\begin{aligned}\int_{\mathcal{M}} wu_z Pw_{zz}\,\mathrm{d}\mathcal{M} &= \int_0^L wu\big]_{-h}^0 Pu_{zz}\,\mathrm{d}x - \int_{\mathcal{M}} uw_z Pu_{zz}\,\mathrm{d}\mathcal{M}\\ &= \int_{\mathcal{M}} uu_z Pu_{zz}\,\mathrm{d}\mathcal{M}.\end{aligned}$$

Finally, the last two terms add up in the following way:

$$\begin{aligned}\int_{\mathcal{M}} (uu_x + wu_z)Q(-u_{zz})\,\mathrm{d}\mathcal{M} &= 2\int_{\mathcal{M}} uu_x Pu_{zz}\,\mathrm{d}\mathcal{M}\\ &= 2\int_{\mathcal{M}} \frac{1}{h}uu_x\big[u_z(x,0) - u_z(x,h)\big]\,\mathrm{d}\mathcal{M},\end{aligned}$$

$$\int_{\mathcal{M}} p_{sx} Q(-u_{zz})\,\mathrm{d}\mathcal{M} = 0 \quad \text{(since } p_{\mathrm{s}} \text{ is independent of } z\text{)},$$

$$-\nu_{\mathbf{v}} \int_{\mathcal{M}} u_{zz} Q(-u_{zz})\,\mathrm{d}\mathcal{M} = \nu_{\mathbf{v}} |Qu_{zz}|_{L^2}^2 = \nu_{\mathbf{v}} |u_{zz}|_{L^2}^2 - \nu_{\mathbf{v}} |Pu_{zz}|_{L^2}^2,$$

$$-\mu_{\mathbf{v}} \int_{\mathcal{M}} u_{xx} Q(-u_{zz})\,\mathrm{d}\mathcal{M} = \mu_{\mathbf{v}} \int_{\mathcal{M}} u_{xx} u_{zz}\,\mathrm{d}\mathcal{M} - \mu_{\mathbf{v}} \int_{\mathcal{M}} (Pu_{xx}) u_{zz}\,\mathrm{d}\mathcal{M}.$$

Using Eq. (3.149) and Lemma 3.1, we see that this expression is equal to

$$\begin{aligned}
&\mu_{\mathbf{v}} |u_{xz}|_{L^2}^2 + \alpha_{\mathbf{v}} \mu_{\mathbf{v}} \int_0^L \big|u_x(x,0)\big|^2\,\mathrm{d}x \\
&\quad - \mu_{\mathbf{v}} \int_0^L \left(\frac{h'(x)^2}{h(x)} + \frac{1}{2} h''(x)\right) \big|u_z\big(x,-h(x)\big)\big|^2\,\mathrm{d}x \\
&\quad - \frac{\alpha_{\mathbf{v}}}{\mu_{\mathbf{v}} \nu_{\mathbf{v}}} \int_0^L u_z(x,-h) u(x,0)\,\mathrm{d}x.
\end{aligned}$$

The other terms are left unchanged, and then gathering all these terms we find

$$\begin{aligned}
&\frac{1}{2} \frac{\mathrm{d}}{\mathrm{d}t} \left(|u_z|_{L^2(\mathcal{M})}^2 + \frac{\alpha_{\mathbf{v}}}{\nu_{\mathbf{v}}} \big|u(x,0)\big|_{L^2(\Gamma_{\mathrm{i}})}^2 \right) + \mu_{\mathbf{v}} |u_{xz}|_{L^2(\mathcal{M})}^2 \\
&\qquad + \alpha_{\mathbf{v}} \mu_{\mathbf{v}} \big|u_x(x,0)\big|_{L^2(\Gamma_{\mathrm{i}})}^2 + \nu_{\mathbf{v}} |u_{zz}|_{L^2(\mathcal{M})}^2 \\
&\quad = \nu_{\mathbf{v}} |Pu_{zz}|_{L^2}^2 + \mu_{\mathbf{v}} \int_0^L \left(\frac{h'^2}{h} + \frac{1}{2} h''\right) \big|u_z\big(x,-h(x)\big)\big|^2\,\mathrm{d}x \\
&\qquad + \frac{\alpha_{\mathbf{v}}}{\mu_{\mathbf{v}} \nu_{\mathbf{v}}} \int_0^L u_z(x,-h) u(x,0)\,\mathrm{d}x \\
&\qquad + \int_{\mathcal{M}} fv Q u_{zz}\,\mathrm{d}\mathcal{M} - g \int_{\mathcal{M}} \left(\int_z^0 \rho_x\,\mathrm{d}z' \right) u_{zz}\,\mathrm{d}\mathcal{M}.
\end{aligned} \tag{3.154}$$

We estimate the right-hand side of Eq. (3.154) as follows, c denoting a constant depending only on $\mathcal{M}$ and on the coefficients $\alpha_{\mathbf{v}}$, $\mu_{\mathbf{v}}$, and $\nu_{\mathbf{v}}$,

$$\begin{aligned}
&Pu_{zz} = \frac{1}{h} \big(u_z(x,0) - u_z(x,-h)\big) = \frac{1}{h} \frac{\alpha_{\mathbf{v}}}{\nu_{\mathbf{v}}} u(x,0) - \frac{1}{h} u_z\big(x,-h(x)\big), \\
&|Pu_{zz}|_{L^2(\mathcal{M})} \le c \big|u(x,0)\big|_{L^2(\Gamma_{\mathrm{i}})} + c \big|u_z\big(x,-h(x)\big)\big|_{L^2(\Gamma_{\mathrm{i}})}.
\end{aligned} \tag{3.155}$$

The last two norms are bounded by the trace theorems:

$$|u|_{L^2(\Gamma_i)} \le c|u|^{1/2}_{L^2(\mathcal{M})}\|u\|^{1/2}, \tag{3.156}$$

$$\begin{aligned}\big|u_z\big(x,-h(x)\big)\big|_{L^2(\Gamma_i)} &\le c|u_z|_{L^2(\Gamma_b)}\\ &\le c|u_z|^{1/2}_{L^2(\mathcal{M})}|\nabla u_z|^{1/2}_{L^2(\mathcal{M})}\\ &\le c\|u\|^{1/2}\big(|u_{zz}|^2_{L^2(\mathcal{M})}+|u_{zx}|^2_{L^2(\mathcal{M})}\big)^{1/4},\end{aligned} \tag{3.157}$$

$$|Pu_{zz}|^2_{L^2(\mathcal{M})} \le c|u|_{L^2(\mathcal{M})}\|u\| + c\|u\|\|u_z\|.$$

We write also

$$\begin{aligned}\left|\int_{\mathcal{M}} fvQu_{zz}\,\mathrm{d}\mathcal{M}\right| &\le c|v|_{L^2(\mathcal{M})}|u_{zz}|_{L^2(\mathcal{M})}\\ &\le c|v|_{L^2(\mathcal{M})}\|u_z\|.\end{aligned}$$

Finally, using again Eqs. (3.156) and (3.157) for the other terms on the right-hand side of Eq. (3.154), we obtain

$$\begin{aligned}&\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left(|u_z|^2_{L^2(\mathcal{M})}+\frac{\alpha_{\mathbf{v}}}{\nu_{\mathbf{v}}}\big|u(x,0)\big|^2_{L^2(\Gamma_i)}\right)+\nu\|u_z\|^2+\alpha_{\mathbf{v}}\mu_{\mathbf{v}}\big|u_x(x,0)\big|^2_{L^2(\Gamma_i)}\\ &\quad\le c|u|_{L^2(\mathcal{M})}\|u\|+c\|u\|\|u_z\|+c|v|_{L^2}\|u_z\|\\ &\qquad+c\big(|T_x|_{L^2}+|S_x|_{L^2}\big)|u_{zz}|_{L^2}+|F_u|_{L^2}|u_{zz}|_{L^2}\\ &\quad\le\frac{\nu}{2}\|u_z\|^2+c\|u\|^2+c|v|^2_{L^2}+c\|U\|^2+c|F_u|^2_{L^2},\end{aligned}$$

where $\nu=\min(\mu_{\mathbf{v}},\nu_{\mathbf{v}})$. Hence,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(|u_z|^2_{L^2(\mathcal{M})}+\frac{\alpha_{\mathbf{v}}}{\nu_{\mathbf{v}}}|u_x(x,0)|^2_{L^2(\Gamma_i)}\right)+\nu\|u_z\|^2\le c\|U\|^2+c|F_u|^2_{L^2(\mathcal{M})}. \tag{3.158}$$

Taking into account the earlier estimates of U in $L^\infty(0,t_1;H)$ and $L^2(0,t_1;V)$, we obtain an a priori bound of u_z in $L^\infty(0,t_0;L^2(\mathcal{M}))$, a first step in proving that t_0 cannot be less than t_1.

Remark 3.2. We recall that the estimates above were made under the simplifying assumption that $g=(g_{\mathbf{v}},g_T)=0$. When this is not the case, we explained that we ought to consider $U'=U-U^*$, U^* defined as in Section 3.1. Then the calculations above are made for the equation for u'. This equation will involve some additional terms such as $u^*\partial u'/\partial x$, $u'\partial u^*/\partial x$; these additional terms are estimated in a similar way, leading to the same conclusions. The same remark applies to the estimates below concerning u_x, v_z., etc.

3.3.5. Estimates for u_x

To show that u_x is bounded in $L^\infty(0,t_0;L^2(\mathcal{M}))\cap L^2(0,t_0;H^1(\mathcal{M}))$, we multiply Eq. (3.129) by $-u_{xx}$, integrate over $\mathcal{M}$, and integrate by parts. At any fixed time t, each term can be written as follows:

$$
\begin{aligned}
-\int_{\mathcal{M}} u_t u_{xx}\,\mathrm{d}\mathcal{M} &= -\int_{\partial\mathcal{M}} u_t u_x n_x\,\mathrm{d}(\partial\mathcal{M}) + \int_{\mathcal{M}} u_{tx}u_x\,\mathrm{d}\mathcal{M}\\
&= \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathcal{M}} u_x^2\,\mathrm{d}\mathcal{M},\\
-\int_{\mathcal{M}} u u_x u_{xx}\,\mathrm{d}\mathcal{M} &= -\frac{1}{2}\int_{\partial\mathcal{M}} u u_x^2 n_x\,\mathrm{d}(\partial\mathcal{M}) + \frac{1}{2}\int_{\mathcal{M}} u_x^3\,\mathrm{d}\mathcal{M}\\
&= \frac{1}{2}\int_{\mathcal{M}} u_x^3\,\mathrm{d}\mathcal{M},\\
-\int_{\mathcal{M}} w u_z u_{xx}\,\mathrm{d}\mathcal{M} &= -\int_{\partial\mathcal{M}} w u_z u_x n_x\,\mathrm{d}(\partial\mathcal{M}) + \int_{\mathcal{M}} w_x u_z u_x\,\mathrm{d}\mathcal{M}\\
&\quad + \int_{\mathcal{M}} w u_{zx} u_x\,\mathrm{d}\mathcal{M}\\
&= \int_{\mathcal{M}} w_x u_z u_x\,\mathrm{d}\mathcal{M} + \frac{1}{2}\int_0^L w u_x^2\big]_{-h}^0\,\mathrm{d}x\\
&\quad - \frac{1}{2}\int_{\mathcal{M}} w_z u_x^2\,\mathrm{d}\mathcal{M}\\
&= \int_{\mathcal{M}} w_x u_z u_x\,\mathrm{d}\mathcal{M} + \frac{1}{2}\int_{\mathcal{M}} u_x^3\,\mathrm{d}\mathcal{M},\\
\nu_{\mathbf{v}}\int_{\mathcal{M}} u_{zz}u_{xx}\,\mathrm{d}\mathcal{M} &= \text{(thanks to Lemma 3.1)}\\
&= \nu_{\mathbf{v}}|u_{zx}|_{L^2}^2 + \alpha_{\mathbf{v}}\nu_{\mathbf{v}}\int_0^L \big|u_x(x,0)\big|^2\,\mathrm{d}x\\
&\quad - \frac{1}{2}\int_0^L h''(x)\big|u_z\big(x,-h(x)\big)\big|^2\,\mathrm{d}x.
\end{aligned}
$$

The other terms are left unchanged and, with $\nu=\min(\mu_{\mathbf{v}},\nu_{\mathbf{v}})$, we arrive at

$$
\begin{aligned}
&\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|u_x|_{L^2}^2 + \nu\|u_x\|^2 + \alpha_{\mathbf{v}}\nu_{\mathbf{v}}\int_0^L \big|u_x(x,0)\big|^2\,\mathrm{d}x\\
&\quad = -\int_{\mathcal{M}} u_x^3\,\mathrm{d}\mathcal{M} - \int_{\mathcal{M}} w_x u_z u_x\,\mathrm{d}\mathcal{M}\\
&\qquad - \int_{\mathcal{M}} f v u_{xx}\,\mathrm{d}\mathcal{M} + \frac{1}{2}\int_0^L h''(x)\big|u_z\big(x,-h(x)\big)\big|^2\,\mathrm{d}x\\
&\qquad - g\int_{\mathcal{M}}\Big(\int_z^0 \rho_x\,\mathrm{d}z'\Big)u_{xx}\,\mathrm{d}\mathcal{M} + \int_{\mathcal{M}} F_u u_{xx}\,\mathrm{d}\mathcal{M}.
\end{aligned}
$$

We write

$$\left| \int_{\mathcal{M}} u_x^3 \, \mathrm{d}\mathcal{M} \right| \leq |u_x|_{L^3}^3$$

$$\leq \text{(by Sobolev embedding and interpolation)}$$

$$\leq c|u_x|_{H^{1/3}}^3 \leq c|u_x|_{L^2}^2 |u_x|_{H^1}$$

$$\leq c\|u\|^2 \|u_x\|$$

$$\leq \frac{\nu}{10}\|u_x\|^2 + c\|u\|^4,$$

$$\left| \int_{\mathcal{M}} w_x u_z u_x \, \mathrm{d}\mathcal{M} \right| \leq |w_x|_{L^2} |u_z|_{L^4} |u_x|_{L^4}$$

$$\leq \text{(by Sobolev embedding and interpolation)}$$

$$\leq c|u_{xx}|_{L^2} |u_z|_{L^2}^{1/2} \|u_z\|^{1/2} |u_x|_{L^2}^{1/2} \|u_x\|^{1/2}$$

$$\leq c|u_z|_{L^2}^{1/2} \|u_z\|^{1/2} |u_x|_{L^2}^{1/2} \|u_x\|^{3/2}$$

$$\leq \frac{\nu}{10}\|u_x\|^2 + c|u_z|_{L^2}^2 \|u_z\|^2 |u_x|_{L^2}^2$$

and

$$\left| \int_{\mathcal{M}} fvu_{xx} \, \mathrm{d}\mathcal{M} \right| \leq |fv|_{L^2} |u_{xx}|_{L^2} \leq \frac{\nu_{\mathbf{v}}}{10}\|u_x\|^2 + c|v|_{L^2}^2.$$

The next terms are bounded as before, and the last term is easy. Hence,

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t}|u_x|_{L^2(\mathcal{M})}^2 + \nu\|u_x\|^2 + \alpha_{\mathbf{v}}\nu_{\mathbf{v}} \int_0^L \left|u_x(x,0)\right|^2 \mathrm{d}x \\ &\quad \leq c\|u\|^4 + c|v|_{L^2}^2 + c|u_z|_{L^2(\mathcal{M})}^2 \|u_z\|^2 |w_z|_{L^2(\mathcal{M})}^2 \\ &\qquad + c\|u\|\|u_z\| + c\|U\|^2 + c|F_u|_{L^2(\mathcal{M})}^2. \end{aligned} \tag{3.159}$$

Remembering that $\|u\|^2 = |u_x|_{L^2(\mathcal{M})}^2 + |u_z|_{L^2(\mathcal{M})}^2$, we see that the right-hand side of Eq. (3.159) is of the form $\xi(t) + \eta(t)|u_x|_{L^2(\mathcal{M})}^2$, where ξ, and η are in $L^1(0, t_1)$; for $\eta = c\|u\|^2 + c|u_z|_{L^2(\mathcal{M})}^2 \|u_z\|^2$, this follows from the previous estimates on U and on u_z; similarly, the contribution of $c\|u\|\|u_z\|$ to ξ is in $L^1(0, t_1)$ due to the previous results on u_z and U. Therefore, the Gronwall lemma applied to Eq. (3.159) provides an a priori bound of u_x in $L^\infty(0, t_0; L^2(\mathcal{M}))$ and $L^2(0, t_0; H^1(\mathcal{M}))$.

3.3.6. Estimates for v, T, and S

We now prove at once that T is bounded in $L^\infty(0, t_0; H^1(\mathcal{M}))$ and $L^2(0, t_0; H^2(\mathcal{M}))$; the proof is similar for v and S, and this will thus conclude the proof of Theorem 3.3.

For this, we multiply each side of Eq. (3.131) by $A_2 T = -\mu_T\, \partial^2 T/\partial x^2 - \nu_T\, \partial^2 T/\partial z^2$. We recall that $g_T = 0$ here (see the end of Section 3.3.2). We have

$$
\begin{aligned}
-\int_{\mathcal{M}} T_t(\mu_T T_{xx} + \nu_T T_{zz})\,\mathrm{d}\mathcal{M} &= -\int_{\partial\mathcal{M}} T_t(\mu_T T_x n_x + \nu_T T_z n_z)\,\mathrm{d}(\partial\mathcal{M}) \\
&\quad + \frac{1}{2}\int_{\mathcal{M}} (\mu_T T_{tx} T_x + \nu_T T_{tz} T_z)\,\mathrm{d}\mathcal{M} \\
&= \big(\text{see the notations in Eq. (2.38) and after Eq. (2.54)}\big) \\
&= -\int_{\partial\mathcal{M}} T_t \frac{\partial T}{\partial \mathbf{n}_T}\,\mathrm{d}(\partial\mathcal{M}) + \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t} a_2(T, T) \\
&= \big(\text{with Eqs. (2.35) and (2.55) and } g_T = 0\big) \\
&= \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t} a_2(T, T) + \int_{\Gamma_\mathrm{i}} \alpha_T T\, T_t\,\mathrm{d}\Gamma_\mathrm{i} \\
&= \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t} a_2(T, T) + \alpha_T \big|T(x, 0)\big|^2_{L^2(\Gamma_\mathrm{i}).}
\end{aligned}
$$

Hence, we find

$$
\begin{aligned}
&\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left(\big|A_2^{1/2} T\big|^2_{L^2} + \alpha_T \big|T(x, 0)\big|^2_{L^2(\Gamma_\mathrm{i})}\right) + |A_2 T|^2_{L^2} \\
&\quad = -\int_{\mathcal{M}} (u T_x + w T_z) A_2 T\,\mathrm{d}\mathcal{M} + \int_{\mathcal{M}} F_T A_2 T\,\mathrm{d}\mathcal{M}. \qquad (3.160)
\end{aligned}
$$

Each term on the right-hand side of Eq. (3.160) is bounded as follows:

$$
\begin{aligned}
\left|\int_{\mathcal{M}} u T_x A_2 T\,\mathrm{d}\mathcal{M}\right| &\le |u|_{L^4} |T_x|_{L^4} |A_2 T|_{L^2} \\
&\le c|u|^{1/2}_{L^2} \|u\|^{1/2} |T_x|^{1/2}_{L^2} \|T_x\|^{1/2} |A_2 T|_{L^2} \\
&\le c|u|^{1/2}_{L^2} \|u\|^{1/2} \big|A_2^{1/2} T\big|^{1/2}_{L^2} |A_2 T|^{3/2}_{L^2} \\
&\le \frac{1}{6}|A_2 T|^2 + c|u|^2_{L^2} \|u\|^2 \big|A_2^{1/2} T\big|^2_{L^2},
\end{aligned}
$$

$$\left|\int_{\mathcal{M}} wT_z A_2 T \,\mathrm{d}\mathcal{M}\right| \le |w|_{L^4}|T_z|_{L^4}|A_2 T|_{L^2}$$
$$\le c|u_x|_{L^4}|T_z|_{L^4}|A_2 T|_{L^2}$$
$$\le c\|u\|^{1/2}\|u_x\|^{1/2}\big|A_2^{1/2}T\big|_{L^2}^{1/2}|A_2 T|^{3/2}$$
$$\le \frac{1}{6}|A_2 T|^2 + c\|u\|^2\|u_x\|^2\big|A_2^{1/2}T\big|_{L^2}^2,$$

and

$$\left|\int_{\mathcal{M}} F_T A_2 T \,\mathrm{d}\mathcal{M}\right| \le |F_T|_{L^2}|A_2 T|_{L^2}$$
$$\le \frac{1}{6}|A_2 T|_{L^2}^2 + c|F_T|_{L^2}^2.$$

Here, we have used the fact (easy to prove) that $|A_2^{1/2}T|_{L^2}$ is a norm equivalent to $\|T\|$ in V_2, and the much more involved result, proved in Section 4.3, that $|A_2 T|_{L^2(\mathcal{M})}$ is, on $D(A_2)$, a norm equivalent to $|T|_{H^2(\mathcal{M})}$; for the application of Theorem 4.3, we required Eq. (3.128), which is the one-dimensional analog of Eq. (4.54). Note that as explained in Remark 4.1, we believe that this purely technical hypothesis can be removed.

With this, we infer from Eq. (3.160) that

$$\frac{\mathrm{d}}{\mathrm{d}t}\big(\big|A_2^{1/2}T\big|_{L^2(\mathcal{M})}^2 + \alpha_T\big|T(x,0)\big|_{L^2(\Gamma_i)}^2\big) + |A_2 T|_{L^2(\mathcal{M})}^2$$
$$\le \xi(t) + \eta(t)\big|A_2^{1/2}T\big|_{L^2(\mathcal{M})}^2, \tag{3.161}$$

with $\xi = c|F_T|_{L^2(\mathcal{M})}^2$ and $\eta = c(|u|_{L^2(\mathcal{M})}^2 + \|u_x\|^2)\|u\|^2$. By assumption $\xi \in L^1(0,t_0)$ and the earlier estimates on U, u_x and u_z show that $\eta \in L^1(0,t_0)$. Then, Gronwall's lemma implies that $|A_2^{1/2}T|_{L^2(\mathcal{M})}$ is in $L^2(0,t_0)$, which means that T is in $L^\infty(0,t_0;H^1(\mathcal{M}))$ and $L^2(0,t_0;H^2(\mathcal{M}))$. This concludes the proof of Theorem 3.3.

3.4. Uniqueness of z-weak solutions for the space periodic case in dimension 2

In this section and in Section 3.5, we (continue to) consider the 2D PEs: all the functions are independent of the x_2-variable, but the velocity v is not zero, so we still model a 3D motion.

Our aim in these two sections is to present some additional existence, uniqueness, and regularity results for the PEs of the ocean in space dimension 2 with periodic boundary conditions. In this section, we prove the existence and uniqueness of the so-called z-weak solutions, and in Section 3.5, we prove the existence and uniqueness of more regular (strong) solutions for the PEs, up to $\mathcal{C}^\infty$-regularity.

For the sake of simplicity, we do not consider the salinity; introducing the salinity would not produce any additional technical difficulty. In this case, ρ is a linear function of T,[12] and, in what follows, ρ is the prognostic variable instead of T.

Because of the hydrostatic equation, it is not possible to produce a solution that is space periodic in all variables; for that reason ρ, p, and T below represent the deviation from a stratified solution $\bar{\rho}$ for which $N^2 = -(g/\rho_0)(\mathrm{d}\bar{\rho}/\mathrm{d}z)$ is a constant, and as usual $\mathrm{d}\bar{p}/\mathrm{d}z = -g\bar{\rho}$ and $\bar{\rho} = \rho_0(1 - \alpha(\overline{T} - T_r))$, ρ_0 and T_r being reference values of ρ and T (of the same order as $\bar{\rho}$ and $\overline{T}$). Furthermore, the periodic (disturbance) solutions that we consider present certain symmetries that are described below (see Eq. (3.163)). We refer the reader to Petcu, Temam and Wirosoetisno [2004b] for more details on the physical background.[13] The equations (for the deviations) read

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x_1} + w\frac{\partial u}{\partial x_3} - fv + \frac{1}{\rho_0}\frac{\partial p}{\partial x_1} = \nu_{\mathbf{v}}\,\Delta u + F_u, \tag{3.162a}$$

$$\frac{\partial v}{\partial t} + u\frac{\partial v}{\partial x_1} + w\frac{\partial v}{\partial x_3} + fu = \nu_{\mathbf{v}}\,\Delta v + F_v, \tag{3.162b}$$

$$\frac{\partial p}{\partial x_3} = -g\rho, \tag{3.162c}$$

$$\frac{\partial u}{\partial x_1} + \frac{\partial w}{\partial x_3} = 0, \tag{3.162d}$$

$$\frac{\partial \rho}{\partial t} + u\frac{\partial \rho}{\partial x_1} + w\frac{\partial \rho}{\partial x_3} - \frac{\rho_0 N^2}{g}w = \nu_\rho\,\Delta\rho + F_\rho. \tag{3.162e}$$

We notice easily that if u, v, ρ, w, and p are solutions of Eqs. (3.162a)–(3.162e) for $F = (F_u, F_v, F_\rho)$, then $\tilde{u}$, $\tilde{v}$, $\tilde{\rho}$, $\tilde{w}$, and $\tilde{p}$ are solutions of Eqs. (3.162a)–(3.162e) for $\widetilde{F}_u$, $\widetilde{F}_v$, and $\widetilde{F}_\rho$, where

$$\begin{aligned}
&\tilde{u}(x, z, t) = u(x, -z, t),\\
&\tilde{v}(x, z, t) = v(x, -z, t),\\
&\tilde{w}(x, z, t) = -w(x, -z, t),\\
&\tilde{p}(x, z, t) = p(x, -z, t),\\
&\tilde{\rho}(x, z, t) = -\rho(x, -z, t),\\
&\widetilde{F}_u(x, z, t) = F_u(x, -z, t),
\end{aligned} \tag{3.163}$$

[12] In fact, ρ is an affine function of T, but the deviation from the density $\overline{\rho}$ considered below is a linear function of the deviation from the temperature $\overline{T}$.

[13] Alternatively ρ is sometimes as in Eq. (2.16) and then Eq. (3.162e) is the combination of Eqs. (2.14) and (2.15) when it is assumed that $\beta_T(\mu_T - 1) = \beta_S(\mu_S - 1)$ and $\mu_\rho = -\beta_T\mu_T + \beta_S\mu_S > 0$.

$$\widetilde{F}_v(x, z, t) = F_v(x, -z, t),$$
$$\widetilde{F}_\rho(x, z, t) = -F_\rho(x, -z, t).$$

Therefore, if we assume that $F_u F_v$ and F_ρ are periodic and F_u and F_v are even in z and, F_ρ is odd in z, then we can anticipate the existence of a solution of Eqs. (3.162a)–(3.162e) such that

$$u, v, w, p, \text{ and } \rho \text{ are periodic in } x \text{ and } z \text{ with periods } L_1 \text{ and } L_3, \tag{3.164}$$

and[14]

$$u, v, \text{ and } p \text{ are even in } z; w \text{ and } \rho \text{ are odd in } z, \tag{3.165}$$

provided the initial conditions satisfy the same symmetry properties.

One motivation for considering periodic boundary conditions is that they are needed in numerical studies of rotating stratified turbulence (see, e.g., BARTELLO [1995] and also for the study of the renormalized equations considered in PETCU, TEMAM and WIROSOETISNO [2004b]). Our aim is to solve the problem (Eqs. (3.162a)–(3.162e)) with initial data

$$u = u_0, \quad v = v_0, \quad \rho = \rho_0 \quad \text{at } t = 0. \tag{3.166}$$

We introduce here the natural spaces for this problem:

$$V = \left\{ (u, v, \rho) \in \left(\dot{H}^1_{\text{per}}(\mathcal{M})\right)^3, u, v \text{ even in } z, \rho \text{ odd in } z, \right.$$
$$\left. \int_{-L_3/2}^{L_3/2} u\left(x, z'\right) \mathrm{d}z' = 0 \right\}, \tag{3.167}$$

$$H = \text{closure of } V \text{ in } \left(\dot{L}^2(\mathcal{M})\right)^3, \tag{3.168}$$

$$V_2 = \text{ the closure of } V \cap (\dot{H}^2_{\text{per}}(\mathcal{M}))^3 \text{ in } (\dot{H}^2_{\text{per}}(\mathcal{M}))^3. \tag{3.169}$$

Here, $\mathcal{M}$ is the limited domain

$$\mathcal{M} = (0, L_1) \times (-L_3/2, L_3/2), \tag{3.170}$$

and, as mentioned, we assume space periodicity with period $\mathcal{M}$, i.e., all functions are taken to satisfy $f(x + L_1, z, t) = f(x, z, t) = f(x, z + L_3, t)$ when extended to $\mathbb{R}^2$. Moreover, we assume that the symmetries (Eq. (3.163)) hold. The dot above $\dot{H}^1_{\text{per}}$ or $\dot{L}^2$ denotes the functions with average in $\mathcal{M}$ equal to zero. These spaces are endowed with the following Hilbert scalar products; in H, the scalar product is

$$\left(U, \widetilde{U}\right)_H = \left(u, \tilde{u}\right)_{L^2} + \left(v, \tilde{v}\right)_{L^2} + \kappa\left(\rho, \tilde{\rho}\right)_{L^2}, \tag{3.171}$$

[14]Alternatively ρ is sometimes as in Eq. (2.16) and then Eq. (3.162e) is the combination of Eqs. (2.14) and (2.15) when it is assumed that $\beta_T(\mu_T - 1) = \beta_S(\mu_S - 1)$ and $\mu_\rho = -\beta_T\mu_T + \beta_S\mu_S > 0$.

and in $\dot{H}^1_{\mathrm{per}}$ and V, the scalar product is (using the same notation when there is no ambiguity)

$$\big((U,\widetilde{U})\big) = \big((u,\tilde{u})\big) + \big((v,\tilde{v})\big) + \kappa\big((\rho,\tilde{\rho})\big), \tag{3.172}$$

where we have written $\mathrm{d}\mathcal{M}$ for $\mathrm{d}x\,\mathrm{d}z$, and

$$\big((\phi,\tilde{\phi})\big) = \int_{\mathcal{M}} \left(\frac{\partial\phi}{\partial x}\frac{\partial\tilde{\phi}}{\partial x} + \frac{\partial\phi}{\partial z}\frac{\partial\tilde{\phi}}{\partial z} \right) \mathrm{d}\mathcal{M}. \tag{3.173}$$

The positive constant κ is defined below. We have

$$|U|_H \leq c_0 \|U\| \quad \forall U \in V, \tag{3.174}$$

where $c_0 > 0$ is a positive constant related to κ and the Poincaré constant in $\dot{H}^1_{\mathrm{per}}(\mathcal{M})$. More generally, the c_i, c_i', and c_i'' will denote various positive constants. Inequality (Eq. (3.174)) implies that $\|U\| = ((U,U))^{1/2}$ is indeed a norm on V.

Let us show how we can express the diagnostic variables w and p in terms of the prognostic variables u, v, and ρ, the situation being slightly different here due to the boundary conditions and the symmetries. For each $U = (u, v, \rho) \in V$, we can determine uniquely $w = w(U)$ from Eq. (3.215),

$$w(U) = w(x,z,t) = -\int_0^z u_x\big(x,z',t\big)\,\mathrm{d}z', \tag{3.175}$$

since $w(x,0) = 0$, w being odd in z. Furthermore, writing that $w(x,-L_3/2,t) = w(x,L_3/2,t)$, we also have

$$\int_{-L_3/2}^{L_3/2} u_x\big(x,z',t\big)\,\mathrm{d}z' = 0. \tag{3.176}$$

As for the pressure, we obtain from Eq. (3.214),

$$p(x,z,t) = p_{\mathrm{s}}(x,t) - \int_0^z \rho\big(x,z',t\big)\,\mathrm{d}z', \tag{3.177}$$

where $p_{\mathrm{s}} = p(x,0,t)$ is the surface pressure. Thus, we can uniquely determine the pressure p in terms of ρ up to p_{s}.

It is appropriate to use Fourier series and we write, e.g., for u,

$$u(x,z,t) = \sum_{(k_1,k_3)\in\mathbb{Z}} u_{k_1,k_3}(t)\mathrm{e}^{\mathrm{i}(k_1'x+k_3'z)}, \tag{3.178}$$

where for notational conciseness we set $k_1' = 2\pi k_1/L_1$ and $k_3' = 2\pi k_3/L_3$. Since u is real and even in z, we have $u_{-k_1,-k_3} = \bar{u}_{k_1,k_3} = \bar{u}_{k_1,-k_3}$, where $\bar{u}$ denotes the complex conjugate of u. Regarding the pressure, we obtain from Eq. (3.177)

$$\begin{aligned} p(x,z,t) &= p(x,0,t) - \int_0^z \sum_{(k_1,k_3)} \rho_{k_1,k_3} \mathrm{e}^{\mathrm{i}(k_1'x + k_3'z')}\,\mathrm{d}z' \\ &= \sum_{k_1} p_{\mathrm{s}k_1} \mathrm{e}^{\mathrm{i}k_1'x} - \sum_{(k_1,k_3),k_3\neq 0} \frac{\rho_{k_1,k_3}}{\mathrm{i}k_3'} \mathrm{e}^{\mathrm{i}k_1'x}\left(\mathrm{e}^{\mathrm{i}k_3'z} - 1\right) \\ &\qquad \text{[using the fact that } \rho_{k_1,0} = 0,\ \rho \text{ being odd in } z\text{]} \\ &= \sum_{k_1}\left(p_{\mathrm{s}k_1} + \sum_{k_3\neq 0} \frac{\rho_{k_1,k_3}}{\mathrm{i}k_3'}\right)\mathrm{e}^{\mathrm{i}k_1'x} - \sum_{(k_1,k_3),k_3\neq 0} \frac{\rho_{k_1,k_3}}{\mathrm{i}k_3'} \mathrm{e}^{\mathrm{i}(k_1'x + k_3'z)} \\ &= \sum_{k_1} p_{\star k_1} \mathrm{e}^{\mathrm{i}k_1'x} - \sum_{(k_1,k_3),k_3\neq 0} \frac{\rho_{k_1,k_3}}{\mathrm{i}k_3'} \mathrm{e}^{\mathrm{i}(k_1'x + k_3'z)}, \end{aligned}$$

where we denoted by p_{s} the surface pressure and $p_\star = \sum_{k_1\in\mathbb{Z}} p_{\star k_1} \mathrm{e}^{\mathrm{i}k_1'x}$, which is the average of p in the vertical direction, is defined by

$$p_{\star,k_1} = p_{\mathrm{s}k_1} + \sum_{k_3\neq 0} \frac{\rho_{k_1,k_3}}{\mathrm{i}k_3'}.$$

Note that p is fully determined by ρ, up to one of the terms p_{s} or $p_\star$, which is connected by the relation above.

The dots above $\dot{H}^1_{\mathrm{per}}$ and $\dot{L}^2$ denote the functions with zero average over $\mathcal{M}$.

The variational formulation of our problem is

Find $U : [0, t_0] \to V$ *such that*

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t}(U, U^\flat)_H + a(U, U^\flat) + b(U, U, U^\flat) + e(U, U^\flat) = (F, U^\flat)_H, \quad \forall\, U^\flat \in V, \\ &U(0) = U_0. \end{aligned} \tag{3.179}$$

In Eq. (3.179), we introduced the following forms:
$a : V \times V \to \mathbb{R}$ bilinear, continuous, coercive:

$$a(U, U^\flat) = \nu((u, u^\flat)) + \nu((v, v^\flat)) + \kappa\mu((\rho, \rho^\flat)), \tag{3.180}$$

with $\kappa = g^2/N^2\rho_0^2$,

$b : V \times V \times V_2 \to \mathbb{R}$ trilinear continuous (see Lemma 2.1):

$$\begin{aligned} b(U, U^\sharp, U^\flat) = &\int_{\mathcal{M}} (u\frac{\partial u^\sharp}{\partial x}u^\flat + v\frac{\partial u^\sharp}{\partial y}u^\flat + w(U)\frac{\partial u^\sharp}{\partial z}u^\flat)\,\mathrm{d}\mathcal{M} \\ &+ \int_{\mathcal{M}} (u\frac{\partial v^\sharp}{\partial x}v^\flat + v\frac{\partial v^\sharp}{\partial y}v^\flat + w(U)\frac{\partial v^\sharp}{\partial z}v^\flat)\,\mathrm{d}\mathcal{M} \\ &+ \kappa\int_{\mathcal{M}} (u\frac{\partial \rho^\sharp}{\partial x}\rho^\flat + v\frac{\partial \rho^\sharp}{\partial y}\rho^\flat + w(U)\frac{\partial \rho^\sharp}{\partial z}\tilde{\rho})\,\mathrm{d}\mathcal{M}, \end{aligned} \tag{3.181}$$

$e : V \times V \to \mathbb{R}$ bilinear, continuous:

$$\begin{aligned} e(U, U^\flat) = &f\int_{\mathcal{M}} (uv^\flat - vu^\flat)\,\mathrm{d}\mathcal{M} + \frac{g}{\rho_0}\int_{\mathcal{M}} \rho w(U^\flat)\mathrm{d}\mathcal{M} \\ &- \frac{g}{\rho_0}\int_{\mathcal{M}} \rho^\flat w(U)\,\mathrm{d}\mathcal{M}, \end{aligned} \tag{3.182}$$

with $e(U, U) = 0$ for all $U \in V$.

Problem (Eq. (3.179)) can also be written as an operator evolution equation in V_2':

$$\begin{aligned} &\frac{\mathrm{d}U}{\mathrm{d}t} + AU + B(U, U) + EU = F, \\ &U(0) = U_0, \end{aligned} \tag{3.183}$$

where we introduced the following operators:

$$\begin{aligned} &A \text{ linear, continuous from } V \text{ into } V' \text{ defined by} \\ &\qquad \langle AU, U^\flat\rangle = a(U, U^\flat), \quad \forall U, U^\flat \in V, \end{aligned} \tag{3.184}$$

$$\begin{aligned} &B \text{ bilinear, continuous from } V \times V \text{ into } V_2' \text{ defined by} \\ &\qquad \langle B(U, U^\flat),\ U^\sharp\rangle = b(U, U^\flat, U^\sharp) \quad \forall\, U, U^\flat \in V,\ \forall\, U^\sharp \in V_2, \end{aligned} \tag{3.185}$$

$$\begin{aligned} &E \text{ linear, continuous from } V \text{ into } V' \text{ defined by} \\ &\qquad \langle EU, U^\flat\rangle = e(U, U^\flat), \quad \forall\, U, U^\flat \in V, \text{ with } \langle EU, U\rangle = 0. \end{aligned} \tag{3.186}$$

In the previous section, we have proved the existence and uniqueness, globally in time, of strong solutions for the 2D PEs, but we could not prove the uniqueness of the weak solution (result that is available for the Navier-Stokes equations). In this section, we prove an intermediate result, i.e., the existence and uniqueness, globally in time, of solutions that are weak in the horizontal direction and strong in the vertical direction (the so-called z-weak solutions). We start by introducing the function spaces necessary for this problem:

$$\mathcal{V} = \left\{U = (u, v, \rho) \in V,\ \frac{\partial U}{\partial x_3} \in (\dot{H}^1_{\mathrm{per}}(\mathcal{M}))^3\right\}, \tag{3.187}$$

which is a Hilbert space when endowed with the following norm:

$$|U|_{\mathcal{V}}^2 = \|U\|^2 + \Big\|\frac{\partial U}{\partial x_3}\Big\|^2. \tag{3.188}$$

Another useful function space is

$$\mathcal{H} = \{U = (u, v, \rho) \in H,\ \frac{\partial U}{\partial x_3} \in (\dot{L}^2_{\text{per}}(\mathcal{M}))^3\}, \tag{3.189}$$

which is a Hilbert space when endowed with the norm:

$$|U|_{\mathcal{H}}^2 = |U|_{L^2}^2 + \Big|\frac{\partial U}{\partial x_3}\Big|_{L^2}^2. \tag{3.190}$$

We now prove the existence and uniqueness, globally in time, of a z-weak solution for Eq. (3.162).

THEOREM 3.4. (z-weak solutions in dimension 2). *Given $U_0 \in \mathcal{H}$ and $F \in L^\infty(0, T; \mathcal{H})$, there exists a unique solution U to problem, (Eq. (3.179)) satisfying the initial condition $U(0) = U_0$ and*

$$U \in \mathcal{C}([0, T]; \mathcal{H}) \cap L^2(0, T; \mathcal{V}). \tag{3.191}$$

PROOF. In Section 2.3, we proved the existence of weak solutions for the PEs the proof is for the general case, both in space dimensions 2 and 3. It remains to prove that starting with an initial data and a forcing more regular (satisfying the hypotheses of Theorem 3.4), the solution is strong in the vertical direction. In order to prove that we need to obtain a priori estimates for $U_{x_3} = \partial U/\partial x_3$, we formally differentiate Eqs. (3.162a), (3.162b), and (3.162e) in x_3 and then multiply, respectively, by u_{x_3}, v_{x_3}, and ρ_{x_3}, add the resulting equation and integrate over $\mathcal{M}$. We find

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|U_{x_3}|_{L^2}^2 &+ \int_{\mathcal{M}} (u_{x_1} + w_{x_3})u_{x_3}^2 \mathrm{d}\mathcal{M} + \frac{1}{\rho_0}\int_{\mathcal{M}} p_{x_1x_3}u_{x_3}\mathrm{d}\mathcal{M} \\ &+ \int_{\mathcal{M}} (u_{x_3}v_{x_1} + v_{x_3}w_{x_3})v_{x_3}\mathrm{d}\mathcal{M} \\ &+ \int_{\mathcal{M}} (u_{x_3}\rho_{x_1} + w_{x_3}\rho_{x_3})\rho_{x_3}\mathrm{d}\mathcal{M} + \nu\|U_{x_3}\|^2 \\ &= (F_{x_3}, U_{x_3})_{L^2}.\end{aligned} \tag{3.192}$$

A term has been omitted in Eq. (3.192), which is zero, because of the mass conservation equation. The pressure term can be estimated using the hydrostatic equation (Eq. (3.162c)) and integrating by parts:

$$\left|\int_{\mathcal{M}} p_{x_1x_3}u_{x_3}\mathrm{d}\mathcal{M}\right| = \left|-g\int_{\mathcal{M}} \rho_{x_1}u_{x_3}\mathrm{d}\mathcal{M}\right| \le g|\rho|\|U_{x_3}\|. \tag{3.193}$$

We also estimate

$$\left|\int_{\mathcal{M}}(u_{x_3}v_{x_1}+v_{x_3}w_{x_3})v_{x_3}\mathrm{d}\mathcal{M}\right| \le |u_{x_3}|_{L^4}|v_{x_1}|_{L^2}|v_{x_3}|_{L^4}+|u_{x_1}|_{L^2}|v_{x_3}|^2_{L^4}$$
$$\le c|U_{x_3}|_{L^2}\|U_{x_3}\|\|U\|, \tag{3.194}$$

and

$$\left|\int_{\mathcal{M}}(u_{x_3}\rho_{x_1}+w_{x_3}\rho_{x_3})\rho_{x_3}\mathrm{d}\mathcal{M}\right| \le |u_{x_3}|_{L^4}|\rho_{x_1}|_{L^2}|\rho_{x_3}|_{L^4}+|u_{x_1}|_{L^2}|\rho_{x_3}|^2_{L^4}$$
$$\le c|U_{x_3}|\|U_{x_3}\|\|U\|. \tag{3.195}$$

Using the above estimates in Eq. (3.192), we find

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|U_{x_3}|^2_{L^2}+\nu\|U_{x_3}\|^2 \le |F|_{L^2}\|U_{x_3}\|+c|\rho|_{L^2}\|U_{x_3}\|$$
$$+c|U_{x_3}|_{L^2}\|U_{x_3}\|\|U\|, \tag{3.196}$$

which, by the Young inequality, implies

$$\frac{\mathrm{d}}{\mathrm{d}t}|U_{x_3}|^2_{L^2}+\nu\|U_{x_3}\|^2 \le c|F|^2_{L^2}+c|U_{x_3}|^2_{L^2}\|U\|^2+c|U|^2_{L^2}. \tag{3.197}$$

Applying the Gronwall lemma to Eq. (3.197) and using the estimates valid for weak solutions (U in $L^2(0,T,V)$, $\forall T$), we find a bound for U_{x_3} in $L^\infty(0,T;L^2(\mathcal{M})^3)$ and $L^2(0,T;\dot{H}^1_{\mathrm{per}}(\mathcal{M})^3)$.

Using all these estimates and the Galerkin method, we can prove the existence of a z-weak solution, i.e., with U and U_{x_3} belonging to $L^\infty(0,T;L^2(\mathcal{M})^3)\cap L^2(0,T;\dot{H}^1_{\mathrm{per}}(\mathcal{M})^3)$.

The forward uniqueness of a z-weak solution is then proved classically: we suppose that U_1 and U_2 are two z-weak solutions for Eq. (3.183), satisfying the same initial condition. Then, $\tilde{U}=U_1-U_2$ satisfies the following equation:

$$\tilde{U}'+A\tilde{U}+E\tilde{U}+B(U_1,\tilde{U})+B(\tilde{U},U_2)=0, \tag{3.198}$$

with $\tilde{U}(0)=0$.

We take the $V'-V$ duality product of Eq. (3.198) with $\tilde{U}$. We find

$$\frac{\mathrm{d}}{\mathrm{d}t}|\tilde{U}|^2_H+c_0\|\tilde{U}\|^2_V+b(U_1,\tilde{U},\tilde{U})+b(\tilde{U},U_2,\tilde{U})\le 0. \tag{3.199}$$

From the orthogonality property, we know that $b(U_1,\tilde{U},\tilde{U})=0$, under the hypotheses of Lemma 2.1. But we note here that in our case, U_1 and $\tilde{U}$ do not satisfy the conditions in Lemma 2.1; however, the same result can be easily obtained for the case $U_1\in V$ and $\tilde{U}\in\mathcal{V}$, using the same kind of reasoning as before. It remains to estimate $b(\tilde{U},U_2,\tilde{U})$:

$$b(\tilde{U},U_2,\tilde{U})=\int_{\mathcal{M}}\tilde{u}\frac{\partial U_2}{\partial x_1}\tilde{U}\mathrm{d}\mathcal{M}+\int_{\mathcal{M}}w(\tilde{U})\frac{\partial U_2}{\partial x_3}\tilde{U}\mathrm{d}\mathcal{M}. \tag{3.200}$$

The first term of Eq. (3.200) is estimated using the Holder inequality and the Sobolev embeddings:

$$\Big|\int_{\mathcal{M}} \tilde{u}\frac{\partial U_2}{\partial x_1}\tilde{U}\mathrm{d}\mathcal{M}\Big| \le |\tilde{u}|_{L^4}\Big|\frac{\partial U_2}{\partial x_1}\Big|_{L^2}|\tilde{U}|_{L^4} \le c|\tilde{U}|_H\|\tilde{U}\|\|U_2\|. \tag{3.201}$$

For the second term, we find

$$\begin{aligned}\Big|\int_{\mathcal{M}} w(\tilde{U})\frac{\partial U_2}{\partial x_3}\tilde{U}\mathrm{d}\mathcal{M}\Big| &\le |w(\tilde{U})|_{L^2}\Big|\frac{\partial U_2}{\partial x_3}\Big|_{L^4}|\tilde{U}|_{L^4}\\ &\le c|\tilde{U}|_H^{1/2}\|\tilde{U}\|^{3/2}\Big|\frac{\partial U_2}{\partial x_3}\Big|_{L^2}^{1/2}\Big\|\frac{\partial U_2}{\partial x_3}\Big\|^{1/2}.\end{aligned} \tag{3.202}$$

Using the above estimates in Eq. (3.199), we find

$$\frac{\mathrm{d}}{\mathrm{d}t}|\tilde{U}|_H^2 + c_0\|\tilde{U}\|_V^2 \le g(t)|\tilde{U}|_H^2, \tag{3.203}$$

where

$$g(t) = c\|U_2\|^2 + c\Big|\frac{\partial U_2}{\partial x_3}\Big|^2\Big\|\frac{\partial U_2}{\partial x_3}\Big\|^2.$$

Since U_2 is a z-weak solution, the function g belongs to $L^1(0,T)$ for any $T>0$. So applying the Gronwall lemma to Eq. (3.203), we find that $\tilde{U}(t)=0$ for all $t>0$.

It remains to prove that the z-weak solution U belongs to $\mathcal{C}([0,T],\mathcal{H})$. We start by proving that $B(U,U)$ belongs to $L^2(0,T,\mathcal{V}')$. Let $\tilde{U}$ be in $\mathcal{V}$; then,

$$< B(U,U),\tilde{U} >_{\mathcal{V}',\mathcal{V}} = b(U,U,\tilde{U}) = \int_{\mathcal{M}} u\frac{\partial U}{\partial x_1}\tilde{U}\mathrm{d}\mathcal{M} + \int_{\mathcal{M}} w(U)\frac{\partial U}{\partial x_3}\tilde{U}\mathrm{d}\mathcal{M}. \tag{3.204}$$

We estimate $b(U,U,\tilde{U})$ using Lemma 3.9 below, which is the 2D analog of Lemma 3.1, and we find

$$\|B(U,U)\|_{\mathcal{V}'} \le c|U|_H\|U\|_{\mathcal{V}} + c|U_{x_1}|_{L^2(\mathcal{M})}\Big|\frac{\partial U}{\partial x_3}\Big|_{L^2(\mathcal{M})}, \tag{3.205}$$

which, taking into account that $U\in L^2(0,T,\mathcal{V})$, implies that $B(U,U)\in L^2(0,T,\mathcal{V}')$.

Then, one can easily conclude from Eq. (3.183) that $U'\in L^2(0,T,\mathcal{V}')$. We know that $U\in L^2(0,T,\mathcal{V})$ and $\mathcal{V}\subset V\subset H\subset V'\subset\mathcal{V}'$ where each space is dense into the other and the injections are continuous. We can then conclude, using a technical result (see Temam [1977] for more details), that U belongs to $\mathcal{C}([0,T],\mathcal{H})$, observing that $\mathcal{H}=[\mathcal{V},\mathcal{V}']_{1/2}$ is the 1/2-interpolate between $\mathcal{V}$ and $\mathcal{V}'$. □

We now state and prove Lemma 3.13, which is the 2D analog of Lemma 3.1.

LEMMA 3.13. *In space dimension 2, the form b is trilinear continuous on* $\mathcal{V} \times \mathcal{V} \times \mathcal{V}$, *and we have*

$$|b(U, \bar{U}, \tilde{U})| \leq c(|U|_{L^2(\mathcal{M})}||\bar{U}||_{\mathcal{V}}||\tilde{U}||_{\mathcal{V}} + ||U|||\bar{U}|_{\mathcal{H}}||\tilde{U}||). \tag{3.206}$$

PROOF. In space dimension 2, we have

$$b(U, \bar{U}, \tilde{U}) = \int_{\mathcal{M}} u \frac{\partial \bar{U}}{\partial x_1} \tilde{U} \mathrm{d}\mathcal{M} + \int_{\mathcal{M}} w(U) \frac{\partial \bar{U}}{\partial x_3} \tilde{U} \mathrm{d}\mathcal{M}. \tag{3.207}$$

The first term is estimated as follows

$$\begin{aligned}
\Big| \int_{\mathcal{M}} u \frac{\partial \bar{U}}{\partial x_1} \cdot \tilde{U} \mathrm{d}\mathcal{M} \Big| &= \Big| \int_0^{L_1} \int_{-L_3/2}^{-L_3/2} u \frac{\partial \bar{U}}{\partial x_1} \cdot \tilde{U} \, \mathrm{d}x_3 \, \mathrm{d}x_1 \Big| \\
&\leq \int_0^{L_1} |u|_{L^2_{x_3}} \Big| \frac{\partial U}{\partial x_1} \Big|_{L^4_{x_3}} |\tilde{U}|_{L^4_{x_3}} \, \mathrm{d}x_1 \\
&\leq \int_0^{L_1} |u|_{L^2_{x_3}} \Big| \frac{\partial \bar{U}}{\partial x_1} \Big|^{1/2}_{L^2_{x_3}} \Big(\Big| \frac{\partial \bar{U}}{\partial x_1} \Big|^{1/2}_{L^2_{x_3}} + \Big| \frac{\partial^2 \bar{U}}{\partial x_1 \partial x_3} \Big|^{1/2}_{L^2_{x_3}} \Big) |\tilde{U}|^{1/2}_{L^2_{x_3}} \\
&\qquad \Big(|\tilde{U}|^{1/2}_{L^2_{x_3}} + \Big| \frac{\partial \tilde{U}}{\partial x_3} \Big|^{1/2}_{L^2_{x_3}} \Big).
\end{aligned} \tag{3.208}$$

Here and below $L^q_{x_1}$ is $L^q(0, L_1)$ and $L^q_{x_3}$ is $L^q(-L_3/2, L_3/2)$.

The most difficult term of Eq. (3.208) is

$$\begin{aligned}
&\int_0^{L_1} |u|_{L^2_{x_3}} \Big| \frac{\partial \bar{U}}{\partial x_1} \Big|^{1/2}_{L^2_{x_3}} \Big| \frac{\partial^2 \bar{U}}{\partial x_1 \partial x_3} \Big|^{1/2}_{L^2_{x_3}} |\tilde{U}|^{1/2}_{L^2_{x_3}} \Big| \frac{\partial \tilde{U}}{\partial x_3} \Big|^{1/2}_{L^2_{x_3}} \, \mathrm{d}x_1 \\
&\quad \leq c|U|_{L^2(\mathcal{M})} \Big| \frac{\partial \bar{U}}{\partial x_1} \Big|^{1/2}_{L^2(\mathcal{M})} \Big| \frac{\partial^2 \bar{U}}{\partial x_1 \partial x_3} \Big|^{1/2}_{L^2(\mathcal{M})} ||\tilde{U}|_{L^2_{x_3}}|^{1/2}_{L^\infty_{x_1}} \Big| \Big| \frac{\partial \tilde{U}}{\partial x_3} \Big|_{L^2_{x_3}} \Big|^{1/2}_{L^\infty_{x_1}} \\
&\quad \leq c|U|_{L^2(\mathcal{M})} \Big| \frac{\partial \bar{U}}{\partial x_1} \Big|^{1/2}_{L^2(\mathcal{M})} \Big| \frac{\partial^2 \bar{U}}{\partial x_1 \partial x_3} \Big|^{1/2}_{L^2(\mathcal{M})} \|\tilde{U}\|^{1/2} \Big\| \frac{\partial \tilde{U}}{\partial x_3} \Big\|^{1/2} \\
&\quad \leq c|U|_{L^2(\mathcal{M})} \|\bar{U}\|_{\mathcal{V}} \|\tilde{U}\|_{\mathcal{V}};
\end{aligned} \tag{3.209}$$

we used the fact that, in dimension 1, we have the Sobolev embedding $H^1_{x_1} \subset L^\infty_{x_1}$, which implies that

$$|\tilde{U}|_{L^\infty_{x_1}(L^2_{x_3})} \leq c|\tilde{U}|_{H^1_{x_1}(L^2_{x_3})} \leq c\|\tilde{U}\|. \tag{3.210}$$

The second term is estimated as follows

$$\begin{aligned}
\Big| \int_{\mathcal{M}} w(U) \frac{\partial \bar{U}}{\partial x_3} \cdot \tilde{U} \mathrm{d}\mathcal{M} \Big| &\leq \int_0^{L_1} |w(U)|_{L^\infty_{x_3}} \Big| \frac{\partial \bar{U}}{\partial x_3} \Big|_{L^2_{x_3}} |\tilde{U}|_{L^2_{x_3}} \, \mathrm{d}x_1 \\
&\leq c \int_0^{L_1} |U_{x_1}|_{L^2_{x_3}} \Big| \frac{\partial \bar{U}}{\partial x_3} \Big|_{L^2_{x_3}} |\tilde{U}|_{L^2_{x_3}} \, \mathrm{d}x_1 \\
&\leq c |U_{x_1}|_{L^2_{x_1}(L^2_{x_3})} \Big| \frac{\partial \bar{U}}{\partial x_3} \Big|_{L^2_{x_1}(L^2_{x_3})} |\tilde{U}|_{L^\infty_{x_1}(L^2_{x_3})} \\
&\leq c |U_{x_1}|_{L^2(\mathcal{M})} \Big| \frac{\partial \bar{U}}{\partial x_3} \Big|_{L^2(\mathcal{M})} |\tilde{U}|_{H^1_{x_1}(L^2_{x_3})} \\
&\leq c |U_{x_1}|_{L^2(\mathcal{M})} \Big| \frac{\partial \bar{U}}{\partial x_3} \Big|_{L^2(\mathcal{M})} \|\tilde{U}\| \\
&\leq c ||U|| \, |\bar{U}|_{\mathcal{H}} ||\tilde{U}||. \qquad (3.211)
\end{aligned}$$

The lemma is proved. □

3.5. The space periodic case in dimension 2: higher regularity[15]

As announced, our aim in this section is to prove the existence and uniqueness of more regular solutions, up to $\mathcal{C}^\infty$–regularity, in space dimension 2. As in Section 3.4, we will use ρ instead of T, but unlike the preceding sections (but this is not important), we consider here the PEs written in nondimensional form, i.e. (see, e.g., Petcu, Temam and Wirosoetisno [2004b]),

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + w \frac{\partial u}{\partial z} - \frac{1}{R_0} v + \frac{1}{R_0} \frac{\partial p}{\partial x} = \nu_{\mathbf{v}} \Delta u + F_u, \qquad (3.212)$$

$$\frac{\partial v}{\partial t} + u \frac{\partial v}{\partial x} + w \frac{\partial v}{\partial z} + \frac{1}{R_0} u = \nu_{\mathbf{v}} \Delta v + F_v, \qquad (3.213)$$

$$\frac{\partial p}{\partial z} = -\rho, \qquad (3.214)$$

$$\frac{\partial u}{\partial x} + \frac{\partial w}{\partial z} = 0, \qquad (3.215)$$

$$\frac{\partial \rho}{\partial t} + u \frac{\partial \rho}{\partial x} + w \frac{\partial \rho}{\partial z} - \frac{N^2}{R_0} w = \nu_\rho \Delta \rho + F_\rho. \qquad (3.216)$$

Here (u, v, w) are the three components of the velocity vector and, as usual, we denote by p and ρ the pressure and density deviations, respectively, from the background

[15]This section essentially reproduces the article by Petcu, Temam and Wirosoetisno [2004b], with the authorization of the publisher of the journal.

state mentioned above. The (dimensionless) parameters are the Rossby number R_0, the Burgers number N, and the inverse (eddy) Reynolds numbers $\nu_{\mathbf{v}}$ and ν_ρ.

In all that follows, we assume that the solutions have the same symmetry properties as in the previous section.

Our aim is to solve the problem (Eqs. (3.212)–(3.216)) with the periodicity and symmetry properties as in Eq. (3.163) and with initial data

$$u = u_0, \quad v = v_0, \quad \rho = \rho_0 \quad \text{at } t = 0.$$

The two spatial directions are $0x$ and $0z$, corresponding to the west–east and vertical directions in the so-called f-plane approximation for geophysical flows ($\Delta = \partial^2/\partial x^2 + \partial^2/\partial z^2$).

The rest of this section is organized as follows: we start in Section 3.5.1 by recalling the variational formulation of problem (Eqs. (3.212)–(3.216)) under suitable assumptions, and we say a few words about (the now standard) proof of existence of weak solutions for the PEs. We continue in Section 3.5.2 by proving the existence and uniqueness of strong solutions, giving another version of Theorem 3.3 for the case with periodic boundary conditions. Finally, in Section 3.5.3, we prove the existence of more regular solutions, up to $\mathcal{C}^\infty$-regularity.

3.5.1. Existence of weak solutions for the PEs

We now obtain the variational formulation of problem (Eqs. (3.212)–(3.216)). For that purpose, we consider a test function $\widetilde{U} = (\tilde{u}, \tilde{v}, \tilde{\rho}) \in V$, and we multiply Eqs. (3.212), (3.213), and (3.216), respectively, by $\tilde{u}$, $\tilde{v}$, and $\kappa\tilde{\rho}$, where the constant κ (which was already introduced in Eqs. (3.171) and (3.172)) will be chosen later. We add the resulting equations and integrate over $\mathcal{M}$. We find

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}\big(U, \widetilde{U}\big)_H + b\big(U, U, \widetilde{U}\big) + a\big(U, \widetilde{U}\big) + \frac{1}{R_0} e\big(U, \widetilde{U}\big)\\ &\quad = \big(F, \widetilde{U}\big)_H \quad \forall \widetilde{U} \in V.\end{aligned} \tag{3.217}$$

Here, we set (compare to Eqs. (3.180)–(3.182))

$$\begin{aligned}
a\big(U, \widetilde{U}\big) &= \nu_{\mathbf{v}}\big((u, \tilde{u})\big) + \nu_{\mathbf{v}}\big((v, \tilde{v})\big) + \kappa\nu_\rho\big((\rho, \tilde{\rho})\big),\\
e\big(U, \widetilde{U}\big) &= \int_{\mathcal{M}} \big(u\tilde{v} - v\tilde{u}\big)\,\mathrm{d}\mathcal{M} + \int_{\mathcal{M}} \big(\rho\tilde{w} - \kappa N^2 w\tilde{\rho}\big)\,\mathrm{d}\mathcal{M},\\
b\big(U, U^\sharp, \widetilde{U}\big)&\\
&= \int_{\mathcal{M}} \left(u\frac{\partial u^\sharp}{\partial x} + w(U)\frac{\partial u^\sharp}{\partial z}\right)\tilde{u}\,\mathrm{d}\mathcal{M} + \int_{\mathcal{M}} \left(u\frac{\partial v^\sharp}{\partial x} + w(U)\frac{\partial v^\sharp}{\partial z}\right)\tilde{v}\,\mathrm{d}\mathcal{M}\\
&\quad + \int_{\mathcal{M}} \left(u\frac{\partial \rho^\sharp}{\partial x} + w(U)\frac{\partial \rho^\sharp}{\partial z}\right)\tilde{\rho}\,\mathrm{d}\mathcal{M}.
\end{aligned}$$

We now choose $\kappa = 1/N^2$, and in this way, we find $e(U, U) = 0$. Also it can be easily seen that

$$\begin{aligned} &a : V \times V \to \mathbb{R} \text{ is bilinear, continuous, coercive, } a(U, U) \geq c_1 \|U\|^2, \\ &e : V \times V \to \mathbb{R} \text{ is bilinear, continuous, } e(U, U) = 0, \\ &b \text{ is trilinear, continuous from } V \times V_2 \times V \text{ into } \mathbb{R}, \text{ and} \\ &\quad \text{from } V \times V \times V_2 \text{ into } \mathbb{R}, \end{aligned} \tag{3.218}$$

where V_2 is the closure of $V \cap (H^2_{\text{per}}(\mathcal{M}))^3$ in $(H^2_{\text{per}}(\mathcal{M}))^3$. Furthermore,

$$\begin{aligned} &b\big(U, \widetilde{U}, U^\sharp\big) = -b\big(U, U^\sharp, \widetilde{U}\big), \\ &b\big(U, \widetilde{U}, \widetilde{U}\big) = 0, \end{aligned} \tag{3.219}$$

when $U, \widetilde{U}, U^\sharp \in V$ with $\widetilde{U}$ or $U^\sharp$ in V_2. We also have the following.

LEMMA 3.14. *There exists a constant $c_2 > 0$ such that for all $U \in V$, $\widetilde{U} \in V_2$, and $U^\sharp \in V$,*

$$\begin{aligned} \big|b\big(U, U^\sharp, \widetilde{U}\big)\big| \leq{}& c_2 |U|_{L^2}^{1/2} \|U\|^{1/2} \big\|U^\sharp\big\| \big|\widetilde{U}\big|_{L^2}^{1/2} \big\|\widetilde{U}\big\|^{1/2} \\ &+ c_2 \|U\| \big\|U^\sharp\big\|^{1/2} \big|U^\sharp\big|_{V_2}^{1/2} \big|\widetilde{U}\big|_{L^2}^{1/2} \big\|\widetilde{U}\big\|^{1/2}. \end{aligned} \tag{3.220}$$

PROOF. We only estimate two typical terms, and the other terms are estimated exactly in the same way. Using the Hölder, Sobolev, and interpolation inequalities, we write

$$\begin{aligned} \left| \int_{\mathcal{M}} u \frac{\partial u^\sharp}{\partial x} \tilde{u} \, \mathrm{d}\mathcal{M} \right| &\leq |u|_{L^4} \left| \frac{\partial u^\sharp}{\partial x} \right|_{L^2} \big|\tilde{u}\big|_{L^4} \\ &\leq c_1' |u|_{L^2}^{1/2} \|u\|^{1/2} \left| \frac{\partial u^\sharp}{\partial x} \right|_{L^2} \big|\tilde{u}\big|_{L^2}^{1/2} \big\|\tilde{u}\big\|^{1/2}, \\ \left| \int_{\mathcal{M}} w(U) \frac{\partial u^\sharp}{\partial z} \tilde{u} \, \mathrm{d}\mathcal{M} \right| &\leq \big|w(U)\big|_{L^2} \left| \frac{\partial u^\sharp}{\partial z} \right|_{L^4} \big|\tilde{u}\big|_{L^4} \\ &\leq c_2' \|u\| \left| \frac{\partial u^\sharp}{\partial z} \right|_{L^2}^{1/2} \left\| \frac{\partial u^\sharp}{\partial z} \right\|^{1/2} \big|\tilde{u}\big|^{1/2} \big\|\tilde{u}\big\|^{1/2}; \end{aligned}$$

Eq. (3.220) follows from these estimates and the analogous estimates for the other terms. □

We now recall the result regarding the existence of weak solutions for the PEs of the ocean; the proof is exactly the same as that of Theorem 2.2 in space dimension 3 (see also Theorem 3.1 for the 2D case with different boundary conditions).

THEOREM 3.5. *Given $U_0 \in H$ and $F \in L^\infty(\mathbb{R}_+; H)$, there exists at least one solution U of Eq. (3.217), $U \in L^\infty(\mathbb{R}_+; H) \cap L^2(0, t_\star; V)$ $\forall t_\star > 0$, with $U(0) = U_0$.*

As for Theorem 2.2, the proof of this theorem is based on the a priori estimates given below, which gives, as in LIONS, TEMAM and WANG [1992b], that $U \in L^\infty(0, t_\star; H)$, $\forall t_\star > 0$; however, as shown below, we have in fact[16]

$$U \in L^\infty(\mathbb{R}_+; H).$$

Taking $\widetilde{U} = U$ in Eq. (3.217), after some simple computations and using Eq. (3.218), we obtain, assuming enough regularity,

$$\frac{\mathrm{d}}{\mathrm{d}t}|U|_H^2 + c_0 c_1 |U|_H^2 \leq \frac{\mathrm{d}}{\mathrm{d}t}|U|_H^2 + c_1 \|U\|^2 \leq c_1' |F|_\infty^2, \tag{3.221}$$

where $|F|_\infty$ is the norm of F in $L^\infty(\mathbb{R}_+; H)$. Using the Gronwall inequality, we infer from Eq. (3.221) that

$$\big|U(t)\big|_H^2 \leq \big|U(0)\big|_H^2 \mathrm{e}^{-c_1 c_0 t} + \frac{c_1'}{c_1 c_0}\big(1 - \mathrm{e}^{-c_1 c_0 t}\big)|F|_\infty^2 \quad \forall t > 0. \tag{3.222}$$

Hence,

$$\limsup_{t \to \infty} \big|U(t)\big|_H^2 \leq \frac{c_1'}{c_1 c_0}|F|_\infty^2 =: r_0^2,$$

and any ball $B(0, r_0')$ in H with $r_0' > r_0$ is an absorbing ball, i.e., for all U_0, there exists $t_0 = t_0(|U_0|_H)$ depending increasingly on $|U_0|_H$ (and depending also on r_0', $|F|_\infty$ and other data) such that $|U(t)|_H \leq r_0'$ $\forall t \geq t_0(|U_0|_H)$. Furthermore, integrating Eq. (3.221) from t to $t + r$, with $r > 0$ arbitrarily chosen, we find

$$\int_t^{t+r} \big\|U(t')\big\|^2 \,\mathrm{d}t' \leq K_1 \quad \text{for all } t \geq t_0(|U_0|_H), \tag{3.223}$$

where K_1 denotes a constant depending on the data but not on U_0. As mentioned before, Eq. (3.222) implies also that

$$U \in L^\infty(\mathbb{R}_+; H), \quad \big|U(t)\big|_H \leq \max\big(|U_0|_H, r_0\big).$$

REMARK 3.3. We notice that in the inviscid case ($\nu_{\mathbf{v}} = \nu_\rho = 0$ with $F = 0$), taking $\widetilde{U} = U$ in Eq. (3.217), we find, at least formally,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(|u|_{L^2}^2 + |v|_{L^2}^2 + \frac{1}{N^2}|\rho|_{L^2}^2\right) = 0. \tag{3.224}$$

The physical meaning of Eq. (3.224) is that the sum of the kinetic energy (given by $\frac{1}{2}(|u|_{L^2}^2 + |v|_{L^2}^2)$) and the available potential energy (given by $\frac{1}{2N^2}|\rho|_{L^2}^2$) is conserved in time. This is the physical justification of the introduction of the constant $\kappa = N^{-2}$ in Eq. (3.171).

[16] The same holds in the previous cases in dimensions 3 and 2, although the result was not stated in this form. At all orders, we present here results uniform in time, $t \in \mathbb{R}_+$.

3.5.2. Existence and uniqueness of strong solutions for the PEs

The solutions given by Theorem 3.5 are called weak solutions as usual. We are now interested in strong solutions (and even more regular solutions in Section 3.5.3). We use here the same terminology as before: weak solutions are those in $L^\infty(L^2) \cap L^2(H^1)$ and strong solutions are those in $L^\infty(H^1) \cap L^2(H^2)$. We notice that as for Theorem 3.1, we cannot obtain directly the global existence of strong solutions for the PEs from a single a priori estimate. Instead, we will proceed as for Theorem 3.1 and derive the necessary a priori estimates by steps: we successively derive estimates in $L^\infty(L^2)$ and $L^2(H^1)$ for u_z, u_x, v_z, v_x, ρ_z, and ρ_x (here the subscripts t, x, and z denote differentiation). Notice that the order in which we obtain these estimates cannot be changed in the calculations below.[17]

Firstly, using Eq. (3.177), we rewrite Eq. (3.212) as

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + w\frac{\partial u}{\partial z} - \frac{1}{R_0}\frac{\partial p_s}{\partial x} + \frac{1}{R_0}\int_0^z \rho_x\big(x, z', t\big)\,\mathrm{d}z' = \nu_{\mathbf{v}}\Delta u + F_u. \tag{3.225}$$

We differentiate Eq. (3.225) with respect to z, and we find, with $w_z = -u_x$,

$$u_{tz} + uu_{xz} + wu_{zz} - \frac{1}{R_0}v_z - \frac{1}{R_0}\rho_x - \nu_{\mathbf{v}}u_{xxz} - \nu_{\mathbf{v}}u_{zzz} = F_{u,z},$$

where $F_{u,z} = \partial_z F_u = \partial F_u/\partial z$. After multiplying this equation by u_z and integrating over $\mathcal{M}$, we find

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|u_z|_{L^2}^2 + \nu_{\mathbf{v}}\|u_z\|^2 + \int_{\mathcal{M}} uu_z u_{xz}\,\mathrm{d}\mathcal{M} + \int_{\mathcal{M}} wu_z u_{zz}\,\mathrm{d}\mathcal{M}$$

$$- \frac{1}{R_0}\int_{\mathcal{M}} v_z u_z\,\mathrm{d}\mathcal{M} - \frac{1}{R_0}\int_{\mathcal{M}} \rho_x u_z\,\mathrm{d}\mathcal{M} = \int_{\mathcal{M}} u_z F_{u,z}\,\mathrm{d}\mathcal{M}.$$

Integrating by parts and taking into account the periodicity and the conservation of mass equation (Eq. (3.215)), we obtain:

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|u_z|_{L^2}^2 + \nu_{\mathbf{v}}\|u_z\|^2 - \frac{1}{R_0}\int_{\mathcal{M}} v_z u_z\,\mathrm{d}\mathcal{M} - \frac{1}{R_0}\int_{\mathcal{M}} \rho_x u_z\,\mathrm{d}\mathcal{M}$$

$$= \int_{\mathcal{M}} u_z F_{u,z}\,\mathrm{d}\mathcal{M}. \tag{3.226}$$

[17] However, as for Theorem 3.3, we could, at once, obtain the estimates for v_x and v_z and then for ρ_x and ρ_z.

In all that follows, K, K', K'', ..., denote constants depending on the data but not on U_0; we use the same symbol for different constants. We easily obtain the following estimates:

$$\frac{1}{R_0}\left|\int_{\mathcal{M}} v_z u_z \,\mathrm{d}\mathcal{M}\right| = \frac{1}{R_0}\left|\int_{\mathcal{M}} v u_{zz} \,\mathrm{d}\mathcal{M}\right| \leq K|v|_{L^2}^2 + \frac{\nu_{\mathbf{v}}}{6}\|u_z\|^2,$$

$$\frac{1}{R_0}\left|\int_{\mathcal{M}} \rho_x u_z \,\mathrm{d}\mathcal{M}\right| = \frac{1}{R_0}\left|\int_{\mathcal{M}} \rho u_{xz} \,\mathrm{d}\mathcal{M}\right| \leq \frac{\nu_{\mathbf{v}}}{6}\|u_z\|^2 + K|\rho|_{L^2}^2,$$

$$\left|\int_{\mathcal{M}} F_{u,z} u_z \,\mathrm{d}\mathcal{M}\right| = \left|\int_{\mathcal{M}} F_u u_{zz} \,\mathrm{d}\mathcal{M}\right| \leq \frac{\nu_{\mathbf{v}}}{6}\|u_z\|^2 + c_1'|F_u|_{L^2}^2.$$

Applied to Eq. (3.226), these give

$$\frac{\mathrm{d}}{\mathrm{d}t}|u_z|_{L^2}^2 + \nu_{\mathbf{v}}\|u_z\|^2 \leq K\big(|v|_{L^2}^2 + |\rho|_{L^2}^2\big) + c_1'|F_u|_{L^2}^2. \tag{3.227}$$

We apply the Poincaré inequality (Eq. (3.174)), and we find

$$\frac{\mathrm{d}}{\mathrm{d}t}|u_z|_{L^2}^2 + c_0\nu_{\mathbf{v}}|u_z|_{L^2}^2 \leq K\big(|v|_{L^2}^2 + |\rho|_{L^2}^2\big) + c_1'|F_u|_{L^2}^2. \tag{3.228}$$

Using the Gronwall lemma, we infer from Eq. (3.228) that

$$\begin{aligned}
\big|u_z(t)\big|_{L^2}^2 &\leq \big|u_z(0)\big|_{L^2}^2 \mathrm{e}^{-c_0\nu_{\mathbf{v}}t} + K\mathrm{e}^{-c_0\nu_{\mathbf{v}}t}\int_0^t \big(\big|v(t')\big|_{L^2}^2 + \big|\rho(t')\big|_{L^2}^2\big)\mathrm{e}^{c_0\nu_{\mathbf{v}}t'}\,\mathrm{d}t' \\
&\qquad + c_2'|F_u|_\infty^2 \\
&\leq \big|u_z(0)\big|_{L^2}^2 \mathrm{e}^{-c_0\nu_{\mathbf{v}}t} + K'\big(1-\mathrm{e}^{-c_0\nu_{\mathbf{v}}t}\big)\big(|v|_\infty^2 + |\rho|_\infty^2\big) + c_2'|F_u|_\infty^2 \\
&\leq \big|u_z(0)\big|_{L^2}^2 \mathrm{e}^{-c_0\nu_{\mathbf{v}}t} + K'\big(|v|_\infty^2 + |\rho|_\infty^2\big) + c_2'|F_u|_\infty^2,
\end{aligned} \tag{3.229}$$

where $|v|_\infty = |v|_{L^\infty(\mathbb{R}_+;L^2(\mathcal{M}))}$, and similarly for ρ and F_u. We obtain an explicit bound for the norm of u_z in $L^\infty(\mathbb{R}_+; H)$,

$$\big|u_z(t)\big|_{L^2}^2 \leq \big|u_z(0)\big|_{L^2}^2 + K'\big(|v|_\infty^2 + |\rho|_\infty^2\big) + c_2'|F_u|_\infty^2. \tag{3.230}$$

For what follows, we recall here the uniform Gronwall lemma (see, e.g., Temam [1997]):

If ξ, η, and y are three positive locally integrable functions on (t_1, ∞) such that y' is locally integrable on (t_1, ∞) and which satisfy

$$\begin{aligned}
&y' \leq \xi y + \eta, \\
&\int_t^{t+r} \xi(s)\,\mathrm{d}s \leq a_1, \qquad \int_t^{t+r} \eta(s)\,\mathrm{d}s \leq a_2, \\
&\int_t^{t+r} y(s)\,\mathrm{d}s \leq a_3 \quad \forall t \geq t_1,
\end{aligned} \tag{3.231}$$

where r, a_1, a_2, and a_3 are positive constants, then

$$y(t+r) \leq \left(\frac{a_3}{r} + a_2\right)\mathrm{e}^{a_1}, \quad t \geq t_1. \tag{3.232}$$

The bound (Eq. (3.230)) depends on the initial data U_0. In order to obtain a bound independent of U_0, we apply the uniform Gronwall lemma to the equation,

$$\frac{\mathrm{d}}{\mathrm{d}t}|u_z|_{L^2}^2 \leq K\big(|v|_{L^2}^2 + |\rho|_{L^2}^2\big) + c_1'|F_u|_{L^2}^2, \tag{3.233}$$

to obtain

$$\big|u_z(t)\big| \leq K'\big(r, r_0'\big) \quad \forall t \geq t_1', \tag{3.234}$$

where $t_1' = t_0(|U_0|_{L^2}) + r$ and $r > 0$ is fixed. Integrating Eq. (3.227) from t to $t + r$ with $r > 0$ as before, we also find

$$\int_t^{t+r} \big\|u_z(s)\big\|^2 \,\mathrm{d}s \leq K''\big(r, r_0'\big) \quad \forall t \geq t_1'. \tag{3.235}$$

We now derive the same kind of estimates for u_x: we differentiate Eq. (3.225) with respect to x, and we obtain

$$\begin{aligned} &u_{tx} + u_x^2 + uu_{xx} + wu_{xz} + w_x u_z \\ &\quad - \frac{1}{R_0}v_x + \frac{1}{R_0}p_{s,xx} + \int_z^0 \rho_{xx}\big(z'\big)\,\mathrm{d}z' - \nu_{\mathbf{v}}u_{xxx} - \nu_{\mathbf{v}}u_{zzx} = F_{u,x}. \end{aligned} \tag{3.236}$$

Multiplying this equation by u_x and integrating over $\mathcal{M}$, we find, using Eq. (3.215)

$$\begin{aligned} &\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|u_x|_{L^2}^2 + \int_{\mathcal{M}} u_x^3 \,\mathrm{d}\mathcal{M} + \int_{\mathcal{M}} w_x u_z u_x \,\mathrm{d}\mathcal{M} \\ &\quad - \frac{1}{R_0}\int_{\mathcal{M}} v_x u_x \,\mathrm{d}\mathcal{M} - \frac{1}{R_0}\int_{\mathcal{M}} p_{s,xx} u_x \,\mathrm{d}\mathcal{M} \\ &\quad + \int_{\mathcal{M}} \left(\int_z^0 \rho_{xx}\big(z'\big)\,\mathrm{d}z'\right) u_x \,\mathrm{d}\mathcal{M} + \nu_{\mathbf{v}}\|u_x\|^2 = \int_{\mathcal{M}} u_x F_{u,x} \,\mathrm{d}\mathcal{M}. \end{aligned} \tag{3.237}$$

Based on the Hölder, Sobolev, and interpolation inequalities, we derive the following estimates:

$$\begin{aligned}
\left|\int_{\mathcal{M}} u_x^3 \,\mathrm{d}\mathcal{M}\right| &\le |u_x|^3_{L^3(\mathcal{M})} \le c_4' |u_x|^3_{H^{1/3}(\mathcal{M})} \\
&\le c_5' |u_x|^2_{L^2} \|u_x\| \le \frac{\nu_{\mathbf{v}}}{12}\|u_x\|^2 + c_6' |u_x|^4_{L^2}, \\
\left|\int_{\mathcal{M}} w_x u_z u_x \,\mathrm{d}\mathcal{M}\right| &\le c_7' |w_x|_{L^2} |u_z|^{1/2}_{L^2} \|u_z\|^{1/2} |u_x|^{1/2}_{L^2} \|u_x\|^{1/2} \\
&\le c_8' |u_{xx}|_{L^2} |u_z|^{1/2}_{L^2} \|u_z\|^{1/2} |u_x|^{1/2}_{L^2} \|u_x\|^{1/2} \\
&\le \frac{\nu_{\mathbf{v}}}{12}\|u_x\|^2 + c_9' |u_z|^2_{L^2} \|u_z\|^2 |u_x|^2_{L^2}.
\end{aligned}$$

By the definition of V, and since p_{s} is independent of z, we find

$$\frac{1}{R_0}\left|\int_{\mathcal{M}} p_{\mathrm{s},xx} u_x \,\mathrm{d}\mathcal{M}\right| = \frac{1}{R_0}\left|\int_0^L p_{\mathrm{s},xx} \int_{-L_3/2}^{L_3/2} u_x \,\mathrm{d}z \,\mathrm{d}x\right| = 0.$$

We can also prove the following estimates:

$$\begin{aligned}
&\frac{1}{R_0}\left|\int_{\mathcal{M}} v_x u_x \,\mathrm{d}\mathcal{M}\right| \le \frac{\nu_{\mathbf{v}}}{12}\|u_x\|^2 + K' |v|^2_{L^2}, \\
&\frac{1}{R_0}\left|\int_{\mathcal{M}} \left(\int_z^0 \rho_{xx}(z') \,\mathrm{d}z'\right) u_x \,\mathrm{d}\mathcal{M}\right| = \frac{1}{R_0}\left|\int_{\mathcal{M}} \left(\int_z^0 \rho_x(z') \,\mathrm{d}z'\right) u_{xx} \,\mathrm{d}\mathcal{M}\right| \\
&\qquad\qquad \le \frac{\nu_{\mathbf{v}}}{12}\|u_x\|^2 + K'' |\rho_x|^2_{L^2}, \\
&\left|\int_{\mathcal{M}} u_x F_{u,x} \,\mathrm{d}\mathcal{M}\right| \le \frac{\nu_{\mathbf{v}}}{12}\|u_x\|^2 + c_{10}' |F_u|^2_{\infty}.
\end{aligned}$$

With these relations, Eq. (3.237) implies

$$\frac{\mathrm{d}}{\mathrm{d}t}|u_x|^2_{L^2} + \nu_{\mathbf{v}}\|u_x\|^2 \le \xi |u_x|^2_{L^2} + \eta, \tag{3.238}$$

where we denoted

$$\xi = \xi(t) = 2c_6' |u_x|^2_{L^2} + 2c_9' |u_z|^2_{L^2} \|u_z\|^2,$$

and

$$\eta = \eta(t) = 2K' |v|^2_{L^2} + 2K'' |\rho_x|^2_{L^2} + 2c_{10}' |F_u|^2_{\infty}.$$

We easily conclude from Eq. (3.238) that

$$u_x \in L^\infty\big(0, t_\star; L^2\big) \cap L^2\big(0, t_\star; H^1\big) \quad \forall t_\star > 0. \tag{3.239}$$

However, for later purposes, Eq. (3.239) is not sufficient, and we need estimates uniform in time.

We will apply the uniform Gronwall lemma to Eq. (3.238) with $t_1 = t_1'$ as in Eq. (3.234). Noting that

$$\begin{aligned}
\int_t^{t+r} \xi\big(t'\big)\,\mathrm{d}t' &= \int_t^{t+r} \big[2c_6'|u_x|_{L^2}^2 + 2c_9'\big|u_z\big(t'\big)\big|_{L^2}^2 \big\|u_z\big(t'\big)\big\|^2\big]\,\mathrm{d}t' \\
&\le 2c_6' \int_t^{t+r} \big|u_x\big(t'\big)\big|_{L^2}^2\,\mathrm{d}t' + 2c_9'|u_z|_\infty^2 \int_t^{t+r} \big\|u_z\big(t'\big)\big\|^2\,\mathrm{d}t' \\
&\le a_1 \qquad \forall t \ge t_1',
\end{aligned} \tag{3.240}$$

$$\begin{aligned}
\int_t^{t+r} \eta\big(t'\big)\,\mathrm{d}t' &= \int_t^{t+r} \big[2K'|v|_{L^2}^2 + 2K''|\rho_x|_{L^2}^2 + 2c_{10}'|F_u|_\infty^2\big]\,\mathrm{d}t' \\
&\le K + 2c_{10}' r |F_u|_\infty^2 \\
&= a_2 \qquad \forall t \ge t_1',
\end{aligned} \tag{3.241}$$

$$\int_t^{t+r} \big|u_x\big(t'\big)\big|_{L^2}^2\,\mathrm{d}t' \le a_3 \quad \forall t \ge t_1', \tag{3.242}$$

Eq. (3.232) then yields

$$\big|u_x(t)\big|_{L^2}^2 \le \left(\frac{a_3}{r} + a_2\right)\mathrm{e}^{a_1} \quad \forall t \ge t_1' + r, \tag{3.243}$$

and thus

$$|u_x|_{L^2} \in L^\infty(\mathbb{R}_+). \tag{3.244}$$

Note that in Eqs. (3.240)–(3.242), we can use bounds on $|u_z|_\infty$ (and other similar terms) independent of U_0 since $t \ge t_0(|U_0|_{L^2}) + r$. Integrating Eq. (3.238) from 0 to $t_1' + r$ where $t_1' = t_1'(|U_0|_{L^2})$, we obtain a bound for u_x in $L^2(0, t_1' + r; H^1)$, which depends on $\|U_0\|$. A bound independent of U_0 is obtained if we work with $t \ge t_1' + r = t_1'' = t_1''(|U_0|_{L^2})$: integrating Eq. (3.238) from t to $t + r$ with r as before, we find

$$\int_t^{t+r} \big\|u_x(s)\big\|^2\,\mathrm{d}s \le K \quad \forall t \ge t_1''. \tag{3.245}$$

We perform similar computations for v_z: we differentiate Eq. (3.213) with respect to z, multiply the resulting equation by v_z, and integrate over $\mathcal{M}$. Using again the conservation of mass relation, we arrive at

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|v_z|_{L^2}^2 + \int_{\mathcal{M}} u_z v_x v_z \,\mathrm{d}\mathcal{M} + \int_{\mathcal{M}} w_z v_z^2 \,\mathrm{d}\mathcal{M} + \frac{1}{R_0}\int_{\mathcal{M}} u_z v_z \,\mathrm{d}\mathcal{M} + \nu_{\mathbf{v}}\|v_z\|^2 = \int_{\mathcal{M}} v_z F_{u,z} \,\mathrm{d}\mathcal{M}. \tag{3.246}$$

We notice the following estimate

$$\begin{aligned}\left|\int_{\mathcal{M}} u_z v_x v_z \,\mathrm{d}\mathcal{M}\right| &\le c'_{11}|u_z|_{L^2}^{1/2}\|u_z\|^{1/2}|v_x|_{L^2}|v_z|_{L^2}^{1/2}\|v_z\|^{1/2}\\ &\le \frac{\nu_{\mathbf{v}}}{8}\|v_z\|^2 + c'_{12}|u_z|_{L^2}^{2/3}\|u_z\|^{2/3}|v_x|_{L^2}^{4/3}|v_z|_{L^2}^{2/3}\\ &\le \frac{\nu_{\mathbf{v}}}{8}\|v_z\|^2 + c'_{12}|u_z|_{L^2}^{2/3}\|u_z\|^{2/3}|v_x|_{L^2}^{4/3}\big(1+|v_z|_{L^2}^2\big).\end{aligned}$$

We also see that

$$\begin{aligned}\left|\int_{\mathcal{M}} w_z v_z v_z \,\mathrm{d}\mathcal{M}\right| &= \left|\int_{\mathcal{M}} u_x v_z v_z \,\mathrm{d}\mathcal{M}\right|\\ &\le c'_{13}|u_x|_{L^2}^{1/2}\|u_x\|^{1/2}|v_z|_{L^2}^{3/2}\|v_z\|^{1/2}\\ &\le \frac{\nu_{\mathbf{v}}}{8}\|v_z\|^2 + c'_{14}|u_x|_{L^2}^{2/3}\|u_x\|^{2/3}|v_z|_{L^2}^2,\\ \frac{1}{R_0}\left|\int_{\mathcal{M}} u_z v_z \,\mathrm{d}\mathcal{M}\right| &= \frac{1}{R_0}\left|\int_{\mathcal{M}} u v_{zz} \,\mathrm{d}\mathcal{M}\right| \le \frac{\nu_{\mathbf{v}}}{8}\|v_z\|^2 + K|u|_{L^2}^2,\\ \left|\int_{\mathcal{M}} F_{v,z} v_z \,\mathrm{d}\mathcal{M}\right| &= \left|\int_{\mathcal{M}} F_v v_{zz} \,\mathrm{d}\mathcal{M}\right| \le \frac{\nu_{\mathbf{v}}}{8}\|v_z\|^2 + c'_{15}|F_v|_\infty^2,\end{aligned}$$

which gives

$$\frac{\mathrm{d}}{\mathrm{d}t}|v_z|_{L^2}^2 + \nu_{\mathbf{v}}\|v_z\|^2 \le \xi|v_z|^2 + \eta, \tag{3.247}$$

where we denoted

$$\eta = \eta(t) = 2c'_{12}|u_z|_{L^2}^{2/3}\|u_z\|^{2/3}|v_x|_{L^2}^{4/3} + 2K|u|^2 + 2c'_{15}|F_v|_\infty^2$$

and

$$\xi = \xi(t) = 2c'_{12}|u_z|_{L^2}^{2/3}\|u_z\|^{2/3}|v_x|_{L^2}^{4/3} + 2c'_{14}|u_x|_{L^2}^{2/3}\|u_x\|^{2/3}.$$

From Eq. (3.247), using the estimates obtained before and applying the classical Gronwall lemma, we obtain bounds depending on the initial data for v_z in $L^\infty_{\mathrm{loc}}(0, t_\star; L^2)$ and $L^2_{\mathrm{loc}}(0, t_\star; H^1)$, valid for any finite interval of time $(0, t_\star)$.

To obtain estimates valid for all time, we apply the uniform Gronwall lemma observing that

$$\begin{aligned}
&\int_t^{t+r} \eta(t')\,\mathrm{d}t' \\
&\quad \le 2c'_{12}|u_z|_\infty^{2/3}\left(\int_t^{t+r}\left\|u_z(t')\right\|\mathrm{d}t'\right)^{1/3}\left(\int_t^{t+r}\left|v_x(t')\right|_{L^2}^2\mathrm{d}t'\right)^{2/3} \\
&\qquad + 2K|u|_\infty^2 r + 2c'_{15}r|F_v|_\infty^2 \\
&\quad \le a_1 \quad \forall t \ge t''_1, \qquad (3.248)\\
&\int_t^{t+r} \xi(t')\,\mathrm{d}t' \\
&\quad \le 2c'_{12}|u_z|_\infty^{2/3}\left(\int_t^{t+r}\left\|u_z(t')\right\|\mathrm{d}t'\right)^{1/3}\left(\int_t^{t+r}\left|v_x(t')\right|_{L^2}^2\mathrm{d}t'\right)^{2/3} \\
&\qquad + 2c'_{14}|u_x|_\infty^{2/3}\int_t^{t+r}\left\|u_x(t')\right\|^{2/3}\mathrm{d}t' \\
&\quad \le a_2 \quad \forall t \ge t''_1, \qquad (3.249)\\
&\int_t^{t+r}\left|v_z(t')\right|^2\mathrm{d}t' \le a_3 \quad \forall t \ge t''_1. \qquad (3.250)
\end{aligned}$$

Then the uniform Gronwall lemma gives

$$\left|v_z(t)\right|_{L^2}^2 \le \left(\frac{a_3}{r} + a_2\right)\mathrm{e}^{a_1} \quad \forall t \ge t''_1 + r, \qquad (3.251)$$

with a_1, a_2, and a_3 as in Eqs. (3.248)–(3.250). Integrating Eq. (3.247) from t to $t+r$ with $r > 0$ as above and $t \ge t''_1 + r$, we find

$$\int_t^{t+r}\left\|v_z(s)\right\|^2\mathrm{d}s \le K \quad \forall t \ge t''_1 + r. \qquad (3.252)$$

The same methods apply to v_x, ρ_z, and ρ_x, noticing that at each step we precisely use the estimates from the previous steps, so the order cannot be changed in these calculations.

With these estimates, the Galerkin method as used for the proof of Theorem 2.2 gives the existence of strong solutions.

THEOREM 3.6. *Given $U_0 \in V$ and $F \in L^\infty(\mathbb{R}_+; H)$, there exists a unique solution U of Eq. (3.217) with $U(0) = U_0$ such that*

$$U \in L^\infty(\mathbb{R}_+; V) \cap L^2\left(0, t_\star; \left(\dot{H}^2(\mathcal{M})\right)^3\right) \quad \forall t_\star > 0. \qquad (3.253)$$

PROOF. As we said, the existence of strong solutions follows from the previous estimates. The proof for the uniqueness follows the same idea as of the Theorem 3.1, so we skip it. □

3.5.3. More regular solutions for the PEs

In this subsection, we show how to obtain estimates on the higher order derivatives from which one can derive the existence of solutions of the PEs in $(\dot{H}^m(\mathcal{M}))^3$ for all $m \in \mathbb{N}$, $m \geq 2$ (hence up to $\mathcal{C}^\infty$-regularity). In all that follows, we work with U_0 in $(\dot{H}^m_{\mathrm{per}}(\mathcal{M}))^3$.

We set $|U|_m = (\sum_{[\alpha]=m} |D^\alpha U|^2_{L^2})^{1/2}$. We fix $m \geq 2$, and proceeding by induction, we assume that for all $0 \leq l \leq m-1$, we have shown that

$$U \in L^\infty\big(\mathbb{R}_+; \big(\dot{H}^l(\mathcal{M})\big)^3\big) \cap L^2\big(0, t_\star; \big(\dot{H}^{l+1}(\mathcal{M})\big)^3\big) \quad \forall t_\star > 0, \tag{3.254}$$

with

$$\int_t^{t+r} \big|U\big(t'\big)\big|^2_{l+1}\, \mathrm{d}t' \leq a_l \quad \forall t \geq t_l(U_0), \tag{3.255}$$

where a_l is a constant depending on the data (and l) but not on U_0, and $r > 0$ is fixed (the same as before). We then want to establish the same results for $l = m$.

In Eq. (3.217), we take $\widetilde{U} = \Delta^m U(t)$ with $m \geq 2$ and t arbitrarily fixed, and we obtain

$$\begin{aligned}&\left(\frac{\mathrm{d}U}{\mathrm{d}t}, \Delta^m U\right)_{L^2} + a\big(U, \Delta^m U\big) + b\big(U, U, \Delta^m U\big) + \frac{1}{R_0} e\big(U, \Delta^m U\big)\\ &\quad = \big(F, \Delta^m U\big)_{L^2}.\end{aligned} \tag{3.256}$$

Integrating by parts, using periodicity and the coercivity of a and the fact that $e(U, U) = 0$, we find

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\big|U(t)\big|^2_m + c_1|U|^2_{m+1} \leq \big|b\big(U, U, \Delta^m U\big)\big| + \big|\big(F, \Delta^m U\big)_{L^2}\big|. \tag{3.257}$$

We need to estimate the terms on the right-hand side of Eq. (3.257). We first notice that

$$\big|\big(F, \Delta^m U\big)_L\big| \leq c|F|^2_{m-1} + \frac{c_1}{2(m+3)}|U|^2_{m+1}, \tag{3.258}$$

and it remains to estimate $|b(U, U, \Delta^m U)|$.

By the definition of b, we have

$$\begin{aligned}b\big(U, U, \Delta^m U\big) = &\int_{\mathcal{M}} \big(uu_x + w(U)u_z\big)\Delta^m u\, \mathrm{d}\mathcal{M}\\ &+ \int_{\mathcal{M}} \big(uv_x + w(U)v_z\big)\Delta^m v\, \mathrm{d}\mathcal{M}\\ &+ \int_{\mathcal{M}} \big(u\rho_x + w(U)\rho_z\big)\Delta^m \rho\, \mathrm{d}\mathcal{M}.\end{aligned} \tag{3.259}$$

The computations are similar for all the terms, and for simplicity, we shall only estimate the first integral on the right-hand side of Eq. (3.259).

We notice that $b(U, U, \Delta^m U)$ is a sum of integrals of the type

$$\int_{\mathcal{M}} u \frac{\partial u}{\partial x} D_1^{2\alpha_1} D_3^{2\alpha_3} u \, \mathrm{d}\mathcal{M}, \qquad \int_{\mathcal{M}} w(U) \frac{\partial u}{\partial z} D_1^{2\alpha_1} D_3^{2\alpha_3} u \, \mathrm{d}\mathcal{M},$$

where $\alpha_i \in \mathbb{N}$ with $\alpha_1 + \alpha_3 = m$. By D_i we denoted the differential operator $\partial/\partial x_i$. Integrating by parts and using periodicity, the integrals take the form

$$\int_{\mathcal{M}} D^\alpha \left(u \frac{\partial u}{\partial x} \right) D^\alpha u \, \mathrm{d}\mathcal{M}, \qquad \int_{\mathcal{M}} D^\alpha \left(w(U) \frac{\partial u}{\partial z} \right) D^\alpha u \, \mathrm{d}\mathcal{M}, \tag{3.260}$$

where $D^\alpha = D_1^{\alpha_1} D_3^{\alpha_3}$. Using Leipzig formula, we see that the integrals are sums of integrals of the form

$$\int_{\mathcal{M}} u D^\alpha \frac{\partial u}{\partial x} D^\alpha u \, \mathrm{d}\mathcal{M}, \qquad \int_{\mathcal{M}} w(U) D^\alpha \frac{\partial u}{\partial z} D^\alpha u \, \mathrm{d}\mathcal{M}, \tag{3.261}$$

and of integrals of the form

$$\int_{\mathcal{M}} \delta^k u \delta^{m-k} \frac{\partial u}{\partial x} D^\alpha u \, \mathrm{d}\mathcal{M}, \qquad \int_{\mathcal{M}} \delta^k w(U) \delta^{m-k} \frac{\partial u}{\partial z} D^\alpha u \, \mathrm{d}\mathcal{M}, \tag{3.262}$$

with $k = 1, \ldots, m$, where δ^k is some differential operator D^α with $[\alpha] = \alpha_1 + \alpha_3 = k$. For each α, after integration by parts, we see that the sum of the two integrals in Eq. (3.261) is zero because of the mass conservation equation (Eq. (3.215)). It remains to estimate the integrals of type (Eq. (3.262)). We use here the Sobolev and interpolation inequalities. For the first term in Eq. (3.262), we write

$$\begin{aligned}
&\left| \int_{\mathcal{M}} \delta^k u \delta^{m-k} \frac{\partial u}{\partial x} D^\alpha u \, \mathrm{d}\mathcal{M} \right| \\
&\quad \le \left| \delta^k u \right|_{L^4} \left| \delta^{m-k} \frac{\partial u}{\partial x} \right|_{L^4} \left| D^\alpha u \right|_{L^2} \\
&\quad \le c_1' \left| \delta^k u \right|_{L^2}^{1/2} \left| \delta^k u \right|_{H^1}^{1/2} \left| \delta^{m-k} \frac{\partial u}{\partial x} \right|_{L^2}^{1/2} \left| \delta^{m-k} \frac{\partial u}{\partial x} \right|_{H^1}^{1/2} \left| D^\alpha u \right|_{L^2} \\
&\quad \le c_1' |U|_k^{1/2} |U|_{k+1}^{1/2} |U|_{m-k+1}^{1/2} |U|_{m-k+2}^{1/2} |U|_m,
\end{aligned} \tag{3.263}$$

where $k = 1, \ldots, m$.

The second term from Eq. (3.262) is estimated as follows:

$$\begin{aligned}
&\left| \int_{\mathcal{M}} \delta^k w(U) \delta^{m-k} \frac{\partial u}{\partial z} D^\alpha u \, \mathrm{d}\mathcal{M} \right| \\
&\quad \le \left| \delta^k w(U) \right|_{L^2} \left| \delta^{m-k} \frac{\partial u}{\partial z} \right|_{L^4} \left| D^\alpha u \right|_{L^4} \\
&\quad \le c_2' \left| \delta^k w(U) \right|_{L^2} \left| \delta^{m-k} \frac{\partial u}{\partial z} \right|_{L^2}^{1/2} \left| \delta^{m-k} \frac{\partial u}{\partial z} \right|_{H^1}^{1/2} \left| D^\alpha u \right|_{L^2}^{1/2} \left| D^\alpha u \right|_{H^1}^{1/2} \\
&\quad \le c_3' |U|_{k+1} |U|_{m-k+1}^{1/2} |U|_{m-k+2}^{1/2} |U|_m^{1/2} |U|_{m+1}^{1/2},
\end{aligned} \tag{3.264}$$

where $k = 1, \ldots, m$.

From Eqs. (3.263) and (3.264), we obtain that

$$\begin{aligned}
&\left|b\left(U, U, \Delta^m U\right)\right| \\
&\quad \leq c_3 \sum_{k=1}^{m} |U|_k^{1/2} |U|_{k+1}^{1/2} |U|_{m-k+1}^{1/2} |U|_{m-k+2}^{1/2} |U|_m \\
&\qquad + c_3 \sum_{k=1}^{m} |U|_{k+1} |U|_{m-k+1}^{1/2} |U|_{m-k+2}^{1/2} |U|_m^{1/2} |U|_{m+1}^{1/2}. \qquad (3.265)
\end{aligned}$$

We now need to bound the terms on the right-hand side of Eq. (3.265). The terms corresponding to $k = 2, \ldots, m-1$ in the first sum do not contain $|U|_{m+1}$, and we leave them as they are. For $k = 1$ and $k = m$, we apply Young's inequality and we obtain

$$\begin{aligned}
&c_3 |U|_1^{1/2} |U|_2^{1/2} |U|_m^{3/2} |U|_{m+1}^{1/2} \\
&\quad \leq \frac{c_1}{2(m+3)} |U|_{m+1}^2 + c_4' |U|_1^{2/3} |U|_2^{2/3} |U|_m^2. \qquad (3.266)
\end{aligned}$$

For the terms in the second sum in Eq. (3.265), we distinguish between $k = 1$, $k = m$, and $k = 2, \ldots, m-1$. The term corresponding to $k = 1$ is bounded by

$$c_3 |U|_2 |U|_m |U|_{m+1} \leq \frac{c_1}{2(m+3)} |U|_{m+1}^2 + c_5' |U|_2^2 |U|_m^2. \qquad (3.267)$$

For $k = m$, we find

$$c_3 |U|_1^{1/2} |U|_2^{1/2} |U|_m^{1/2} |U|_{m+1}^{3/2} \leq \frac{c_1}{2(m+3)} |U|_{m+1}^2 + c_6' |U|_1^2 |U|_2^2 |U|_m^2. \qquad (3.268)$$

For the terms corresponding to $k = 2, \ldots, m-1$, we apply Young's inequality in the following way:

$$\begin{aligned}
&c_3 |U|_{k+1} |U|_{m-k+1}^{1/2} |U|_{m-k+2}^{1/2} |U|_m^{1/2} |U|_{m+1}^{1/2} \\
&\quad \leq \frac{c_1}{2(m+3)} |U|_{m+1}^2 + c_7' |U|_{k+1}^{4/3} |U|_{m-k+1}^{2/3} |U|_{m-k+2}^{2/3} |U|_m^{2/3}. \qquad (3.269)
\end{aligned}$$

Gathering all the estimates above, we find

$$\frac{\mathrm{d}}{\mathrm{d}t} |U|_m^2 + c_1 |U|_{m+1}^2 \leq \xi + \eta |U|_m^2,$$

where the expressions of ξ and η are easily derived from Eqs. (3.257) and (3.266)–(3.269). Using the Gronwall lemma and the induction hypotheses (Eqs. (3.254) and (3.255)), we obtain a bound for U in $L^\infty(0, t_\star; H^m)$ and $L^2(0, t_\star; H^{m+1})$, and for all fixed $t_\star > 0$, this bound depending also on $|U_0|_m$. We also see that because of the induction hypotheses (Eqs. (3.254) and (3.255)), we can apply the uniform Gronwall lemma, and we obtain U bounded in $L^\infty(\mathbb{R}_+; H^m)$ with a bound independent of $|U_0|_m$ when $t \geq t_m(U_0)$; we also obtain an analog of Eq. (3.255). The details regarding the way we apply the uniform Gronwall lemma and derive these bounds are similar to the developments in Section 3.5.2.

In summary, we have proven the following result.

THEOREM 3.7. *Given* $m \in \mathbb{N}$, $m \geq 1$, $U_0 \in V \cap (\dot{H}^m_{\text{per}}(\mathcal{M}))^3$, *and* $F \in L^\infty(\mathbb{R}_+; H \cap (\dot{H}^{m-1}_{\text{per}}(\mathcal{M}))^3)$, Eq. (3.217) *has a unique solution* U *such that*

$$U \in L^\infty\big(\mathbb{R}_+; \big(\dot{H}^m_{\text{per}}(\mathcal{M})\big)^3\big) \cap L^2\big(0, t_\star; \big(\dot{H}^{m+1}_{\text{per}}(\mathcal{M})\big)^3\big) \quad \forall t_\star > 0. \tag{3.270}$$

REMARK 3.4. Since $\bigcap_{m\geq 0} \dot{H}^m_{\text{per}}(\mathcal{M}) = \dot{\mathcal{C}}^\infty_{\text{per}}(\mathcal{M})$, given $U_0 \in (\dot{\mathcal{C}}^\infty_{\text{per}}(\mathcal{M}))^3$ and $F \in L^\infty(\mathbb{R}_+; (\dot{\mathcal{C}}^\infty_{\text{per}}(\mathcal{M}))^3)$, Eq. (3.217) has a unique solution U belonging to $L^\infty(\mathbb{R}_+; (\dot{H}^m_{\text{per}}(\mathcal{M}))^3)$ for all $m \in \mathbb{N}$, i.e., U is in $L^\infty(\mathbb{R}_+; (\dot{\mathcal{C}}^\infty_{\text{per}}(\mathcal{M}))^3)$. Regularity (differentiability) in time can be also derived if F is also $\mathcal{C}^\infty$ in time. However, the arguments above do not provide the existence of an absorbing set in $(\dot{\mathcal{C}}^\infty_{\text{per}}(\mathcal{M}))^3$.

3.6. *The space periodic case in dimension 3: higher Sobolev regularity and Gevrey regularity*

In this section, we consider the PEs of the ocean, in a 3D domain, with periodic boundary conditions. PETCU and WIROSOETISNO [2005] proved the short-time existence and uniqueness of the strong solutions in space dimension 3 and also the local existence of very regular solutions, up to the $\mathcal{C}^\infty$–regularity. They also studied the Gevrey-type regularity for the solutions of the PEs, the result being obtained on a limited interval of time. In the present section, we extend these results, proving that they remain true for all time. Of course, this section is also related to Section 3.5 in which similar results are proved in space dimension 2, but the methods needed for dimension 3 are different.

This section is structured in two parts: in the first part, using the results of CAO and TITI [2007] and KOBELKOV [2006] and improvements due to JU [2007], we prove the long-time existence of the solutions with values in the Sobolev spaces H^m, for all $m \geq 0$, and derive from this the $\mathcal{C}^\infty$–regularity of the solution. In the second part, we study the Gevrey regularity for the solutions (analyticity in space).

We start by recalling the model and the functional settings of the problem. The PEs in their dimensional form read

$$\begin{aligned}
&\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x_1} + v\frac{\partial u}{\partial x_2} + w\frac{\partial u}{\partial x_3} - fv + \frac{1}{\rho_0}\frac{\partial p}{\partial x_1} = \nu\Delta_3 u + F_u,\\
&\frac{\partial v}{\partial t} + u\frac{\partial v}{\partial x_1} + v\frac{\partial v}{\partial x_2} + w\frac{\partial v}{\partial x_3} + fu + \frac{1}{\rho_0}\frac{\partial p}{\partial x_2} = \nu\Delta_3 v + F_v,\\
&\frac{\partial p}{\partial x_3} = -\rho g,\\
&\frac{\partial u}{\partial x_1} + \frac{\partial v}{\partial x_2} + \frac{\partial w}{\partial x_3} = 0,\\
&\frac{\partial T}{\partial t} + u\frac{\partial T}{\partial x_1} + v\frac{\partial T}{\partial x_2} + w\frac{\partial T}{\partial x_3} = \mu\Delta_3 T + F_T.
\end{aligned} \tag{3.271}$$

In this model, $(u,\ v,\ w)$ are the three components of the velocity vector, and p, ρ, and T are, respectively, the perturbations of the pressure, density and temperature from the reference (average) constant states p_0, ρ_0, and T_r. The relation between the full temperature and the full density is given by the equation of state, and we consider here a version of this equation linearized around the reference state ρ_0 and T_r,

$$\rho_{\text{full}} = \rho_0(1 - \beta_T(T_{\text{full}} - T_r)), \tag{3.272}$$

so that for the perturbations ρ and T

$$\rho = -\beta_T \rho_0 T. \tag{3.273}$$

The constant g is the gravitational acceleration, and f is the Coriolis parameter; ν and μ are the eddy diffusivity coefficients, $(F_u,\ F_v)$ represent body forces per unit of mass, and F_T represents a heating source.

We work in a bounded domain:

$$\mathcal{M} = (0, L_1) \times (0, L_2) \times (-L_3/2, L_3/2), \tag{3.274}$$

and we assume space periodicity with period $\mathcal{M}$. We also assume that the functions have the same symmetries as in Eq. (3.163).

The variational formulation of the problem

We start by introducing the natural function spaces for this problem:

$$V = \{U = (u,\ v,\ T) \in (\dot{H}^1_{\text{per}}(\mathcal{M}))^3,\ u,\ v \text{ even in } x_3,\ T \text{ odd in } x_3, \tag{3.275}$$

$$\int_{-L_3/2}^{L_3/2} (u_{x_1}(x_1, x_2, x_3') + v_{x_2}(x_1, x_2, x_3'))\, \mathrm{d}x_3' = 0\},$$

$$H = \text{closure of } V \text{ in } (\dot{L}^2(\mathcal{M}))^3,$$

$$V_2 = \text{ the closure of } V \cap (\dot{H}^2_{\text{per}}(\mathcal{M}))^3 \text{ in } (\dot{H}^2_{\text{per}}(\mathcal{M}))^3. \tag{3.276}$$

As before, we endow these spaces with the following scalar products:
On H, we consider

$$(U,\ \tilde{U})_H = (u,\ \tilde{u})_{L^2} + (v,\ \tilde{v})_{L^2} + \kappa(T,\ \tilde{T})_{L^2}, \tag{3.277}$$

and on V,

$$((U,\ \tilde{U}))_V = ((u,\ \tilde{u})) + ((v,\ \tilde{v})) + \kappa((T,\ \tilde{T})). \tag{3.278}$$

Here, the dots above $\dot{H}^1_{\text{per}}$ and $\dot{L}^2$ denote the functions with zero average over $\mathcal{M}$. Since we work with functions with zero average over $\mathcal{M}$, we can use the generalized Poincaré inequality:

$$c_0|U|_H \le \|U\|_V,\ \forall\, U \in V, \tag{3.279}$$

so we know that the norm $\|\cdot\|_V$ defined above is equivalent to the usual H^1-norm, where c_0 is a constant related to the Poincaré constant.

We briefly recall that the unknown functions are of two types: the prognostic variables u, v, and T, for which the initial values are prescribed, and the diagnostic variables ρ, w, and p, which can be defined, at each instant of time, as functions of the prognostic variables. More details regarding the way the diagnostic variables can be determined are available in PETCU and WIROSOETISNO [2005].

The variational formulation of this problem:

Find $U : [0, t_0] \to V$ such that

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t}(U, U^\flat)_H + a(U, U^\flat) + b(U, U, U^\flat) + e(U, U^\flat) = (F, U^\flat)_H, \quad \forall\, U^\flat \in V, \\ &U(0) = U_0. \end{aligned} \tag{3.280}$$

In Eq. (3.280), we introduced the following forms:

$a : V \times V \to \mathbb{R}$ bilinear, continuous:

$$a(U, U^\flat) = \nu((u, u^\flat)) + \nu((v, v^\flat)) + \kappa\mu((T, T^\flat)) - g\beta_T \int_{\mathcal{M}} Tw(U^\flat)\mathrm{d}\mathcal{M}, \tag{3.281}$$

$b : V \times V \times V_2 \to \mathbb{R}$ trilinear:

$$\begin{aligned} b(U, U^\sharp, U^\flat) = &\int_{\mathcal{M}} (u\frac{\partial u^\sharp}{\partial x_1}u^\flat + v\frac{\partial u^\sharp}{\partial x_2}u^\flat + w(U)\frac{\partial u^\sharp}{\partial x_3}u^\flat)\mathrm{d}\mathcal{M} \\ &+ \int_{\mathcal{M}} (u\frac{\partial v^\sharp}{\partial x_1}v^\flat + v\frac{\partial v^\sharp}{\partial x_2}v^\flat + w(U)\frac{\partial v^\sharp}{\partial x_3}v^\flat)\mathrm{d}\mathcal{M} \\ &+ \int_{\mathcal{M}} (u\frac{\partial T^\sharp}{\partial x_1}T^\flat + v\frac{\partial T^\sharp}{\partial x_2}T^\flat + w(U)\frac{\partial T^\sharp}{\partial x_3}\tilde{T})\mathrm{d}\mathcal{M}, \end{aligned} \tag{3.282}$$

$e : V \times V \to \mathbb{R}$ bilinear, continuous:

$$e(U, U^\flat) = f\int_{\mathcal{M}} (uv^\flat - vu^\flat)\mathrm{d}\mathcal{M}. \tag{3.283}$$

We choose κ sufficiently large so that a is coercive on V.

The trilinear form b has the properties proved in Lemma 2.1.

Problem (Eq. (3.280)) can also be written as an operator evolution equation in V_2':

$$\begin{aligned} &\frac{\mathrm{d}U}{\mathrm{d}t} + AU + B(U, U) + EU = F, \\ &U(0) = U_0, \end{aligned} \tag{3.284}$$

where we introduced the following operators:

$$A \text{ linear, continuous from } V \text{ into } V' \text{ defined by} \quad \langle AU, U^\flat \rangle = a(U, U^\flat), \quad \forall U, U^\flat \in V, \tag{3.285}$$

$$B \text{ bilinear, continuous from } V \times V \text{ into } V_2' \text{ defined by} \quad \langle B(U, U^\flat),\ U^\sharp \rangle = b(U, U^\flat, U^\sharp) \quad \forall\, U, U^\flat \in V,\ \forall\, U^\sharp \in V_2, \tag{3.286}$$

$$E \text{ linear, continuous from } V \text{ into } V' \text{ defined by} \quad \langle EU, U^\flat \rangle = e(U, U^\flat), \quad \forall\, U, U^\flat \in V. \tag{3.287}$$

3.6.1. Existence of strong solutions for all time

The classical result about the existence of weak solutions can be found in Section 2. We briefly recall this result and the main lines of the proof since the estimates obtained are used later.

THEOREM 3.8. *Given $U_0 \in H$ and $F \in L^\infty(\mathbb{R}_+; H)$, there exists at least one solution U to problem Eq. (3.280) such that*

$$U \in L^\infty(\mathbb{R}_+; H) \cap L^2(0, t'; V), \quad \forall\, t' > 0. \tag{3.288}$$

PROOF. We start by estimating U in $L^2(\mathcal{M})$:

$$\frac{\mathrm{d}|U|_{L^2}^2}{\mathrm{d}t} + c_1 \|U\|^2 \le c_1' |F|_{L^2}^2, \tag{3.289}$$

where c_1 is the constant of coercivity for a.

Using the Poincaré inequality and the Gronwall lemma, we obtain

$$|U(t)|_{L^2}^2 \le \mathrm{e}^{-c_1 c_0 t} |U_0|_{L^2}^2 + \frac{c_1'}{c_1 c_0} |F|_\infty^2, \tag{3.290}$$

where $|\cdot|_\infty$ is the norm in $L^\infty(\mathbb{R}_+; H)$. We then deduce

$$\limsup_{t \to \infty} |U(t)|_{L^2}^2 \le \frac{c_1'}{c_0 c_1} |F|_\infty^2 := r_0^2. \tag{3.291}$$

From Eq. (3.291), we find that any ball $\mathcal{B}(0, r_0')$ in H, with $r_0' > r_0$, is an absorbing ball. In fact, for all U_0 in H, there exists $t_0 = t_0(|U_0|_H)$ depending increasingly on $|U_0|_H$ such that $|U(t)|_H \le r_0'$ for all $t \ge t_0(|U_0|_H)$. Furthermore, we integrate Eq. (3.289) over $(t, t + r)$ an arbitrary interval of time, with $t \ge t_0(|U_0|_H)$, and using Eq. (3.290), we find

$$\begin{aligned} &\int_t^{t+r} \|U(s)\|^2 \,\mathrm{d}s \le |U(t)|_{L^2}^2 + \frac{r c_1'}{c_0 c_1} |F|_\infty^2, \\ &\int_t^{t+r} \|U(s)\|^2 \,\mathrm{d}s \le K_1,\ \forall\, t \ge t_0(|U_0|_H), \end{aligned} \tag{3.292}$$

where K_1 is a constant depending on the data but not on U_0.

All the estimates above lead us to the existence, globally in time, of a weak solution, as well as to the existence of an absorbing set for $U = (\mathbf{v}, T)$ in H. □

The long-time existence of the strong solutions was established by Cao and Titi [2007] and Kobelkov (see Kobelkov [2006] or Section 3.2 of this chapter), and the result was improved in Ju [2007], showing the existence of the global attractor proved for the strong solution to the PEs.

For subsequent utilization in the proof of the higher order regularity of the solution, we briefly recall the main steps for proving the existence and uniqueness of a global strong solution to the PEs, following Cao and Titi [2007] and Ju [2007].

The main idea is to split the velocity profile $\mathbf{v}$ into two parts: the vertical average $\bar{\mathbf{v}}$ and the remaining part, $\tilde{\mathbf{v}} = \mathbf{v} - \bar{\mathbf{v}}$. As a general notation, we have

$$\bar{\phi}(x_1, x_2) = \frac{1}{L_3} \int_{-L_3/2}^{L_3/2} \phi(x_1, x_2, x_3)\, \mathrm{d}x_3, \qquad \tilde{\phi} = \phi - \bar{\phi}. \tag{3.293}$$

We can find the equation for $\bar{\mathbf{v}}$ by integrating the equations for the horizontal velocity in the vertical direction:

$$\begin{aligned}
&\frac{\partial \bar{\mathbf{v}}}{\partial t} - \nu \Delta \bar{\mathbf{v}} + (\bar{\mathbf{v}} \cdot \nabla)\bar{\mathbf{v}} + \overline{(\tilde{\mathbf{v}} \cdot \nabla)\tilde{\mathbf{v}} + (\nabla \cdot \tilde{\mathbf{v}})\tilde{\mathbf{v}}} + f\vec{\kappa} \times \bar{\mathbf{v}} \\
&+ \nabla \Big[p_s(x_1, x_2, t) + \beta_T \rho_0 g \frac{1}{L_3} \int_{-L_3/2}^{L_3/2} \int_{-L_3/2}^{x_3} T(x_1, x_2, s, t)\, \mathrm{d}s\, \mathrm{d}x_3 \Big] = 0, \\
&\nabla \cdot \bar{\mathbf{v}} = 0, \qquad \text{in } (0, L_1) \times (0, L_2).
\end{aligned} \tag{3.294}$$

For $\tilde{\mathbf{v}}$, we have the following equation:

$$\begin{aligned}
&\frac{\partial \tilde{\mathbf{v}}}{\partial t} - \nu \Delta_3 \tilde{\mathbf{v}} + (\tilde{\mathbf{v}} \cdot \nabla)\tilde{\mathbf{v}} - \left(\int_{-L_3}^{x_3} \nabla \cdot \tilde{\mathbf{v}}(x_1, x_2, s, t)\, \mathrm{d}s \right) \frac{\partial \tilde{\mathbf{v}}}{\partial x_3} \\
&\quad + \overline{(\tilde{\mathbf{v}} \cdot \nabla)\tilde{\mathbf{v}} + (\nabla \cdot \tilde{\mathbf{v}})\tilde{\mathbf{v}}} + (\tilde{\mathbf{v}} \cdot \nabla)\bar{\mathbf{v}} + (\tilde{\mathbf{v}} \cdot \nabla)\tilde{\mathbf{v}} + f\vec{\kappa} \times \tilde{\mathbf{v}} \\
&\quad + \nabla \Bigg[\kappa_1 \int_{-L_3/2}^{x_3} T(x_1, x_2, s, t)\mathrm{d}s \\
&\quad - \kappa_1 \frac{1}{L_3} \int_{-L_3/2}^{L_3/2} \int_{-L_3/2}^{x_3} T(x_1, x_2, s, t)\, \mathrm{d}s\, \mathrm{d}x_3 \Bigg] = 0,
\end{aligned} \tag{3.295}$$

where κ_1 denotes $\beta_T g$.

The idea of Cao and Titi [2007] is to notice that in Eq. (3.295), the term for the surface pressure disappears and we can obtain an a priori estimate of $\tilde{\mathbf{v}}$ in $L^6(\mathcal{M})$. We rather recall here the improved estimates from Ju [2007] for T and $\tilde{\mathbf{v}}$ in $L^6(\mathcal{M})$: the

estimate of T in $L^6(\mathcal{M})$ is obtained by multiplying the equation for the temperature by T^5 and integrating over $\mathcal{M}$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\|T\|_{L^6}^6 + \frac{5}{9}\mu|\nabla_3(T^3)|^2 \le |F_T||T^5| = |F_T|\|T^3\|_{L^{10/3}}^{5/3}$$

$$\le \frac{2}{9}\mu|\nabla_3(T^3)|^2 + c'|F_T|^2\|T\|_{L^6}^4 + c''|F_T|\|T\|_{L^6}^5, \tag{3.296}$$

where we used the following inequality:

$$\|f\|_{L^{10/3}} \le c|f|^{2/5}|\nabla_3 f|^{3/5} + c|f|. \tag{3.297}$$

Using the fact that $\|T\|_{L^6} \le c\|T\|_V$ in space dimension 3, we can apply the uniform Gronwall lemma to Eq. (3.296) and we obtain, using Eq. (3.292), a time-uniform bound on $\|T\|_{L^6}$ and also the existence of an absorbing ball for T in $L^6(\mathcal{M})$.

For $\tilde{\mathbf{v}}$, multiplying Eq. (3.295) by $|\tilde{\mathbf{v}}(t)|^4\tilde{\mathbf{v}}(t)$ and integrating over $\mathcal{M}$, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\tilde{\mathbf{v}}\|_{L^6}^6 + \nu||\nabla_3\tilde{\mathbf{v}}||\tilde{\mathbf{v}}|^2|^2 \le (c_1 + c_2\|\mathbf{v}\|^2)\|\tilde{\mathbf{v}}\|_{L^6}^6 + c_3|\mathbf{v}|^2\|\mathbf{v}\|^2\|\tilde{\mathbf{v}}\|_{L^6}^2 + c_4\|T\|_{L^6}^6. \tag{3.298}$$

From the previous estimates, we obtain, after applying the uniform Gronwall lemma, that

$$\|\tilde{\mathbf{v}}\|_{L^6}^6 \le K_2, \quad \forall\, t \ge t_0(|U_0|_H) + r, \tag{3.299}$$

where K_2 is a constant independent of the initial data, and $t_0(|U_0|_H)$ and r were defined above.

Using the estimates obtained above and the fact that $\|\tilde{\mathbf{v}}\|_{L^6} \le c\|\mathbf{v}\|_{H^1}$, we find

$$\int_t^{t+r} ||\tilde{\mathbf{v}}(s)|^2|\nabla_3\tilde{\mathbf{v}}||^2\,\mathrm{d}s \le c_1 \sup_t \|\tilde{\mathbf{v}}\|_{L^6}^6 + c_2 \sup_t |\mathbf{v}|^2\|\tilde{\mathbf{v}}\|_{L^6}^6 \int_t^{t+r} |\nabla_3\mathbf{v}(s)|^2\,\mathrm{d}s$$

$$+ c_3 \sup_t \|\tilde{\mathbf{v}}\|_{L^6}^6 \int_t^{t+r} |\nabla_3\tilde{\mathbf{v}}(s)|^2\,\mathrm{d}s + c_4 \sup_t \|T\|_{L^6}^6. \tag{3.300}$$

H^1 estimates The next step in obtaining the a priori estimates for the strong solution is to estimate the H^1 norm of $\bar{\mathbf{v}}$. We notice that Eq. (3.294) mainly behaves as a 2D Navier-Stokes equation with rotation, so the estimates are classical. Multiplying Eq. (3.294) by $-\Delta\bar{\mathbf{v}}$ and integrating over $(0, L_1) \times (0, L_2)$, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}|\nabla\bar{\mathbf{v}}|^2 + \nu|\Delta\bar{\mathbf{v}}|^2 \le c_1|\mathbf{v}|^2|\nabla_3\mathbf{v}|^2|\nabla_2\bar{\mathbf{v}}|^2$$

$$+ c_2(|\nabla_3\mathbf{v}|^2 + |\mathbf{v}|^2 + ||\tilde{\mathbf{v}}|^2|\nabla_3\tilde{\mathbf{v}}||^2). \tag{3.301}$$

Applying the uniform Gronwall lemma to Eq. (3.301) and using the previous estimates, we find a time-uniform estimate for $|\nabla\bar{\mathbf{v}}|$ and also the existence of an absorbing ball for $|\nabla\bar{\mathbf{v}}|$ in L^2. Since

$$\|\bar{\mathbf{v}}\|_{L^6} \leq c|\nabla\bar{\mathbf{v}}|,$$

we also have the uniform boundedness of $\mathbf{v}$ in $L^6(\mathcal{M})$ and the existence of an absorbing ball for $\mathbf{v}$ in $L^6(\mathcal{M})$.

After this additional step, one can now obtain H^1 estimates for $\mathbf{v}$, which are no longer local in time. We first estimate $\mathbf{v}_{x_3}$ and then $\nabla_3\mathbf{v} = (\nabla\mathbf{v}, \mathbf{v}_{x_3})$. For $\mathbf{v}_{x_3}$, we find

$$\frac{\mathrm{d}}{\mathrm{d}t}|\mathbf{v}_{x_3}|^2 + \nu\|\mathbf{v}_{x_3}\|^2 \leq c_1\|\mathbf{v}\|_{L^6}^6|\mathbf{v}_{x_3}|^2 + c_2|T|^2. \tag{3.302}$$

As before, we obtain uniform estimates on $\mathbf{v}_{x_3}$. We can write

$$\frac{\mathrm{d}}{\mathrm{d}t}|\nabla_3\mathbf{v}|^2 + \nu|\Delta_3\mathbf{v}|^2 \leq \xi(t)|\nabla_3\mathbf{v}|^2 + \eta(t), \tag{3.303}$$

where

$$\xi(t) = c_1(\|\mathbf{v}(t)\|_{L^6}^4 + |\nabla_3\mathbf{v}(t)|^2|\mathbf{v}_{x_3}(t)|^2), \quad \eta(t) = c_2|\nabla_3 T(t)|^2.$$

We find, by the uniform Gronwall lemma, an uniform bound for $\mathbf{v}$ in $H^1(\mathcal{M})$ and also the existence of an absorbing ball for $\mathbf{v}$ in $H^1(\mathcal{M})$.

In order to be able to use the Gronwall uniform lemma for higher order estimates, notice that we also deduce that

$$\int_t^{t+r} |\Delta_3\mathbf{v}|^2 \,\mathrm{d}t \leq K_1, \ \forall\, t \geq t_1(U_0) = t_0(|U_0|_H) + 2r,$$

where K_1 is a constant depending on the data but not on the initial data U_0.

The H^1-estimates for T are now immediate. Using the results above, we find the existence globally of a strong solution U and also the existence of an absorbing ball for U in V.

The proof of Kobelkov [2006] presented in Section 3.2 is different and based on a suitable a priori estimate of the surface pressure, but it leads mainly to the same existence and uniqueness result (as in Theorem 3.2).

Theorem 3.9. *Given $U_0 \in V$ and $F \in L^\infty(\mathbb{R}_+; H)$, there exists a unique solution $U = U(t)$ of Eq. (3.280) on $\mathbb{R}_+$ such that*

$$U \in \mathcal{C}(\mathbb{R}_+; V) \cap L^2(0, t'; (\dot{H}^2_{\mathrm{per}}(\mathcal{M}))^3), \ \forall\, t' > 0. \tag{3.304}$$

Remark 3.5. Using these a priori estimates, Ning Ju proved the existence of the global attractor for the strong solutions of the PEs. In what follows, we prove the existence of absorbing balls for the solution U in all the $H^m(\mathcal{M})$ Sobolev spaces and the existence of the global attractor in $H^m(\mathcal{M})$, for all $m \geq 1$ will follow by the general theory of the global attractors (more details regarding the existence of global attractors can be found in a general context in Temam [1997], Chepyzhov and Vishik [2002], Hale [2001] or in the context of the Navier-Stokes equations in Constantin, Foias and Temam [1985]).

3.6.2. Higher order regularity results, H^2-estimates

We are now interested in proving the long-time existence of a solution in $H^2(\mathcal{M})$. In order to simplify the computations, we introduce the following notations: we set $|U|_m = (\sum_{[\alpha]=m} |D^\alpha U|_H^2)^{1/2}$, where D^α is the differential operator $D^\alpha = D_1^{\alpha_1} D_2^{\alpha_2} D_3^{\alpha_3}$, $D_i = \partial/\partial x_i$, α is a multi-index, $\alpha = (\alpha_1, \alpha_2, \alpha_3)$, $\alpha_i \in \mathbb{N}$, and $[\alpha] = \alpha_1 + \alpha_2 + \alpha_3$.

We multiply the equation for the velocity by $(-\Delta_3)^2\mathbf{v}$ and integrate over $\mathcal{M}$. We find

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|\Delta_3\mathbf{v}|^2 &+ \int_{\mathcal{M}} (\mathbf{v}\cdot\nabla)\mathbf{v}\cdot(-\Delta_3)^2\mathbf{v}\,\mathrm{d}\mathcal{M} + \int_{\mathcal{M}} w(\mathbf{v})\frac{\partial\mathbf{v}}{\partial x_3}(-\Delta_3)^2\mathbf{v}\,\mathrm{d}\mathcal{M}\\ &+ f\int_{\mathcal{M}} (\kappa\times\mathbf{v})\cdot(-\Delta_3)^2\mathbf{v}\,\mathrm{d}\mathcal{M} + \frac{1}{\rho_0}\int_{\mathcal{M}} \nabla p_s\cdot(-\Delta_3)^2\mathbf{v}\,\mathrm{d}\mathcal{M}\\ &+ \kappa_1\int_{\mathcal{M}} \nabla\left(\int_0^{x_3} T(z)\,\mathrm{d}z\right)\cdot(-\Delta_3)^2\mathbf{v}\,\mathrm{d}\mathcal{M} + \nu|(-\Delta_3)^{3/2}\mathbf{v}|^2\\ &= \int_{\mathcal{M}} F_{\mathbf{v}}\cdot(-\Delta_3)^2\mathbf{v}\,\mathrm{d}\mathcal{M}, \end{aligned} \tag{3.305}$$

where by $F_{\mathbf{v}}$ we understand the function (F_u, F_v).

We can easily check that

$$f\int_{\mathcal{M}} (\kappa\times\mathbf{v})\cdot(-\Delta_3)^2\mathbf{v}\,\mathrm{d}\mathcal{M} = 0.$$

For the integral from Eq. (3.305) containing the surface term p_s, integrating by parts, using the boundary conditions and the conservation of mass equation, we find

$$\begin{aligned}\int_{\mathcal{M}} \nabla p_s\cdot(-\Delta_3)^2\mathbf{v}\,\mathrm{d}\mathcal{M} &= \int_0^{L_1}\int_0^{L_2} \nabla p_{\mathrm{s}}\cdot\left(\int_{-L_3/2}^{L_3/2}(-\Delta_3)^2\mathbf{v}\,\mathrm{d}x_3\right)\mathrm{d}x_2\,\mathrm{d}x_1\\ &= -\int_0^{L_1}\int_0^{L_2} p_{\mathrm{s}}\left(\int_{-L_3/2}^{L_3/2}(-\Delta_3)^2(\nabla\cdot\mathbf{v})\,\mathrm{d}x_3\right)\mathrm{d}x_2\,\mathrm{d}x_1\\ &= \int_0^{L_1}\int_0^{L_2} p_{\mathrm{s}}\left(\int_{-L_3/2}^{L_3/2}(-\Delta_3)^2 w_{x_3}\,\mathrm{d}x_3\right)\mathrm{d}x_2\,\mathrm{d}x_1 = 0. \end{aligned} \tag{3.306}$$

The integral containing T is also easy to estimate:

$$\begin{aligned}\left|\kappa_1\int_{\mathcal{M}} \nabla\left(\int_0^{x_3} T(z)\,\mathrm{d}z\right)\cdot(-\Delta_3)^2\mathbf{v}\,\mathrm{d}\mathcal{M}\right| &\le c\int_{\mathcal{M}} |\Delta_3 T||(-\Delta_3)^{3/2}\mathbf{v}|\mathrm{d}\mathcal{M}\\ &\le \frac{\nu}{6}|(-\Delta_3)^{3/2}\mathbf{v}|^2 + c|\Delta_3 T|^2. \end{aligned} \tag{3.307}$$

It remains to estimate the integrals obtained from the nonlinear terms. We note that in fact we need to estimate integrals of the type:

$$\int_{\mathcal{M}} u\frac{\partial u}{\partial x_1}D_1^{2\alpha_1}D_2^{2\alpha_2}D_3^{2\alpha_3}u\mathrm{d}\mathcal{M}, \quad \int_{\mathcal{M}} v\frac{\partial u}{\partial x_2}D_1^{2\alpha_1}D_2^{2\alpha_2}D_3^{2\alpha_3}u\mathrm{d}\mathcal{M},$$
$$\int_{\mathcal{M}} w(\mathbf{v})\frac{\partial u}{\partial x_3}D_1^{2\alpha_1}D_2^{2\alpha_2}D_3^{2\alpha_3}u\mathrm{d}\mathcal{M}, \tag{3.308}$$

where $\alpha_i \in \mathbb{N}$ with $[\alpha] = \alpha_1 + \alpha_2 + \alpha_3 = 2$.

Integrating by parts and using periodicity, the integrals become

$$\int_{\mathcal{M}} D^\alpha\left(u\frac{\partial u}{\partial x_1}\right)D^\alpha u\mathrm{d}\mathcal{M}, \quad \int_{\mathcal{M}} D^\alpha\left(v\frac{\partial u}{\partial x_2}\right)D^\alpha u\mathrm{d}\mathcal{M},$$
$$\int_{\mathcal{M}} D^\alpha\left(w(U)\frac{\partial u}{\partial x_3}\right)D^\alpha u\mathrm{d}\mathcal{M}. \tag{3.309}$$

Using Leibniz formula, we see that the integrals can be written in the form

$$\int_{\mathcal{M}} uD^\alpha\frac{\partial u}{\partial x_1}D^\alpha u\mathrm{d}\mathcal{M}, \quad \int_{\mathcal{M}} vD^\alpha\frac{\partial u}{\partial x_2}D^\alpha u\mathrm{d}\mathcal{M}, \quad \int_{\mathcal{M}} w(\mathbf{v})D^\alpha\frac{\partial u}{\partial x_3}D^\alpha u\mathrm{d}\mathcal{M}, \tag{3.310}$$

or in the form

$$\int_{\mathcal{M}} \delta^k u\delta^{2-k}\frac{\partial u}{\partial x_1}D^\alpha u\mathrm{d}\mathcal{M}, \quad \int_{\mathcal{M}} \delta^k v\delta^{2-k}\frac{\partial u}{\partial x_2}D^\alpha u\mathrm{d}\mathcal{M},$$
$$\int_{\mathcal{M}} \delta^k w(\mathbf{v})\delta^{2-k}\frac{\partial u}{\partial x_3}D^\alpha u\mathrm{d}\mathcal{M}, \tag{3.311}$$

where $k = 1, 2$, and δ^k is some differential operator D^α with $[\alpha] = k$.

Note that for each α, after integration by parts, the sum of the integrals of type (Eq. (3.310)) is zero because of the mass conservation equation.

It remains to estimate the integrals of type (Eq. (3.311)). The first two integrals in (Eq. (3.311)) lead to the same kind of estimates, so in fact we only need to estimate the first and last integrals.

The first integrals are easiest to estimate: for $[k] = 1$, we write

$$\left|\int_{\mathcal{M}} \delta^k u\,\delta^{2-k}\frac{\partial u}{\partial x_1}D^\alpha u\mathrm{d}\mathcal{M}\right| \le |\delta^k u|_{L^6}\left|\delta^{2-k}\frac{\partial u}{\partial x_1}\right|_{L^3}|D^\alpha u|_{L^2}$$
$$\le c_1|\mathbf{v}|_2^{5/2}|\mathbf{v}|_3^{1/2} \le \frac{\nu}{6}|\mathbf{v}|_3^2 + c_2|\mathbf{v}|_2^{10/3}, \tag{3.312}$$

while for $[k] = 2$, we find

$$\left|\int_{\mathcal{M}} \delta^k u\,\frac{\partial u}{\partial x_1}D^\alpha u\mathrm{d}\mathcal{M}\right| \le |\delta^k u|_{L^3}\left|\frac{\partial u}{\partial x_1}\right|_{L^3}|D^\alpha u|_{L^2}$$
$$\le c_1|\mathbf{v}|_2^{5/2}|\mathbf{v}|_3^{1/2} \le \frac{\nu}{6}|\mathbf{v}|_3^2 + c_2|\mathbf{v}|_2^{10/3}. \tag{3.313}$$

The last type of integrals from Eq. (3.311) is estimated as follows: for $[k] = 2$, we find

$$\begin{aligned}\left|\int_{\mathcal{M}} \delta^k w(\mathbf{v}) \frac{\partial u}{\partial x_3} D^\alpha u \mathrm{d}\mathcal{M}\right| &\leq |\delta^k w(\mathbf{v})|_{L^2} \left|\frac{\partial u}{\partial x_3}\right|_{L^6} |D^\alpha u|_{L^3} \\ &\leq c_1 \left|\frac{\partial u}{\partial x_3}\right|_{L^6} |\mathbf{v}|_2^{1/2} |\mathbf{v}|_3^{3/2} \leq \frac{\nu}{6}|\mathbf{v}|_3^2 + c_2 |\mathbf{v}_{x_3}|_{L^6}^4 |\mathbf{v}|_2^2,\end{aligned} \tag{3.314}$$

where $\mathbf{v}_{x_3} = \frac{\partial \mathbf{v}}{\partial x_3}$, and for $[k] = 1$, we have

$$\begin{aligned}\left|\int_{\mathcal{M}} \delta^k w(\mathbf{v}) \delta^{2-k} \frac{\partial u}{\partial x_3} D^\alpha u \mathrm{d}\mathcal{M}\right| &\leq |\delta^k w(\mathbf{v})|_{L^3} \left|\frac{\partial u}{\partial x_3}\right|_{L^6} |D^\alpha u|_{L^2} \\ &\leq c_1 |\mathbf{v}|_2^{3/2} |\mathbf{v}|_3^{1/2} |\mathbf{v}_{x_3}|_{H^1} \\ &\leq \frac{\nu}{6}|\mathbf{v}|_3^2 + c_2 |\mathbf{v}|_2^2 |\mathbf{v}_{x_3}|_{H^1}^{4/3}.\end{aligned} \tag{3.315}$$

Gathering all the estimates above, Eq. (3.305) leads to

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|\mathbf{v}|_2^2 + \nu|\mathbf{v}|_3^2 \leq \frac{5\nu}{6}|\mathbf{v}|_3^2 + \alpha(t)|\mathbf{v}|_2^2 + \beta(t), \tag{3.316}$$

where by α and β, we understand the following functions:

$$\begin{aligned}\alpha(t) &= c_1 |T|_2^2 + c_2, \\ \beta(t) &= c_3 |\mathbf{v}_{x_3}|_{L^6}^4 + |\mathbf{v}_{x_3}|_{H^1}^{4/3}.\end{aligned}$$

Taking into account the estimates derived in the previous section, we can apply the uniform Gronwall lemma to Eq. (3.316) and we find a uniform bound for $\mathbf{v}$ in $H^2(\mathcal{M})$ if we estimate the norm of $\mathbf{v}_{x_3}$ in L^6. The same kind of estimates can be deduced for the temperature T.

We start deducing the required estimates on $\mathbf{v}_{x_3}$ in L^6. The equation for $\mathbf{v}_{x_3}$ reads

$$\begin{aligned}&\frac{\partial \mathbf{v}_{x_3}}{\partial t} - \nu\Delta_3 \mathbf{v}_{x_3} + \nabla_{\mathbf{v}} \mathbf{v}_{x_3} - \Big(\int_0^{x_3} \nabla \cdot \mathbf{v}(\xi)\,\mathrm{d}\xi\Big)\frac{\partial \mathbf{v}_{x_3}}{\partial x_3} \\ &\quad + \nabla_{\mathbf{v}_{x_3}} \mathbf{v} - (\nabla \cdot \mathbf{v})\mathbf{v}_{x_3} + f\kappa \mathbf{v}_{x_3} + \kappa_1 \nabla T = 0\end{aligned} \tag{3.317}$$

We multiply Eq. (3.317) by $|\mathbf{v}_{x_3}|^4 \mathbf{v}_{x_3}$ and integrate over $\mathcal{M}$. We find

$$\begin{aligned}&\frac{1}{6}\frac{\mathrm{d}}{\mathrm{d}t}|\mathbf{v}_{x_3}|_{L^6}^6 + \nu \int_{\mathcal{M}} (|\nabla_3 \mathbf{v}_{x_3}|^2 |\mathbf{v}_{x_3}|^4 + |\nabla_3 |\mathbf{v}_{x_3}|^2|^2 |\mathbf{v}_{x_3}|^2) \mathrm{d}\mathcal{M} \\ &\quad + f\int_{\mathcal{M}} \kappa \times \mathbf{v}_{x_3} \cdot |\mathbf{v}_{x_3}|^4 \mathbf{v}_{x_3} \mathrm{d}\mathcal{M} + \int_{\mathcal{M}} \nabla_{\mathbf{v}} \mathbf{v}_{x_3} \cdot |\mathbf{v}_{x_3}|^4 \mathbf{v}_{x_3} \mathrm{d}\mathcal{M} \\ &\quad - \int_{\mathcal{M}} \Big(\int_0^{x_3} \nabla \cdot \mathbf{v}(\xi)\,\mathrm{d}\xi\Big)\frac{\partial \mathbf{v}_{x_3}}{\partial x_3} \cdot |\mathbf{v}_{x_3}|^4 \mathbf{v}_{x_3} \mathrm{d}\mathcal{M} \\ &\quad + \int_{\mathcal{M}} \nabla_{\mathbf{v}_{x_3}} \mathbf{v} \cdot |\mathbf{v}_{x_3}|^4 \mathbf{v}_{x_3} \mathrm{d}\mathcal{M} - \int_{\mathcal{M}} (\nabla \cdot \mathbf{v}) \mathbf{v}_{x_3} \cdot |\mathbf{v}_{x_3}|^4 \mathbf{v}_{x_3} \mathrm{d}\mathcal{M} \\ &\quad + \kappa_1 \int_{\mathcal{M}} \nabla T \cdot |\mathbf{v}_{x_3}|^4 \mathbf{v}_{x_3} \mathrm{d}\mathcal{M} = 0.\end{aligned} \tag{3.318}$$

Integrating by parts, we find

$$\int_{\mathcal{M}} \nabla_{\mathbf{v}} \mathbf{v}_{x_3} \cdot |\mathbf{v}_{x_3}|^4 \mathbf{v}_{x_3} \mathrm{d}\mathcal{M} - \int_{\mathcal{M}} \Big(\int_0^{x_3} \nabla \cdot \mathbf{v}(\xi)\, \mathrm{d}\xi \Big) \frac{\partial \mathbf{v}_{x_3}}{\partial x_3} \cdot |\mathbf{v}_{x_3}|^4 \mathbf{v}_{x_3} \mathrm{d}\mathcal{M} = 0. \tag{3.319}$$

It is also immediate that

$$f \int_{\mathcal{M}} \kappa \times \mathbf{v}_{x_3} \cdot |\mathbf{v}_{x_3}|^4 \mathbf{v}_{x_3} \mathrm{d}\mathcal{M} = 0.$$

The following term is estimated using the integration by parts:

$$\begin{aligned}
\Big| \int_{\mathcal{M}} \nabla_{\mathbf{v}_{x_3}} \mathbf{v} \cdot |\mathbf{v}_{x_3}|^4 \mathbf{v}_{x_3} \mathrm{d}\mathcal{M} \Big| &\le c \int_{\mathcal{M}} |\nabla \mathbf{v}_{x_3}| |\mathbf{v}_{x_3}|^5 |\mathbf{v}| \mathrm{d}\mathcal{M} \\
&\le c \Big(\int_{\mathcal{M}} |\nabla \mathbf{v}_{x_3}|^2 |\mathbf{v}_{x_3}|^4 \mathrm{d}\mathcal{M} \Big)^{1/2} \Big(\int_{\mathcal{M}} |\mathbf{v}_{x_3}|^6 \mathrm{d}\mathcal{M} \Big)^{1/2} |\mathbf{v}|_{L^\infty} \\
&\le \frac{\nu}{6} \int_{\mathcal{M}} |\nabla \mathbf{v}_{x_3}|^2 |\mathbf{v}_{x_3}|^4 \mathrm{d}\mathcal{M} + c |\mathbf{v}|_{L^\infty} \int_{\mathcal{M}} |\mathbf{v}_{x_3}|^6 \mathrm{d}\mathcal{M} \\
&\quad \text{(using Agmon's inequality: see for example TEMAM [1997])} \\
&\le \frac{\nu}{6} \int_{\mathcal{M}} |\nabla \mathbf{v}_{x_3}|^2 |\mathbf{v}_{x_3}|^4 \mathrm{d}\mathcal{M} + c |\mathbf{v}|_{H^2} |\mathbf{v}_{x_3}|_{L^6}^6.
\end{aligned} \tag{3.320}$$

Similarly, we find

$$\begin{aligned}
\Big| \int_{\mathcal{M}} (\nabla \cdot \mathbf{v}) \mathbf{v}_{x_3} \cdot |\mathbf{v}_{x_3}|^4 \mathbf{v}_{x_3} \mathrm{d}\mathcal{M} \Big| &\le c \int_{\mathcal{M}} |\nabla \mathbf{v}_{x_3}| |\mathbf{v}_{x_3}|^5 |\mathbf{v}| \mathrm{d}\mathcal{M} \\
&\le \frac{\nu}{6} \int_{\mathcal{M}} |\nabla \mathbf{v}_{x_3}|^2 |\mathbf{v}_{x_3}|^4 \mathrm{d}\mathcal{M} + c |\mathbf{v}|_{H^2} |\mathbf{v}_{x_3}|_{L^6}^6.
\end{aligned} \tag{3.321}$$

For the last integral in Eq. (3.318), we find by integration by parts:

$$\begin{aligned}
\Big| \kappa_1 \int_{\mathcal{M}} \nabla T \cdot |\mathbf{v}_{x_3}|^4 \mathbf{v}_{x_3} \mathrm{d}\mathcal{M} \Big| &\le c \int_{\mathcal{M}} |T| |\nabla \mathbf{v}_{x_3}| |\mathbf{v}_{x_3}|^5 \mathrm{d}\mathcal{M} \\
&\le \frac{\nu}{6} \int_{\mathcal{M}} |\nabla \mathbf{v}_{x_3}|^2 |\mathbf{v}_{x_3}|^4 \mathrm{d}\mathcal{M} + c |T|_{L^\infty} \int_{\mathcal{M}} |\mathbf{v}_{x_3}|^6 \mathrm{d}\mathcal{M} \\
&\le \frac{\nu}{6} \int_{\mathcal{M}} |\nabla \mathbf{v}_{x_3}|^2 |\mathbf{v}_{x_3}|^4 \mathrm{d}\mathcal{M} + c |T|_{H^2} |\mathbf{v}_{x_3}|_{L^6}^6.
\end{aligned} \tag{3.322}$$

Gathering all the above estimates, we find

$$\frac{1}{6} \frac{\mathrm{d}}{\mathrm{d}t} |\mathbf{v}_{x_3}|_{L^6}^6 + \frac{\nu}{2} \int_{\mathcal{M}} |\nabla \mathbf{v}_{x_3}|^2 |\mathbf{v}_{x_3}|^4 \mathrm{d}\mathcal{M} \le c(|\mathbf{v}|_{H^2} + |T|_{H^2}) |\mathbf{v}_{x_3}|_{L^6}^6, \tag{3.323}$$

and by the uniform Gronwall lemma, we obtain a uniform estimate for $|\mathbf{v}_{x_3}|_{L^6}$. We can now return to Eq. (3.316) and apply the Gronwall lemma.

In order to be able to use the uniform Gronwall lemma for further a priori estimates, we also need to integrate Eq. (3.316) from t to $t+r$. We thus obtain

$$\int_t^{t+r} |\mathbf{v}(t')|_3^2 \, \mathrm{d}t' \leq K_2, \qquad \forall\, t \geq t_2(U_0) = t_0(|U_0|_H) + 3r, \tag{3.324}$$

where K_2 is a constant depending on the data but not on the initial data, U_0, $r > 0$ is fixed as before, and t_2 is a time depending on the initial data.

The same computations can be done in order to estimate the temperature T:

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|T|_2^2 + \mu|T|_3^2 \leq \frac{\mu}{2}|T|_3^2 + c(|\mathbf{v}|_3^{4/3} + |\mathbf{v}|_2^{4/3})|T|_2^2 + |F_T|_2^2. \tag{3.325}$$

Using the uniform bounds obtained above, we prove the existence of an absorbing ball for T in $H^2(\mathcal{M})$. Integrating from t to $t+r$, we also compute an uniform bound for the time average of $|T|_{H^3}$, which will be used later for the H^m estimates, with $m > 2$.

Combining the results on $\mathbf{v}$ and T, we found uniform bounds for $U = (u, v, T)$ in $H^2(\mathcal{M})$, and we also proved the existence of an absorbing ball for U in $H^2(\mathcal{M})$.

Estimates in H^m, for $m > 2$ In order to obtain the a priori estimates in H^m with $m > 2$, we can work directly with the variational formulation of the PEs. In Eq. (3.284), we take the test function $U^\flat = (-\Delta_3)^m U(t)$ where t is a fixed, arbitrary moment in time. The computations follow exactly the same steps as in Petcu and Wirosoetisno [2005]:

$$\frac{\mathrm{d}}{\mathrm{d}t}(U, (-\Delta_3)^m U)_H + a(U, (-\Delta_3)^m U) + b(U, U, (-\Delta_3)^m U) + e(U, (-\Delta_3)^m U) = (F, (-\Delta_3)^m U)_H. \tag{3.326}$$

We also note that

$$a(U, (-\Delta_3)^m U) = a((-\Delta_3)^{m/2} U, (-\Delta_3)^{m/2} U) \geq c_1 |U(t)|_{m+1}^2, \tag{3.327}$$

where we used the coercivity of a.

Integrating by parts and using the periodicity, we find

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|U(t)|_m^2 + c_1|U(t)|_{m+1}^2 \leq |b(U, U, (-\Delta_3)^m U)| + |(F, (-\Delta_3)^m U)_H|. \tag{3.328}$$

We then need to estimate the terms in the right hand side of Eq. (3.328). Analysing the behavior of the integrals of the type (Eq. (3.311)), we find

$$|b(U, U, (-\Delta_3)^m U)| \leq c \sum_{k=1}^{m} |U|_{m-k+2} |U|_k^{1/2} |U|_{k+1}^{1/2} |U|_m + c \sum_{k=1}^{m-1} |U|_{k+1} |U|_{m-k+1}^{1/2} |U|_{m-k+2}^{1/2} |U|_{m+1} + c|U|_2 |U|_m^{1/2} |U|_{m+1}^{3/2}. \tag{3.329}$$

Gathering all the estimates above, we obtain the differential inequality:

$$\frac{\mathrm{d}}{\mathrm{d}t}|U|_m^2 + c_1|U|_{m+1}^2 \leq \alpha(t) + \beta(t)|U|_m^2, \tag{3.330}$$

where the expressions of the functions $\alpha(t)$ and $\beta(t)$ can be easily derived from the estimates above. The functions α and β are formed from sums involving the terms $|U|_k$, with $k \leq m$. Taking into account the estimates from all the previous steps, we can apply the uniform Gronwall lemma to Eq. (3.330), and we find a uniform bound for U in $H^m(\mathcal{M})$, as well as the existence of an absorbing set for U, in $H^m(\mathcal{M})$, for all $m \geq 2$.

We conclude with the following result.

THEOREM 3.10. *Given $m \in \mathbb{N}$, $m \geq 2$, $U_0 \in V \cap (\dot{H}^m_{\mathrm{per}}(\mathcal{M}))^3$, and $F \in L^\infty(\mathbb{R}_+; H \cap (\dot{H}^{m-1}_{\mathrm{per}}(\mathcal{M}))^3)$, there exists a unique solution U of Eq. (3.284) on $\mathbb{R}_+$ such that*

$$U \in \mathcal{C}(\mathbb{R}_+; (\dot{H}^m_{\mathrm{per}}(\mathcal{M}))^3) \cap L^2(0, t'; (\dot{H}^{m+1}_{\mathrm{per}}(\mathcal{M}))^3), \quad \forall\, t' > 0. \tag{3.331}$$

Moreover, if $U_0 \in V$ and $F \in L^\infty(\mathbb{R}_+; H \cap (\dot{H}^{m-1}_{\mathrm{per}}(\mathcal{M}))^3)$, then the solution U of Eq. (3.284) belongs to $\mathcal{C}(\mathbb{R}_+^\star; \dot{H}^m_{\mathrm{per}}(\mathcal{M}))^3)$.

PROOF. The proof for the existence of the solution is based on the a priori estimates on higher orders, which were obtained above, and on the Galerkin method, using the Fourier series for the Galerkin basis. The uniqueness of the solution can be easily obtained by classical methods: we consider U_1 and U_2 two solutions of the PEs (Eq. (3.280)) and estimating the H^1-norm of $U = U_1 - U_2$, and we find that the solutions coincide.

In order to prove the second part of the theorem, regarding the regularity of the solution starting with an initial condition in V, we use the same reasoning as in PETCU and WIROSOETISNO [2005].

If U_0 belongs to V, the solution U of problem (Eq. (3.284)) belongs to $L^2(0, t; \dot{H}^2_{\mathrm{per}})$, for all $t > 0$. This means that $U(t) \in \dot{H}^2_{\mathrm{per}}(\mathcal{M})$ almost everywhere on $\mathbb{R}_+$, so there exists a t_1 arbitrarily small such that $U(t_1) \in \dot{H}^2_{\mathrm{per}}(\mathcal{M})$. Using the first part of the theorem, we know that the solution U is such that

$$U \in \mathcal{C}([t_1, +\infty); \dot{H}^2_{\mathrm{per}}(\mathcal{M})) \cap L^2(t_1, t; \dot{H}^3_{\mathrm{per}}(\mathcal{M})), \qquad \forall\, t > 0.$$

By the same argument as before, we find a t_2 arbitrarily close to t_1 such that $U(t_2) \in \dot{H}^3_{\mathrm{per}}(\mathcal{M})$. Using the first part of the theorem, we obtain that the solution U is such that

$$U \in \mathcal{C}([t_2, +\infty); \dot{H}^3_{\mathrm{per}}(\mathcal{M})) \cap L^2(t_2, t; \dot{H}^4_{\mathrm{per}}(\mathcal{M})), \qquad \forall\, t > 0.$$

Recurrently, we find

$$U \in \mathcal{C}([t_{m-1}, +\infty); \dot{H}^m_{\mathrm{per}}(\mathcal{M})) \cap L^2(t_{m-1}, t; \dot{H}^{m+1}_{\mathrm{per}}(\mathcal{M})), \qquad \forall\, t > 0.$$

where t_{m-1} is arbitrarily close to zero. From this relation, the result follows immediately:

$$U \in \mathcal{C}((\mathbb{R}_+^\star; \dot{H}^m_{\mathrm{per}}(\mathcal{M})).$$

□

3.6.3. Gevrey regularity results

In PETCU and WIROSOETISNO [2005] they proved that the solutions of the 3D PEs are the restriction on a positive finite interval in time of some complex analytic function in time with values in a Gevrey space. In view of the above results of long-time existence of strong solutions, we can actually prove that the solution is analytic in time on every interval of time $(0, t)$, with $t > 0$, i.e., on $(0, \infty)$.

This result follows the main steps as in PETCU and WIROSOETISNO [2005], so we introduce the main notations, and we briefly recall the methods used in PETCU and WIROSOETISNO [2005].

We introduce the following notation:

$$[U_k]_\kappa^2 = |u_k|^2 + |v_k|^2 + \kappa |T_k|^2. \tag{3.332}$$

We define the Gevrey space $D(e^{\tau(-\Delta_3)^{1/2}})$, $\tau > 0$, as the set of functions U in H satisfying

$$|\mathcal{M}| \sum_{k\in\mathbb{Z}^3} e^{2\tau|k'|}[U_k]_\kappa^2 = |e^{\tau(-\Delta_3)^{1/2}}U|_H^2 < \infty. \tag{3.333}$$

The Hilbert norm of $D(e^{\tau(-\Delta_3)^{1/2}})$ is given by

$$|U|_\tau := |U|_{D(e^{\tau(-\Delta_3)^{1/2}})} = |e^{\tau(-\Delta_3)^{1/2}}U|_H, \text{ for } U \in D(e^{\tau(-\Delta_3)^{1/2}}), \tag{3.334}$$

and the associated scalar product is

$$(U, V)_\tau := (U, V)_{D(e^{\tau(-\Delta_3)^{1/2}})} = (e^{\tau(-\Delta_3)^{1/2}}U, e^{\tau(-\Delta_3)^{1/2}}V)_H, \quad \text{for } U, V \in D(e^{\tau(-\Delta_3)^{1/2}}). \tag{3.335}$$

Another Gevrey space that we will use is $D((-\Delta_3)^{m/2}e^{\tau(-\Delta_3)^{1/2}})$, $m \geq 1$ integer, which is a Hilbert space when endowed with the inner product:

$$(U, V)_{D((-\Delta_3)^{m/2}e^{\tau(-\Delta_3)^{1/2}})} = ((-\Delta_3)^{m/2}e^{\tau(-\Delta_3)^{1/2}}U, (-\Delta_3)^{m/2}e^{\tau(-\Delta_3)^{1/2}}V)_H. \tag{3.336}$$

In order to be able to estimate the norm of the solution U into a Gevrey space, we first need to be able to estimate the nonlinear term. The result we give here was proved by PETCU and WIROSOETISNO [2005].

LEMMA 3.15. *Let U, $U^\sharp$, and $U^\flat$ be given in $D((-\Delta_3)^{3/2}e^{\tau(-\Delta_3)^{1/2}})$, for $\tau \geq 0$. Then the following inequality holds:*

$$\begin{aligned}
&|((-\Delta_3)^{1/2}B(U, U^\sharp), (-\Delta_3)^{3/2}U^\flat)_\tau| \\
&\quad \leq c|\Delta_3 U|_\tau |\Delta_3 U^\sharp|_\tau^{1/2} |(-\Delta_3)^{3/2}U^\sharp|_\tau^{1/2} |(-\Delta_3)^{3/2}U^\flat|_\tau \\
&\qquad + c|\Delta_3 U|_\tau^{1/2} |(-\Delta_3)^{3/2}U|_\tau^{1/2} |\Delta_3 U^\sharp|_\tau |(-\Delta_3)^{3/2}U^\flat|_\tau.
\end{aligned} \tag{3.337}$$

PROOF. We first write the trilinear form b in Fourier modes. In order to simplify the writing, we define, for each $j = (j_1, j_2, j_3) \in \mathbb{Z}^3$, $\delta_{j,n}$ as j'_n/j'_3 when $j'_3 \neq 0$ and as 0 when $j'_3 = 0$, for $n = 1, 2$. With obvious notations, the trilinear form is then written as

$$
\begin{aligned}
b(U, U^\sharp, U^\flat) &= \sum_{j+l+k=0} i(l'_1 u_j + l'_2 v_j + l'_3 w_j) u_l^\sharp u_k^\flat \\
&\quad + \sum_{j+l+k=0} i(l'_1 u_j + l'_2 v_j + l'_3 w_j) v_l^\sharp v_k^\flat + \sum_{j+l+k=0} i(l'_1 u_j + l'_2 v_j + l'_3 w_j) T_l^\sharp T_k^\flat \\
&= \text{(using the fact that, by definition, } w_j = 0 \text{ for } j_3 = 0) \text{ } (w \text{ is odd in } x_3) \\
&= \sum_{j+l+k=0} i\left[(l'_1 - \delta_{j,1} l'_3) u_j + (l'_2 - \delta_{j,2} l'_3) v_j\right] (u_l^\sharp u_k^\flat + v_l^\sharp v_k^\flat + \kappa T_l^\sharp T_k^\flat). \qquad (3.338)
\end{aligned}
$$

We then compute

$$
\begin{aligned}
&((-\Delta_3)^{1/2} B(U, U^\sharp), (-\Delta_3)^{3/2} U^\flat)_\tau \\
&\quad = \sum_{j+l+k=0} i[(l'_1 - \delta_{j,1} l'_3) u_j + (l'_2 - \delta_{j,2} l'_3) v_j] \mathrm{e}^{2\tau|k'|} |k'|^4 (u_l^\sharp u_k^\flat + v_l^\sharp v_k^\flat + \kappa T_l^\sharp T_k^\flat).
\end{aligned} \qquad (3.339)
$$

We associate to each function u, a function $\check{u}$ defined by

$$
\check{u} = \sum_{j \in \mathbb{Z}^3} \check{u}_j \mathrm{e}^{i(j'_1 x + j'_2 y + j'_3 z)}, \text{ where } \check{u}_j = \mathrm{e}^{\tau|j'|} |u_j|; \qquad (3.340)
$$

we also use similar notations for the other functions.

Since all the terms are similar, we need only to estimate the first sum from Eq. (3.339), denoted by I. We find

$$
|I| \leq c \sum_{j+k+l=0} |l'||j'||k'|^4 e^{2\tau|k'|} |u_j| |u_l^\sharp| |u_k^\flat|, \qquad (3.341)
$$

where we used the estimate $|l'_1 - \delta_{j',1} l'_3| \leq (L_3/2\pi)|j'||l'|$. Since $j + k + l = 0 \Longleftrightarrow j' + k' + l' = 0$, we find $|k'| - |l'| - |j'| \leq 0$ and we have

$$
\begin{aligned}
|I| &\leq \sum_{j+k+l=0} |l'||j'|(|l'| + |j'|)|k'|^3 \check{u}_j \check{u}_l^\sharp \check{u}_k^\flat \\
&\leq \sum_{j+k+l=0} |j'||l'|^2 |k'|^3 \check{u}_j \check{u}_l^\sharp \check{u}_k^\flat + \sum_{j+k+l=0} |j'|^2 |l'||k'|^3 \check{u}_j \check{u}_l^\sharp \check{u}_k^\flat \\
&= \frac{1}{|\mathcal{M}|} \int_{\mathcal{M}} q_1(x) q_2^\sharp(x) q_3^\flat(x) \mathrm{d}\mathcal{M} + \frac{1}{|\mathcal{M}|} \int_{\mathcal{M}} q_2(x) q_1^\sharp(x) q_3^\flat(x) \mathrm{d}\mathcal{M}, \qquad (3.342)
\end{aligned}
$$

where we wrote

$$q_1(x) = \sum_{j\in\mathbb{Z}^3} |j'|\check{u}_j e^{i(j'_1x+j'_2y+j'_3z)}, \quad q_2(x) = \sum_{j\in\mathbb{Z}^3} |j'|^2\check{u}_j e^{i(j'_1x+j'_2y+j'_3z)},$$
$$q_3(x) = \sum_{j\in\mathbb{Z}^3} |j'|^3\check{u}_j e^{i(j'_1x+j'_2y+j'_3z)}, \tag{3.343}$$

and the definitions for $q_i^\sharp$ and $q_i^\flat$ for $i = 1, 2, 3$ are similar.

Using the Hölder and the imbedding inequalities, we find

$$\begin{aligned}
|I| &\le |q_1|_{L^6}|q_2^\sharp|_{L^3}|q_3^\flat|_{L^2} + |q_2|_{L^3}|q_1^\sharp|_{L^6}|q_3^\flat|_{L^2}\\
&\le c|q_1|_{H^1}|q_2^\sharp|_{L^2}^{1/2}|q_2^\sharp|_{H^1}^{1/2}|q_3^\flat|_{L^2} + c|q_2|_{L^2}^{1/2}|q_2|_{H^1}^{1/2}|q_1^\sharp|_{H^1}|q_3^\flat|_{L^2}\\
&\le c|\Delta_3 U|_\tau|\Delta_3 U^\sharp|_\tau^{1/2}|(-\Delta_3)^{3/2}U^\sharp|_\tau^{1/2}|(-\Delta_3)^{3/2}U^\flat|_\tau\\
&\quad + c|\Delta_3 U|_\tau^{1/2}|(-\Delta_3)^{3/2}U|_\tau^{1/2}|\Delta_3 U^\sharp|_\tau|(-\Delta_3)^{3/2}U^\flat|_\tau.
\end{aligned} \tag{3.344}$$

Analog estimates for the other terms yield Lemma 3.15. □

A priori estimates As announced, we want to prove that the solution is analytic in time with values in a certain Gevrey space and that it coincides with the restriction of a complex function in time to the real axis. We thus consider Eq. (3.284) with a complex time $\zeta \in \mathbb{C}$ and U a complex function. We take the complexified spaces H and V denoted as $H_\mathbb{C}$ and $V_\mathbb{C}$,[18] so Eq. (3.183) is rewritten as

$$\begin{aligned}
&\frac{dU}{d\zeta} + AU + B(U, U) + E(U) = F,\\
&U(0) = U_0,
\end{aligned} \tag{3.345}$$

where $\zeta \in \mathbb{C}$ is the complex time.

We consider ζ of the form $\zeta = se^{i\theta}$, where $s > 0$ and $\cos\theta > 0$ so that the real part of ζ is positive. We apply $e^{\varphi(s\cos\theta)(-\Delta_3)^{1/2}}\Delta_3$ to Eq. (3.345) and take the scalar product in $H_\mathbb{C}$ with $e^{\varphi(s\cos\theta)(-\Delta_3)^{1/2}}\Delta_3 U$. We then multiply the resulting equation by $e^{i\theta}$ and take the real part. We find

$$\begin{aligned}
&\mathrm{Re}\, e^{i\theta}\left(e^{\varphi(s\cos\theta)(-\Delta_3)^{1/2}}\Delta_3\frac{dU}{d\zeta},\ \Delta_3 e^{\varphi(s\cos\theta)(-\Delta_3)^{1/2}}U\right)_H\\
&\quad = \frac{1}{2}\frac{d}{ds}|e^{\varphi(s\cos\theta)(-\Delta_3)^{1/2}}\Delta_3 U|_H^2\\
&\qquad + \varphi'(s\cos\theta)\cos\theta\,\mathrm{Re}\, e^{i\theta}(\Delta_3 e^{\varphi(s\cos\theta)(-\Delta_3)^{3/2}}U,\ \Delta_3 e^{\varphi(s\cos\theta)(-\Delta_3)^{1/2}}U)_H\\
&\quad \ge \frac{1}{2}\frac{d}{ds}|\Delta_3 U|_{\varphi(s\cos\theta)}^2 - \cos\theta|(-\Delta_3)^{3/2}U|_{\varphi(s\cos\theta)}|\Delta_3 U|_{\varphi(s\cos\theta)}.
\end{aligned} \tag{3.346}$$

[18] For the scalar products and the norms, we use the same notations as in the real case.

Since a is coercive for our choice of κ and e is antisymmetric, we also find

$$\begin{aligned}
&\mathrm{Re}\, \mathrm{e}^{i\theta}(\mathrm{e}^{\varphi(s\cos\theta)(-\Delta_3)^{1/2}}\Delta_3 AU, \mathrm{e}^{\varphi(s\cos\theta)(-\Delta_3)^{1/2}}\Delta_3 U)_H \\
&\quad + \mathrm{Re}\, \mathrm{e}^{i\theta}(\mathrm{e}^{\varphi(s\cos\theta)(-\Delta_3)^{1/2}}\Delta_3 EU, \mathrm{e}^{\varphi(s\cos\theta)(-\Delta_3)^{1/2}}\Delta_3 U)_H \\
&\quad \geq c_1 \cos\theta |\mathrm{e}^{\varphi(s\cos\theta)(-\Delta_3)^{1/2}}(-\Delta_3)^{3/2}U|_H^2 \\
&\quad = c_1 \cos\theta |(-\Delta_3)^{3/2}U|^2_{\varphi(s\cos\theta)}. \qquad (3.347)
\end{aligned}$$

For the forcing term, we write

$$\begin{aligned}
&|\mathrm{Re}\, \mathrm{e}^{i\theta}(\mathrm{e}^{\varphi(s\cos\theta)(-\Delta_3)^{1/2}}\Delta_3 F, \mathrm{e}^{\varphi(s\cos\theta)(-\Delta_3)^{1/2}}\Delta_3 U)_H| \\
&\quad \leq |(-\Delta_3)^{1/2}F|_{\varphi(s\cos\theta)} |(-\Delta_3)^{3/2}U|_{\varphi(s\cos\theta)} \\
&\quad \leq \frac{c_1}{6}\cos\theta |(-\Delta_3)^{3/2}U|^2_{\varphi(s\cos\theta)} + \frac{1}{c_1\cos\theta}|(-\Delta_3)^{1/2}F|^2_{\varphi(s\cos\theta)}. \qquad (3.348)
\end{aligned}$$

For the bilinear term B, we use Lemma 3.15 and the Young inequality:

$$\begin{aligned}
|\mathrm{Re}\, \mathrm{e}^{i\theta}(\Delta_3 B(U,U),\Delta_3 U)_{\varphi(s\cos\theta)}| &\leq c_2 |\Delta_3 U|^{3/2}_{\varphi(s\cos\theta)} |(-\Delta_3)^{3/2}U|^{3/2}_{\varphi(s\cos\theta)} \\
&\leq \frac{c_1}{6}\cos\theta |(-\Delta_3)^{3/2}U|^2_{\varphi(s\cos\theta)} + \frac{c_3}{(\cos\theta)^3}|\Delta_3 U|^6_{\varphi(s\cos\theta)}. \qquad (3.349)
\end{aligned}$$

Gathering all the estimates above, we find the following differential inequality:

$$\begin{aligned}
\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}s}|\Delta_3 U|^2_{\varphi(s\cos\theta)} + \frac{c_1}{2}\cos\theta |(-\Delta_3)^{3/2}U|^2_{\varphi(s\cos\theta)} &\leq \frac{1}{c_1\cos\theta}|(-\Delta_3)^{1/2}F|^2_{\varphi(s\cos\theta)} \\
&+ \frac{\cos\theta}{c_1}|\Delta_3 U|^2_{\varphi(s\cos\theta)} + \frac{c_3}{(\cos\theta)^3}|\Delta_3 U|^6_{\varphi(s\cos\theta)}. \qquad (3.350)
\end{aligned}$$

We restrict θ so that $\sqrt{2}/2 \leq \cos\theta \leq 1$ (in fact we can restrict θ to any domain such that $\cos\theta \geq c > 0$). Writing

$$y(s) = 1 + |\Delta_3 U(s)|^2_{\varphi(s\cos\theta)},$$

the differential inequality (Eq. (3.350)) becomes:

$$\frac{\mathrm{d}y(s)}{\mathrm{d}s} \leq c(F)y^3(s), \quad 0 < s < t_1, \qquad (3.351)$$

where $c(F)$ is a constant depending as before on the forcing term F. Therefore, by similar reasonings as for the real case, we find that there exists a time $t'_\star$, $0 < t'_\star \leq t_1$, $t'_\star = t'_\star(F, U_0)$ such that

$$|\mathrm{e}^{\varphi(s\cos\theta)(-\Delta_3)^{1/2}}\Delta_3 U(s\mathrm{e}^{i\theta})|^2_H \leq 1 + 2|\Delta_3 U_0|^2_H, \quad \forall\, 0 \leq s \leq t'_\star(F, U_0). \qquad (3.352)$$

Considering the complex region

$$\mathcal{D}(U_0, F, \sigma_1) = \{\zeta = se^{i\theta},\ |\theta| \le \pi/4,\ 0 < s < t'_\star(F, U_0)\}, \tag{3.353}$$

estimate (Eq. (3.352)) gives a bound for $U(\zeta)$, when $\zeta \in \mathcal{D}(U_0, F, \sigma_1)$.

In the previous section, we proved the existence of an absorbing set for U in $\dot{H}^2_{\text{per}}$, so the previous arguments can be reiterated for all times, and we find

$$|e^{\varphi(s\cos\theta)(-\Delta_3)^{1/2}}\Delta_3 U(se^{i\theta})|^2_H \le 1 + 2M, \quad \forall\, s \ge 0, \tag{3.354}$$

where M is the uniform bound for U in H^2 and $\varphi(t) = \min(t, \sigma_1, t'_\star(F, M))$.

We can now state the main result of this section:

THEOREM 3.11. *Let U_0 be given in $\dot{H}^2_{\text{per}}(\mathcal{M})$ and let F be a given function analytic in time with values in $D(e^{\sigma_1(-\Delta_3)^{1/2}}(-\Delta_3)^{1/2})$ for some $\sigma_1 > 0$. Then the unique solution U of Eq. (3.284) is analytic in time on $(0, t^\star)$ with values in $D(e^{\varphi(t)(-\Delta_3)^{1/2}}(-\Delta_3)^{1/2})$ where $\varphi(t) = \min(t, \sigma_1, t'_\star(F, M))$ and $t^\star = \min(\sigma_1, t'_\star(F, M))$ and on $(t^\star, +\infty)$ with values in $D(e^{\sigma_1(-\Delta_3)^{1/2}}(-\Delta_3)^{1/2})$.*

PROOF. The proof is based on the a priori estimates obtained above and the use of the Galerkin–Fourier method (see, e.g., FOIAS and TEMAM [1989]). The solutions of the Galerkin approximation satisfy rigorously the estimates formally derived above, and the bounds are independent of the order j of the Galerkin approximation. We can then pass to the limit $j \to \infty$, using classical results on the convergence of analytic functions. □

3.7. On the backward uniqueness of the PEs

In this section, we consider the PEs of the ocean, in a 2D and then in a 3D domain, with periodic boundary conditions. The question to which we want to respond is as follows: for which kind of solutions can we prove the backward uniqueness, i.e., if two solutions of the PEs (with same forcings F) coincide at a given time $t_* \in \mathbb{R}$, then they coincide at all times $t < t_*$ for which they are defined. Lions and Malgrange (1960) treated the problem of the backward uniqueness for certain parabolic problems and later in BARDOS and TARTAR [1973] they proved in particular that the weak solutions for the 2D Navier-Stokes equations have this property. In this section, we will prove that the 2D PEs possess the backward uniqueness property for a special class of weak solutions that we called the z-weak solutions (see Section 3.4). We also prove that the strong solutions of the 3D PEs possess the backward uniqueness property as well (for more details, see also PETCU [2007]).

We will first treat the backward uniqueness for the 2D PEs, and then we consider the 3D case. We start by proving the existence and uniqueness of z-strong solutions (the strong solutions for which the z derivative is bounded in H^1 for all finite time). The result is needed in the proof of the backward uniqueness for the 2D PEs, but we give it here in general, for both the 2D and 3D cases.

The equations we are working with read

$$\begin{aligned}
&\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x_1} + v\frac{\partial u}{\partial x_2} + w\frac{\partial u}{\partial x_3} - fv + \frac{1}{\rho_0}\frac{\partial p}{\partial x_1} = \nu\Delta_3 u + F_u,\\
&\frac{\partial v}{\partial t} + u\frac{\partial v}{\partial x_1} + v\frac{\partial v}{\partial x_2} + w\frac{\partial v}{\partial x_3} + fu + \frac{1}{\rho_0}\frac{\partial p}{\partial x_2} = \nu\Delta v + F_v,\\
&\frac{\partial p}{\partial x_3} = -\rho g,\\
&\frac{\partial u}{\partial x_1} + \frac{\partial v}{\partial x_2} + \frac{\partial w}{\partial x_3} = 0,\\
&\frac{\partial \rho}{\partial t} + u\frac{\partial \rho}{\partial x_1} + v\frac{\partial \rho}{\partial x_2} + w\frac{\partial \rho}{\partial x_3} - \frac{\rho_0 N^2}{g}w = \mu\Delta_3\rho + F_\rho.
\end{aligned} \tag{3.355}$$

The function spaces for this problem are the same as in the previous section, as well as the symmetry conditions imposed on the solutions.

3.7.1. Existence and uniqueness of z-strong solutions in dimension 3

We recall that in Section 3.4, we proved the existence and uniqueness of what we called the z-weak solutions of the PEs. In what follows, we also need the existence globally in time as well as the uniqueness of z-strong solutions in dimensions 2 and 3, a concept defined below (see the statement of Theorem 3.12). We can prove the following result.

THEOREM 3.12. (z-strong solution in dimensions 2 and 3) *Given $U_0 \in \mathcal{V}$ and $F \in L^\infty(0, T; \mathcal{V})$, there exists a unique solution U of problem (Eq.* (3.355)*), satisfying the initial condition $U(0) = U_0$ and*

$$U \in L^\infty([0, T]; \mathcal{V}) \cap L^2(0, T; \dot{H}^2_{\text{per}}(\mathcal{M})), \quad \frac{\partial U}{\partial x_3} \in L^2(0, T; \dot{H}^2_{\text{per}}(\mathcal{M})). \tag{3.356}$$

PROOF. We start by mentioning that the following reasoning is related to dimension 3; the dimension 2 is similar and much easier.

In Section 3.1, the existence and uniqueness of a strong solution, locally in time, was proved, using the Galerkin approximation. We are now interested in obtaining a priori estimates for the z-strong solution, using the Galerkin method to prove the existence locally in time and the uniqueness of a z-strong solution.

We assume U is a smooth solution to the PEs, and we first derive here some a priori estimates on U_{x_3}. At the end of the proof, we explain how these estimates provide the existence of the z-strong solution, globally in time.

We start by differentiating the evolution equation (Eq. (3.183)) in x_3; we find

$$U'_{x_3} + AU_{x_3} + EU_{x_3} + (B(U, U))_{x_3} = F_{x_3}. \tag{3.357}$$

Multiplying Eq. (3.357) by $-\Delta_3 U_{x_3}$ and integrating over $\mathcal{M}$, we find

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|U_{x_3}\|^2 + c_0|\Delta_3 U_{x_3}|^2 \le & \Big|\int_{\mathcal{M}} u_{x_3}U_{x_1}\cdot\Delta_3 U_{x_3}\mathrm{d}\mathcal{M}\Big| + \Big|\int_{\mathcal{M}} v_{x_3}U_{x_2}\cdot\Delta_3 U_{x_3}\mathrm{d}\mathcal{M}\Big| \\ & + \Big|\int_{\mathcal{M}} w(U)_{x_3}U_{x_3}\cdot\Delta_3 U_{x_3}\mathrm{d}\mathcal{M}\Big| + \Big|\int_{\mathcal{M}} uU_{x_1x_3}\cdot\Delta_3 U_{x_3}\mathrm{d}\mathcal{M} \\ & + \int_{\mathcal{M}} vU_{x_2x_3}\cdot\Delta_3 U_{x_3}\mathrm{d}\mathcal{M} + \int_{\mathcal{M}} w(U)U_{x_3x_3}\cdot\Delta_3 U_{x_3}\mathrm{d}\mathcal{M}\Big| + \Big|\int_{\mathcal{M}} F_{x_3}\Delta_3 U_{x_3}\mathrm{d}\mathcal{M}\Big|.\end{aligned} \tag{3.358}$$

We need to estimate the terms from the right-hand side of Eq. (3.358). The first three terms are similar, so we will estimate just one of them:

$$\begin{aligned}\Big|\int_{\mathcal{M}} u_{x_3}U_{x_1}\cdot\Delta_3 U_{x_3}\mathrm{d}\mathcal{M}\Big| &\le |U_{x_3}|_{L^4}|U_{x_1}|_{L^4}|\Delta_3 U_{x_3}|_{L^2} \\ &\le c|\Delta_3 U_{x_3}||U_{x_1}|^{1/4}\|U_{x_1}\|^{3/4}|U_{x_3}|^{1/4}\|U_{x_3}\|^{3/4} \\ &\le c|\Delta_3 U_{x_3}|\|U\|^{1/4}|U|_{H^2}^{3/4}|U_{x_3}|^{1/4}\|U_{x_3}\|^{3/4} \\ &\le \frac{c_0}{8}|\Delta_3 U_{x_3}|^2 + c|U|_{H^1}^{1/2}|U|_{H^2}^{3/2}|U_{x_3}|^{1/2}\|U_{x_3}\|^{3/2}.\end{aligned} \tag{3.359}$$

By integration by parts, we also find for the other terms:

$$\begin{aligned}\int_{\mathcal{M}}(uU_{x_1x_3}+vU_{x_2x_3}+w(U)U_{x_3x_3})\cdot\Delta_3 U_{x_3}\mathrm{d}\mathcal{M} = & -\int_{\mathcal{M}} u\nabla_3 U_{x_1x_3}\cdot\nabla_3 U_{x_3}\mathrm{d}\mathcal{M} \\ & -\int_{\mathcal{M}} v\nabla_3 U_{x_2x_3}\cdot\nabla_3 U_{x_3}\mathrm{d}\mathcal{M} - \int_{\mathcal{M}} w(U)\nabla_3 U_{x_3x_3}\cdot\nabla_3 U_{x_3}\mathrm{d}\mathcal{M} \\ & -\int_{\mathcal{M}}[(\nabla_3 u\cdot\nabla_3)U_{x_3}]\cdot U_{x_1x_3}\mathrm{d}\mathcal{M} - \int_{\mathcal{M}}[(\nabla_3 v\cdot\nabla_3)U_{x_3}]\cdot U_{x_2x_3}\mathrm{d}\mathcal{M} \\ & -\int_{\mathcal{M}}[(\nabla_3 w(U)\cdot\nabla_3)U_{x_3}]\cdot U_{x_3x_3}\mathrm{d}\mathcal{M}.\end{aligned} \tag{3.360}$$

We first notice that by integration by parts and using the mass conservation, we find

$$\begin{aligned}\int_{\mathcal{M}} u\nabla_3 U_{x_1x_3}\cdot\nabla_3 U_{x_3}\mathrm{d}\mathcal{M} + \int_{\mathcal{M}} v\nabla_3 U_{x_2x_3}\cdot\nabla_3 U_{x_3}\mathrm{d}\mathcal{M} \\ + \int_{\mathcal{M}} w(U)\nabla_3 U_{x_3x_3}\cdot\nabla_3 U_{x_3}\mathrm{d}\mathcal{M} = 0.\end{aligned} \tag{3.361}$$

We need to estimate the remaining terms, which are of two types: containing or not containing $w(U)$. We find

$$\left|\int_{\mathcal{M}}[(\nabla_3 u\cdot\nabla_3)U_{x_3}]\cdot U_{x_1x_3}\mathrm{d}\mathcal{M}\right| \le |\nabla_3 U|_{L^2}|\nabla_3 U_{x_3}|_{L^4}^2 \le c\|U\|\|U_{x_3}\|^{1/2}|U_{x_3}|_{H^2}^{3/2}$$
$$\le \frac{c_0}{8}|\Delta_3 U_{x_3}|_{L^2}^2 + c\|U\|^4\|U_{x_3}\|^2, \tag{3.362}$$

and

$$\int_{\mathcal{M}}[(\nabla_3 w(U)\cdot\nabla_3)U_{x_3}]\cdot U_{x_3x_3}\mathrm{d}\mathcal{M} = \int_{\mathcal{M}'}\int_{-L_3/2}^{L_3/2}[(\nabla_3 w(U)\cdot\nabla_3)U_{x_3}]\cdot U_{x_3x_3}\,\mathrm{d}x_3\mathrm{d}\mathcal{M}'$$
$$\le \int_{\mathcal{M}'}|\nabla_3 w(U)|_{L^\infty_{x_3}}|\nabla_3 U_{x_3}|_{L^2_{x_3}}^2\,\mathrm{d}\mathcal{M}' \le c\int_{\mathcal{M}'}|\Delta_3 U|_{L^2_{x_3}}|\nabla_3 U_{x_3}|_{L^2_{x_3}}^2\,\mathrm{d}\mathcal{M}'$$
$$\le c||\Delta_3 U|_{L^2_{x_3}}|_{L^2(\mathcal{M}')}||\nabla_3 U_{x_3}|_{L^2_{x_3}}|_{L^4(\mathcal{M}')}^2$$
$$\le c|\Delta_3 U|_{L^2(\mathcal{M})}||\nabla_3 U_{x_3}|_{L^2_{x_3}}|_{L^2(\mathcal{M}')}||\nabla_3 U_{x_3}|_{L^2_{x_3}}|_{H^1(\mathcal{M}')}, \tag{3.363}$$

where $\mathcal{M}' = (0, L_1)\times(0, L_2)$.

One can easily show, by direct differentiation and classical estimates, that

$$||\nabla_3 U_{x_3}|_{L^2_{x_3}}|_{H^1(\mathcal{M}')}^2 \le c(|\nabla_3 U_{x_3}|_{L^2(\mathcal{M})}^2 + |U_{x_3}|_{H^2(\mathcal{M})}^2) \le c|U_{x_3}|_{H^2(\mathcal{M})}^2. \tag{3.364}$$

Using Eq. (3.364) in Eq. (3.363), we find

$$\int_{\mathcal{M}}[(\nabla_3 w(U)\cdot\nabla_3)U_{x_3}]\cdot U_{x_3x_3}\mathrm{d}\mathcal{M} \le c|\Delta_3 U||\nabla_3 U_{x_3}||\Delta_3 U_{x_3}|$$
$$\le \frac{c_0}{8}|\Delta_3 U_{x_3}|^2 + c|\Delta_3 U|^2|\nabla_3 U_{x_3}|^2. \tag{3.365}$$

The forcing term is easy to estimate, and gathering all the above estimates, we find

$$\frac{\mathrm{d}}{\mathrm{d}t}\|U_{x_3}\|^2 + c_0|\Delta_3 U_{x_3}|^2 \le f(t)\|U_{x_3}\|^2 + g(t), \tag{3.366}$$

with

$$f(t) = c(\|U\|^2 + |\Delta_3 U|_{L^2}^2), \quad g(t) = |F_{x_3}|_{L^2}^2.$$

Using these a priori estimates and the Galerkin method, we prove that the z-weak solution exists on an interval $(0, t_\star)$, with $t_\star \le T$. But the recent improvements from Cao and Titi [2007] and Kobelkov [2006] showed the existence of a global strong solution (meaning $t_\star = T$), and since the estimates in Eq. (3.366) depend only on ΔU, we conclude that the z-strong solution exists globally in time. □

3.7.2. Backward uniqueness for the z-weak solutions in dimension 2

In what follows, we prove that the z-weak solutions to the 2D PEs have the backward uniqueness property. This means that if two z-weak solutions U_1 and U_2 defined on the interval $[0, T]$ coincide at a point $t_\star \in (0, T)$, then we can conclude that the solutions coincide on the whole interval $[0, t_\star]$. The arguments we use are similar to the case of Navier-Stokes equations considered in LIONS and MALGRANGE [1960] and in BARDOS and TARTAR [1973].

In fact, we can prove the following result.

THEOREM 3.13. (z-weak solutions in dimension 2) *Let F be in $L^2(0, T, \mathcal{V})$ and let U_1, and U_2 be two z-weak solutions to the PEs (Eq. (3.355)), U_1 and U_2 belonging to $\mathcal{C}([0, T]; \mathcal{H}) \cap L^2(0, T, \mathcal{V})$ such that $U_1(t_\star) = U_2(t_\star)$. Then, $U_1 = U_2$ on the interval $[0, t_\star]$.*

Before starting to prove Theorem 3.13, we give the following useful result

PROPOSITION 3.1. *Let F be in $L^2(0, T; \mathcal{V})$ and U_0 in $\mathcal{V}$. Let us also consider U solution to the linear PEs:*

$$\begin{aligned} &U'(t) + AU(t) + EU(t) = F, \\ &U(0) = U_0. \end{aligned} \tag{3.367}$$

For all time t such that $U(t) \neq 0$, we define the following function:

$$\phi(t) = \frac{((A + E)U(t), U(t))_{\mathcal{H}}}{|U(t)|^2_{\mathcal{H}}}. \tag{3.368}$$

Then, ϕ is differentiable for almost every t where it is defined (meaning where $U(t) \neq 0$) and

$$\phi'(t) \leq \frac{|F(t)|^2_{\mathcal{H}}}{|U(t)|^2_{\mathcal{H}}}. \tag{3.369}$$

PROOF. By classical methods, one can immediately show (compared with Theorem 3.12) that the solutions U of the linear PEs satisfy

$$U \in L^\infty(0, T; \mathcal{V}) \cap L^2(0, T; \dot{H}^2_{\text{per}}(\mathcal{M})), \quad \frac{\partial U}{\partial x_3} \in L^2(0, T; \dot{H}^2_{\text{per}}(\mathcal{M})), \quad U \in \mathcal{C}([0, T], \mathcal{H}).$$

We first note that the function ϕ is defined on the open subset of $(0, T)$ where $|U(t)|_{\mathcal{H}} > 0$; the set where $|U(t)|_{\mathcal{H}} > 0$ is open because $U \in \mathcal{C}([0, T], \mathcal{H})$.

Then, all the computations below, performed formally, can be fully justified by using a Galerkin approximation. We first note that since E is an skewsymmetric operator, we have

$$\phi(t) = \frac{((A + E)U(t), U(t))_{\mathcal{H}}}{|U(t)|^2_{\mathcal{H}}} = \frac{(AU(t), U(t))_{\mathcal{H}}}{|U(t)|^2_{\mathcal{H}}}.$$

We find

$$\begin{aligned}\phi'(t) &= 2\frac{< AU'(t), U(t) >_{V',V} + < AU'_{x_3}(t), U_{x_3}(t) >_{V',V}}{|U(t)|^2_{\mathcal{H}}} \\ &\quad - 2\frac{(AU(t), U(t))_{\mathcal{H}}}{|U(t)|^4_{\mathcal{H}}}\{< U'(t), U(t) >_{V',V} + < U'_{x_3}(t), U_{x_3}(t) >_{V',V}\} \\ &= 2\frac{(F - AU(t) - EU(t), AU(t))_{\mathcal{H}}}{|U(t)|^2_{\mathcal{H}}} \\ &\quad - 2\frac{(AU(t), U(t))_{\mathcal{H}}}{|U(t)|^4_{\mathcal{H}}}(F - AU(t) - EU(t), U(t))_{\mathcal{H}} \\ &= 2\frac{(F, AU(t))_{\mathcal{H}}}{|U(t)|^2_{\mathcal{H}}} - 2\frac{|AU(t)|^2_{\mathcal{H}}}{|U(t)|^2_{\mathcal{H}}} - 2\frac{(AU(t), U(t))_{\mathcal{H}}}{|U(t)|^4_{\mathcal{H}}}(F, U(t))_{\mathcal{H}} \\ &\quad + 2\frac{|(AU(t), U(t))_{\mathcal{H}}|^2}{|U(t)|^4_{\mathcal{H}}}, \end{aligned} \tag{3.370}$$

where, in the computations above, we used the fact that

$$< AU(t), EU(t) >_{V',V} = 0.$$

The relation above can be formally checked as follows (rigorous justifications can be derived):

$$\begin{aligned}< AU(t), EU(t) >_{V',V} &= -f\int_{\mathcal{M}}(u\Delta v - v\Delta u)\mathrm{d}\mathcal{M} \\ &\quad - \frac{g}{\rho_0}\int_{\mathcal{M}}\rho w(\Delta U)\mathrm{d}\mathcal{M} + \frac{g}{\rho_0}\int_{\mathcal{M}}\Delta\rho w(U)\mathrm{d}\mathcal{M} \\ &= -\frac{g}{\rho_0}\sum_{l+m=0, l_3\neq 0}|l|^2\rho_l\frac{m_1}{m_3}u_m + \frac{g}{\rho_0}\sum_{l+m=0, m_3\neq 0}\rho_l\frac{m_1}{m_3}|m|^2u_m \\ &= 0, \end{aligned} \tag{3.371}$$

where we used the definition of $w(U)$ as $-k_1/k_3u_k$ for $k_3 \neq 0$, and 0 when $k_3 = 0$.

We have the following relation:

$$\begin{aligned}&|(AU(t),U(t))_{\mathcal{H}}|^2 - (AU(t), U(t))_{\mathcal{H}}(F, U(t))_{\mathcal{H}} + \frac{1}{4}|(F, U(t))_{\mathcal{H}}|^2 \\ &\quad = |(AU(t) - F/2, U(t))_{\mathcal{H}}|^2 \leq |AU(t) - F/2|^2_{\mathcal{H}}|U|^2_{\mathcal{H}}. \end{aligned} \tag{3.372}$$

Continuing to estimate ϕ' in Eq. (3.370), we can conclude

$$\begin{aligned}\phi'(t) \leq & 2\frac{(F, AU(t))_{\mathcal{H}}}{|U(t)|^2_{\mathcal{H}}} - 2\frac{|AU(t)|^2_{\mathcal{H}}}{|U(t)|^2_{\mathcal{H}}} + 2\frac{|AU(t) - F/2|^2_{\mathcal{H}}}{|U(t)|^2_{\mathcal{H}}} - \frac{1}{2}\frac{|(F, U(t))_{\mathcal{H}}|^2}{|U(t)|^4_{\mathcal{H}}} \\ \leq & 2\frac{(F, AU(t))_{\mathcal{H}}}{|U(t)|^2_{\mathcal{H}}} - 2\frac{|AU(t)|^2_{\mathcal{H}}}{|U(t)|^2_{\mathcal{H}}} - \frac{1}{2}\frac{|(F, U(t))_{\mathcal{H}}|^2}{|U(t)|^4_{\mathcal{H}}} \\ & + \frac{2}{|U|^2_{\mathcal{H}}}\{|AU|^2_{\mathcal{H}} - (F, AU(t))_{\mathcal{H}} + \frac{1}{4}|F|^2_{\mathcal{H}}\} \\ \leq & \frac{|F(t)|^2_{\mathcal{H}}}{|U(t)|^2_{\mathcal{H}}}. \end{aligned} \tag{3.373}$$

□

We can now start to prove the main result of this section.

REMARK 3.6. A similar result is also true in dimension 3 but in other spaces. More exactly, let F be in $L^2(0, T; V)$ and U_0 in V. Let us also consider U as the solution to the linear PEs:

$$\begin{aligned} & U'(t) + AU(t) + EU(t) = F, \\ & U(0) = U_0. \end{aligned} \tag{3.374}$$

For all time t such that $U(t) \neq 0$, we define the following function:

$$\phi(t) = \frac{((A + E)U(t), U(t))_H}{|U(t)|^2_H}.$$

Then, ϕ is differentiable for almost all t where it is defined and

$$\phi'(t) \leq \frac{|F(t)|^2_H}{|U(t)|^2_H}. \tag{3.375}$$

Proof of Theorem 3.13. We notice that since U_1 and U_2 are z-weak solutions, U_1 and U_2 belong to $L^2(0, T, \mathcal{V})$, and we can thus find a δ arbitrarily small such that $U_1(\delta)$ and $U_2(\delta)$ belong to $\mathcal{V}$. Considering the PEs having $U_1(\delta)$ and $U_2(\delta)$ as initial condition at $t = \delta$, one obtains, using Theorem 3.12 (for the dimension 2), the existence of z-strong solutions $\tilde{U}_1$ and $\tilde{U}_2$.

By the uniqueness of the solution, we conclude that $\tilde{U}_1 = U_1$ and $\tilde{U}_2 = U_2$ on the interval $[\delta, T]$, so U_1 and U_2 belong to $L^\infty(\delta, T, \mathcal{V}) \cap L^2(\delta, T, \dot{H}^2_{\text{per}}(\mathcal{M}))$, and $\partial U_1/\partial x_3$ and $\partial U_2/\partial x_3$ belong to $L^2(\delta, T, \dot{H}^2_{\text{per}}(\mathcal{M}))$ for $\delta > 0$ arbitrarily small.

We write $U^\sharp = U_1 - U_2$ and $\tilde{U} = U_1 + U_2$. Combining the equations for U_1 and U_2, we find that $U^\sharp$ satisfies the following equation:

$$U^{\sharp\prime} + AU^\sharp + EU^\sharp + \frac{1}{2}B(\tilde{U}, U^\sharp) + \frac{1}{2}B(U^\sharp, \tilde{U}) = 0, \tag{3.376}$$

with $U^\sharp(t_\star) = 0$.

We define the following operator:

$$M(t)U^\sharp = \frac{1}{2}B(\tilde{U}, U^\sharp) + \frac{1}{2}B(U^\sharp, \tilde{U})$$
$$= \frac{1}{2}\left(\tilde{u}\frac{\partial U^\sharp}{\partial x_1} + w(\tilde{U})\frac{\partial U^\sharp}{\partial x_3}\right) + \frac{1}{2}\left(u^\sharp\frac{\partial \tilde{U}}{\partial x_1} + w(U^\sharp)\frac{\partial \tilde{U}}{\partial x_3}\right). \tag{3.377}$$

In what follows, the task is to prove that $\|M(t)\|_{\mathcal{L}(\mathcal{V},\mathcal{H})}$ belongs to $L^2(\delta, T)$. We thus compute

$$\left|w(\tilde{U})\frac{\partial U^\sharp}{\partial x_3}\right|_{L^2} \le |w(\tilde{U})|_{L^4}\left|\frac{\partial U^\sharp}{\partial x_3}\right|_{L^4} \le c|w(\tilde{U})|_{L^2}^{1/2}\|w(\tilde{U})\|^{1/2}\left|\frac{\partial U^\sharp}{\partial x_3}\right|_{L^2}^{1/2}\left\|\frac{\partial U^\sharp}{\partial x_3}\right\|^{1/2}$$
$$\le c\|\tilde{U}\|^{1/2}|\tilde{U}|_{H^2}^{1/2}\|U^\sharp\|_{\mathcal{V}},$$
$$\left|\tilde{u}\frac{\partial U^\sharp}{\partial x_1}\right|_{L^2} \le |\tilde{U}|_{L^\infty}\|U^\sharp\| \le c|\tilde{U}|_{H^2}^{1/2}\|\tilde{U}\|^{1/2}\|U^\sharp\|_{\mathcal{V}}. \tag{3.378}$$

We also find

$$\left|w(U^\sharp)\frac{\partial \tilde{U}}{\partial x_3}\right|_{L^2} \le |w(U^\sharp)|_{L^2}\left|\frac{\partial \tilde{U}}{\partial x_3}\right|_{H^2} \le c\|U^\sharp\|_{\mathcal{V}}\left|\frac{\partial \tilde{U}}{\partial x_3}\right|_{H^2},$$
$$|u^\sharp\frac{\partial \tilde{U}}{\partial x_1}|_{L^2} \le \left|U^\sharp\right|_{L^4}\left|\frac{\partial \tilde{U}}{\partial x_1}\right|_{L^4} \le c\|U^\sharp\|_{\mathcal{V}}\|\tilde{U}\|^{1/2}|\tilde{U}|_{H^2}^{1/2}. \tag{3.379}$$

Gathering the above estimates, we find

$$|M(t)U^\sharp|_{L^2} \le c\|\tilde{U}\|^{1/2}|\tilde{U}|_{H^2}^{1/2}\|U^\sharp\|_{\mathcal{V}} + c\left|\frac{\partial \tilde{U}}{\partial x_3}\right|_{H^2}\|U^\sharp\|_{\mathcal{V}}. \tag{3.380}$$

We now need to estimate the L^2-norm of the x_3-derivative of $M(t)U^\sharp$; in fact, we need to estimate the following expression:

$$2(M(t)U^\sharp)_{x_3} = \tilde{u}_{x_3}\frac{\partial U^\sharp}{\partial x_1} + w(\tilde{U})_{x_3}\frac{\partial U^\sharp}{\partial x_3} + \tilde{u}\frac{\partial^2 U^\sharp}{\partial x_1\partial x_3} + w(\tilde{U})\frac{\partial^2 U^\sharp}{\partial x_3^2}$$
$$+ u^\sharp_{x_3}\frac{\partial \tilde{U}}{\partial x_1} + w(U^\sharp)_{x_3}\frac{\partial \tilde{U}}{\partial x_3} + u^\sharp\frac{\partial^2 \tilde{U}}{\partial x_1\partial x_3} + w(U^\sharp)\frac{\partial^2 \tilde{U}}{\partial x_3^2}, \tag{3.381}$$

and we separately bound each of the terms.

We easily find

$$\left|\tilde{u}_{x_3}\frac{\partial U^\sharp}{\partial x_1}\right|_{L^2} \le |\tilde{u}_{x_3}|_{L^\infty}\left|\frac{\partial U^\sharp}{\partial x_1}\right|_{L^2} \le c|\tilde{u}_{x_3}|_{H^2}\|U^\sharp\|_{\mathcal{V}},$$
$$\left|w(\tilde{U})_{x_3}\frac{\partial U^\sharp}{\partial x_3}\right|_{L^2} \le |\tilde{u}_{x_1}|_{L^4}\left|\frac{\partial U^\sharp}{\partial x_3}\right|_{L^4} \le c|\tilde{U}_{x_1}|_{L^2}^{1/2}\|\tilde{U}_{x_1}\|^{1/2}\|U^\sharp\|_{\mathcal{V}}, \tag{3.382}$$
$$\left|\tilde{u}\frac{\partial^2 U^\sharp}{\partial x_1\partial x_3}\right|_{L^2} \le |\tilde{u}|_{L^\infty}\left|\frac{\partial^2 U^\sharp}{\partial x_1\partial x_3}\right|_{L^2} \le c|\tilde{u}|_{H^2}\|U^\sharp\|_{\mathcal{V}},$$

and

$$\begin{aligned}
&\left|u^\sharp \frac{\partial^2 \tilde{U}}{\partial x_1 \partial x_3}\right|_{L^2} \le c|U^\sharp|_{H^1}\left|\frac{\partial^2 \tilde{U}}{\partial x_1 \partial x_3}\right|_{H^1} \le c\|U^\sharp\|_{\mathcal{V}}\left|\frac{\partial \tilde{U}}{\partial x_3}\right|_{H^2},\\
&\left|u^\sharp_{x_3} \frac{\partial \tilde{U}}{\partial x_1}\right|_{L^2} \le c|U^\sharp_{x_3}|_{H^1}|\tilde{U}|_{H^2} \le c\|U^\sharp\|_{\mathcal{V}}|\tilde{U}|_{H^2},\\
&\left|w(U^\sharp)_{x_3} \frac{\partial \tilde{U}}{\partial x_3}\right|_{L^2} \le |U^\sharp_{x_1}|_{L^2}\left|\frac{\partial \tilde{U}}{\partial x_3}\right|_{L^\infty} \le c\|U^\sharp\|_{\mathcal{V}}\left|\frac{\partial \tilde{U}}{\partial x_3}\right|_{H^2}.
\end{aligned} \tag{3.383}$$

We remain with some more delicate terms to estimate, which need anisotropic estimates,

$$\begin{aligned}
\left|w(\tilde{U})\frac{\partial^2 U^\sharp}{\partial x_3^2}\right|_{L^2} &\le \left||w(\tilde{U})|_{L^\infty_{x_3}}\left|\frac{\partial^2 U^\sharp}{\partial x_3^2}\right|_{L^2_{x_3}}\right|_{L^2_{x_1}} \le c||\tilde{U}_{x_1}|_{L^2_{x_3}}|_{L^\infty_{x_1}}\left|\frac{\partial^2 U^\sharp}{\partial x_3^2}\right|_{L^2}\\
&\le c|\tilde{U}|_{H^2}\left|\frac{\partial^2 U^\sharp}{\partial x_3^2}\right|_{L^2} \le c|\tilde{U}|_{H^2}\|U^\sharp\|_{\mathcal{V}},\\
\left|w(U^\sharp)\frac{\partial^2 \tilde{U}}{\partial x_3^2}\right|_{L^2} &\le \left||w(U^\sharp)|_{L^\infty_{x_3}}\left|\frac{\partial^2 \tilde{U}}{\partial x_3^2}\right|_{L^2_{x_3}}\right|_{L^2_{x_1}} \le c||U^\sharp_{x_1}|_{L^2_{x_1}(L^2_{x_3})}\left|\left|\frac{\partial^2 \tilde{U}}{\partial x_3^2}\right|_{L^\infty_{x_1}(L^2_{x_3})}\right|_{L^2_{x_1}}\\
&\le c\|U^\sharp\|_{\mathcal{V}}\left\|\frac{\partial^2 \tilde{U}}{\partial x_3^2}\right\|_{H^2}.
\end{aligned} \tag{3.384}$$

From the computations above, we can now conclude that

$$|M(t)U^\sharp|_{\mathcal{H}} \le c\{\|\tilde{U}\|^{1/2}|\tilde{U}|^{1/2}_{H^2} + \left|\frac{\partial \tilde{U}}{\partial x_3}\right|_{H^2} + |\tilde{U}|_{H^2}\}\|U^\sharp\|_{\mathcal{V}}. \tag{3.385}$$

Thus, $\|M(t)\|_{\mathcal{L}(\mathcal{V},\mathcal{H})}$ is bounded by the expression between brackets in Eq. (3.385) and we conclude, taking into account the properties of $\tilde{U}$, that $\|M(t)\|_{\mathcal{L}(\mathcal{V},\mathcal{H})}$ belongs to $L^2(\delta, T)$ for $\delta > 0$ arbitrarily small.

We now need to prove that if $|U(t_\star)|_{\mathcal{H}} = 0$, then $|U(t)|_{\mathcal{H}} = 0$ for all $t \in [\delta, t_\star]$, $0 < \delta < t_\star$. The equivalent relation that we prove is that if there exists a time $t \in (\delta, t_\star)$ such that $|U^\sharp(t)|_{\mathcal{H}} > 0$, then $|U^\sharp(t_\star)|_{\mathcal{H}} > 0$. Since we proved that $U^\sharp \in \mathcal{C}([0, T], \mathcal{H})$, it is enough to show that $\log |U^\sharp(t)|_{\mathcal{H}}$ is bounded from below on $[\delta, t_\star]$.

Writing Eq. (3.376) as

$$U^{\sharp\prime} + AU^\sharp + EU^\sharp + M(t)U^\sharp = 0,$$

we can use Proposition 3.1, where ϕ is defined as in Eq. (3.368), for $U^\sharp$. We find

$$\phi'(t) \le \frac{|M(t)U^\sharp(t)|^2_{\mathcal{H}}}{|U^\sharp(t)|^2_{\mathcal{H}}} \le \|M(t)\|^2_{\mathcal{L}(\mathcal{V},\mathcal{H})}\frac{|U^\sharp(t)|^2_{\mathcal{V}}}{|U^\sharp(t)|^2_{\mathcal{H}}} \le \frac{1}{c_0}\|M(t)\|^2_{\mathcal{L}(\mathcal{V},\mathcal{H})}\phi(t); \tag{3.386}$$

in Eq. (3.386), we used the fact that

$$((A+E)U^\sharp(t), U^\sharp(t))_{\mathcal{H}} = (AU^\sharp(t), U^\sharp(t))_{\mathcal{H}} \geq c_0\|U^\sharp\|^2_{\mathcal{V}}.$$

Since $\|M(t)\|_{\mathcal{L}(\mathcal{V},\mathcal{H})}$ belongs to $L^2(\delta, T)$, we can apply the Gronwall lemma to Eq. (3.386) and find

$$\phi(t) \leq \phi(\delta)\exp\left(\int_\delta^t c_0^{-1}\|M(s)\|^2_{\mathcal{L}(\mathcal{V},\mathcal{H})}\,\mathrm{d}s\right) \leq K, \tag{3.387}$$

with K a constant independent of t.

Considering the function $\log|U^\sharp(t)|^2_{\mathcal{H}}$, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(\log|U^\sharp(t)|^2_{\mathcal{H}}) &= 2\frac{(U^\sharp, U^{\sharp\prime})_{\mathcal{H}}}{|U^\sharp(t)|^2_{\mathcal{H}}} = -2\frac{(U^\sharp, (A+E)U^\sharp)_{\mathcal{H}}}{|U^\sharp(t)|^2_{\mathcal{H}}} - 2\frac{(U^\sharp, M(t)U^\sharp)_{\mathcal{H}}}{|U^\sharp(t)|^2_{\mathcal{H}}}\\ &\geq -2\phi(t) - 2c'\|M(t)\|_{\mathcal{L}(\mathcal{V},\mathcal{H})}\phi(t)\end{aligned} \tag{3.388}$$

since we can estimate

$$\begin{aligned}(U^\sharp, M(t)U^\sharp)_{\mathcal{H}} &\leq |U^\sharp|_{\mathcal{H}}|M(t)U^\sharp|_{\mathcal{H}}\\ &\leq |U^\sharp|_{\mathcal{H}}\|M(t)\|_{\mathcal{L}(\mathcal{V},\mathcal{H})}\|U^\sharp\|_{\mathcal{V}}\\ &\leq c'|U^\sharp(t)|^2_{\mathcal{H}}\|M(t)\|_{\mathcal{L}(\mathcal{V},\mathcal{H})}\phi(t).\end{aligned} \tag{3.389}$$

Using Eq. (3.387) in Eq. (3.388), we find that

$$\frac{\mathrm{d}}{\mathrm{d}t}(\log|U^\sharp(t)|^2_{\mathcal{H}}) \geq -2K(1+c'\|M(t)\|_{\mathcal{L}(\mathcal{V},\mathcal{H})}), \tag{3.390}$$

and since $\|M(t)\|_{\mathcal{L}(\mathcal{V},\mathcal{H})}$ is in $L^1(\delta, T)$, we find that

$$\log|U^\sharp(t)|^2_{\mathcal{H}} \geq -2K(t_\star - t) + \log|U^\sharp(\delta)|^2_{\mathcal{H}} \geq K_1, \quad \forall t \in [\delta, t_\star],$$

with K_1 a constant independent of t. This gives that $|U^\sharp(t_\star)|^2_{\mathcal{H}} \neq 0$, which implies that if $|U^\sharp(t_\star)|^2_{\mathcal{H}} = 0$, then $|U^\sharp(t)|^2_{\mathcal{H}} = 0$ on the interval $[\delta, t_\star]$. But we know that this relation can be proved for almost all δ in $[0, t_\star]$, and from the fact that $U^\sharp \in \mathcal{C}([0, T], \mathcal{H})$, the desired result follows.

3.7.3. Backward uniqueness for the strong solutions of the 3D PEs

The purpose of this section is to prove the backward uniqueness for the strong solutions of the 3D PEs (Eq. (3.271)). In Section 3.6, we have shown the existence and uniqueness of the strong solutions, as well as the existence and (forward) uniqueness of very strong solutions (solutions with values in $H^m, m \geq 2$). These results will be used in what follows (see also Petcu [2006]).

The result we will prove here is the following one.

THEOREM 3.14. *Let F be in $L^2(0, T, V)$ and let U_1 and U_2 be two strong solutions to the PEs (Eq.* (3.355)*), U_1 and U_2 belonging to $\mathcal{C}([0, T]; V) \cap L^2(0, T, \dot{H}^2_{\rm per}(\mathcal{M}))$ such that $U_1(t_\star) = U_2(t_\star)$. Then, $U_1 = U_2$ on the interval $[0, t_\star]$.*

The proof of the theorem follows the main steps as in Theorem 3.13, so we only emphasize the points that are different.

Proof of Theorem 3.14: Let U_1 and U_2 be two strong solutions. We can then find a δ arbitrarily small such that $U_1(\delta)$ and $U_2(\delta)$ belong to $\dot{H}^2_{\rm per}(\mathcal{M})$. This implies, with the results of PETCU [2006], that

$$U_1, U_2 \in \mathcal{C}(\delta, T, \dot{H}^2_{\rm per}(\mathcal{M})) \cap L^2(\delta, T, \dot{H}^3_{\rm per}(\mathcal{M})).$$

As in the previous section, we write $U^\sharp = U_1 - U_2$ and $\tilde{U} = U_1 + U_2$. Combining the equations for U_1 and U_2, we find that $U^\sharp$ satisfies the same equation as Eq. (3.376) with $U^\sharp(t_\star) = 0$.

We need again to prove that the operator $M(t)$ defined by

$$\begin{aligned} M(t)U^\sharp &= \frac{1}{2}B(\tilde{U}, U^\sharp) + \frac{1}{2}B(U^\sharp, \tilde{U}) \\ &= \frac{1}{2}\left(\tilde{u}\frac{\partial U^\sharp}{\partial x_1} + w(\tilde{U})\frac{\partial U^\sharp}{\partial x_3}\right) + \frac{1}{2}\left(u^\sharp\frac{\partial \tilde{U}}{\partial x_1} + w(U^\sharp)\frac{\partial \tilde{U}}{\partial x_3}\right) \end{aligned} \tag{3.391}$$

has the property that $|M(t)|_{\mathcal{L}(V,H)}$ belongs to $L^2(\delta, T)$.

Here, we estimate each term of Eq. (3.391) as follows:

$$\begin{aligned} &\left|\tilde{u}\frac{\partial U^\sharp}{\partial x_1}\right|_H \le |\tilde{u}|_{L^\infty}\left|\frac{\partial U^\sharp}{\partial x_1}\right|_H \le c|\tilde{U}|_{H^2}|U^\sharp|_V, \\ &\left|\tilde{v}\frac{\partial U^\sharp}{\partial x_2}\right|_H \le c|\tilde{U}|_{H^2}|U^\sharp|_V, \\ &\left|w(\tilde{U})\frac{\partial U^\sharp}{\partial x_3}\right|_H \le |w(\tilde{U})|_{L^\infty}\left|\frac{\partial U^\sharp}{\partial x_3}\right|_H \le c|\tilde{U}|_{H^3}|U^\sharp|_V, \end{aligned} \tag{3.392}$$

and also

$$\begin{aligned} &\left|u^\sharp\frac{\partial \tilde{U}}{\partial x_1}\right|_H \le |u^\sharp|_{L^4}\left|\frac{\partial \tilde{U}}{\partial x_1}\right|_{L^4} \le c|\tilde{U}|_{H^2}|U^\sharp|_V, \\ &\left|v^\sharp\frac{\partial \tilde{U}}{\partial x_2}\right|_H \le c|\tilde{U}|_{H^2}|U^\sharp|_V, \\ &\left|w(U^\sharp)\frac{\partial \tilde{U}}{\partial x_3}\right|_H \le |w(U^\sharp)|_{L^2}\left|\frac{\partial \tilde{U}}{\partial x_3}\right|_{L^\infty} \le c|\tilde{U}|_{H^3}|U^\sharp|_V. \end{aligned} \tag{3.393}$$

Gathering the estimates above, we find

$$|M(t)U^\sharp|_H \leq c|\tilde{U}|_{H^3}|U^\sharp|_V, \tag{3.394}$$

which implies that $|M(t)|_{\mathcal{L}(V,H)}$ belongs to $L^2(\delta, T)$. We can now perform the same kind of reasoning as in Theorem 3.13 in order to prove the desired result. □

4. Regularity for the elliptic linear problems in geophysical fluid dynamics

We have used many times, in particular, in Section 3 the result of H^2-regularity of the solutions to certain linear elliptic problems. Following the general results from AGMON, DOUGLIS and NIRENBERG [1959, 1964], we know that the solutions to second-order elliptic problems are in H^{m+2} if the right-hand sides of the equations are in H^m, $m \geq 0$, and the other data are in suitable spaces (see also LIONS and MAGENES [1972] for $m < 0$). Results of this type are proved in this section.

There are several specific aspects and several specific difficulties that justify the lengthy and technical developments of this section, which do not allow us to directly refer to the general results of AGMON, DOUGLIS and NIRENBERG [1959, 1964]:

For the whole atmosphere (not studied in detail here) and for the space periodic case studied in Section 3.5, the domains are smooth, making the results of this section easy.

(i) The (linear, stationary) geophysical fluid dynamics (GFD)–Stokes problem (see Section 4.4.1) involves an integral equation (the second equation in Eq. (4.96)), which prevents from a purely local treatment, like for the classical Stokes problem of incompressible fluid mechanics.
(ii) The boundary conditions of the problem can be a combination of Dirichlet, Neumann and/or Robin boundary conditions.
(iii) The domains that we have considered and that we consider in this section are not smooth; they have angles in two dimensions and edges in three dimensions. This is automatically the case for the ocean and for regional atmosphere or ocean problems. For this reason, technique pertaining to the theory of elliptic problems in nonsmooth domains (see GRISVARD [1985], KOZLOV, MAZYA and ROSSMANN [1997], MAZYA and ROSSMANN [1994]) are needed and used here.
(iv) Because the domain is not smooth, only the H^2 regularity is proved here, $m = 0$. The H^3 regularity, $m \geq 1$, is not expected in general.
(v) Another aspect of the study in this section concerns the shape of the ocean or the atmosphere (shallowness). A "small" parameter ε is introduced, the depth being called εh instead of h, $0 < \varepsilon < 1$, and we want to see how the regularity constants (which depend on the domain) depend on ε.

The small depth hypothesis was considered by HU, TEMAM and ZIANE [2003] and is not considered in this chapter. Introducing the parameter ε makes the proofs of this section generally more involved than needed for this chapter. However, these results usefully complement the article by HU, TEMAM and ZIANE [2002] used in HU, TEMAM and ZIANE [2003].

Many of the results presented in this section are new although some related results appeared in ZIANE [1995, 1997]. The results are fairly general, except for the orthogonality condition appearing in Eq. (4.54) (Γ_b orthogonal to Γ_ℓ). This condition is not physically desirable, and it is not either mathematically needed (most likely) as no such condition appears for the regularity theory of elliptic problems in angles or edges (GRISVARD [1985], MAZYA and ROSSMANN [1994], KOZLOV, MAZYA and ROSSMANN [1997]). We believe that it can be removed, but this problem is open. Let us recall also that all the results in Section 3 are valid whenever the necessary H^2-regularity results can be proved.

For the notations, the basic domain under consideration is $\mathcal{M}_\varepsilon$:

$$\mathcal{M}_\varepsilon = \big\{(x, y, z), (x, y) \in \Gamma_i, -\varepsilon h(x, y) < z < 0\big\}.$$

For $\varepsilon = 1$, we recover the domain $\mathcal{M}(\mathcal{M}_1 = \mathcal{M})$ used in Sections 2 and 3. Below, for technical reasons, we introduce auxiliary domains $\widehat{\mathcal{M}}_\varepsilon$ and Q_ε. A number of unspecified constants independent of ε are generically denoted by C_0.

Note also that this section is closer to PDE theory than to geoscience and therefore we stay closer to the PDE notations than to the geoscience notations. Hence, the notations are not necessarily the same as in the rest of the chapter; in particular, we do not use bold faces for vectors and the current point of $\mathbb{R}^2$ or $\mathbb{R}^3$ is denoted as $x = (x_1, x_2)$ or $x = (x_1, x_2, x_3)$ instead of (x, y) or (x, y, z).

4.1. Regularity of solutions of elliptic boundary value problems in cylinder-type domains

We study in this section the H^2 regularity of solutions of elliptic problems of the second order in a cylinder-type domain; the boundary condition is either of Dirichlet or of Neumann type on all the boundary.

Since the domain contains wedges, it is not smooth, and we rely heavily on the results of GRISVARD [1985] about regularity for elliptic problems in nonsmooth domains. However, a convexity assumption of the domain is essential in GRISVARD [1985] that we want to avoid: this section is mainly devoted to the implementation of a suitable technique, corresponding to a tubular (cylindrical) covering of the domain under consideration.

Let $\widehat{\mathcal{M}}_\varepsilon = \{(x_1, x_2, x_3) \in \mathbb{R}^3;\ (x_1, x_2) \in \Gamma_i, -\varepsilon h(x_1, x_2) < x_3 < \varepsilon h(x_1, x_2)\}$, where Γ_i is a bounded open subset of $\mathbb{R}^2$, with $\partial\Gamma_i$ a C^2-curve, and $h : \overline{\Gamma}_i \to \mathbb{R}_+$ is a positive function, $h \in C^4(\overline{\Gamma}_i)$, and there exist two positive constants $\underline{h}$ and $\bar{h}$ such that $\underline{h} \le h(x_1, x_2) \le \bar{h}$ for all (x_1, x_2) in $\overline{\Gamma}_i$. Define the elliptic operator A by

$$Au = -\sum_{k,\ell=1}^{3} \frac{\partial}{\partial x_k}\left(a^{k\ell}(x)\frac{\partial u}{\partial x_\ell}\right) + \sum_{k=1}^{3} b^k(x)\frac{\partial u}{\partial x_k} + c(x)u, \tag{4.1}$$

where the coefficients $a^{k,\ell}$, $k, \ell = 1, 2, 3$ are of class $C^2(\overline{\widehat{\mathcal{M}}_\varepsilon})$, b^k, $k = 1, 2, 3$, are of class $C^1(\overline{\widehat{\mathcal{M}}_\varepsilon})$, and c is of class $C^0(\overline{\widehat{\mathcal{M}}_\varepsilon})$. Furthermore, we assume that A is uniformly

strongly elliptic, i.e., there exists a positive constant α independent of x and ε such that

$$\sum_{k,\ell=1}^{3} a^{k\ell}(x)\xi_k\xi_\ell \geqslant \alpha|\xi|^2 \quad \forall x \in \overline{\widehat{\mathcal{M}}_\varepsilon},\ \forall \xi \in \mathbb{R}^3. \tag{4.2}$$

We also assume that the functions $a^{k\ell}, b^k, c, k, \ell = 1, 2, 3$ are independent of ε. We aim to study the regularity and the dependence on ε of the solutions to the Dirichlet problem

$$\begin{cases} Au = f & \text{in } \widehat{\mathcal{M}}_\varepsilon, \\ u = 0 & \text{in } \partial\widehat{\mathcal{M}}_\varepsilon \end{cases} \tag{4.3}$$

and the solutions of the Neumann problem

$$\begin{cases} Au = f & \text{in } \widehat{\mathcal{M}}_\varepsilon, \\ \frac{\partial u}{\partial n_A} = 0 & \text{on } \partial\hat{\mathcal{M}}_\varepsilon, \end{cases} \tag{4.4}$$

where $\frac{\partial}{\partial n_A}$ denotes the conormal boundary operator defined by

$$\frac{\partial u}{\partial n_A} = \sum_{k,\ell} a^{k\ell}\frac{\partial u}{\partial x_\ell}n_k, \tag{4.5}$$

and $n = (n_1, n_1, n_3)$ denotes the unit vector in the direction of the outward normal to $\partial\widehat{\mathcal{M}}_\varepsilon$. Our goal is to prove the H^2 regularity of solutions to the Dirichlet problem Eq. (4.3) or the Neumann problem Eq. (4.4), and to obtain the dependence on ε of the constant C_ε appearing in the inequality

$$\sum_{k,\ell}\left|\frac{\partial^2 u}{\partial x_k\,\partial x_\ell}\right|^2_{L^2(\widehat{\mathcal{M}}_\varepsilon)} \leq C_\varepsilon|Au|^2_{L^2(\widehat{\mathcal{M}}_\varepsilon)}. \tag{4.6}$$

In fact, we will show that $C_\varepsilon = C_0$ is independent of ε and, more precisely, we will prove the following.

THEOREM 4.1. *Let* $\widehat{\mathcal{M}}_\varepsilon = \{(x_1, x_2, x_3) \in \mathbb{R}^3;\ (x_1, x_2) \in \Gamma_{\mathrm{i}}, -\varepsilon h(x_1, x_2) < x_3 < \varepsilon h(x_1, x_2)\}$, *where* Γ_{i} *is a bounded subset of* $\mathbb{R}^2$, *with* $\partial\Gamma_{\mathrm{i}}$ *a* C^2*-curve, and* $h \in C^4(\Gamma_{\mathrm{i}})$, *and there exists two positive constants* $\underline{h}, \bar{h}$ *such that* $\underline{h} \leq h(x_1, x_2) \leq \bar{h}$ *for all* (x_1, x_2) *in* $\overline{\Gamma}_{\mathrm{i}}$. *Let* $f \in L^2(\widehat{\mathcal{M}}_\varepsilon)$ *and* $u \in H^1_0(\widehat{\mathcal{M}}_\varepsilon)$, *with* $\|u\|_{H^1(\widehat{\mathcal{M}}_\varepsilon)} \leq C_0|f|_{L^2(\widehat{\mathcal{M}}_\varepsilon)}$ *with* C_0 *independent of* ε.

If u satisfies

$$Au = -\sum_{k,\ell=1}^{3}\frac{\partial}{\partial x_k}\left(a^{k\ell}(x)\frac{\partial u}{\partial x_\ell}\right) + \sum_{k=1}^{3} b^k(x)\frac{\partial u}{\partial x_k} + c(x)u = f,$$

where a^{ij}, b^i are in $C^2(\overline{\widehat{\mathcal{M}}_\varepsilon})$, c in $C^1(\overline{\widehat{\mathcal{M}}_\varepsilon})$, and A is uniformly strongly elliptic in the sense of Eq. (4.2), *then $u \in H^2(\widehat{\mathcal{M}}_\varepsilon)$ and there exists a constant C_0 independent of ε such that*

$$\sum_{i,j=1} \left| \frac{\partial^2 u}{\partial x_i \, \partial x_j} \right|^2_{L^2(\widehat{\mathcal{M}}_\varepsilon)} \leq C_0 |f|^2_{L^2(\widehat{\mathcal{M}}_\varepsilon)}. \tag{4.7}$$

Proof. The proof of Theorem 4.1 is divided into four steps.

Step 1 (flattening the top and bottom boundaries) We straighten the bottom and top boundaries and transform the domain $\widehat{\mathcal{M}}_\varepsilon$ into the cylinder $Q_\varepsilon = \Gamma_{\mathrm{i}} \times (-\varepsilon, \varepsilon)$, and the operator A is transformed to an operator $\widetilde{A}$ of the same form and satisfying the same assumptions as A. In fact, let

$$\begin{aligned} &\Psi : (x_1, x_2, x_3) \mapsto (y_1, y_2, y_3), \\ &y_1 = x_1, \quad y_2 = x_2, \quad y_3 = \frac{x_3}{h(x_1, x_2)}. \end{aligned} \tag{4.8}$$

Since $h \in C^4(\overline{\Gamma}_{\mathrm{i}})$, we have $\Psi \in C^4(\overline{\widehat{\mathcal{M}}_\varepsilon})$. Furthermore, we note that Au may be written as

$$Au = -\sum_{k,\ell=1}^{3} a^{k\ell}(x) \frac{\partial^2 u}{\partial x_k \, \partial x_\ell} + \sum_{k=1}^{3} \mathrm{d}^k(x) \frac{\partial u}{\partial x_k} + c(x)u, \tag{4.9}$$

where

$$\mathrm{d}^k(x) = b^k(x) - \sum_{\ell=1}^{3} \frac{\partial a^{\ell k}(x)}{\partial x_\ell} \in C^1(\overline{\widehat{\mathcal{M}}_\varepsilon}). \tag{4.10}$$

Now let $\tilde{u} : Q_\varepsilon \to \mathbb{R}$, $\tilde{u}(y_1, y_2, y_3) = u(x_1, x_2, x_3)$. Since $\Psi \in C^4(\overline{\widehat{\mathcal{M}}_\varepsilon})$ and Ψ is independent of ε, the H^2 norm of $\tilde{u}$ is equivalent to the H^2 norm of u with constants independent of ε. More precisely, there exists a constant C_0 independent of ε such that

$$\begin{aligned} C_0^{-1} \sum_{k,\ell} \left| \frac{\partial^2 \tilde{u}}{\partial y_k \, \partial y_\ell} \right|^2_{L^2(Q_\varepsilon)} &\leq \sum_{k,\ell} \left| \frac{\partial^2 u}{\partial x_k \, \partial x_\ell} \right|^2_{L^2(\widehat{\mathcal{M}}_\varepsilon)} \\ &\leq C_0 \sum_{k,\ell} \left| \frac{\partial^2 \tilde{u}}{\partial y_k \, \partial y_\ell} \right|^2_{L^2(Q_\varepsilon)}. \end{aligned} \tag{4.11}$$

Furthermore, we can easily check that

$$\widetilde{A}\tilde{u}(y) = -\sum_{k,\ell} \tilde{a}^{k\ell}(y) \frac{\partial^2 \tilde{u}}{\partial y_k \, \partial y_\ell} + \sum_{k} \tilde{\mathrm{d}}^k(y) \frac{\partial \tilde{u}}{\partial y_k} + \tilde{c}(y)\tilde{u}(y) = \tilde{f}(y), \tag{4.12}$$

where

$$\begin{cases} \tilde{a}^{k\ell}(y) = \sum_{r,s} \frac{\partial \Psi_k}{\partial x_r} \frac{\partial \Psi_\ell}{\partial x_s} a^{rs}(\Psi^{-1}(y)), \\ \tilde{\mathrm{d}}^k(y) = -\sum_{r,s} \frac{\partial^2 \Psi_k}{\partial x_r \partial x_s} a^{sr}(\Psi^{-1}(y)) + \sum_r \frac{\partial \Psi_k}{\partial x_r} \mathrm{d}^r(\Psi^{-1}(y)), \\ \tilde{c}(y) = c(\Psi^{-1}(y)) \quad \text{and} \quad \tilde{f}(y) = f(\Psi^{-1}(y)). \end{cases} \tag{4.13}$$

It is clear, since $\Psi \in C^4(\overline{\widehat{\mathcal{M}}_\varepsilon})$, that $\tilde{a}^{k,\ell} \in C^2(\overline{Q})$, $\tilde{d}^k \in C^1(\overline{Q})$, and $\tilde{c} \in C^0(\overline{Q}_\varepsilon)$. Finally, if $u = 0$ on $\partial\widehat{\mathcal{M}}_\varepsilon$, then $\tilde{u} = 0$ on ∂Q_ε and if $\partial u/\partial n_A = 0$ on $\partial\widehat{\mathcal{M}}_\varepsilon$, then $\partial\tilde{u}/\partial n_{\widetilde{A}} = 0$ on ∂Q_ε, where, as in Eq. (4.5),

$$\frac{\partial \tilde{u}}{\partial n_{\widetilde{A}}} = \sum_{k,\ell} \tilde{a}^{k\ell} \frac{\partial \tilde{u}}{\partial y_\ell} n_k. \tag{4.14}$$

This is classical, but we include the verification here at the bottom boundary $x_3 = -\varepsilon h$ or equivalently $y_3 = -\varepsilon$. By Eq. (4.5), we have

$$\begin{aligned} &(1 + \varepsilon^2 |\nabla h|^2)^{1/2} \frac{\partial u}{\partial n_A} \\ &\quad = \varepsilon \sum_{j=1}^3 a^{1j} \frac{\partial u}{\partial x_j} \frac{\partial h}{\partial x_1} + \varepsilon \sum_{j=1}^3 a^{2j} \frac{\partial u}{\partial x_j} \frac{\partial h}{\partial x_2} + \sum_{j=1}^3 a^{3j} \frac{\partial u}{\partial x_j}, \end{aligned} \tag{4.15}$$

but $\dfrac{\partial u}{\partial x_j} = \sum_{r=1}^3 \dfrac{\partial \tilde{u}}{\partial y_r} \dfrac{\partial \Psi_r}{\partial x_j}$, and thus

$$\begin{aligned} &\left(1 + \varepsilon^2 |\nabla h|^2\right)^{1/2} \frac{\partial u}{\partial n_A} \\ &\quad = \sum_{r=1}^3 \sum_{j=1}^3 \left(a^{1j} \left(\varepsilon \frac{\partial h}{\partial x_1} \right) + a^{2j} \left(\varepsilon \frac{\partial h}{\partial x_2} \right) + a^{3j} \right) \frac{\partial \Psi_r}{\partial x_j} \frac{\partial \tilde{u}}{\partial y_r}. \end{aligned} \tag{4.16}$$

On the other hand, by definition, we have

$$\frac{\partial \tilde{u}}{\partial n_{\widetilde{A}}} = -\sum_{r=1}^3 \tilde{a}^{3r}(-\varepsilon) \frac{\partial \tilde{u}}{\partial y_r} \quad \text{(the normal is in the direction of } y_3 < 0\text{)}, \tag{4.17}$$

but

$$\tilde{a}^{3r} = \sum_{m,n} \frac{\partial \Psi_3}{\partial x_m} \frac{\partial \Psi_r}{\partial x_n} a^{mn} = \sum_{n=1}^3 \frac{\partial \Psi_r}{\partial x_n} \left(\sum_{m=1}^3 a^{mn} \frac{\partial \Psi_3}{\partial x_m} \right), \tag{4.18}$$

and since $\Psi_3(x_1, x_2, x_3) = x_3/h(x_1, x_2)$, we have

$$\tilde{a}^{3r}(-\varepsilon) = \frac{1}{h} \sum_{j=1}^3 \left(a^{1j} \left(\varepsilon \frac{\partial h}{\partial x_1} \right) + a^{2j} \left(\varepsilon \frac{\partial h}{\partial x_2} \right) + a^{3j} \right) \frac{\partial \Psi_r}{\partial x_j}. \tag{4.19}$$

Hence,

$$\frac{\partial \tilde{u}}{\partial n_{\widetilde{A}}} = 0 \quad \text{at } y_3 = -\varepsilon. \tag{4.20}$$

A similar computation yields $\partial \tilde{u}/\partial n_{\widetilde{A}} = 0$ at $y_3 = \varepsilon$. Now we check the Neumann condition at the lateral boundary. First, write

$$\frac{\partial u}{\partial n_A} = \sum_{k=1}^{2}\sum_{\ell=1}^{3} a^{k\ell}\frac{\partial u}{\partial x_\ell}n_k = \sum_{k=1}^{2}\sum_{\ell,r=1}^{3} a^{k\ell}\frac{\partial \tilde{u}}{\partial y_r}\frac{\partial \Psi_r}{\partial x_\ell}n_k \tag{4.21}$$

and

$$\frac{\partial \tilde{u}}{\partial n_{\widetilde{A}}} = \sum_{k=1}^{2}\sum_{\ell=1}^{3} \tilde{a}^{k\ell}\frac{\partial \tilde{u}}{\partial y_\ell}n_k = \sum_{k=1}^{2}\sum_{\ell,r,s=1}^{3} a^{sr}\frac{\partial \Psi_k}{\partial x_s}\frac{\partial \Psi_\ell}{\partial x_r}\frac{\partial \tilde{u}}{\partial y_\ell}n_k, \tag{4.22}$$

but since for $k = 1, 2$, $\Psi_k(x_1, x_2, x_3) = x_k$, we have $\partial\Psi_k/\partial x_s = \delta_{sk}$ (the Kronecker symbol). Hence,

$$\frac{\partial \tilde{u}}{\partial n_{\widetilde{A}}} = \sum_{k,\ell,r=1}^{3} a^{kr}\frac{\partial \Psi_\ell}{\partial x_r}\frac{\partial \tilde{u}}{\partial y_\ell}n_k. \tag{4.23}$$

Interchanging ℓ and r, we obtain $\partial \tilde{u}/\partial n_{\widetilde{A}} = \partial u/\partial n_A = 0$ on the lateral boundary. From now on, we concentrate on the Dirichlet boundary condition and the Neumann condition follows in the same manner.

STEP 2 (interior regularity) Let B_R be an open ball, with $B_R \subset\subset \Gamma_i$; without loss of generality, we assume that B_R is centered at 0. By Step 1, we may assume that $\widehat{\mathcal{M}}_\varepsilon$ is a right cylinder, i.e., $\widehat{\mathcal{M}}_\varepsilon = Q_\varepsilon = \Gamma_i \times (-\varepsilon, \varepsilon)$. Now let $\theta \in C_0^\infty(B_R)$ (θ independent of x_3), with $\theta \equiv 1$ in $B_{R/2}$ and $0 \le \theta \le 1$. Then

$$\widetilde{A}(\theta\tilde{u}) = \theta\tilde{f} + E_\theta\tilde{u},$$

where E_θ is a first-order differential operator, which implies that $|E_\theta\tilde{u}|_{L^2(Q_\varepsilon)} \le C_0|f|_\varepsilon$, where C_0 is independent on ε, and $|f|_\varepsilon$ is an alternate notation for $|f|_{L^2(\widehat{\mathcal{M}}_\varepsilon)}$. Hence,

$$\widetilde{A}(\theta\tilde{u}) = f_{\text{loc}}, \quad \text{with } |f_{\text{loc}}|_{L^2(B_R\times(-\varepsilon,\varepsilon))} \le C_0|f|_\varepsilon,$$

and for either boundary condition (Dirichlet, Neumann), $\theta\tilde{u} = 0$ on $\partial(B_R) \times (-\varepsilon, \varepsilon)$, and in the Dirichlet case, $\theta\tilde{u} = 0$ on $B_R \times (-\varepsilon, \varepsilon)$, and in the Neumann case, $\partial(\theta\tilde{u})/\partial z = 0$ on $B_R \times (-\varepsilon, \varepsilon)$.

We quote the following theorems from GRISVARD [1985], and we start first with the case of the Dirichlet boundary condition.

THEOREM 4.2 (GRISVARD [1985], Theorem 3.2.1.2). *Let Ω be a convex, bounded, and open subset of $\mathbb{R}^n$. Then, for each $f \in L^2(\Omega)$, there exists a unique $u \in H^2(\Omega)$, the solution of $Au = f$ in Ω, $u = 0$ on $\partial\Omega$.*

The proof of the theorem, given in [Grisvard, 1985, pp. 148–149], is based on a priori bounds for solutions in $H^2(\Omega)$. These bounds depend in the case of general domains on the curvature of $\partial\Omega$; however in the case of a convex domain, the curvature is negative and the constants in the bounds are, therefore, independent of the domain.

Similarly, in the case of Neumann boundary condition, we have the following theorem.

Theorem 4.3 (Grisvard [1985], Theorem 3.2.1.3). *Let Ω be a convex, bounded, and open subset of $\mathbb{R}^n$. Then for each $f \in L^2(\Omega)$ and for each $\lambda > 0$, there exists a unique $u \in H^2(\Omega)$, the solution of*

$$-\sum_{k,\ell=1}^{3} \int_\Omega a^{k\ell}(x) \frac{\partial u}{\partial x_\ell} \frac{\partial v}{\partial x_k} \, \mathrm{d}x + \lambda \int_\Omega uv \, \mathrm{d}x = \int_\Omega fv \, \mathrm{d}x \tag{4.24}$$

for all $v \in H^1(\Omega)$.

We note that Eq. (4.24) is the weak form of the Neumann problem for the equation

$$-\sum_{k,\ell=1}^{3} \frac{\partial}{\partial x_k}\left(a^{k\ell}(x)\frac{\partial u}{\partial x_\ell}\right) + \lambda u = f \quad \text{in } \Omega$$

together with the boundary condition $\frac{\partial u}{\partial n_A} = 0$ on $\partial\Omega$. Again, here, the convexity of the domain implies that the curvature of the boundary of the domain is negative and therefore the constants in the bounds on the L^2 norm of the mixed second derivatives in terms of the L^2 norm of f are independent of the domain (for more details, see Grisvard [1985]).

Now, we use Theorem A above by first rewriting $\widetilde{A}$ in a divergence form and moving the extra terms to the right-hand side. As above, since $\theta\tilde{u} \in H^1(B_R \times (-\varepsilon, \varepsilon))$, the extra terms are in $L^2(B_R \times (-\varepsilon, \varepsilon))$ and

$$\sum_{i,j} \left| \frac{\partial^2(\theta\tilde{u})}{\partial y_i \, \partial y_j} \right|^2_{L^2(B_R \times (-\varepsilon,\varepsilon))} \leq C_0 |f|^2_{L^2(\widehat{\mathcal{M}}_\varepsilon)}.$$

Finally, since $\theta \equiv 1$ in $B_{R/2}$, we have $\tilde{u} \in H^2(B_{R/2} \times (-\varepsilon, \varepsilon))$ and

$$\sum_{i,j} \left| \frac{\partial^2\tilde{u}}{\partial y_i \, \partial y_j} \right|^2_{L^2(B_{R/2} \times (-\varepsilon,\varepsilon))} \leq C_0 |f|^2_{L^2(\widehat{\mathcal{M}}_\varepsilon)}.$$

Step 3 (boundary regularity) Let $\mathbb{R}^2_+ = \{x \in \mathbb{R}^2;\, x_2 > 0\}$ and let $B_r^+ = \{x \in \mathbb{R}^2_+;\, |x| < r\}$ be the open half-ball with center at the origin and radius r contained in $\mathbb{R}^2$. By the assumption on Γ_{i}, for all $x^0 \in \partial\Gamma_{\mathrm{i}}$, there exists a neighborhood V of x^0 in $\mathbb{R}^2$ and a diffeomorphism $\widetilde{\Psi}$ such that

$$\widetilde{\Psi}(V \cap \Gamma_{\mathrm{i}}) = B_r^+,$$
$$\widetilde{\Psi}(x^0) = 0.$$

Using the diffeomorphism $\widetilde{\Psi}$, we can construct a (tubular) diffeomorphism Ψ in $\mathbb{R}^3$ such that

$$\Psi(V \times (-\varepsilon, \varepsilon) \cap Q_\varepsilon) = B_r^+ \times (-\varepsilon, \varepsilon),$$

by setting $\Psi_i(y_1, y_2, y_3) = \widetilde{\Psi}_i(y_1, y_2)$, $i = 1, 2$, and $\Psi_3(y_1, y_2, y_3) = y_3$. Following the same procedure as in Steps 1 and 2, let W be an open set of $\mathbb{R}^2$ containing x^0 such that $\overline{W} \subset V$ and let $\theta \in C_0^\infty(V)$ be such that $0 \le \theta \le 1$ and $\theta \equiv 1$ in W. Then,

$$\widetilde{A}(\theta\tilde{u}) = f_{\text{loc}}, \quad \text{with } |f_{\text{loc}}|_{L^2(V\times(-\varepsilon,\varepsilon)\cap Q_\varepsilon)} \le C_0 |f|_\varepsilon,$$

and in the case of the Dirichlet boundary condition, $\theta\tilde{u} = 0$ on $\partial(V \times (-\varepsilon, \varepsilon) \cap Q_\varepsilon)$. Next we use the transformation Ψ that is independent of ε and transforms the domain $V \times (-\varepsilon, \varepsilon) \cap Q_\varepsilon$ into $B_r^+ \times (-\varepsilon, \varepsilon)$, $\theta\tilde{u}$ into u^*, and $\widetilde{A}$ into A^*, with A^*u^* given as in Step 1. Now, $u^* = 0$ on $\partial(B_r^+ \times (-\varepsilon, \varepsilon))$ and $B_r^+ \times (-\varepsilon, \varepsilon)$ is convex; hence rewriting A^* in a divergence form, we obtain using Grisvard [1985], $u^* \in H^2(B_r^+ \times (-\varepsilon, \varepsilon))$ and

$$\sum_{i,j} \left| \frac{\partial^2 u^*}{\partial z_i \, \partial z_j} \right|^2_{L^2(B_R^+\times(-\varepsilon,\varepsilon))} \le C_0 |f|_\varepsilon^2.$$

Going back to $V \times (-\varepsilon, \varepsilon) \cap Q_\varepsilon$ using Ψ^{-1}, we obtain

$$\sum_{i,j} \left| \frac{\partial^2 (\theta\tilde{u})}{\partial y_i \, \partial y_j} \right|^2_{L^2(V\times(-\varepsilon,\varepsilon)\cap Q_\varepsilon)} \le C_0 |f|_\varepsilon^2.$$

Hence,

$$\sum_{i,j} \left| \frac{\partial^2 \tilde{u}}{\partial y_i \, \partial y_j} \right|^2_{L^2(V\times(-\varepsilon,\varepsilon)\cap Q_\varepsilon)} \le C_0 |f|_\varepsilon^2.$$

Step 4 (partition of unity and conclusion) Let $V_0, V_1, \ldots, V_N$ and $W_0, \ldots, W_N$ be two finite open coverings of $\overline{\Gamma_i}$ satisfying $\overline{V_0} \subset \Gamma_i$; V_k, $k \ge 1$ is contained in the domain of a local map $\widetilde{\Psi}_{(k)}$ such that

$$\widetilde{\Psi}_{(k)}(V_k \cap \Gamma_i) = B_r^+,$$
$$W_0 = V_0,$$
$$\overline{W}_k \subset V_k, \quad \text{for all } k \ge 1.$$

Finally, let $\{\varphi_k\}_k$ be a partition of unity subordinated to the covering $\{W_k\}_k$ of Γ_i. Then, $\tilde{u} = \sum_{k=1}^N \varphi_k \tilde{u}$ and by Steps 1–3, $\varphi_k\tilde{u} \in H^2(Q_\varepsilon)$ and

$$\sum_{k=0}^{N} \sum_{i,j=1}^{3} \left| \frac{\partial^2 (\varphi_k \tilde{u})}{\partial y_i \, \partial y_j} \right|^2_{L^2(Q_\varepsilon)} \le C_0 |f|_\varepsilon^2.$$

Therefore, $\tilde{u} \in H^2(Q_\varepsilon)$ and

$$\sum_{i,j=1}^{3} \left| \frac{\partial^2 \tilde{u}}{\partial y_i \, \partial y_j} \right|^2_{L^2(Q_\varepsilon)} \leq C_0 |f|^2_\varepsilon.$$

Finally, we go back to the domain $\widehat{\mathcal{M}}_\varepsilon$ and conclude that $u \in H^2(\widehat{\mathcal{M}}_\varepsilon)$ and

$$\sum_{i,j=1}^{3} \left| \frac{\partial^2 u}{\partial x_i \, \partial x_j} \right|^2_{L^2(\widehat{\mathcal{M}}_\varepsilon)} \leq C_0 |f|^2_\varepsilon.$$

Theorem 4.1 is proved. □

4.2. Regularity of solutions of a Dirichlet–Robin mixed boundary value problem

We now want to derive a result similar to that of Section 4.1 for a boundary value problem with mixed Dirichlet–Robin boundary conditions, the elliptic operator being the same as in Eq. (4.1). The proof consists in reducing the boundary condition to a full Dirichlet boundary condition and then use Theorem 4.1.

From now on, we will consider the actual domain

$$\mathcal{M}_\varepsilon = \big\{(x_1, x_2, x_3) \in \mathbb{R}^3;\, (x_1, x_2) \in \Gamma_{\mathrm{i}},\, -\varepsilon h(x_1, x_2) < x_3 < 0\big\};$$

$|\cdot|_\varepsilon$ will denote the norm in $L^2(\mathcal{M}_\varepsilon)$ (or product of such spaces), and $|\cdot|_i$ will denote the norm in $L^2(\Gamma_{\mathrm{i}})$ (or product of such spaces), $\|g\|_i = |\nabla g|_i$.

We will prove the following theorem.

THEOREM 4.4. *Assume that Γ_{i} is an open bounded set of $\mathbb{R}^2$, with C^3-boundary $\partial\Gamma_{\mathrm{i}}$, and $h \in C^4(\Gamma_{\mathrm{i}})$. Then, for $f \in L^2(\mathcal{M}_\varepsilon)$ and $g \in H^1_0(\Gamma_{\mathrm{i}})$, there exists a unique $\Psi \in H^2(\mathcal{M}_\varepsilon)$ solution of*

$$\begin{cases} -\Delta_3 \Psi = f & \text{in } \mathcal{M}_\varepsilon, \\ \frac{\partial \Psi}{\partial x_3} + \alpha \Psi = g & \text{on } \Gamma_{\mathrm{i}}, \\ \Psi = 0 & \text{on } \Gamma_{\mathrm{b}} \cup \Gamma_\ell. \end{cases} \tag{4.25}$$

Furthermore, there exists a constant $C = C(h, \Gamma_{\mathrm{i}}, \alpha)$ independent of ε such that

$$\sum_{i,j=1}^{3} \left| \frac{\partial^2 \Psi}{\partial x_i \, \partial x_j} \right|^2_\varepsilon \leq C(h, \Gamma_{\mathrm{i}}, \alpha)\big[|f|^2_\varepsilon + \|g\|^2_{H^1(\Gamma_{\mathrm{i}})}\big]. \tag{4.26}$$

PROOF. The proof is divided into several steps.

First we construct a function Ψ^* satisfying the boundary conditions in Eq. (4.25) and find the precise dependence on ε of the L^2 norm of the second-order derivatives of Ψ^* (see Lemma 4.1). Then, we set

$$\widehat{\Phi} = \mathrm{e}^{\alpha x_3}(\Psi - \Psi^*) \tag{4.27}$$

and verify that $\widehat{\Phi}$ satisfies the homogeneous Neumann condition on Γ_i ($\partial\widehat{\Phi}/\partial x_3 = 0$ on Γ_i) and the homogeneous Dirichlet boundary condition on $\Gamma_\ell \cup \Gamma_b$ ($\widehat{\Phi} = 0$ on $\Gamma_\ell \cup \Gamma_b$). By a reflection argument, we extend f to $x_3 > 0$ to be an even function and consider a homogeneous Dirichlet problem on the extended domain $\widehat{\mathcal{M}}_\varepsilon = \{(x_1, x_2, x_3);\ (x_1, x_2) \in \Gamma_i,\ -\varepsilon h(x_1, x_2) < x_3 < \varepsilon h(x_1, x_2)\}$, the solution of which, $\widehat{W}$, coincides with $\widehat{\Phi}$ on $\mathcal{M}_\varepsilon$. Finally, we invoke Theorem 4.1 to conclude the H^2 regularity of $\widehat{W}$ and thus of $\widehat{\Phi}$, along with an estimate of the type (Eq. (4.26)) for $\widehat{\Phi}$; we, therefore, obtain the H^2 regularity of Ψ and Eq. (4.26) by simply using Eq. (4.27).

Thus the whole proof of Theorem 4.4 hinges on the following lifting lemma.

LEMMA 4.1. *Let $h \in C^4(\overline{\Gamma}_i)$ and $g \in H_0^1(\Gamma_i)$. There exists $\Psi^* \in H^2(\mathcal{M}_\varepsilon)$ such that $\frac{\partial\Psi^*}{\partial x_3} + \alpha\Psi^* = g$ on Γ_i, $\Psi^* = 0$ on $\Gamma_\ell \cup \Gamma_b$, and there exists a constant $C = C(h, \Gamma_i)$ independent of ε such that for $0 < \varepsilon \leq 1$,*

$$\sum_{i,j} \left| \frac{\partial^2 \Psi^*}{\partial x_k\, \partial x_j} \right|_\varepsilon^2 \leq C(h, \Gamma_i) \|g\|^2_{H^1(\Gamma_i)}. \tag{4.28}$$

PROOF. First, we construct a function $\widetilde{\Psi}$ as a solution of the heat equation with $-x_3$ corresponding to time

$$\begin{cases} \frac{\partial\widetilde{\Psi}}{\partial x_3} = -\Delta\widetilde{\Psi} & \text{in } \Gamma_i \times (-\infty, 0), \\ \widetilde{\Psi} = 0 & \text{on } \partial\Gamma_i \times (-\infty, 0), \\ \widetilde{\Psi}(x_1, x_2, 0) = g(x_1, x_2) & \text{on } \Gamma_i. \end{cases} \tag{4.29}$$

Here, $\Delta = \Delta_2 = (\partial^2/\partial x_1^2 + \partial^2/\partial x_2^2)$ and, below, $\nabla = \nabla_2 = (\partial/\partial x_1, \partial/\partial x_2)$. The function Ψ^* is then constructed as

$$\Psi^*(x_1, x_2, x_3) = \mathrm{e}^{-\alpha x_3} \int_{-\varepsilon h(x_1, x_2)}^{x_3} \widetilde{\Psi}(x_1, x_2, z)\, \mathrm{d}z. \tag{4.30}$$

It is clear that $\Psi^* \equiv 0$ on $\Gamma_\ell \cup \Gamma_b$, and $\partial\Psi^*/\partial x_3 + \alpha\Psi^* = \mathrm{e}^{-\alpha x_3}\widetilde{\Psi}(x_1, x_2, x_3)$ in $\Gamma_i \times (-\infty, 0)$, which implies that $\partial\Psi^*/\partial x_3 + \alpha\Psi^* = g$ on Γ_i. We only need to check that $\Psi^* \in H^2(\mathcal{M}_\varepsilon)$ and that the inequality (Eq. (4.28)) is valid. This will be done using the classical energy estimates on the solution of the heat equation, which are recalled in Lemmas 4.2 and 4.3.

We note that for $k = 1, 2$,

$$\mathrm{e}^{\alpha x_3} \frac{\partial\Psi^*}{\partial x_k} = \int_{-\varepsilon h(x_1, x_2)}^{x_3} \frac{\partial\widetilde{\Psi}}{\partial x_k}(x_1, x_2, z)\, \mathrm{d}z + \varepsilon \frac{\partial h}{\partial x_k} \widetilde{\Psi}\big(x_1, x_2, -\varepsilon h(x_1, x_2)\big), \tag{4.31}$$

and for $k, j = 1, 2$,

$$\begin{aligned}
\mathrm{e}^{\alpha x_3}\frac{\partial^2\Psi^*}{\partial x_k\,\partial x_j} &= \int_{-\varepsilon h(x_1,x_2)}^{x_3}\frac{\partial^2\widetilde{\Psi}}{\partial x_k\,\partial x_j}(x_1,x_2,z)\,\mathrm{d}z \\
&\quad + \varepsilon\frac{\partial h}{\partial x_j}\frac{\partial\widetilde{\Psi}}{\partial x_k}(x_1,x_2,-\varepsilon h(x_1,x_2)) \\
&\quad + \varepsilon\frac{\partial h}{\partial x_k}\frac{\partial\widetilde{\Psi}}{\partial x_j}(x_1,x_2,-\varepsilon h(x_1,x_2)) \\
&\quad + \varepsilon\frac{\partial^2 h}{\partial x_k\,\partial x_j}\widetilde{\Psi}(x_1,x_2,-\varepsilon h(x_1,x_2)) \\
&\quad - \varepsilon^2\frac{\partial h}{\partial x_k}\frac{\partial h}{\partial x_j}\frac{\partial\widetilde{\Psi}}{\partial x_3}(x_1,x_2,-\varepsilon h(x_1,x_2)).
\end{aligned} \tag{4.32}$$

Here, we need bounds on the L^2 norm (on Γ_i) of $(\partial\widetilde{\Psi}/\partial x_k)(x_1, x_2, -\varepsilon h(x_1, x_2))$ and $(\partial\widetilde{\Psi}/\partial x_3)(x_1, x_2, -\varepsilon h(x_1, x_2))$, which are provided by Lemma 4.3. We have

$$\begin{aligned}
&\int_{\Gamma_\mathrm{i}}\left|\widetilde{\Psi}(x_1,x_2,-\varepsilon h(x_1,x_2))\right|^2\mathrm{d}x_1\,\mathrm{d}x_2 \le C_0\|g\|_i^2, \\
&\int_{\Gamma_\mathrm{i}}\left|\nabla\widetilde{\Psi}(x_1,x_2,-\varepsilon h(x_1,x_2))\right|^2\mathrm{d}x_1\,\mathrm{d}x_2 \le C_0\|g\|_i^2, \\
&\int_{\Gamma_\mathrm{i}}\left|\frac{\partial\widetilde{\Psi}}{\partial x_3}(x_1,x_2,-\varepsilon h(x_1,x_2))\right|^2\mathrm{d}x_1\,\mathrm{d}x_2 \le \frac{1}{\varepsilon}C_0\|g\|_i^2.
\end{aligned} \tag{4.33}$$

Now, using Eqs. (4.30)–(4.32) and (4.33), we obtain

$$\left|\frac{\partial^2\Psi^*}{\partial x_k\,\partial x_j}\right|_\varepsilon^2 \le C_0\|g\|_i^2, \quad k, j = 1, 2. \tag{4.34}$$

Similar relations hold for Ψ^* and $\nabla\Psi^*$, using Eqs. (4.30) and (4.31). Furthermore, since $\partial\Psi^*/\partial x_3 = -\alpha\Psi^* + \mathrm{e}^{-\alpha x_3}\widetilde{\Psi}$, we have

$$\left|\frac{\partial\Psi^*}{\partial x_3}\right|_\varepsilon^2 \le 2\alpha^2\left|\Psi^*\right|_\varepsilon^2 + 2\mathrm{e}^{2\alpha\bar{h}}\left|\widetilde{\Psi}\right|_\varepsilon^2 \le \varepsilon^2 C(h)\|g\|_i^2, \tag{4.35}$$

$$\left|\frac{\partial}{\partial x_3}\nabla\Psi^*\right|_\varepsilon^2 \le 2\alpha^2\left|\nabla\Psi^*\right|_\varepsilon^2 + 2\mathrm{e}^{2\alpha\bar{h}}\left|\nabla\widetilde{\Psi}\right|_\varepsilon^2 \le \varepsilon^2 C(h)\|g\|_i^2. \tag{4.36}$$

Finally,

$$\frac{\partial^2\Psi^*}{\partial x_3^2} = -\alpha\frac{\partial\Psi^*}{\partial x_3} - \alpha\mathrm{e}^{-\alpha x_3}\widetilde{\Psi} + \mathrm{e}^{-\alpha x_3}\frac{\partial\widetilde{\Psi}}{\partial x_3} \tag{4.37}$$

implies

$$\left|\frac{\partial^2 \Psi^*}{\partial x_3^2}\right|_\varepsilon^2 \le C(h)\|g\|_i^2. \tag{4.38}$$

The proof of Lemma 4.1 is complete. □

The proof of Theorem 4.4 relied on estimates given by Lemmas 4.2 and 4.3, which we now state and prove.

LEMMA 4.2 (estimates on solutions of the heat equation). *Let* Ψ *be the solution of the heat equation*

$$\begin{aligned} &\frac{\partial \Psi}{\partial x_3} = -\Delta \Psi \quad \textit{in } \Gamma_{\mathrm{i}} \times (-\infty, 0), \\ &\Psi(x_1, x_2, 0) = g(x_1, x_2) \quad \textit{on } \Gamma_{\mathrm{i}}, \end{aligned} \tag{4.39}$$

with either Dirichlet or Neumann boundary condition, and

$$g \in H_0^1(\Gamma_{\mathrm{i}}) \quad \textit{and} \quad \Psi = 0 \quad \textit{on } \partial\Gamma_{\mathrm{i}} \times (-\infty, 0)$$

or

$$g \in H^1(\Gamma_{\mathrm{i}}) \quad \textit{and} \quad \frac{\partial \Psi}{\partial n_{\Gamma_{\mathrm{i}}}} = 0 \quad \textit{on } \partial\Gamma_{\mathrm{i}} \times (-\infty, 0).$$

Then,

$$\begin{aligned} &\frac{1}{2}|x_3|^k \left|\frac{\partial^j \Psi}{\partial x_3^j}\right|_i^2 (x_3) + \int_{x_3}^0 |z|^k \left|\nabla \frac{\partial^j \Psi}{\partial x_3^j}\right|_i^2 (z)\, \mathrm{d}z \\ &\quad \le \begin{cases} \frac{1}{2}|g|_i^2 & \textit{for } k = j = 0, \\ C|x_3|^{k-2j+1}\|g\|_i^2 & \textit{for } k \ge 2j-1,\ j \ge 1, \end{cases} \end{aligned} \tag{4.40}$$

and

$$\begin{aligned} &\frac{1}{2}|x_3|^k \left|\nabla \frac{\partial^j \Psi}{\partial x_3^j}\right|_i^2 (x_3) + \int_{x_3}^0 |z|^k \left|\frac{\partial^{j+1} \Psi}{\partial x_3^{j+1}}\right|_i^2 (z)\, \mathrm{d}z \\ &\quad \le \begin{cases} \frac{1}{2}\|g\|_i^2 & \textit{for } k = 0,\ j = 0, \\ C|x_3|^{k-2j}\|g\|_i^2 & \textit{for } k \ge 2j,\ j \ge 1. \end{cases} \end{aligned} \tag{4.41}$$

□

In Eqs. (4.40) and (4.41), C is a constant depending on k, j, and h. As before, $\nabla = \nabla_2 = (\partial/\partial x_1, \partial/\partial x_2)$.

PROOF OF LEMMA 4.2. Denote by $e_{k,j}$ and $f_{k,j}$ the left-hand sides of Eqs. (4.40) and (4.41).

We differentiate Eq. (4.39) j times in x_3, multiply the resulting equation by $x_3^k\,\partial^j\Psi/\partial x_3^j$, and integrate over Γ_{i}. Using Stokes formula and observing that $\partial^j\Psi/\partial x_3^j$ satisfies the same boundary condition on $\partial\Gamma_{\mathrm{i}}$ as Ψ, we obtain after multiplication by $(-1)^k$:

$$e_{k,j} = \begin{cases} \frac{k}{2}\int_{x_3}^0 |z|^{k-1}\big|\frac{\partial^j\Psi}{\partial x_3^j}\big|_i^2(z)\,\mathrm{d}z & \text{for } k \geq 1, \\ \frac{1}{2}\big|\frac{\partial^j\Psi}{\partial x_3^j}\big|_i^2(0) = \frac{1}{2}\big|\Delta^j\Psi\big|_i^2(0) = \frac{1}{2}\big|\Delta^j g\big|_i^2 & \text{for } k = 0. \end{cases} \tag{4.42}$$

Similarly, if we differentiate Eq. (4.39) j times in x_3, multiply the resulting equation by $x_3^k\,\partial^{j+1}\Psi/\partial x_3^{j+1}$, and integrate over Γ_{i}, we find

$$f_{k,j} = \begin{cases} \frac{k}{2}\int_{x_3}^0 |z|^{k-1}\big|\nabla\frac{\partial^j\Psi}{\partial x_3^j}\big|_i^2(z)\,\mathrm{d}z & \text{for } k \geq 1, \\ \frac{1}{2}\big|\nabla\frac{\partial^j\Psi}{\partial x_3^j}\big|_i^2(0) = \frac{1}{2}\big|\nabla\Delta^j g\big|_i^2 & \text{for } k = 0. \end{cases} \tag{4.43}$$

Using Eq. (4.42), Eq. (4.43) with $k = j = 0$, Eq. (4.42) with $k = j = 1$, and Eq. (4.43) with $k = 2$, $j = 1$, we find some of the relations (Eqs. (4.40), (4.41)), namely

$$\begin{aligned} &\frac{1}{2}\big|\Psi(x_3)\big|_i^2 + \int_{x_3}^0 |\nabla\Psi|_i^2(z)\,\mathrm{d}z \leq \frac{1}{2}|g|_i^2, \\ &\frac{1}{2}\big|\nabla\Psi(x_3)\big|_i^2 + \int_{x_3}^0 \left|\frac{\partial\Psi}{\partial x_3}\right|_i^2(z)\,\mathrm{d}z \leq \frac{1}{2}\|g\|_i^2, \\ &\frac{1}{2}|x_3|\left|\frac{\partial\Psi}{\partial x_3}\right|_i^2(x_3) + \int_{x_3}^0 |z|\left|\nabla\frac{\partial\Psi}{\partial x_3}\right|_i^2(z)\,\mathrm{d}z \leq \frac{1}{4}\|g\|_i^2, \\ &\frac{1}{2}|x_3|^2\left|\nabla\frac{\partial\Psi}{\partial x_3}\right|_i^2(x_3) + \int_{x_3}^0 z^2\left|\frac{\partial^2\Psi}{\partial x_3^2}\right|_i^2(z)\,\mathrm{d}z \leq \frac{1}{4}\|g\|_i^2. \end{aligned} \tag{4.44}$$

To derive the other relations (Eqs. (4.40) and (4.41)), we first integrate Eq. (4.42) from x_3 to 0, with $x_3 < 0$, $k \geq 1$ and $j \geq 0$; we obtain

$$\begin{aligned} \int_{x_3}^0 |z|^k\left|\frac{\partial^j\Psi}{\partial x_3^j}\right|_i^2(z)\,\mathrm{d}z &\leq k\int_{x_3}^0\int_t^0 |z|^{k-1}\left|\frac{\partial^j\Psi}{\partial x_3^j}\right|_i^2(z)\,\mathrm{d}z\,\mathrm{d}t \\ &\leq k\int_{x_3}^0 (z - x_3)|z|^{k-1}\left|\frac{\partial^j\Psi}{\partial x_3^j}\right|_i^2(z)\,\mathrm{d}z. \end{aligned}$$

Thus, for $k \geq 1$, $j \geq 0$,

$$\int_{x_3}^0 |z|^k\left|\frac{\partial^j\Psi}{\partial x_3^j}\right|_i^2(z)\,\mathrm{d}z \leq \frac{k}{k+1}|x_3|\int_{x_3}^0 |z|^{k-1}\left|\frac{\partial^j\Psi}{\partial x_3^j}\right|_i^2(z)\,\mathrm{d}z, \tag{4.45}$$

and by induction on k and using Eq. (4.41),

$$\int_{x_3}^{0} |z|^k \left| \frac{\partial^j \Psi}{\partial x_3^j} \right|_i^2 (z)\, \mathrm{d}z \le \frac{|x_3|^k}{k+1} \int_{x_3}^{0} \left| \frac{\partial^j \Psi}{\partial x_3^j} \right|_i^2 (z)\, \mathrm{d}z,$$

$$\le \begin{cases} \frac{|x_3|^{k+1}}{2(k+1)} |g|_i^2 & \text{for } j = 0, \\ \frac{|x_3|^k}{2(k+1)} \|g\|_i^2 & \text{for } j = 1 (k \ge 1). \end{cases} \tag{4.46}$$

Now, combining Eqs. (4.42) and (4.43), we find for $k \ge 2$, $j \ge 1$,

$$e_{k,j} \le \frac{k(k-1)}{2^2} e_{k-2,j-1}.$$

Hence, by induction

$$e_{k,j} \le \frac{k!}{2^{2r}(k-2r)!} e_{k-2r,j-r},$$

$$e_{k,j} \le \frac{k!}{2^{2j-2}(k-2j+2)!} e_{s,1},$$

$s = k - 2j + 2$, for $k \ge 2j - 2$, $j \ge 1$. For $s \ge 1$, i.e., $k \ge 2j - 1$ and $j \ge 1$, we have, thanks to Eqs. (4.42) and (4.46),

$$\begin{aligned} e_{s,1} &= \frac{1}{2} \int_{x_3}^{0} |z|^{s-1} \left| \frac{\partial \Psi}{\partial x_3} \right|_i^2 (z)\, \mathrm{d}z \\ &\le \frac{1}{2} |x_3|^{s-1} \|g\|_i^2, \\ e_{k,j} &\le \frac{k!}{2^{2j-1}(k-2j+2)!} |x_3|^{k-2j+1} \|g\|_i^2, \end{aligned} \tag{4.47}$$

for $k \ge 2j - 1$, $j \ge 1$. The relations (Eq. (4.40)) are proven; the relations (Eq. (4.41)) follow from Eq. (4.44) for $j = 0$ and from $f_{k,j} \le (k/2) e_{k-1,j}$ for $j \ge 1$.

Lemma 4.2 is proved. □

From Lemma 4.2, we easily infer the following lemma.

LEMMA 4.3. *Under the hypotheses of Lemma* 4.2,

$$\int_{\Gamma_i} \left| \nabla \frac{\partial^j \Psi}{\partial x_3^j} \right|^2 \big(x_1, x_2, -\varepsilon h(x_1, x_2)\big)\, \mathrm{d}x_1\, \mathrm{d}x_2 \le C\varepsilon^{-2j} \|g\|_{H^1(\Gamma_i)}^2 \quad \textit{for } j \ge 0, \tag{4.48}$$

$$\int_{\Gamma_i} \left| \frac{\partial^j \Psi}{\partial x_3^j} \right|^2 \big(x_1, x_2, -\varepsilon h(x_1, x_2)\big)\, \mathrm{d}x_1\, \mathrm{d}x_2$$

$$\le \begin{cases} C\|g\|_{H^1(\Gamma_i)}^2 & \textit{for } j = 0, \\ C\varepsilon^{-2j+1} \|g\|_{H^1(\Gamma_i)}^2 & \textit{for } j \ge 1, \end{cases} \tag{4.49}$$

$$\int_{\Gamma_i} \left|\nabla^j \Psi\right|^2 \big(x_1, x_2, -\varepsilon h(x_1, x_2)\big)\, \mathrm{d}x_1\, \mathrm{d}x_2 \le C\varepsilon^{-j+1} \|g\|^2_{H^1(\Gamma_i)}, \quad j = 2, 3, \tag{4.50}$$

$$\int_{\Gamma_i} \left|\frac{\partial}{\partial x_3} \nabla^2 \Psi\right|^2 \big(x_1, x_2, -\varepsilon h(x_1, x_2)\big)\, \mathrm{d}x_1\, \mathrm{d}x_2 \le C\varepsilon^{-3} \|g\|^2_{H^1(\Gamma_i)}, \tag{4.51}$$

where $\|g\|^2_{H^1(\Gamma_i)} = |g|_i^2 + \|g\|_i^2$, *and* C *is a constant depending on* j *and* h *and independent of* ε.

Proof. We write

$$\left(\nabla \frac{\partial^j \Psi}{\partial x_3^j}\right)^2 (x_1, x_2, -\varepsilon h(x_1, x_2))$$
$$= \left(\nabla \frac{\partial^j \Psi}{\partial x_3^j}\right)^2 (x_1, x_2, -\varepsilon \bar{h}) + 2 \int_{-\varepsilon \bar{h}}^{-\varepsilon h(x_1, x_2)} \left(\nabla \frac{\partial^j \Psi}{\partial x_3^j}\right)\left(\nabla \frac{\partial^{j+1} \Psi}{\partial x_3^{j+1}}\right)(x_1, x_2, z)\, \mathrm{d}z.$$

Integrating in x_1, x_2 on Γ_i, we find

$$\int_{\Gamma_i} \left(\nabla \frac{\partial^j \Psi}{\partial x_3^j}\right)^2 (x_1, x_2, -\varepsilon h(x_1, x_2))\, \mathrm{d}x_1\, \mathrm{d}x_2$$
$$\le \left|\nabla \frac{\partial^j \Psi}{\partial x_3^j}\right|_i^2 (-\varepsilon \bar{h})$$
$$+ 2\left(\int_{-\varepsilon \bar{h}}^{-\varepsilon \underline{h}} \left|\nabla \frac{\partial^j \Psi}{\partial x_3^j}\right|_i^2 (z)\, \mathrm{d}z\right)^{1/2} \left(\int_{-\varepsilon \bar{h}}^{-\varepsilon \underline{h}} \left|\nabla \frac{\partial^{j+1} \Psi}{\partial x_3^{j+1}}\right|_i^2 (z)\, \mathrm{d}z\right)^{1/2}.$$

By Eq. (4.41),

$$\left|\nabla \frac{\partial^j \Psi}{\partial x_3^j}\right|_i^2 (-\varepsilon \bar{h}) \le C(\varepsilon \bar{h})^{-2j} \|g\|^2_{H^1(\Gamma_i)} \quad (j \ge 0)$$

and by integration of Eq. (4.41),

$$\int_{-\varepsilon \bar{h}}^{-\varepsilon \underline{h}} \left|\nabla \frac{\partial^j \Psi}{\partial x_3^j}\right|_i^2 (z)\, \mathrm{d}z \le C(\varepsilon \bar{h})^{-k} (\varepsilon \bar{h})^{k-2j+1} \|g\|^2_{H^1(\Gamma_i)}$$
$$\le C\varepsilon^{-2j+1} \|g\|^2_{H^1(\Gamma_i)} \quad (j \ge 0),$$

Eq. (4.48) follows. The proof of Eq. (4.49) is similar.

For Eqs. (4.50) and (4.51), we observe that by the regularity property for the Neumann problem in Γ_i,

$$\sum_{k,\ell=1}^{2} \left| \frac{\partial^2 \Psi}{\partial x_k \, \partial x_\ell} \right|^2_{L^2(\Gamma_i)} (x_3) \le C(|\Psi|^2_{L^2(\Gamma_i)}(x_3) + |\Delta \Psi|^2_{L^2(\Gamma_i)}(x_3))$$

$$\le C\Big(|\Psi|^2_{L^2(\Gamma_i)}(x_3) + \left| \frac{\partial \Psi}{\partial x_3} \right|^2_{L^2(\Gamma_i)} (x_3) \Big). \tag{4.52}$$

By repeating the argument above, it appears that the bounds for the left-hand sides of Eqs. (4.50) and (4.51) are the same as those of Eqs. (4.48) and (4.49), for $j = 1$ and 2, respectively; Eqs. (4.50) and (4.51) are proved. □

4.3. Regularity of solutions of a Neumann–Robin boundary value problem

We now want to derive a result similar to that of Sections 4.1 and 4.2 for a mixed Neumann–Robin-type boundary condition, i.e., for the problem (4.53) below. Our result is quite general except for the restrictions (Eq. (4.54)) below.

We will prove the following theorem.

THEOREM 4.5. *Assume* Γ_i *is an open bounded set of* $\mathbb{R}^2$, *with a* C^3*-boundary* $\partial\Gamma_i$, *and* $h \in C^4(\overline{\Gamma}_i)$. *For* $f \in L^2(\mathcal{M}_\varepsilon)$ *and* $g \in H^1(\Gamma_i)$, *there exists a unique* $\Psi \in H^2(\mathcal{M}_\varepsilon)$ *solution of*

$$\begin{aligned} &-\Delta_3 \Psi = f, \\ &\frac{\partial \Psi}{\partial x_3} + \alpha \Psi = g \quad \textit{on } \Gamma_i, \\ &\frac{\partial \Psi}{\partial n} = 0 \quad \textit{on } \Gamma_b \cup \Gamma_\ell. \end{aligned} \tag{4.53}$$

Furthermore, if

$$\nabla h \cdot n_{\Gamma_i} = 0 \quad \textit{on } \partial\Gamma_i, \tag{4.54}$$

then there exists $C = C(h, \Gamma_i, \alpha)$ *such that*

$$\sum_{k,\ell=1}^{3} \left| \frac{\partial^2 \Psi}{\partial x_k \, \partial x_\ell} \right|^2_{\varepsilon} \le C\big[|f|^2_\varepsilon + \|g\|^2_{H^1} \big].$$

REMARK 4.1. Condition in Eq. (4.54) means that Γ_b and Γ_ℓ meet at a wedge angle of $\pi/2$. This is a technical condition needed in the method of proof used below; this condition is not required by the theory of regularity of elliptic problems in nonsmooth 3D problems GRISVARD [1985]. It might be possible to remove this condition with a different proof. Note that this condition is needed for the dependence on ε and not for the sole H^2-regularity.

PROOF OF THEOREM 4.5. The proof is divided into several steps. First we reduce the problem to the case $\alpha = 0$, then we reduce the problem to the case where $g = 0$ and $\alpha = 0$. Thus, we obtain a homogeneous Neumann problem. Then by a reflection around $x_3 = 0$, we can use Theorem 2.1 and conclude the proof.

STEP 1 (reduction to the case $g = 0$) Our goal now is to find an explicit function $\overline{T}$ that satisfies the nonhomogeneous boundary conditions imposed on the temperature. Since we are interested in obtaining a sharp dependence on the thickness ε, we have to construct $\overline{T}$ explicitly instead of the classical method of lifting by localization and straightening the boundary, which yields constants that do not have the right dependence on ε. We will carry the computations only in the case where $\nabla h \cdot n_{\Gamma_i} = 0$ on $\partial\Gamma_i$.

LEMMA 4.4. *Let $h \in C^4(\overline{\Gamma}_i)$ with $\nabla h \cdot n_{\Gamma_i} = 0$ on $\partial\Gamma_i$, and let $g \in H^1(\Gamma_i)$. There exists a function $\overline{T} \in H^2(\mathcal{M}_\varepsilon)$ such that*

$$\begin{aligned} &\frac{\partial \overline{T}}{\partial x_3} + \alpha \overline{T} = g \quad \text{on } \Gamma_i, \\ &\frac{\partial \overline{T}}{\partial n} = 0 \quad \text{on } \Gamma_\ell \cup \Gamma_b. \end{aligned} \tag{4.55}$$

Furthermore,

$$\sum_{k,l=1}^{3} \left| \frac{\partial^2 \overline{T}}{\partial x_k \, \partial x_l} \right|_\varepsilon^2 \le C \|g\|^2_{H^1(\Gamma_i)}, \tag{4.56}$$

where C is a constant independent of ε. □

PROOF. Let $\overline{\Psi}$ be a solution of the heat equation, where $-x_3$ corresponds to time:

$$\begin{cases} \dfrac{\partial \overline{\Psi}}{\partial x_3} = -\Delta \overline{\Psi} & \text{in } \Gamma_i \times (-\infty, 0), \\ \dfrac{\partial \overline{\Psi}}{\partial n_{\Gamma_i}} = 0 & \text{on } \partial\Gamma_i \times (-\infty, 0), \\ \overline{\Psi}(x_1, x_2, 0) = g(x_1, x_2), \end{cases} \tag{4.57}$$

and define $\overline{T}$ by

$$\begin{aligned} \overline{T}(x_1, x_2, x_3) = \mathrm{e}^{-\alpha x_3} &\int_{-\varepsilon h(x_1, x_2)}^{x_3} \overline{\Psi}(x_1, x_2, z) \, \mathrm{d}z \\ &- \left(x_3 - \frac{1}{\alpha} \right) \theta_1(x_1, x_2) + x_3^2 (x_3 + \varepsilon h) \theta_2(x_1, x_2), \end{aligned} \tag{4.58}$$

where

$$\theta_1(x_1, x_2) = \mathrm{e}^{\alpha\varepsilon h(x_1,x_2)}\overline{\Psi}(x_1, x_2, -\varepsilon h(x_1, x_2))(1 + \varepsilon^2|\nabla h|^2) \tag{4.59}$$

and

$$\theta_2(x_1, x_2) = \frac{-(\varepsilon h + \frac{1}{\alpha})}{\varepsilon h^2(1 + \varepsilon^2|\nabla h|^2)}\nabla\theta_1 \cdot \nabla h. \tag{4.60}$$

Then,

$$\begin{aligned}\frac{\partial\overline{T}}{\partial x_3} &= -\alpha\mathrm{e}^{-\alpha x_3}\int_{-\varepsilon h}^{x_3}\overline{\Psi}(x_1, x_2, z)\,\mathrm{d}z + \mathrm{e}^{-\alpha x_3}\overline{\Psi}(x_1, x_2, x_3)\\ &\quad - \theta_1(x_1, x_2) + 2x_3(x_3 + \varepsilon h)\theta_2(x_1, x_2) + x_3^2\theta_2(x_1, x_2),\end{aligned} \tag{4.61}$$

$$\left.\frac{\partial\overline{T}}{\partial x_3} + \alpha\overline{T}\right|_{x_3=0} = \overline{\Psi}(x_1, x_2, 0) = g, \tag{4.62}$$

$$\begin{aligned}\nabla\overline{T} &= \mathrm{e}^{-\alpha x_3}\int_{-\varepsilon h}^{x_3}\nabla\overline{\Psi}(x_1, x_2, z)\,\mathrm{d}z + \varepsilon\mathrm{e}^{-\alpha x_3}\overline{\Psi}(x_1, x_2, -\varepsilon h)\nabla h\\ &\quad - \left(x_3 - \frac{1}{\alpha}\right)\nabla\theta_1 + x_3^2(x_3 + \varepsilon h)\nabla\theta_2 + \varepsilon x_3^2\theta_2\nabla h,\end{aligned} \tag{4.63}$$

$$\begin{aligned}\varepsilon\nabla\overline{T}\cdot\nabla h\big|_{x_3=-\varepsilon h} &= \varepsilon^2\mathrm{e}^{\alpha\varepsilon h}\overline{\Psi}(x_1, x_2, -\varepsilon h)|\nabla h|^2\\ &\quad + \varepsilon\left(\varepsilon h + \frac{1}{\alpha}\right)\nabla\theta_1\cdot\nabla h + \varepsilon^4h^2\theta_2|\nabla h|^2,\end{aligned} \tag{4.64}$$

$$\left.\frac{\partial\overline{T}}{\partial x_3}\right|_{x_3=-\varepsilon h} = \mathrm{e}^{\alpha\varepsilon h}\overline{\Psi}(x_1, x_2, -\varepsilon h) - \theta_1(x_1, x_2) + \varepsilon^2h^2\theta_2,$$

and

$$\begin{aligned}\left.\frac{\partial\overline{T}}{\partial x_3} + \varepsilon\nabla\overline{T}\cdot\nabla h\right|_{x_3=-\varepsilon h} &= \mathrm{e}^{\alpha\varepsilon h}\overline{\Psi}(x_1, x_2, -\varepsilon h)(1 + \varepsilon^2|\nabla h|^2)\\ &\quad + \varepsilon^2h^2\theta_2(1 + \varepsilon^2|\nabla h|^2) - \theta_1\\ &\quad + \varepsilon\left(\varepsilon h + \frac{1}{\alpha}\right)\nabla\theta_1\cdot\nabla h.\end{aligned}$$

Hence, with θ_1 and θ_2 as in Eqs. (4.59) and (4.60), we have

$$\frac{\partial\overline{T}}{\partial x_3} + \varepsilon\nabla\overline{T}\cdot\nabla h = 0 \quad \text{on } \Gamma_{\mathrm{b}}.$$

Now, we use the assumption (Eq. (4.54)) and prove that

$$\nabla\theta_1 \cdot n_{\Gamma_i} = 0, \nabla\theta_2 \cdot n_{\Gamma_i} = 0 \qquad \text{on } \partial\Gamma_i, \tag{4.65}$$

which implies that $\nabla\overline{T} \cdot n_{\Gamma_i} = 0$ on $\partial\Gamma_i$.

First, by working in local coordinates (s, t), where s is the coordinate in the normal direction of $\partial\Gamma_i$ and t the coordinate in the tangential direction, the condition $\nabla h \cdot n_{\Gamma_i} = 0$ on Γ_i implies (since $\partial\Gamma_i$ is smooth)

$$\frac{\partial h}{\partial s} = 0 \quad \text{and} \quad \frac{\partial^2 h}{\partial s\,\partial t} = 0 \quad \text{on } \partial\Gamma_i. \tag{4.66}$$

Therefore,

$$\frac{\partial}{\partial s}|\nabla h|^2 = \frac{\partial}{\partial s}\left(\left|\frac{\partial h}{\partial s}\right|^2 + \left|\frac{\partial h}{\partial t}\right|^2\right) = 2\left(\frac{\partial h}{\partial s}\frac{\partial^2 h}{\partial s^2} + \frac{\partial h}{\partial t}\frac{\partial^2 h}{\partial s\,\partial t}\right) = 0. \tag{4.67}$$

Now,

$$\begin{aligned}\nabla\theta_1 \cdot n_{\Gamma_i} = {} & \alpha\varepsilon e^{\alpha\varepsilon h}\overline{\Psi}(x_1, x_2, -\varepsilon h)(1 + \varepsilon^2|\nabla h|^2)\nabla h \cdot n_{\Gamma_i} \\ & + e^{\alpha\varepsilon h}(1 + \varepsilon^2|\nabla h|^2)\nabla\overline{\Psi} \cdot n_{\Gamma_i} \\ & - \varepsilon e^{\alpha\varepsilon h}\frac{\partial\overline{\Psi}}{\partial x_3}(x_1, x_2, -\varepsilon h)(1 + \varepsilon^2|\nabla h|^2)\nabla h \cdot n_{\Gamma_i} \\ & + e^{\alpha\varepsilon h}\overline{\Psi}(x_1, x_2, -\varepsilon h)\nabla(|\nabla h|^2) \cdot n_{\Gamma_i},\end{aligned} \tag{4.68}$$

and since $\nabla h \cdot n_{\Gamma_i} = 0$, $\nabla\overline{\Psi} \cdot n_{\Gamma_i} = 0$, and $\nabla(|\nabla h|^2) \cdot n_{\Gamma_i} = 0$ on $\partial\Gamma_i$, we have $\nabla\theta_1 \cdot n_{\Gamma_i} = 0$ on $\partial\Gamma_i$.

Next, we check that $\nabla\theta_2 \cdot n_{\Gamma_i} = 0$ on $\partial\Gamma_i$. Here, we only need to show that $\nabla(\nabla\theta_1 \cdot \nabla h) \cdot n_{\Gamma_i} = 0$ on $\partial\Gamma_i$. Again, this can be done by working in local coordinates. We have

$$\nabla\theta_1 \cdot \nabla h = \frac{\partial\theta_1}{\partial s}\frac{\partial h}{\partial s} + \frac{\partial\theta_1}{\partial t}\frac{\partial h}{\partial t}, \tag{4.69}$$

and therefore

$$\begin{aligned}\nabla(\nabla\theta_1 \cdot \nabla h) \cdot n_{\Gamma_i} &= \frac{\partial}{\partial n_{\Gamma_i}}(\nabla\theta_1 \cdot \nabla h) \\ &= \frac{\partial\theta_1}{\partial s}\frac{\partial^2 h}{\partial s^2} + \frac{\partial^2\theta_1}{\partial s^2}\frac{\partial h}{\partial s} + \frac{\partial^2\theta_1}{\partial t\,\partial s}\frac{\partial h}{\partial t} + \frac{\partial\theta_1}{\partial t}\frac{\partial^2 h}{\partial s\,\partial t},\end{aligned} \tag{4.70}$$

but since $\partial h/\partial s = 0$ and $\partial\theta_1/\partial s = 0$, we have $\partial^2 h/\partial s\,\partial t = 0$ and $\partial^2\theta_1/\partial s\,\partial t = 0$. Thus,

$$\nabla(\nabla\theta_1 \cdot \nabla h) \cdot n_{\Gamma_i} = 0 \quad \text{on } \partial\Gamma_i. \tag{4.71}$$

Finally, since θ_2 is the product of functions each of which has normal derivative to $\partial\Gamma_i$ vanishing on $\partial\Gamma_i$, we have

$$\nabla\theta_2 \cdot n_{\Gamma_i} = 0 \quad \text{on } \partial\Gamma_i. \tag{4.72}$$

This concludes the verification of $\overline{T}$ satisfying the boundary conditions. We now use estimates on solutions of the heat equation and the explicit expression of $\overline{T}$ to establish the inequality (Eq. (4.56)) of the lemma. Here, we fully rely on the estimates for $\overline{\Psi}$ provided by Lemmas 4.2 and 4.3.

With $\partial\overline{T}/\partial x_3$ and $\nabla\overline{T}$ given by Eqs. (4.61) and (4.63), we write

$$\begin{aligned}\frac{\partial^2\overline{T}}{\partial x_3^2} &= \alpha^2 \mathrm{e}^{-\alpha x_3}\int_{-\varepsilon h}^{x_3}\overline{\Psi}(x_1,x_2,z)\,\mathrm{d}z\\ &\quad - 2\alpha\mathrm{e}^{-\alpha x_3}\overline{\Psi}(x_1,x_2,x_3) + \mathrm{e}^{-\alpha x_3}\frac{\partial\overline{\Psi}}{\partial x_3}(x_1,x_2,x_3)\\ &\quad + 2(x_3+\varepsilon h)\theta_2(x_1,x_2) + 4x_3\theta_2(x_1,x_2),\end{aligned} \tag{4.73}$$

and for $k = 1, 2$,

$$\begin{aligned}&\frac{\partial^2\overline{T}}{\partial x_k\,\partial x_3}\\ &\quad = -\alpha\mathrm{e}^{-\alpha x_3}\int_{-\varepsilon h}^{x_3}\frac{\partial\overline{\Psi}}{\partial x_k}(x_1,x_2,z)\,\mathrm{d}z - \alpha\varepsilon\mathrm{e}^{-\alpha x_3}\overline{\Psi}(x_1,x_2,-\varepsilon h)\frac{\partial h}{\partial x_k}\\ &\qquad + \mathrm{e}^{-\alpha x_3}\frac{\partial\overline{\Psi}}{\partial x_k} - \frac{\partial\theta_1}{\partial x_k} + 2\varepsilon x_3\theta_2\frac{\partial h}{\partial x_k} + 2x_3(x_3+\varepsilon h)\frac{\partial\theta_2}{\partial x_k} + x_3^2\frac{\partial\theta_2}{\partial x_k}.\end{aligned} \tag{4.74}$$

Finally, for $k, \ell = 1, 2$,

$$\begin{aligned}\frac{\partial^2\overline{T}}{\partial x_k\,\partial x_\ell} &= \mathrm{e}^{-\alpha x_3}\int_{-\varepsilon h(x_1,x_2)}^{x_3}\frac{\partial^2\overline{\Psi}}{\partial x_k\,\partial x_\ell}(x_1,x_2,z)\,\mathrm{d}z\\ &= +\varepsilon\mathrm{e}^{-\alpha x_3}\frac{\partial\overline{\Psi}}{\partial x_\ell}(x_1,x_2,-\varepsilon h(x_1,x_2))\frac{\partial h}{\partial x_k}\\ &\quad + \varepsilon\mathrm{e}^{-\alpha x_3}\frac{\partial\overline{\Psi}}{\partial x_k}(x_1,x_2,-\varepsilon h(x_1,x_2))\frac{\partial h}{\partial x_\ell}\\ &\quad - \varepsilon^2\mathrm{e}^{-\alpha x_3}\frac{\partial\overline{\Psi}}{\partial x_3}(x_1,x_2,-\varepsilon h)\frac{\partial h}{\partial x_k}\frac{\partial h}{\partial x_\ell}\\ &\quad + \varepsilon\mathrm{e}^{-\alpha x_3}\overline{\Psi}(x_1,x_2,-h)\frac{\partial^2 h}{\partial x_k\,\partial x_\ell} - \left(x_3-\frac{1}{\alpha}\right)\frac{\partial^2\theta_1}{\partial x_k\,\partial x_\ell}\\ &\quad + x_3^2(x_3+\varepsilon h)\frac{\partial^2\theta_2}{\partial x_k\,\partial x_\ell} + \varepsilon x_3^2\frac{\partial h}{\partial x_k}\frac{\partial\theta_2}{\partial x_\ell} + \varepsilon x_3^2\theta_2\frac{\partial^2 h}{\partial x_k\,\partial x_\ell}.\end{aligned} \tag{4.75}$$

To estimate the L^2-norms of the second derivatives of $\overline{T}$, we need to bound the L^2-norms of θ_1, θ_2 and their derivatives, which we do in Lemma 4.4.

Using Lemma 4.4, we estimate as follows the norm in $L^2(\mathcal{M}_\varepsilon)$ of $\partial^2\overline{T}/\partial x_k\,\partial x_\ell$, $k, \ell = 1, 2$ as given by Eq. (4.75).

The first term in the right-hand side of Eq. (4.75) is bounded by a constant times the norm of $\partial^2\overline{\Psi}/\partial x_k\,\partial x_\ell$ in $Q_\varepsilon = \Gamma_{\rm i}\times(-\varepsilon\bar{h},0)$; using Eqs. (4.51) and (4.41), this term is bounded as in Eq. (4.56). We then use Lemma 4.3 to estimate the four subsequent terms, and the bounds are consistent with Eq. (4.56). The remaining terms involve θ_1 and θ_2 and their derivatives; the integration over $\Gamma_{\rm i}$ of these functions provides the bounds given by Lemma 4.4, and for each of these terms, there is a factor of $C\varepsilon^m$, $m\geq 2$, which is due to the integration in x_3. The bound (Eq. (4.56)) follows.

We proceed similarly for $\partial^2\overline{T}/\partial x_3^2$ given by Eq. (4.75) and for $\partial^2\overline{T}/\partial x_k\,\partial x_3$, $k=1,2$, given by Eq. (4.74). Lemma 4.4 follows. □

We now conclude the proof of Lemma 4.4 by proving Lemma 4.4.

LEMMA 4.5. *The functions θ_1 and θ_2 introduced in Eqs.* (4.59) *and* (4.60) *are bounded as follows*:

$$|\theta_1|_{L^2(\Gamma_{\rm i})}+|\nabla\theta_1|_{L^2(\Gamma_{\rm i})}+\varepsilon^{1/2}\big|\nabla^2\theta_1\big|_{L^2(\Gamma_{\rm i})}\leq C\|g\|^2_{H^1(\Gamma_{\rm i})},\tag{4.76}$$

$$|\theta_2|_{L^2(\Gamma_{\rm i})}+\varepsilon^{1/2}\big|\nabla\theta_2\big|_{L^2(\Gamma_{\rm i})}+\varepsilon\big|\nabla^2\theta_2\big|_{L^2(\Gamma_{\rm i})}\leq C\varepsilon^{-1}\|g\|^2_{H^1(\Gamma_{\rm i})},\tag{4.77}$$

where C is a constant independent of ε.

PROOF. The proof strongly relies on the definitions (Eqs. (4.59) and (4.60)) of θ_1 and θ_2 and on the estimates on $\overline{\Psi}$ given by Lemmas 4.2 and 4.3.

We write $\widetilde{\Psi}(x_1,x_2)=\overline{\Psi}(x_1,x_2,-\varepsilon h(x_1,x_2))$ and observe that, pointwise,

$$\begin{aligned}
\nabla\widetilde{\Psi}&=\nabla\overline{\Psi}+\varepsilon\xi\frac{\partial\overline{\Psi}}{\partial x_3},\\
\nabla^2\widetilde{\Psi}&=\nabla^2\overline{\Psi}+\varepsilon\xi\frac{\partial\nabla\overline{\Psi}}{\partial x_3}+\varepsilon\xi\frac{\partial\overline{\Psi}}{\partial x_3}+\varepsilon^2\xi\frac{\partial^2\overline{\Psi}}{\partial x_3^2},\\
\nabla^3\widetilde{\Psi}&=\nabla^3\overline{\Psi}+\varepsilon\xi\frac{\partial\nabla^2\overline{\Psi}}{\partial x_3}+\varepsilon\xi\frac{\partial\nabla\overline{\Psi}}{\partial x_3}+\varepsilon\xi\frac{\partial^2\overline{\Psi}}{\partial x_3^2}+\varepsilon\xi\frac{\partial\overline{\Psi}}{\partial x_3}\\
&\quad+\varepsilon^2\xi\frac{\partial^2\nabla\overline{\Psi}}{\partial x_3^2}+\varepsilon^3\xi\frac{\partial^2\overline{\Psi}}{\partial x_3^3}+\varepsilon\xi\frac{\partial\nabla\overline{\Psi}}{\partial x_3}+\varepsilon^2\xi\frac{\partial^2\overline{\Psi}}{\partial x_3^2}.
\end{aligned}\tag{4.78}$$

Here, the ξ are (different) continuous (scalar, vector or tensor) functions bounded on $\Gamma_{\rm i}$ independently of ε ($\varepsilon\leq 1$), $\widetilde{\Psi}$ and its derivatives are evaluated at $(x_1,x_2)\in\Gamma_{\rm i}$, and $\overline{\Psi}$ and its derivatives are evaluated at $(x_1,x_2,-\varepsilon h(x_1,x_2))$.

It follows then from Lemma 4.3 that

$$\begin{aligned}
&\big|\widetilde{\Psi}\big|_{L^2(\Gamma_{\rm i})}\leq C, && \big|\nabla\widetilde{\Psi}\big|_{L^2(\Gamma_{\rm i})}\leq C,\\
&\big|\nabla^2\widetilde{\Psi}\big|_{L^2(\Gamma_{\rm i})}\leq C\varepsilon^{1/2}, && \big|\nabla^3\widetilde{\Psi}\big|_{L^2(\Gamma_{\rm i})}\leq C\varepsilon^{-1}.
\end{aligned}\tag{4.79}$$

Now, similarly, θ_1 and its first, second, and third derivatives are of the following forms:

$$\theta_1 = \xi\widetilde{\Psi}, \qquad \nabla\theta_1 = \nabla\xi\cdot\widetilde{\Psi} + \xi\nabla\widetilde{\Psi},$$
$$\nabla^2\theta_1 = \nabla^2\xi\cdot\widetilde{\Psi} + 2\nabla\xi\otimes\nabla\widetilde{\Psi} + \xi\nabla^2\widetilde{\Psi},$$
$$\nabla^3\theta_1 = \nabla^3\xi\cdot\widetilde{\Psi} + 3\nabla^2\xi\otimes\nabla\tilde{\psi} + 3\nabla\xi\otimes\nabla^2\widetilde{\Psi} + \xi\nabla^3\widetilde{\Psi},$$

where ξ and its first, second, and third derivatives are uniformly bounded on Γ_i (for $\varepsilon \le 1$), hence Eq. (4.76) using Eq. (4.78). To obtain Eq. (4.77), we observe that, with a different ξ, θ_2 is of the form $\varepsilon^{-1}\xi\cdot\nabla\theta_1$.

Lemma 4.4 is proved. □

STEP 2 (reduction to the case $\alpha = 0$ (and $g = 0$)) Let $\widetilde{T}$ be the solution of

$$\begin{aligned} &-\Delta_3\widetilde{T} = f_2 && \text{in } \mathcal{M}_\varepsilon,\\ &\frac{\partial\widetilde{T}}{\partial x_3} + \alpha\widetilde{T} = 0 && \text{on } \Gamma_i, \text{ and}\\ &\frac{\partial\widetilde{T}}{\partial n} = 0 && \text{on } \Gamma_b\cup\Gamma_\ell. \end{aligned} \tag{4.80}$$

We first note that

$$\left|\nabla\widetilde{T}\right|_\varepsilon^2 + \left|\frac{\partial\widetilde{T}}{\partial x_3}\right|_\varepsilon^2 + \alpha\left|\widetilde{T}\right|_i^2 \le |f_2|_\varepsilon\left|\widetilde{T}\right|_\varepsilon. \tag{4.81}$$

Also by a density argument and since

$$\begin{aligned} \left|\widetilde{T}(x_1,x_2,x_3)\right| &\le \left|\widetilde{T}(x_1,x_2,0)\right| + \int_0^{x_3}\left|\frac{\partial\widetilde{T}}{\partial x_3}\right|\mathrm{d}x_3'\\ &\le \left|\widetilde{T}(x_1,x_2,0)\right| + \sqrt{\varepsilon\bar{h}}\left(\int_{-\varepsilon h(x_1,x_2)}^0\left|\frac{\partial\widetilde{T}}{\partial x_3}\right|^2\mathrm{d}x_3\right)^{1/2}, \end{aligned} \tag{4.82}$$

and

$$\int_{M_\varepsilon}\left|\widetilde{T}(x_1,x_2,x_3)\right|^2\mathrm{d}x_1\,\mathrm{d}x_2\,\mathrm{d}x_3 \le 2\varepsilon\bar{h}\left|\widetilde{T}\right|_i^2 + 2\varepsilon\bar{h}\left|\frac{\partial\widetilde{T}}{\partial x_3}\right|_\varepsilon^2. \tag{4.83}$$

We infer from Eq. (4.81) that

$$\begin{aligned} \left|\nabla\widetilde{T}\right|_\varepsilon^2 + \left|\frac{\partial\widetilde{T}}{\partial x_3}\right|_\varepsilon^2 + \alpha\left|\widetilde{T}\right|_i^2 &\le \sqrt{2\varepsilon\bar{h}}|f_2|_\varepsilon\left|\widetilde{T}\right|_i + \sqrt{2\varepsilon\bar{h}}|f_2|_\varepsilon\left|\frac{\partial\widetilde{T}}{\partial x_3}\right|_\varepsilon\\ &\le \frac{\alpha}{2}\left|\widetilde{T}\right|_i^2 + \frac{\varepsilon\bar{h}}{2\alpha}|f_2|_\varepsilon^2 + \frac{1}{2}\left|\frac{\partial\widetilde{T}}{\partial x_3}\right|_\varepsilon^2 + \varepsilon\bar{h}|f_2|_\varepsilon^2. \end{aligned} \tag{4.84}$$

Therefore,

$$\left|\nabla\widetilde{T}\right|_{\varepsilon}^{2}+\frac{1}{2}\left|\frac{\partial\widetilde{T}}{\partial x_3}\right|_{\varepsilon}^{2}+\frac{\alpha}{2}\left|\widetilde{T}\right|_{i}^{2}\leq\varepsilon\bar{h}|f_2|_{\varepsilon}^{2}\left(\frac{1}{2\alpha}+1\right). \tag{4.85}$$

Hence,

$$\left|\widetilde{T}\right|_{i}^{2}\leq C\varepsilon|f_2|_{\varepsilon}^{2} \tag{4.86}$$

and

$$\left|\frac{\partial\widetilde{T}}{\partial x_3}\right|_{\varepsilon}^{2}\leq C\varepsilon|f_2|_{\varepsilon}^{2},$$

where C is a constant independent of ε. Then by Eq. (4.82),

$$\left|\widetilde{T}\right|_{\varepsilon}^{2}\leq C\varepsilon^{2}|f_2|_{\varepsilon}^{2}. \tag{4.87}$$

Next, transform Eq. (4.80) into a homogeneous Neumann condition. Let $T^{*}=\eta\widetilde{T}$, where

$$\begin{aligned}&\eta(x_1,x_2,x_3)\\&\quad=\exp\left[\alpha x_3+\frac{\alpha x_3^{2}}{2\varepsilon h(x_1,x_2)}+\alpha x_3^{2}(x_3+\varepsilon h(x_1,x_2))\varphi(x_1,x_2)\right]\end{aligned} \tag{4.88}$$

with

$$\varphi=\frac{|\nabla h|^{2}}{2h^{2}(1+\varepsilon^{2}|\nabla h|^{2})}. \tag{4.89}$$

Noting that

$$\begin{aligned}&\frac{\partial T^{*}}{\partial x_3}=\frac{\partial\eta}{\partial x_3}\widetilde{T}+\eta\frac{\partial\widetilde{T}}{\partial x_3},\qquad\frac{\partial T^{*}}{\partial n}=\frac{\partial\eta}{\partial n}\widetilde{T}+\eta\frac{\partial\widetilde{T}}{\partial n},\\&\frac{\partial\eta}{\partial x_3}=\eta\left[\alpha+\frac{\alpha x_3}{\varepsilon h}+2\alpha x_3(x_3+\varepsilon h)\varphi+\alpha x_3^{2}\varphi\right],\end{aligned} \tag{4.90}$$

and at $x_3=0$, $\partial\eta/\partial x_3=\alpha\eta$ and $\eta=1$, which implies that

$$\frac{\partial T^{*}}{\partial x_3}=\eta\left(\frac{\partial\widetilde{T}}{\partial x_3}+\alpha\widetilde{T}\right)=0.$$

Furthermore, at $x_3=-\varepsilon h(x_1,x_2)$, we have

$$\frac{\partial\eta}{\partial x_3}=\alpha\eta\varepsilon^{2}h^{2}(x_1,x_2)\varphi(x_1,x_2). \tag{4.91}$$

Now we compute $\nabla\eta$, where $\nabla=(\partial/\partial x_1,\partial/\partial x_2)$:

$$\nabla\eta=\eta\left[-\frac{\alpha x_3^{2}}{2\varepsilon h^{2}}\nabla h+\varepsilon\alpha x_3^{2}\varphi\nabla h+\alpha x_3^{2}(x_3+\varepsilon h)\nabla\varphi\right]. \tag{4.92}$$

Hence, at $x_3 = -\varepsilon h(x_1, x_2)$, $\nabla\eta = \eta[-\frac{1}{2}\alpha\varepsilon\xi\nabla h + \varepsilon^3\alpha h^2\xi\nabla h]$ and

$$\begin{aligned}\frac{\partial\eta}{\partial n} &= \frac{\partial\eta}{\partial x_3} + \varepsilon\nabla h\nabla\eta \\ &= \eta\left[\alpha\varepsilon^2h^2\varphi - \frac{\varepsilon^2\alpha}{2}|\nabla h|^2 + \varepsilon^4\alpha h^2\varphi|\nabla h|^2\right] \\ &= \varepsilon^2\eta\alpha\left[h^2\varphi(1+\varepsilon^2|\nabla h|^2) - \frac{1}{2}|\nabla h|^2\right] = 0. \end{aligned} \tag{4.93}$$

That is, $\partial\eta/\partial n|_{\Gamma_b} = 0$, and $\partial T^*/\partial n|_{\Gamma_b} = 0$. Assume now Eq. (4.54), i.e.,

$$\nabla h \cdot n_{\Gamma_i} = 0 \quad \text{on } \partial\Gamma_i. \tag{4.94}$$

One can easily check that $\nabla\varphi \cdot n_{\Gamma_i} = 0$: using a system (s, t) of local coordinates on $\partial\Gamma_i$, with s normal to $\partial\Gamma_i$ and t tangential, and using Eqs. (4.66) and (4.67). Hence,

$$\nabla h \cdot n_{\Gamma_i} = 0 \quad \text{and} \quad \nabla\varphi \cdot n_{\Gamma_i} = 0 \quad \text{on } \partial\Gamma_i, \tag{4.95}$$

and we have immediately

$$\nabla\eta \cdot n_{\Gamma_i} = 0 \quad \text{on } \partial\Gamma_i,$$

and therefore

$$\frac{\partial T^*}{\partial n_{\Gamma_i}} = 0 \quad \text{on } \partial\Gamma_i.$$

The conclusion of these computations and of Step 2 is summarized by the following lemma.

LEMMA 4.6. *Assume that* $\partial\Gamma_i$ *is of class* C^2, $h : \overline{\Gamma}_i \to \mathbb{R}_+$ *belong to* $C^4(\overline{\Gamma}_i)$ *and*

$$\nabla h \cdot n_{\Gamma_i} = 0 \quad \textit{on } \partial\Gamma_i.$$

Let $\widetilde{T}$ *be a solution of Eq.* (4.80) *and* $T^* = \eta\widetilde{T}$, *with*

$$\eta = \exp\left[\alpha x_3 + \frac{\alpha x_3^2}{2\varepsilon h(x_1, x_2)} + \alpha x_3^2\big(x_3 + \varepsilon h(x_1, x_2)\big)\varphi(x_1, x_2)\right],$$

where

$$\varphi(x_1, x_2) = \frac{|\nabla h(x_1, x_2)|^2}{2h^2(x_1, x_2)(1+\varepsilon^2|\nabla h(x_1, x_2)|^2)}.$$

Then

$$-\Delta T^* = f^*,$$
$$\frac{\partial T^*}{\partial x_3} = 0 \quad \textit{on } \Gamma_{\mathrm{i}},$$
$$\frac{\partial T^*}{\partial n} = 0 \quad \textit{on } \Gamma_{\mathrm{b}} \cup \Gamma_{\ell},$$

where $f^* = \eta f_2 - 2\nabla_3\eta \cdot \nabla_3\widetilde{T} - \widetilde{T}\Delta_3\eta$ *and* $|f^*|_\varepsilon \le C_0|f_2|_\varepsilon$, *with* C_0 *independent of* ε.

Proof. It remains only to estimate the L^2 norm of f^*. First note that there exists a constant C_0 independent of ε (depending on α and h) such that

$$\frac{1}{C_0} \le \eta(x_1, x_2, x_3) \le C_0 \qquad \text{for } (x_1, x_2, x_3) \in \mathcal{M}_\varepsilon,$$

and using Eqs. (4.90) and (4.92), there exists another constant still denoted as C_0 such that

$$\left|\frac{\partial \eta}{\partial x_3}\right|_{L^\infty(\mathcal{M}_\varepsilon)} \le C_0 \quad \text{and} \quad |\nabla\eta|_{L^\infty(\mathcal{M}_\varepsilon)} \le C_0.$$

Now, we compute $\Delta\eta(\Delta = \partial^2/\partial x_1^2 + \partial^2/\partial x_2^2)$:

$$\begin{aligned}\Delta\eta &= \operatorname{div}(\nabla\eta)\\ &= \eta\left[\frac{\alpha x_3^2}{2\varepsilon}\Delta\left(\frac{1}{h}\right) + \varepsilon\alpha x_3^2 \operatorname{div}(\varphi\nabla h) + \alpha x_3^2 \operatorname{div}((x_3 + \varepsilon h)\nabla\varphi)\right],\end{aligned}$$

and therefore since $h \in C^4(\overline{\Gamma}_{\mathrm{i}})$, we have

$$|\Delta\eta|_{L^\infty(\mathcal{M}_\varepsilon)} \le C_0, \quad \text{with } C_0 \text{ independent of } \varepsilon.$$

Finally, we compute $\partial^2\eta/\partial x_3^2$:

$$\begin{aligned}\frac{\partial^2\eta}{\partial x_3^2} &= \frac{\partial\eta}{\partial x_3}\left[\alpha + \frac{\alpha x_3}{\varepsilon h} + 2\alpha x_3(x_3 + \varepsilon h)\varphi + \alpha x_3^2\varphi\right]\\ &\quad + \eta\left[\frac{\alpha}{\varepsilon h} + 2\alpha(x_3 + \varepsilon h)\varphi + 6\alpha x_3\varphi\right].\end{aligned}$$

Therefore,

$$\left|\frac{\partial^2\eta}{\partial x_3^2}\right|_{L^\infty(\mathcal{M}_\varepsilon)} \le \frac{C}{\varepsilon},$$

where C is independent of ε. Now,

$$\left|f^*\right|_\varepsilon \le |\eta|_{L^\infty}|f_2|_\varepsilon + |\nabla_3\eta|_{L^\infty}\left|\nabla_3\widetilde{T}\right|_\varepsilon + |\Delta\eta|_{L^\infty}\left|\widetilde{T}\right|_\varepsilon + \left|\frac{\partial^2\eta}{\partial x_3^2}\right|_{L^\infty}\left|\widetilde{T}\right|_\varepsilon,$$

and since by Eqs. (4.84) and (4.87),

$$\left|\nabla\widetilde{T}\right|_\varepsilon \le C|f_2|_\varepsilon \quad \text{and} \quad \left|\widetilde{T}\right|_\varepsilon \le C\varepsilon|f_2|_\varepsilon,$$

we have

$$\left|f^*\right|_\varepsilon \le C|f_2|_\varepsilon,$$

where C is independent of ε. □

4.4. Regularity of the velocity

In this section, we study the H^2 regularity of the velocity, solution of the GFD–Stokes problem (we use either x_3 or z to denote the vertical variable):

$$\begin{cases} -\left(\Delta v + \frac{\partial^2 v}{\partial x_3^2}\right) + \nabla p = f_v & \text{in } \mathcal{M}_\varepsilon, \\ \operatorname{div}\int_{-\varepsilon h}^0 v\,\mathrm{d}z = 0 & \text{in } \Gamma_{\mathrm{i}}, \\ v = 0 & \text{on } \Gamma_\ell \cup \Gamma_{\mathrm{b}}, \\ \dfrac{\partial v}{\partial x_3} + \alpha_v v = g_v & \text{on } \Gamma_{\mathrm{i}}. \end{cases} \tag{4.96}$$

The H^2 regularity of problems similar to Eq. (4.96) is given by Ziane [1995], where $\varepsilon = 1$ and $g_v = 0$, and by Hu, Temam and Ziane [2002], in the case of constant depth function and under a convexity condition of $\mathcal{M}_\varepsilon$. We study here the H^2 regularity of solutions to Eq. (4.96) and give the dependence on ε of the constant appearing in the Cattabriga–Solonnikov-type inequality associated with the H^2 regularity of solutions. By contrast with the articles quoted above, our analysis here will be carried out in the case where $\mathcal{M}_\varepsilon$ is not necessarily convex and with a varying-bottom topography. This regularity result is discussed in Section 4.4.2, and we start in Section 4.4.1 with a discussion of the weak formulation of the GFD–Stokes problem (Eq. (4.96)).

4.4.1. Weak formulation of the GFD–Stokes problem

In this section we drop the index ε that is irrelevant ($\varepsilon = 1$, $\mathcal{M}_\varepsilon = \mathcal{M}$). For the weak formulation of Eq. (4.96), we consider the space

$$V = \left\{ v \in H^1(\mathcal{M})^2, \operatorname{div}\int_{-h}^0 v\,\mathrm{d}z = 0 \text{ on } \Gamma_{\mathrm{i}},\ v = 0 \text{ on } \Gamma_\ell \cup \Gamma_{\mathrm{b}} \right\};$$

thanks to the Poincaré inequality, this space is Hilbert for the scalar product

$$\big((v,\tilde{v})\big) = \sum_{i=1}^2 \sum_{j=1}^3 \int_{\mathcal{M}} \frac{\partial v_i}{\partial x_j}\frac{\partial \tilde{v}_i}{\partial x_j}\,\mathrm{d}\mathcal{M}.$$

To obtain the weak formulation, we multiply the first equation (Eq. (4.96)) by a test function $\tilde{v} \in V$ and integrate over $\mathcal{M}$; assuming regularity for v, p, and $\tilde{v}$, the term involving p (independent of x_3) disappears:

$$\begin{aligned}
&\int_{\mathcal{M}} \nabla p\tilde{v}\,\mathrm{d}\mathcal{M} \\
&\quad = \int_{\partial\mathcal{M}} p\tilde{v}\cdot n\,\mathrm{d}(\partial\mathcal{M}) - \int_{\mathcal{M}} p\,\mathrm{div}\,\tilde{v}\,\mathrm{d}\mathcal{M} \quad \text{(by Stokes formula)} \\
&\quad = -\int_{\mathcal{M}} p\,\mathrm{div}\,\tilde{v}\,\mathrm{d}\mathcal{M} \\
&\quad = -\int_{\Gamma_\mathrm{i}} p\left(\int_{-h}^{0} \mathrm{div}\,\tilde{v}\,\mathrm{d}x_3\right)\mathrm{d}x_1\,\mathrm{d}x_2 \\
&\quad = -\int_{\Gamma_\mathrm{i}} p\left(\mathrm{div}\int_{-h}^{0} \tilde{v}\,\mathrm{d}x_3\right)\mathrm{d}x_1\,\mathrm{d}x_2 \\
&\quad = 0 \quad \text{(by the properties of } \tilde{v}\text{)}.
\end{aligned}$$

We also have with Stokes formula and since $\tilde{v}$ vanishes on $\Gamma_\mathrm{b} \cup \Gamma_\ell$,

$$\begin{aligned}
-\int_{\mathcal{M}} \left(\Delta v + \frac{\partial^2 v}{\partial x_3^2}\right)\tilde{v}\,\mathrm{d}\mathcal{M} &= -\int_{\Gamma_\mathrm{i}} \frac{\partial v}{\partial x_3}\tilde{v}\,\mathrm{d}\Gamma_\mathrm{i} + ((v,\tilde{v})) \\
&= \alpha_v \int_{\Gamma_\mathrm{i}} v\tilde{v}\,\mathrm{d}\Gamma_\mathrm{i} - \int_{\Gamma_\mathrm{i}} g_v\tilde{v}\,\mathrm{d}\Gamma_\mathrm{i} + ((v,\tilde{v})),
\end{aligned}$$

hence the weak formulation of Eq. (4.96).

To find $v \in V$ such that

$$a(v,\tilde{v}) = \ell(\tilde{v}) \quad \forall \tilde{v} \in V, \tag{4.97}$$

with

$$\begin{aligned}
a(v,\tilde{v}) &= ((v,\tilde{v})) + \alpha_v \int_{\Gamma_\mathrm{i}} v\tilde{v}\,\mathrm{d}\Gamma_\mathrm{i}, \\
\ell(\tilde{v}) &= (f_v,\tilde{v})_H + \int_{\Gamma_\mathrm{i}} g_v\tilde{v}\,\mathrm{d}\Gamma_\mathrm{i}.
\end{aligned} \tag{4.98}$$

Existence and uniqueness of a solution $v \in V$ of Eq. (4.97) follow promptly from the Lax–Milgram theorem. More delicate is the question of showing that, conversely, v is, in some sense, solution of Eq. (4.96). The second equation (4.96) and $v = 0$ on $\Gamma_\ell \cup \Gamma_\mathrm{b}$ follow from the fact that $v \in V$; hence, we need to show that there exists a distribution p independent of x_3 such that

$$-\left(\Delta v + \frac{\partial^2 v}{\partial x_3^2}\right) + \nabla p = f_v \quad \text{in } \mathcal{M} \tag{4.99}$$

and also that

$$\frac{\partial v}{\partial x_3} + \alpha_v v = g_v \quad \text{on } \Gamma_i. \tag{4.100}$$

For Eq. (4.99), consider a test function $\varphi \in \mathcal{C}_0^\infty(\mathcal{M})$ ($\mathcal{C}^\infty$ with compact support in $\mathcal{M}$), and observe that $\tilde{v} = (\tilde{v}_1, 0)$, $\tilde{v}_1 = \partial\varphi/\partial x_3$, belongs to V. Writing Eq. (4.97) with this $\tilde{v}$, we conclude that

$$\frac{\partial}{\partial x_3}(\Delta_3 v_1 + f_{v1}) = 0.$$

In the same way, we prove that

$$\frac{\partial}{\partial x_3}(\Delta_3 v_2 + f_{v2}) = 0, \tag{4.101}$$

showing that each component of $\Delta v + f_v$ is a distribution on $\mathcal{M}$ independent of x_3.

Distributions independent of x_3 Now we can identify a distribution G on $\mathcal{M}$ independent of x_3, with a distribution on Γ_i as follows: let θ be any $\mathcal{C}^\infty$ scalar function with compact support in $(-\underline{h}, 0)$, and such that

$$\int_{-\underline{h}}^{0} \theta(z)\,\mathrm{d}z = 1. \tag{4.102}$$

Then, if $\varphi \in \mathcal{C}_0^\infty(\Gamma_i)$ is a $\mathcal{C}^\infty$ scalar function with a compact support in Γ_i, $\varphi\theta \in \mathcal{C}_0^\infty(\mathcal{M}_\varepsilon)$, and we associate to G a distribution $\widetilde{G}$ on Γ_i by setting

$$\langle \widetilde{G}, \varphi \rangle_{\Gamma_i} = \langle G, \varphi\theta \rangle_{\mathcal{M}}. \tag{4.103}$$

The right-hand side of Eq. (4.103) is independent of θ; indeed, if θ_1 and θ_2 are two such functions, then $\langle G, \varphi\theta_1 \rangle = \langle G, \varphi\theta_2 \rangle$ because

$$\int_{-\underline{h}}^{0} (\theta_1 - \theta_2)(z)\,\mathrm{d}z = 0$$

so that $\theta_0(x_3) = \int_{x_3}^{0} (\theta_1 - \theta_2)(z)\,\mathrm{d}z$ is a $\mathcal{C}^\infty$ function with compact support in $(-\underline{h}, 0)$, and

$$\langle G, \varphi(\theta_1 - \theta_2) \rangle_{\mathcal{M}} = -\left\langle G, \frac{\partial}{\partial x_3}(\varphi\theta_0) \right\rangle_{\mathcal{M}} = \left\langle \frac{\partial G}{\partial x_3}, \varphi\theta_0 \right\rangle_{\mathcal{M}} = 0.$$

It is then easy to see that Eq. (4.103) defines $\widetilde{G}$ as a distribution on Γ_i.

Now, conversely, assume that $\widetilde{G}$ is a distribution on Γ_i, and let $\varphi \in \mathcal{C}_0^\infty(\mathcal{M})$. It is clear that $\tilde{\varphi} = \int_{-\underline{h}}^{0} \varphi\,\mathrm{d}x_3$ belongs to $\mathcal{C}_0^\infty(\Gamma_i)$, and we associate to $\widetilde{G}$ a distribution G on $\mathcal{M}$ by setting

$$\langle G, \varphi \rangle_{\mathcal{M}} = \langle \widetilde{G}, \tilde{\varphi} \rangle_{\Gamma_i} \quad \forall \varphi \in \mathcal{C}_0^\infty(\mathcal{M}).$$

Introduction of p Thanks to the previous discussion, we can now consider $\Delta v + f_v$ as a distribution on Γ_i. As in the theory of Navier-Stokes equations, consider now a vector function $v^* \in \mathcal{V}(\Gamma_i)$, i.e., v^* is (2D) $\mathcal{C}^\infty$ with compact support in Γ_i, and $\operatorname{div} v^* = 0$. It is clear that $\tilde{v} = v^*\theta$ belongs to V, where θ is a function as above (see Eq. (4.102)). Writing Eq. (4.97) with this $\tilde{v}$, we obtain

$$\begin{aligned} &((v, v^*\theta)) = (f_v, v^*\theta)_H, \\ &\langle \Delta v + f_v, v^*\theta \rangle_{\mathcal{M}} = 0, \\ &\langle \Delta v + f_v, v^* \rangle_{\Gamma_i} = 0 \quad \forall v^* \in \mathcal{V}(\Gamma_i). \end{aligned} \tag{4.104}$$

The last equation, which is well known in the theory of Navier-Stokes equations (see Lions [1969], Temam [1977]), implies that there exists a distribution p on Γ_i such that

$$\Delta v + f_v = \nabla p \quad \text{in } \Gamma_i \text{ (or } \mathcal{M}),$$

and Eq. (4.99) is proved.

A trace theorem The proof of Eq. (4.100) necessitates establishing first a trace theorem: we need to show that for a function v in V such that Eq. (4.99) holds, one can define the trace of $\partial v/\partial x_3$ on Γ_i as an element of $(H_{00}^{1/2}(\Gamma_i))'$, the dual of $H_{00}^{1/2}(\Gamma_i)$ (i.e., the $1/2$ interpolate between $H_0^1(\Gamma_i)$ and $L^2(\Gamma_i)$).

Observe first that the trace on Γ_i of any function Φ in $H^1(\mathcal{M})$, which vanishes on Γ_ℓ, belongs to $H_{00}^{1/2}(\Gamma_i)$. Indeed by odd symmetry and truncation, one can extend such a Φ as a function Φ^* in $H_0^1(\Gamma_i \times \mathbb{R})$, vanishing for $|x_3|$ sufficiently large, and the trace of such a function on any plane $x_3 = c_0$ belongs to $H_{00}^{1/2}(\Gamma_i)$. Conversely, if φ belongs to $H_{00}^{1/2}(\Gamma_i)$, there exists Φ^* in $H_0^1(\Gamma_i \times \mathbb{R})$ such that the trace of Φ^* on Γ_i is φ, the mapping $\varphi \to \Phi^*$ being linear continuous (lifting operator). From the remark above, we infer that the trace on Γ_i of a function in V belongs to $H_{00}^{1/2}(\Gamma_i)$.

We then show that the traces on Γ_i of the functions of V are all in $H_{00}^{1/2}(\Gamma_i)^2$. Indeed, let $\varphi \in H_{00}^{1/2}(\Gamma_i)^2$. Using the previous lifting operator, there exists $\widetilde{\Phi} \in H_0^1(\Gamma_i \times \mathbb{R})^2$ such that $\widetilde{\Phi}|_{\Gamma_i} = \varphi$; by truncation, we can assume that $\widetilde{\Phi} \in H^1(\Gamma_i \times (-\underline{h}, 0))^2$ and $\widetilde{\Phi}$ vanishes on $\partial\Gamma_i$ and at $x_3 = -\underline{h}$. Let

$$\xi = \operatorname{div} \int_{-\underline{h}}^{0} \widetilde{\Phi}\, \mathrm{d}x_3,$$

and observe that $\xi \in L^2(\Gamma_i)$ and

$$\int_{\Gamma_i} \xi \, \mathrm{d}\Gamma_i = \int_{Q_{\underline{h}}} \operatorname{div} \widetilde{\Phi} \, \mathrm{d}\mathcal{M} = \int_{\partial Q_\varepsilon} \widetilde{\Phi} \cdot n_h \, \mathrm{d}(\partial\mathcal{M}) = 0, \tag{4.105}$$

where n_h is the horizontal component of the unit outward normal n on $\partial\mathcal{M}$. Because of Eq. (4.105), we can solve in $Q_{\underline{h}} = \Gamma_i \times (-\underline{h}, 0)$ the (usual) Stokes problem

$$\begin{cases} \Delta\Phi^* + \nabla\pi = 0 & \text{in } Q_{\underline{h}}, \\ \operatorname{div}_3 \Phi^* = \frac{1}{\underline{h}}\xi & \text{in } Q_{\underline{h}}, \\ \Phi^* = 0 & \text{on } \partial Q_{\underline{h}}, \end{cases} \tag{4.106}$$

and $\Phi^* \in H^1(Q_{\underline{h}})^3$, $\pi \in L^2(Q_{\underline{h}})$. Now,

$$\int_{-\underline{h}}^{0} \operatorname{div}_3 \Phi^* \, \mathrm{d}x_3 = \operatorname{div} \int_{-\underline{h}}^{0} (\phi_1^*, \phi_2^*) \, \mathrm{d}x_3 = \xi,$$

and it is easy to see that the function $\Phi = \widetilde{\Phi} - (\Phi_1^*, \Phi_2^*)$ extended by 0 in $\mathcal{M}\backslash Q_{\underline{h}}$ belongs to V, and its trace on Γ_i is precisely φ. We can, furthermore, observe that with the construction above, the mapping $\varphi \mapsto \Phi$ is linear continuous from $H_{00}^{1/2}(\Gamma_i)$ into V.

Finally, Eq. (4.100) follows promptly from Eqs. (4.97), (4.99), and the following proposition.

PROPOSITION 4.1. *Let v be a function in $H^1(\mathcal{M})^2$, which vanishes on $\Gamma_b \cup \Gamma_\ell$, and assume that $-\Delta v + \nabla p \in L^2(\mathcal{M})^2$, for some distribution p independent of x_3.*

Then, there exists $\gamma_1 v \in (H_{00}^{1/2}(\Gamma_i))^2$ such that

$$\gamma_1 v = \left.\frac{\partial v}{\partial x_3}\right|_{\Gamma_i} \quad \text{if } v \in \mathcal{C}^2\big(\overline{\mathcal{M}}\big)^2, \tag{4.107}$$

and $\gamma_1 v$ is defined by

$$\langle \gamma_1 v, \varphi \rangle = \big((v, \Phi)\big) - \int_{\mathcal{M}} (-\Delta v + \nabla p)\Phi \, \mathrm{d}\mathcal{M}, \tag{4.108}$$

where φ is arbitrary in $H_{00}^{1/2}(\Gamma_i)$ and Φ is any function of V such that $\Phi|_{\Gamma_i} = \varphi$.

PROOF. We first show that the right-hand side $X(\Phi)$ of Eq. (4.108) depends on φ and not on Φ. Indeed, let Φ_1 and Φ_2 be two functions of V such that $\Phi_1|_{\Gamma_i} = \Phi_2|_{\Gamma_i} = \varphi$. Then, $\Phi^* = \Phi_1 - \Phi_2$ belongs to $H_0^1(\mathcal{M})^2$ and $\operatorname{div} \int_{-\underline{h}}^{0} \Phi^* \, \mathrm{d}x_3 = 0$. It was shown by LIONS, TEMAM and WANG [1992a] that Φ^* is limit in $H_0^1(\mathcal{M})^2$ of $\mathcal{C}^\infty$ functions Φ_n^* with compact support in $\mathcal{M}$ such that $\operatorname{div} \int_{-\underline{h}}^{0} \Phi_n^* \, \mathrm{d}x_3 = 0$. It is easy to see that $X(\Phi_n^*) = 0$ and, by continuity, $X(\Phi^*) = 0$, i.e., $X(\Phi_1) = X(\Phi_2)$.

After this observation, we choose Φ as constructed above so that the mapping $\varphi \to \Phi$ is linear continuous from $H_{00}^{1/2}(\Gamma_i)^2$ into V. It then appears that the right-hand side of Eq. (4.108) is a linear form continuous on $H_{00}^{1/2}(\Gamma_i)^2$, and thus $\gamma_1 v$ is defined and belongs to $(H_{00}^{1/2}(\Gamma_i))^2$. Finally, Eq. (4.107) follows from the fact that Eq. (4.108) is easy when v and Φ are smooth and $\gamma_1 v$ is replaced by $\partial v/\partial x_3|_{\Gamma_i}$. □

REMARK 4.2. We have shown the complete equivalence of Eq. (4.96) with its variational formulation (Eq. (4.97)).

4.4.2. H^2 regularity for the GFD–Stokes problem

For convenience, we use hereafter the classical notation $\mathbb{L}^2$, $\mathbb{H}^1$, etc., for spaces of vector functions with components in L^2, H^1, etc.

The main result of this section is the following theorem.

THEOREM 4.6. *Assume that h is a positive function in $\mathcal{C}^4(\overline{\Gamma}_i)$, $h \geq \underline{h} > 0$ and that $f_v \in \mathbb{L}^2(\mathcal{M}_\varepsilon)$ and $g_v \in \mathbb{H}^1_0(\Gamma_\mathrm{i})$. Let $(v, p) \in \mathbb{H}^1(M_\varepsilon) \times L^2(\Gamma_\mathrm{i})$ be a weak solution of Eq.* (4.96). *Then,*

$$(v, p) \in \mathbb{H}^2(\mathcal{M}_\varepsilon) \times H^1(\mathcal{M}_\varepsilon). \tag{4.109}$$

Moreover, the following inequality holds:

$$|v|^2_{\mathbb{H}^2(\mathcal{M}_\varepsilon)} + \varepsilon |p|^2_{H^1(\Gamma_\mathrm{i})} \leq C\big[|f_v|^2_\varepsilon + |g_v|^2_{L^2(\Gamma_\mathrm{i})} + \varepsilon |\nabla g_v|^2_{L^2(\Gamma_\mathrm{i})}\big]. \tag{4.110}$$

The approach to the proof of the H^2 regularity in Theorem 4.6 is the same as in the articles by ZIANE [1995, 1997] and HU, TEMAM and ZIANE [2002]; it is based on the following observation: the weak solution of Eq. (4.96) satisfies $p \in L^2(\Gamma_\mathrm{i})$; assume further that the solution v of Eq. (4.96) satisfies $\frac{\partial v}{\partial x_3}|_{\Gamma_\mathrm{i}} \in \mathbb{L}^2(\Gamma_\mathrm{i})$, $\frac{\partial v}{\partial x_3}|_{\Gamma_\mathrm{b}} \in \mathbb{L}^2(\Gamma_\mathrm{b})$, and $\frac{\partial v}{\partial x_k}|_{\Gamma_\mathrm{b}} \in \mathbb{L}^2(\Gamma_\mathrm{b})$, $k = 1, 2$, then an integration of the first equation in Eq. (4.96) with respect to x_3 over $(-\varepsilon h, 0)$ yields a 2D Stokes problem on the smooth domain Γ_i with a homogeneous boundary condition. By the classical regularity theory of the 2D Stokes problem in smooth domains (see, for instance, GHIDAGLIA [1984], TEMAM [1977], and CONSTANTIN and FOIAS [1988]), p belongs to $H^1(\Gamma_\mathrm{i})$. Then, by moving the pressure term to the right-hand side, problem Eq. (4.96) reduces to an elliptic problem of the type studied in Section 4.2, and the H^2 regularity of v follows. The estimates on the L^2-norms of the second derivatives are then obtained using the trace theorem and the estimates in Section 4.2.

We start this proof by showing that $\frac{\partial v}{\partial x_3}|_{\Gamma_\mathrm{i}} \in \mathbb{L}^2(\Gamma_\mathrm{i})$, $\frac{\partial v}{\partial x_3}|_{\Gamma_\mathrm{b}} \in \mathbb{L}^2(\Gamma_\mathrm{b})$, and $\frac{\partial v}{\partial x_k}|_{\Gamma_\mathrm{i}} \in \mathbb{L}^2(\Gamma_\mathrm{i})$, $k = 1, 2$. The following lemma is just a rewriting of Theorem 4.4.

LEMMA 4.7. *Assume that $h \in \mathcal{C}^2(\overline{\Gamma}_i)$. For $f \in L^2(\mathcal{M}_\varepsilon)$ and $g \in H^1_0(\Gamma_\mathrm{i})$, there exists a unique $\Psi \in H^2(\mathcal{M}_\varepsilon)$ solution of*

$$\begin{cases} -\Delta_3 \Psi = f & \text{in } \mathcal{M}_\varepsilon, \\ \frac{\partial \Psi}{\partial x_3} + \alpha \Psi = g & \text{on } \Gamma_\mathrm{i}, \\ \Psi = 0 & \text{on } \Gamma_\mathrm{b} \cup \Gamma_\mathrm{l}. \end{cases} \tag{4.111}$$

Furthermore, there exists a constant $C(h, \alpha)$ depending only on α and h (and Γ_i) such that

$$\sum_{k,j=1}^{3} \left| \frac{\partial^2 \Psi}{\partial x_k \, \partial x_j} \right|^2_\varepsilon \leq C(h, \alpha)\big[|f|^2_\varepsilon + |g|^2_i + |\nabla g|^2_i\big].$$

As we said, this lemma is just a rewriting of Theorem 4.4. We will also need the following intermediate result simply obtained by interpolation between H^1 and H^2.

LEMMA 4.8. *Under the assumptions of Lemma* 4.7, *and with* $g = 0$: *if* $f \in H^{-1/2+\delta}(\mathcal{M}_\varepsilon)$ *with* $-\frac{1}{2} < \delta < \frac{1}{2}$, $\delta \neq 0$, *then* $\Psi \in H^{3/2+\delta}(\mathcal{M}_\varepsilon)$.

Before we start the proof of the main result of this section (Theorem 4.6), we first prove the following.

LEMMA 4.9. *Assume that* $h \in \mathcal{C}^3(\overline{\Gamma}_i)$. *For* $f \in L^2(\mathcal{M}_\varepsilon)$, $g \in H_0^1(\Gamma_\mathrm{i})$, *and* $\psi_i \in H_0^{1+\gamma}(\Gamma_\mathrm{b})$, $-\frac{1}{2} < \gamma < \frac{1}{2}$, $\gamma \neq 0$, *there exists a unique* $\Psi_i \in H^{3/2+\gamma}(M_\varepsilon)$ *solution of*

$$\begin{cases} -\triangle_3 \Psi_i = f & \text{in } \mathcal{M}_\varepsilon, \\ \frac{\partial \Psi_i}{\partial x_3} + \alpha_v \Psi_i = g - \alpha_v \psi_i & \text{on } \Gamma_\mathrm{i}, \\ \Psi_i = -\psi_i & \text{on } \Gamma_\mathrm{b}, \\ \Psi_i = 0 & \text{on } \Gamma_\mathrm{l}. \end{cases} \tag{4.112}$$

PROOF. Using Lemma 4.7, we reduce the problem to the case $f = 0$ and $g = 0$ by replacing Ψ_i with $\Psi_i - \Psi$, where Ψ is the function constructed in Lemma 4.7. Thus, without loss of generality, we will assume from now on that $f = 0$ and $g = 0$. Our next step is to construct a function $\tilde{v}_p$ that agrees with Ψ_i on $\partial\mathcal{M}_\varepsilon$. This will be done by first constructing an auxiliary function v_p on a straight cylinder and then the explicit expression of $\tilde{v}_p$ will be given.

Let Q_ε be the cylinder $Q_\varepsilon = \Gamma_\mathrm{i} \times (-\varepsilon, 0)$, and let v_p be the unique solution of

$$\begin{cases} \Delta_3 v_p = 0 & \text{in } Q_\varepsilon, \\ v_p = 0 & \text{on } \partial\Gamma_\mathrm{i} \times (-\varepsilon, 0), \\ v_p = -\psi_i & \text{on } \Gamma_\mathrm{i} \times \{-\varepsilon\}, \\ v_p = \varepsilon h \alpha_v \psi_i & \text{on } \Gamma_\mathrm{i} \times \{0\}. \end{cases} \tag{4.113}$$

We will show that $v_p \in H^{3/2+\gamma}(Q_\varepsilon)$ for all $-\frac{1}{2} < \gamma < \frac{1}{2}$, $\gamma \neq 0$. Then, setting

$$\tilde{v}_p(x_1, x_2, x_3) = -\frac{x_3}{\varepsilon h(x_1, x_2)} v_p\left(x_1, x_2, \frac{x_3}{h(x_1, x_2)}\right) \quad \text{for } (x_1, x_2, x_3) \in \mathcal{M}_\varepsilon, \tag{4.114}$$

it is obvious that $\tilde{v}_p \in H^{3/2+\gamma}(\mathcal{M}_\varepsilon)$, $\tilde{v}_p(x_1, x_2, -\varepsilon h(x_1, x_2)) = -\psi_i(x_1, x_2)$, and $\frac{\partial \tilde{v}_p}{\partial x_3} + \alpha_v \tilde{v}_p = -\alpha_v \psi_i$ on Γ_i. Therefore, setting $\widetilde{V} = \Psi_i - \tilde{v}_p$, we have

$$\begin{aligned} &\Delta_3 \widetilde{V} = -\Delta_3 \tilde{v}_p \in H^{-1/2+\gamma}(\mathcal{M}_\varepsilon), \\ &\widetilde{V} = 0 \quad \text{on } \Gamma_\mathrm{l} \cup \Gamma_\mathrm{b}, \\ &\frac{\partial \widetilde{V}}{\partial x_3} + \alpha_v \widetilde{V} = 0 \quad \text{on } \Gamma_\mathrm{i}. \end{aligned} \tag{4.115}$$

Hence, thanks to Lemmas 4.7 and 4.8, we see that $\widetilde{V}$ and thus Ψ_i are in $H^{3/2+\gamma}(M_\varepsilon)$ for $-\frac{1}{2} < \gamma < \frac{1}{2}$, $\gamma \neq 0$.

To complete the proof of Lemma 4.9, it remains only to show that $v_p \in H^{3/2+\gamma}(Q_\varepsilon)$ for all $-\frac{1}{2} < \gamma < \frac{1}{2}$, $\gamma \neq 0$. To this end, let $\widehat{Q}_\varepsilon$ be any C^2-domain containing Q_ε such that $\Gamma_\mathrm{i} \times \{-\varepsilon, 0\} \subset \partial\widehat{Q}_\varepsilon$. Since ψ_i (respectively $h\alpha_v\psi_i$) is in $H_0^{1+\gamma}(\Gamma_\mathrm{i} \times \{-\varepsilon\})$ (respectively $H_0^{1+\gamma}(\Gamma_\mathrm{i} \times \{0\})$), we can define a function $V_i \in H^1(\partial\widehat{Q}_\varepsilon)$ by setting $V_i = -\psi_i$ on $\Gamma_\mathrm{i} \times \{-\varepsilon\}$, $V_i = \varepsilon h\alpha_v\psi_i$ on $\Gamma_\mathrm{i} \times \{0\}$, and $V_i = 0$ on $\partial\widehat{Q}_\varepsilon \backslash \Gamma_\mathrm{i} \times \{-\varepsilon, 0\}$. Now, let V_p be the unique solution of $\Delta_3 V_p = 0$ in $\widehat{Q}_\varepsilon$ and $V_p = V_i$ on $\partial\widehat{Q}_\varepsilon$. Since $\partial\widehat{Q}_\varepsilon$ is of class C^2, the classical regularity results for elliptic problems (see Lions and Magenes [1972]) yield $V_p \in H^{3/2+\gamma}(\widehat{Q}_\varepsilon)$ for $-\frac{1}{2} < \gamma < \frac{1}{2}$, $\gamma \neq 0$. Now let $\widetilde{V}_i$ be the trace of V_p on $\partial\Gamma_\mathrm{i} \times (-\varepsilon, 0)$. It is easy to see that $\widetilde{V}_i \in H_0^{1+\gamma}(\partial\Gamma_\mathrm{i} \times (-\varepsilon, 0))$. Let $\widetilde{V}_p = V_p - v_p$, we have

$$\begin{aligned} &\Delta_3 \widetilde{V}_p = 0 \quad \text{in } Q_\varepsilon, \\ &\widetilde{V}_p = 0 \qquad \text{on } \Gamma_\mathrm{i} \times \{-\varepsilon, 0\}, \\ &\widetilde{V}_p = \widetilde{V}_i \qquad \text{on } \partial\Gamma_\mathrm{i} \times (-\varepsilon, 0). \end{aligned} \tag{4.116}$$

Using a reflection argument around $x_3 = 0$ (respectively $x_3 = -\varepsilon$) by extending $\widetilde{V}_i$ in a "symmetrically" odd function defined on $\partial\Gamma_\mathrm{i} \times (-\varepsilon, \varepsilon)$ (respectively $\partial\Gamma_\mathrm{i} \times (-2\varepsilon, 0)$) and using the classical local regularity theory (see Lions and Magenes [1972]), we conclude that $\widetilde{V}_p \in H^{3/2+\gamma}(Q_\varepsilon)$ for $-\frac{1}{2} < \gamma < \frac{1}{2}$, $\gamma \neq 0$. Therefore, since $V_p \in H^{3/2+\gamma}(Q_\varepsilon)$, we have $v_p = V_p - \widetilde{V}_p \in H^{3/2+\gamma}(Q_\varepsilon)$. □

Lemma 4.10. *Assume that $h \in C^3(\overline{\Gamma}_\mathrm{i})$, with $h \geq \underline{h}_1 > 0$ on $\overline{\Gamma}_\mathrm{i}$. Let (v, p) be the weak solution of Eq.* (4.96), *then $v \in \mathbb{H}^{2-\delta}(\mathcal{M}_\varepsilon)$ for $0 < \delta < \frac{1}{2}$ and consequently,*

$$\left.\frac{\partial v}{\partial x_3}\right|_{\Gamma_\mathrm{i}} \in \mathbb{L}^2(\Gamma_\mathrm{i}), \quad \nabla v|_{\Gamma_\mathrm{b}} \text{ and } \left.\frac{\partial v}{\partial x_3}\right|_{\Gamma_\mathrm{b}} \in \mathbb{L}^2(\Gamma_\mathrm{b}). \tag{4.117}$$

Proof. We saw in Section 4.1.1 that Eq. (4.97) has a unique solution $v \in H^1(\mathcal{M})^2$, and that there exists p such that (v, p) satisfy Eq. (4.96). By Eq. (4.96), ∇p belongs to $H^{-1}(\mathcal{M}_\varepsilon)$ and thus to $H^{-1}(\Gamma_\mathrm{i})$ since p is independent of x_3 (see Section 4.4.1). Let $v_i \in H_0^1(\Gamma_\mathrm{i})$ be the unique solution of the 2D Dirichlet problem on Γ_i:

$$\begin{cases} \Delta v_i = \nabla p & \text{in } \Gamma_\mathrm{i}, \\ v_i = 0 & \text{on } \partial\Gamma_\mathrm{i}. \end{cases} \tag{4.118}$$

Let $\tilde{v} = v - v_i$, then $\tilde{v}$ satisfies

$$\begin{cases} \Delta_3 \tilde{v} = f_v & \text{in } \mathcal{M}_\varepsilon, \\ \tilde{v} = 0 & \text{on } \Gamma_\ell, \\ \tilde{v} = -v_i & \text{on } \Gamma_\mathrm{b}, \\ \frac{\partial \tilde{v}}{\partial x_3} + \alpha_v \tilde{v} = g_v - \alpha_v v_i & \text{on } \Gamma_\mathrm{i}. \end{cases} \tag{4.119}$$

Thanks to Lemma 4.7, with $g = g_v$, $\psi_i = v_i$, and $\gamma = -\delta$ for some $0 < \delta < \frac{1}{2}$, we have $\tilde{v} \in H^{3/2-\delta}(\mathcal{M}_\varepsilon)$. Hence,

$$g_i = -\frac{1}{\varepsilon h}\int_{-\varepsilon h}^{0} \operatorname{div} \tilde{v}\, \mathrm{d}x_3 \in H^{1/2-\delta}(\Gamma_\mathrm{i}). \tag{4.120}$$

Therefore, since $\operatorname{div} v_i = g_i$, we rewrite the equation for v_i in the form of a 2D Stokes problem

$$\begin{cases} -\Delta v_i + \nabla p = 0 & \text{in } \Gamma_\mathrm{i}, \\ \operatorname{div} v_i = g_i \in H^{1/2-\delta}(\Gamma_\mathrm{i}), & \\ v_i = 0 & \text{on } \partial\Gamma_\mathrm{i}, \end{cases} \tag{4.121}$$

and thanks to the classical regularity result for the nonhomogeneous Stokes problem on Γ_i (see Ghidaglia [1984], Temam [1977]), we have $v_i \in H^{3/2-\delta}(\Gamma_\mathrm{i}) \cap H_0^1(\Gamma_\mathrm{i}) = H_0^{3/2-\delta}(\Gamma_\mathrm{i})$. With this new information on the regularity of v_i, we return to problem (Eq. (4.119)), and using Lemma 4.9 with $\psi_i = v_i$ and $\gamma = \frac{1}{2} - \delta$, $0 < \gamma < \frac{1}{2}$, we conclude that $\tilde{v} \in H^{2-\delta}(\mathcal{M}_\varepsilon)$. Therefore, g_i given by Eq. (4.120) belongs to $H^{1-\delta}(\Gamma_\mathrm{i})$. This in turn implies by the classical regularity of the 2D Stokes problem that the solution v_i of Eq. (4.121) is in $H^{2-\delta}(\Gamma_\mathrm{i})$. Therefore, $v = \tilde{v} + v_i$ belongs to $H^{2-\delta}(\mathcal{M}_\varepsilon)$. Consequently, the trace on Γ_i of the normal derivative $\partial v/\partial x_3|_{\Gamma_\mathrm{i}}$ belongs to $H^{1/2-\delta}(\Gamma_\mathrm{i})$, hence to $L^2(\Gamma_\mathrm{i})$, taking, e.g., $\delta = 1/4$. Similarly, the traces on Γ_b of v and its normal derivative $\partial v/\partial n$ belong to $H^{3/2-\delta}(\Gamma_\mathrm{b})$ and $H^{1/2-\delta}(\Gamma_\mathrm{b})$, respectively, from which we infer that $\nabla v|_{\Gamma_\mathrm{b}}$ and $\partial v/\partial x_3|_{\Gamma_\mathrm{b}}$ are in $H^{1/2-\delta}(\Gamma_\mathrm{b})$ and therefore in $L^2(\Gamma_\mathrm{b})$. The proof of the lemma is now complete. □

Proof of Theorem 4.6. The proof is divided into two steps. In Step 1, we prove the H^2 regularity of solutions, i.e., $v \in H^2(\mathcal{M}_\varepsilon)$ and $p \in H^1(\Gamma_\mathrm{i})$. Then, in Step 2, we establish the Cattabriga–Solonnikov-type inequality on the solutions, i.e., establish the bounds (Eq. (2.97)) on the L^2-norms of the second derivatives of v and the H^1 norm on the pressure, in particular, we establish their (non)dependence on ε.

Step 1 (The H^2-regularity of solutions). Let $\bar{v} = \int_{-\varepsilon h}^{0} v\, \mathrm{d}z$; we have

$$\frac{\partial^2}{\partial x_k^2}\bar{v}(x_1, x_2, x_3) = \int_{-\varepsilon h}^{0} \frac{\partial^2}{\partial x_k^2} v(x_1, x_2, z)\, \mathrm{d}z + I_k(v), \tag{4.122}$$

$$I_k(v) = 2\varepsilon \frac{\partial h}{\partial x_k}\frac{\partial v}{\partial x_3}\big(x_1, x_2, -\varepsilon h(x_1, x_2)\big) - \varepsilon^2\left(\frac{\partial h}{\partial x_k}\right)^2 v\big(x_1, x_2, -\varepsilon h(x_1, x_2)\big), \quad k = 1, 2. \tag{4.123}$$

Therefore, by integrating the first equation in Eq. (4.96) with respect to x_3, we obtain the 2D Stokes problem:

$$\begin{cases} -\Delta \bar{v} + \nabla(\varepsilon h p) = \bar{f} & \text{in } \Gamma_\mathrm{i}, \\ \operatorname{div} \bar{v} = 0 \quad \text{in } \Gamma_\mathrm{i}, & v = 0 \quad \text{on } \partial\Gamma_\mathrm{i}, \end{cases} \tag{4.124}$$

where

$$\bar{f} = \int_{-\varepsilon h}^{0} f_v \, \mathrm{d}z + \left.\frac{\partial v}{\partial x_3}\right|_{x_3=0} - \left.\frac{\partial v}{\partial x_3}\right|_{x_3=-\varepsilon h} + I_1(v) + I_2(v) + \varepsilon p \nabla h. \tag{4.125}$$

Thanks to Lemma 4.10, each term on the right-hand side of Eq. (4.125) is in $L^2(\Gamma_\mathrm{i})$, which implies $\bar{f} \in L^2(\Gamma_\mathrm{i})$: this is stated in Eq. (4.117) for $\partial v/\partial x_3|_{\Gamma_\mathrm{i}}$ and $\partial v/\partial x_3|_{\Gamma_\mathrm{b}}$; similarly, each term in I_1 and I_2 belongs to $L^2(\Gamma_\mathrm{b})$ (and thus $L^2(\Gamma_\mathrm{i})$) because $v \in H^{2-\delta}(\mathcal{M}_\varepsilon)$, $0 < \delta < 1/2$; finally for p, we recall from Eq. (4.118) that $\nabla p \in H^{-1}(\Gamma_\mathrm{i})$ and thus $p \in L^2(\Gamma_\mathrm{i})$. Therefore, from the classical regularity theory of the 2D Stokes problem, we conclude that $\nabla(hp) \in L^2(\Gamma_\mathrm{i})$, and then $\nabla p \in L^2(\Gamma_\mathrm{i})$. We return to problem (Eq. (4.96)) and move the gradient of the pressure to the right-hand side and obtain, thanks to Lemma 4.7, $v \in H^2(\mathcal{M}_\varepsilon)$ and

$$\sum_{k,j=1}^{3} \left| \frac{\partial^2 v}{\partial x_k \, \partial x_j} \right|_\varepsilon^2 \leq C(h, \alpha_v)\big[|f_v|_\varepsilon^2 + |g_v|_i^2 + |\nabla g_v|_i^2\big] + C(h, \alpha)\varepsilon |\nabla p|_i^2. \tag{4.126}$$

Note that we have the pressure term on the right-hand side of Eq. (4.126). Removing that term is done in the second step below.

Step 2 (the Cattabriga–Solonnikov-type inequality). Our aim is now to bound $|\nabla p|_i$ properly and to derive Eq. (4.110) from Eq. (4.126). First, we homogenize the boundary condition in Eq. (4.96). Let $v_\ell = (\Psi_1, \Psi_2)$, where Ψ_1 and Ψ_2 are constructed using Lemma 4.7, i.e.,

$$\begin{cases} -\Delta_3 \Psi_k = f_{v,k} & \text{in } \mathcal{M}_\varepsilon, k = 1, 2, \\ \frac{\partial \Psi_k}{\partial x_3} + \alpha_v \Psi_k = g_{v,k} & \text{on } \Gamma_\mathrm{i}, k = 1, 2, \\ \Psi_k = 0 & \text{on } \Gamma_\mathrm{b} \cup \Gamma_\ell, k = 1, 2, \end{cases}$$

where $f_v = (f_{v,1}, f_{v,2})$, $g_v = (g_{v,1}, g_{v,2})$. Thanks to Lemma 4.7,

$$\sum_{k,j=1}^{3} \left| \frac{\partial^2 v_l}{\partial x_k \, \partial x_j} \right|_\varepsilon^2 \leq C(h, \alpha_v)(|f_v|_\varepsilon^2 + |g_v|_i^2 + |\nabla g_v|_i^2). \tag{4.127}$$

Setting $v^* = v - v_\ell$, it suffices to establish Eq. (4.127) with v_l replaced by v^*. We have

$$\begin{cases} -\big(\Delta v^* + \frac{\partial^2 v^*}{\partial x_3^2}\big) + \nabla p = 0 & \text{in } \mathcal{M}_\varepsilon, \\ \operatorname{div} \int_{-\varepsilon h}^{0} v^* \, \mathrm{d}z = g^* & \text{on } \Gamma_\mathrm{i}, \\ v^* = 0 & \text{on } \Gamma_\ell \cup \Gamma_\mathrm{b}, \\ \frac{\partial v^*}{\partial x_3} + \alpha_v v^* = 0 & \text{on } \Gamma_\mathrm{i}, \end{cases} \tag{4.128}$$

where

$$g^* = -\operatorname{div} \int_{-\varepsilon h}^{0} v_l \, \mathrm{d}x_3.$$

Note that inequality (Eq. (4.127)) together with the Cauchy–Schwarz inequality implies

$$\left\|g^*\right\|^2_{H^1(\Gamma_\mathrm{i})} \le C(h,\alpha_v)\varepsilon\left[|f_1|^2_\varepsilon + |g_v|^2_i + |\nabla g_v|^2_i\right]. \tag{4.129}$$

Define

$$V^* = \int_{-\varepsilon h}^{0} v^*\,\mathrm{d}x_3;$$

V^* is the solution of the 2D Stokes problem

$$\begin{cases} -\Delta V^* + \nabla(\varepsilon p) = F^* & \text{in } \Gamma_\mathrm{i}, \\ \operatorname{div} V^* = g^*, & \\ V^* = 0 & \text{on } \partial\Gamma_\mathrm{i}, \end{cases} \tag{4.130}$$

where

$$F^* = \left.\frac{\partial v^*}{\partial x_3}\right|_{x_3=0} - \left.\frac{\partial v^*}{\partial x_3}\right|_{x_3=-\varepsilon h} + I_1(v^*) + I_2(v^*),$$

with I_1 and I_2 as in Eq. (4.123). Hence,

$$\left|F^*\right|^2_{L^2(\Gamma_\mathrm{i})} \le C(h)\left[\left|\frac{\partial v^*}{\partial x_3}\right|^2_{L^2(\Gamma_\mathrm{i})} + \left|\frac{\partial v^*}{\partial x_3}\right|^2_{L^2(\Gamma_\mathrm{b})}\right]. \tag{4.131}$$

Now, since $v^* = 0$ on Γ_b, we have $\frac{\partial v^*}{\partial x_k} = \varepsilon\frac{\partial h}{\partial x_k}\frac{\partial v^*}{\partial x_3}$ on Γ_b, and by the Poincaré inequality and the boundary condition satisfied by v^* on Γ_i, we have $|\frac{\partial v^*}{\partial x_3}|^2_{L^2(\Gamma_\mathrm{i})} \le 2\alpha_v^2\varepsilon\bar{h}|\frac{\partial v^*}{\partial x_3}|^2_\varepsilon$. Furthermore, we can write

$$\begin{aligned} \left|\frac{\partial v^*}{\partial x_3}\right|^2_{L^2(\Gamma_\mathrm{b})} &\le \left|\frac{\partial v^*}{\partial x_3}\right|^2_{L^2(\Gamma_\mathrm{i})} + 2\left|\frac{\partial v^*}{\partial x_3}\right|_\varepsilon\left|\frac{\partial^2 v^*}{\partial x_3^2}\right|_\varepsilon \\ &\le 2\alpha_v^2\varepsilon\bar{h}\left|\frac{\partial v^*}{\partial x_3}\right|^2_\varepsilon + \theta\varepsilon\left|\frac{\partial^2 v^*}{\partial x_3^2}\right|^2_\varepsilon + \frac{C\varepsilon}{\theta}\left|\frac{\partial v^*}{\partial x_3}\right|^2_\varepsilon, \end{aligned} \tag{4.132}$$

where θ is a positive constant independent of ε, which will be chosen below.

Therefore,

$$\left|F^*\right|^2_{L^2(\Gamma_\mathrm{i})} \le C\left|\frac{\partial v^*}{\partial x_k}\right|^2_\varepsilon + C(h)\theta\varepsilon\left|\frac{\partial^2 v^*}{\partial x_3^2}\right|^2_\varepsilon. \tag{4.133}$$

We estimate the H^1 norm of v^*, using $v^* = v - v_l$, and the H^1 estimates of v and v_l. Therefore, we can easily obtain

$$\left|F^*\right|^2_{L^2(\Gamma_\mathrm{i})} \le C(h)\varepsilon\left[|f_1|^2_\varepsilon + |g_v|^2_i + |\nabla g_v|^2_i\right] + C(h)\theta\varepsilon\left|\frac{\partial^2 v^*}{\partial x_3^2}\right|^2_\varepsilon. \tag{4.134}$$

Now, using the Cattabriga–Solonnikov inequality for the 2D Stokes problem (Eq. (4.130)), there exists a constant C independent of ε such that

$$\big|V^*\big|^2_{H^2(\Gamma_\mathrm{i})} + \varepsilon^2\big|\nabla(hp)\big|^2_{L^2(\Gamma_\mathrm{i})} \le C\big|F^*\big|^2_{L^2(\Gamma_\mathrm{i})}. \tag{4.135}$$

From this, we obtain

$$\begin{aligned}
\varepsilon^2\big|\nabla(hp)\big|^2_{L^2(\Gamma_\mathrm{i})} &\le C(h,\theta)\varepsilon\big[|f_v|^2_\varepsilon + |g_v|^2_i + |\nabla g_v|^2_i\big] + C(h)\theta\varepsilon\left|\frac{\partial^2 v^*}{\partial x_3^2}\right|^2_\varepsilon,\\
\varepsilon^2|\nabla p|^2_{L^2(\Gamma_\mathrm{i})} &\le C(h,\theta)\varepsilon^2|p|^2_{L^2(\Gamma_\mathrm{i})}\\
&\quad + C(h,\theta)\varepsilon\big[|f_v|^2_\varepsilon + |g_v|^2_i + |\nabla g_v|^2_i\big] + C|h|\theta\varepsilon\left|\frac{\partial^2 v^*}{\partial x_3^2}\right|^2_\varepsilon.
\end{aligned} \tag{4.136}$$

From Eq. (4.96) and the weak formulation Eq. (4.97) of Eq. (4.96), we see that

$$\begin{aligned}
|p|_{L^2(\Gamma_\mathrm{i})/\mathbb{R}} &\le C|\nabla p|_{H^{-1}(\Gamma_\mathrm{i})}\\
&\le C\|v\|_{H^1(\mathcal{M}_\varepsilon)} \le C|f_v|_\varepsilon
\end{aligned}$$

so that we actually have the same type of estimate (Eq. (4.136)) for ∇p as for $\nabla(hp)$. Finally, since $\Delta_3 v^* = \nabla p$, in $\mathcal{M}_\varepsilon$, $v^* = 0$ on $\Gamma_\mathrm{b}\cup\Gamma_\ell$ and $\frac{\partial v^*}{\partial x_3} + \alpha^* v = 0$ on Γ_i, we have, thanks to Lemma 4.7,

$$\begin{aligned}
\sum_{k,j=1}^{3}\left|\frac{\partial^2 v^*}{\partial x_k\,\partial x_j}\right|^2_\varepsilon &\le C(h,\alpha_v)\varepsilon|\nabla p|^2_{L^2(\Gamma_\mathrm{i})}\\
&\le C(h,\alpha_v)\big[|f_1|^2_\varepsilon + |g_v|^2_i + |\nabla g_v|^2_i\big] + C(h,\alpha_v)\theta\left|\frac{\partial^2 v^*}{\partial x_3^2}\right|^2_\varepsilon,
\end{aligned}$$

and therefore for θ small enough so that $C(h,\alpha_v)\theta \le \frac{1}{2}$, we conclude that

$$\sum_{k,j=1}^{3}\left|\frac{\partial^2 v^*}{\partial x_k\,\partial x_j}\right|^2_\varepsilon \le C(h,\alpha)\varepsilon|\nabla p|^2_{L^2(\Gamma_\mathrm{i})} \le C(h,\alpha)\big[|f_1|^2_\varepsilon + |g_v|^2_i + |\nabla g_v|^2_i\big].$$

The proof of Theorem 4.6 is now complete. □

By interpolation, it is easy to derive from Theorem 4.6 the following result.

THEOREM 4.7. *Assume that h is a positive function in $\mathcal{C}^3(\overline{\Gamma}_i)$. Let $(v,p)\in\mathbb{H}^1(M_\varepsilon)\times L^2(\Gamma_\mathrm{i})$ be a weak solution of*

$$\begin{cases}
-\big(\Delta v + \frac{\partial^2 v}{\partial x_3^2}\big) + \nabla p = f_v & \text{in } \mathcal{M}_\varepsilon,\\
div\int_{-\varepsilon h}^{0} v\,\mathrm{d}z = 0 & \text{on } \Gamma_\mathrm{i},\\
v = 0 & \text{on } \Gamma_\ell\cup\Gamma_b,\\
\frac{\partial v}{\partial x_3} + \alpha_v v = g_v & \text{on } \Gamma_\mathrm{i}.
\end{cases} \tag{4.137}$$

Then, if $f_v \in \mathbb{L}^2(\mathcal{M}_\varepsilon)$ *and* $g_v \in \mathbb{H}^s(\Gamma_\mathrm{i})$, $0 \le s \le 1$,

$$(v,\, p) \in \mathbb{H}^{s+1}(\mathcal{M}_\varepsilon) \times H^s(\mathcal{M}_\varepsilon). \tag{4.138}$$

Moreover, the following inequality holds:

$$|v|^2_{\mathbb{H}^{1+s}(\mathcal{M}_\varepsilon)} + \varepsilon |p|^2_{H^s(\Gamma_\mathrm{i})} \le C_0\big[|f_v|^2_\varepsilon + \varepsilon^{1-s}\|g_v\|^2_{H^s(\Gamma_\mathrm{i})}\big], \tag{4.139}$$

where C_0 *is a constant depending on the data but not on* ε.

4.5. Regularity of the coupled system

In this section, we prove the H^2-regularity of the solution of a coupled system of equations corresponding to the linear part of the primitive equations of the CAO. We will concentrate on the velocity part; the temperature and salinity parts follow in the same manner.

The unknown is $v = (v^\mathrm{a} v^\mathrm{s})$, with v^a and v^s corresponding to the horizontal velocities in the air and in the ocean.[19] These functions satisfy the following equations and boundary conditions:

$$\begin{cases} -\Delta v^\mathrm{a} - \frac{\partial^2 v^\mathrm{a}}{\partial x_3^2} + \nabla p^\mathrm{a} = f_v^\mathrm{a} & \text{in } \mathcal{M}_\varepsilon^\mathrm{a}, \\ \operatorname{div} \int_0^L v^\mathrm{a}(x_1, x_2, z)\, \mathrm{d}z = 0, & (x_1, x_2) \in \mathbb{R}^2, \\ v^\mathrm{a} = 0 & \text{on } \Gamma_\mathrm{u} \cup \Gamma_\ell^\mathrm{a}, \\ \frac{\partial v^\mathrm{a}}{\partial x_3} + \alpha_v (v^\mathrm{a} - v^\mathrm{s}) = g_v & \text{on } \Gamma_\mathrm{i}, \\ \frac{\partial v^\mathrm{a}}{\partial x_3} + \alpha_v v^\mathrm{a} = g_v & \text{on} \Gamma_\mathrm{e}, \end{cases} \tag{4.140}$$

and

$$\begin{cases} -\Delta v^\mathrm{s} - \frac{\partial^2 v^\mathrm{s}}{\partial x_3^2} + \nabla p^\mathrm{s} = f_v^\mathrm{s} & \text{in } \mathcal{M}_\varepsilon^\mathrm{s}, \\ \operatorname{div} \int_{-\varepsilon h}^0 v^\mathrm{s}\, \mathrm{d}z = 0 & \text{in } \Gamma_\mathrm{i}, \\ v^\mathrm{s} = 0 & \text{on } \Gamma_\ell^\mathrm{s} \cup \Gamma_\mathrm{b}, \\ -\frac{\partial v^\mathrm{s}}{\partial x_3} + \alpha_{v^\mathrm{a}} (v^\mathrm{a} - v^\mathrm{s}) = g_v & \text{on } \Gamma_\mathrm{i}. \end{cases} \tag{4.141}$$

The domain $\mathcal{M}_\varepsilon^\mathrm{s}$ is the domain occupied by the ocean, while $\mathcal{M}_\varepsilon^\mathrm{a}$ is the domain occupied by the atmosphere and $\mathcal{M}_\varepsilon = \mathcal{M}_\varepsilon^\mathrm{a} \cup \mathcal{M}_\varepsilon^\mathrm{s}$:

$$\mathcal{M}_\varepsilon^\mathrm{s} = \big\{(x_1, x_2, x_3) \in \mathbb{R}^3;\, (x_1, x_2) \in \Gamma_\mathrm{i},\, -\varepsilon h(x_1, x_2) < x_3 < 0\big\},$$
$$\mathcal{M}_\varepsilon^\mathrm{a} = \big\{(x_1, x_2, x_3) \in \mathbb{R}^3;\, (x_1, x_2) \in \Gamma,\, 0 < x_3 < \varepsilon L\big\}.$$

Here, Γ, which is a bounded domain in the plane $x_3 = 0$, is the lower boundary of the atmosphere; it consists of the interface Γ_i with the ocean and the interface Γ_e with the

[19] We recall that we use the superscript "s" as sea, instead of "o" as ocean, which can be confused with a zero.

earth, $\Gamma = \Gamma_i \cup \Gamma_e$ ($\Gamma_i \cap \Gamma_e = \emptyset$) (see Section 2.5); furthermore, and as in Section 2.5,

$$\Gamma_b = \{(x', x_3);\ x' \in \Gamma_i,\ x_3 = -\varepsilon h(x')\},$$

$$\Gamma_\ell^a = \{(x', x_3),\ x' \in \partial\Gamma,\ 0 < x_3 < \varepsilon L\},$$

$$\Gamma_\ell^s = \{(x', x_3);\ x' \in \partial\Gamma_i,\ -\varepsilon h(x') < x_3 < 0\},$$

$$\Gamma_u = \{(x', x_3);\ x' \in \Gamma_i,\ x_3 = \varepsilon L\},$$

$$\Gamma_e = \Gamma\backslash\Gamma_i,\ \Gamma \quad \text{and} \quad \Gamma_i \text{ as above.}$$

The coefficient $\alpha_{\mathbf{v}}$ is a positive number, and $g_{\mathbf{v}}$ is a function defined on Γ.

Problem (Eqs. (4.140)–(4.141)) is the stationary linearized form of the PEs of the CAO. Besides its intrinsic interest, the study of this problem is needed for the study of the full nonlinear (stationary or time dependent) CAO system.

4.5.1. Weak formulation of the coupled system

As in Section 4.4.1, we start with the weak formulation of Eqs. (4.140) and (4.141). In this section, we drop the index ε that is irrelevant ($\varepsilon = 1$).

We are given $f_{\mathbf{v}}$ in $\mathbb{L}^2(\mathcal{M})$ and $g_{\mathbf{v}}$ in $\mathbb{H}^{1/2}(\Gamma)$.

For the weak formulation of Eqs. (4.140) and (4.141), we consider the space

$$V = \left\{ v = (v^a, v^s) \in \mathbb{H}^1(\mathcal{M}^a) \times \mathbb{H}^1(\mathcal{M}^s),\ \text{div} \int_{-h}^{0} v^s \, dz = 0, \right.$$
$$\left. v^a = 0 \text{ on } \Gamma_u \cup \Gamma_\ell^a,\ v^s = 0 \text{ on } \Gamma_b \cup \Gamma_\ell^s \right\}.$$

Here, $v^a = v|_{\mathcal{M}^a}$ and $v^s = v|_{\mathcal{M}^s}$; note that the traces of v^a and v^s on Γ_i are not necessarily equal as explained in Remark 2.7(iii). We set, with obvious notations:

$$((v, \tilde{v})) = ((v^a, \tilde{v}^a))_a + ((v^s, \tilde{v}^s))_s$$
$$= \sum_{i=1}^{2} \sum_{j=1}^{3} \int_{\mathcal{M}} \frac{\partial v_i}{\partial x_j} \frac{\partial \tilde{v}_i}{\partial x_j} \, d\mathcal{M};$$

because of the Poincaré inequality, $\|v\| = ((v, v))^{1/2}$ is a Hilbert norm on V.

To obtain the weak formulation, we consider a test function $\tilde{v} = (\tilde{v}^a, \tilde{v}^s) \in V$; we multiply the first equation (Eq. (4.140)) by $\tilde{v}^a$ and the first equation in Eq. (4.141) by $\tilde{v}^s$. We integrate over $\mathcal{M}^a$ and $\mathcal{M}^s$, respectively, and add the resulting equations; we proceed exactly as in Section 4.4.1, using the boundary condition in Eqs. (4.140) and (4.141) and we arrive at the following: to find $v \in V$ such that

$$a(v, \tilde{v}) = \ell(\tilde{v}) \quad \forall \tilde{v} \in V, \tag{4.142}$$

with

$$a(v,\tilde{v}) = ((v,\tilde{v})) + \int_{\Gamma_i} \alpha_{\mathbf{v}}(v^a - v^s)(\tilde{v}^a - \tilde{v}^s)\,d\Gamma_i + \int_{\Gamma_e} \alpha_v v^a \tilde{v}^a\,d\Gamma_e,$$

$$\ell(\tilde{v}) = \int_{\mathcal{M}^a} f_v^a \tilde{v}^a\,d\mathcal{M}^a + \int_{\mathcal{M}^s} f_v^a \tilde{v}^s\,d\mathcal{M}^s + \int_{\Gamma_i} g_v(\tilde{v}^a - \tilde{v}^s)\,d\Gamma_i + \int_{\Gamma_e} g_v \tilde{v}^a\,d\Gamma_e. \tag{4.143}$$

The existence and uniqueness of a solution $v \in V$ of Eq. (4.142) is elementary and follows from the Lax–Milgram theorem. The more delicate question of showing that $v = (v^a, v^s)$ actually satisfies all the Eqs. (4.140) and (4.141) is handled as follows: we find p^a and p^s such that the first equation (Eqs. (4.140) and (4.141)) are valid exactly as we did in Section 4.4.1 for the ocean and the atmosphere. Using also Proposition 4.1 for the ocean and the atmosphere, we obtain the boundary conditions on Γ_i and Γ_e; the other equations and boundary conditions follow from $v \in V$.

REMARK 4.3. Setting $\tilde{v} = v$ in Eq. (4.142), we find

$$\begin{aligned}&\left|\nabla v^a\right|^2_{L^2(\mathcal{M}^a_\varepsilon)^4} + \left|\nabla v^s\right|^2_{L^2(\mathcal{M}^s_\varepsilon)^4} + \int_{\Gamma_i} \alpha_v \left|v^a - v^s\right|^2 d\Gamma_i + \int_{\Gamma_e} \alpha_v \left|v^a\right|^2 d\Gamma_e \\ &\quad = \int_{\mathcal{M}^a} f_v^a v^a\,d\mathcal{M}^a_\varepsilon + \int_{\mathcal{M}^s} f_v^s v^s\,d\mathcal{M}^s_\varepsilon + \int_{\Gamma_i} g_v(v^a - v^s)\,d\Gamma_i + \int_{\Gamma_e} g_v v^a\,d\Gamma_e.\end{aligned} \tag{4.144}$$

4.5.2. H^2-regularity for the coupled system

Having established the complete equivalence of Eqs. (4.140)–(4.141) with Eq. (4.142), we now want to show that the solution of this system possesses the H^2-regularity, namely

$$(v^a, p^a) \in \mathbb{H}^2(\mathcal{M}^a) \times H^1(\mathcal{M}^a) \quad \text{and} \quad (v^s, p^s) \in \mathbb{H}^2(\mathcal{M}^s) \times H^1(\mathcal{M}^s), \tag{4.145}$$

whenever $f_v^a \in \mathbb{L}^2(\mathcal{M}^a)$, $f_v^s \in \mathbb{L}^2(\mathcal{M}^s)$, and $g_v \in \mathbb{H}^{1/2}(\Gamma)$.

More precisely, we will prove the following theorem.

THEOREM 4.8. *Assume that h is a positive function in $\mathcal{C}^3(\overline{\Gamma}_i)$. Let $(v^a, p^a) \in \mathbb{H}^1(\mathcal{M}^a_\varepsilon) \times L^2(\Gamma_i \cup \Gamma_e)$ and $(v^s, p^s) \in \mathbb{H}^1(\mathcal{M}^s_\varepsilon) \times L^2(\Gamma_i)$ be a weak solution of Eqs.* (4.140) *and* (4.141) (*or Eq.* (4.142)). *If $f_v^a \in \mathbb{L}^2(\mathcal{M}^a_\varepsilon)$, $f_v^s \in \mathbb{L}^2(\mathcal{M}^s_\varepsilon)^2$, and $g_v \in \mathbb{H}^1(\Gamma)$, $g_v = 0$ on $\partial\Gamma_e$, then*

$$\left(v^a, p^a\right) \in \mathbb{H}^2\left(\mathcal{M}^a_\varepsilon\right) \times H^1(\Gamma_i \cup \Gamma_e) \quad \text{and} \quad \left(v^s, p^s\right) \in \mathbb{H}^2\left(\mathcal{M}^s_\varepsilon\right) \times H^1(\Gamma_i). \tag{4.146}$$

Moreover, the following inequality holds:

$$\begin{aligned}&\left|v^a\right|^2_{\mathbb{H}^2(\mathcal{M}^a_\varepsilon)} + \left|v^s\right|^2_{\mathbb{H}^2(\mathcal{M}^s_\varepsilon)} + \varepsilon\left|p^a\right|^2_{H^1(\Gamma_i)} + \varepsilon\left|p^s\right|^2_{H^1(\Gamma_i)} \\ &\quad \le C_0\left[\left|f_v^a\right|^2_\varepsilon + \left|f_v^s\right|^2_\varepsilon + \left|\nabla g_v\right|^2_{\mathbb{L}^2(\Gamma)}\right].\end{aligned} \tag{4.147}$$

PROOF. Since $v^{\mathrm{a}} \in \mathbb{H}^1(\mathcal{M}_\varepsilon^{\mathrm{a}})$ and $v^{\mathrm{s}} \in \mathbb{H}^1(\mathcal{M}_\varepsilon^{\mathrm{s}})$, $v^{\mathrm{a}}|_{\Gamma_{\mathrm{i}}}$ and $v^{\mathrm{s}}|_{\Gamma_{\mathrm{i}}}$ belong to $\mathbb{H}^{1/2-\delta}(\Gamma_{\mathrm{i}})$ for all δ, $0 < \delta < 1/2$, and there exists a constant C_0 independent of ε such that

$$\left|v^{\mathrm{a}}\right|^2_{\mathbb{H}^{1/2-\delta}(\Gamma_{\mathrm{i}})} + \left|v^{\mathrm{s}}\right|^2_{\mathbb{H}^{1/2-\delta}(\Gamma_{\mathrm{i}})} \leq C_0\big[\left\|v^{\mathrm{a}}\right\|^2_{\mathbb{H}^1(\mathcal{M}_\varepsilon^{\mathrm{a}})} + \left\|v^{\mathrm{s}}\right\|^2_{\mathbb{H}^1(\mathcal{M}_\varepsilon^{\mathrm{s}})}\big]. \tag{4.148}$$

Furthermore, Eq. (4.144) implies that the right-hand side of Eq. (4.148) can be bounded by an expression identical to the right-hand side of Eq. (4.146).

The boundary conditions on Γ_{i} imply then that

$$\frac{\partial v^{\mathrm{a}}}{\partial x_3} + \alpha_v v^{\mathrm{a}} \quad \text{and} \quad -\frac{\partial v^{\mathrm{s}}}{\partial x_3} + \alpha_v v^{\mathrm{s}}$$

belong to $\mathbb{H}^{1/2-\delta}(\Gamma_{\mathrm{i}})$, and their norm in these spaces are bounded similarly.

Therefore, by Theorem 4.7 applied separately to $\mathcal{M}_\varepsilon^{\mathrm{a}}$ and $\mathcal{M}_\varepsilon^{\mathrm{s}}$, we conclude that

$$(v^{\mathrm{a}}, p^{\mathrm{a}}) \in \mathbb{H}^{3/2-\delta}(M_\varepsilon^{\mathrm{a}}) \times H^{1/2-\delta}(\Gamma),$$
$$(v^{\mathrm{s}}, p^{\mathrm{s}}) \in \mathbb{H}^{3/2-\delta}(M_\varepsilon^{\mathrm{s}}) \times H^{1/2-\delta}(\Gamma_{\mathrm{i}}),$$

and

$$\left|v^{\mathrm{a}}\right|^2_{\mathbb{H}^{3/2-\delta}(\mathcal{M}_\varepsilon^{\mathrm{a}})} + \left|v^{\mathrm{s}}\right|^2_{\mathbb{H}^{3/2-\delta}(\mathcal{M}_\varepsilon^{\mathrm{s}})} + \varepsilon\left|p^{\mathrm{a}}\right|^2_{H^{1/2-\delta}(\Gamma_{\mathrm{i}})} + \varepsilon\left|p^{\mathrm{s}}\right|^2_{H^{1/2-\delta}(\Gamma_{\mathrm{i}})} \leq \tilde{\kappa}, \tag{4.149}$$

where $\tilde{\kappa}$ is the right-hand side of Eq. (4.139) with a possibly different constant C_0.

Using the trace theorem again, we see that

$$\frac{\partial v^{\mathrm{a}}}{\partial x_3} + \alpha_v v^{\mathrm{a}} \quad \text{and} \quad -\frac{\partial v^{\mathrm{s}}}{\partial x_3} + \alpha_v v^{\mathrm{s}} \quad \text{belong to } \mathbb{H}_0^{1-\delta}(\Gamma_{\mathrm{i}})^2 \ \forall \delta,$$

and there exists a constant C_0 independent of ε such that

$$\left\|v^{\mathrm{a}}\right\|^2_{\mathbb{H}^{1-\delta}(\Gamma)} + \left\|v^{\mathrm{s}}\right\|^2_{\mathbb{H}^{1-\delta}(\Gamma_{\mathrm{i}})} \leq C_0\big(\left\|v^{\mathrm{a}}\right\|^2_{\mathbb{H}^{3/2-\delta}(\mathcal{M}_\varepsilon^{\mathrm{a}})} + \left\|v^{\mathrm{s}}\right\|^2_{\mathbb{H}^{3/2-\delta}(\mathcal{M}_\varepsilon^{\mathrm{s}})}\big). \tag{4.150}$$

Therefore, by Theorem 4.7, we conclude that

$$(v^{\mathrm{a}}, p^{\mathrm{a}}) \in \mathbb{H}^{2-\delta}(\mathcal{M}_\varepsilon^{\mathrm{a}}) \times H^{1-\delta}(\Gamma),$$

$$(v^{\mathrm{s}}, p^{\mathrm{s}}) \in \mathbb{H}^{2-\delta}(\mathcal{M}_\varepsilon^{\mathrm{s}}) \times H^{1-\delta}(\Gamma_{\mathrm{i}})$$

and

$$\left\|v^{\mathrm{a}}\right\|^2_{\mathbb{H}^{2-\delta}(\mathcal{M}_\varepsilon^{\mathrm{a}})} + \left\|v^{\mathrm{s}}\right\|^2_{\mathbb{H}^{2-\delta}(\mathcal{M}_\varepsilon^{\mathrm{s}})} + \varepsilon\left|p^{\mathrm{a}}\right|^2_{H^{1-\delta}(\Gamma)} + \varepsilon\left|p^{\mathrm{s}}\right|^2_{H^{1-\delta}(\Gamma_{\mathrm{i}})} \leq \tilde{\kappa},$$

$\tilde{\kappa}$ as above. A final application of the trace theorem and of Theorem 4.7 to $\mathcal{M}_\varepsilon^{\mathrm{a}}$ and $\mathcal{M}_\varepsilon^{\mathrm{s}}$ yields

$$(v^{\mathrm{a}}, p^{\mathrm{a}}) \in \mathbb{H}^{2-\delta}(\mathcal{M}_\varepsilon^{\mathrm{a}}) \times H^{1-\delta}(\Gamma),$$
$$(v^{\mathrm{s}}, p^{\mathrm{s}}) \in \mathbb{H}^{2-\delta}(\mathcal{M}_\varepsilon^{\mathrm{s}}) \times H^{1-\delta}(\Gamma_{\mathrm{i}}),$$

and

$$\left|v^{\mathrm{a}}\right|^2_{\mathbb{H}^2(\mathcal{M}^{\mathrm{a}}_\varepsilon)} + \left|v^{\mathrm{s}}\right|^2_{\mathbb{H}^2(\mathcal{M}^{\mathrm{s}}_\varepsilon)} + \varepsilon\left|p^{\mathrm{a}}\right|^2_{H^1(\Gamma_{\mathrm{i}})} + \varepsilon\left|p^{\mathrm{s}}\right|^2_{H^1(\Gamma_{\mathrm{i}})} \leq \tilde{\kappa},$$

$\tilde{\kappa}$ as above. The proof is complete. □

5. Acknowledgments

This work was partially supported by the National Science Foundation under the grants NSF-DMS-0074334, NSF-DMS-0204863, NSF-DMS-0505974, and NSF-DMS-0604235 and by the Research Fund of Indiana University.

References

Agmon, S., Douglis, A., Nirenberg, L. (1959). Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions. *I. Comm. Pure Appl. Math.* **12**, 209–230.

Agmon, S., Douglis, A., Nirenberg, L. (1964). Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions. *II. Comm. Pure Appl. Math.* **17**, 35–92.

Azérad, P., Guillén, F. (2001). Mathematical justification of the hydrostatic approximation in the primitive equations of geophysical fluid dynamics. *SIAM J. Math. Anal.* **33**, 847–859.

Bardos, C., Tartar, L. (1973). Sur l'unicité rétrograde des équations paraboliques et quelques questions voisines. *Arch. Rational Mech. Anal.* **50**, 10–25.

Bartello, P. (1995). Geostrophic adjustment and inverse cascades in rotating stratified turbulence. *J. Atmos. Sci.* **52** (24), 4410–4428.

Bewley, T., Temam, R., Ziane, M. (2000). A general framework for robust control in fluid mechanics. *Phys. D.* **138**, 360–392.

Bewley, T., Temam, R., Ziane, M. (2000). Existence and uniqueness of optimal control to the Navier–Stokes equations. *C. R. Acad. Sci. Paris Sér. I.* **330**, 1007–1011.

Bresch, D., Kazhikhov, A., Lemoine, J. (2004). On the two-dimensional hydrostatic Navier-Stokes equations. *SIAM J. Math. Anal.* **36** (1), 796–814.

Constantin, P., Foias, C., Temam, R. (1985). Attractors representing turbulent flows. *Mem. Am. Math. Soc.* **53**.

Constantin, P., Foias, C. (1988). *Navier–Stokes Equations* (Univ. Chicago Press, Chicago, IL).

Chepyzhov, V.V., Vishik, M.I. (2002). *Attractors for Equations of Mathematical Physics (American Mathematical Society Colloquium Publications, vol. 49)* (American Mathematical Society, Providence, RI).

Cao, C., Titi, E.S. (2007). Global well-posedness of the three-dimensional viscous primitive equations of large scale ocean and atmosphere dynamics, arXiv: Math. AP/0503028 2 16 Nov 2005 and *Ann. of Math.* **166**, 245–267.

Ewald, B., Temam, R. (2001). Maximum principles for the primitive equations of the atmosphere. *Discrete Contin. Dyn. Syst.* **7** (2), 343–362.

Foias, C., Temam, R. (1989). Gevrey class regularity for the solutions of the Navier–Stokes equations. *J. Funct. Anal.* **87**, 359–369.

Ghidaglia, J.M. (1984). Régularité des solutions de certains problèmes aux limites linéaires liés aux équations d'Euler (in French) [Regularity of the solutions of certain linear boundary value problems associated with the Euler equations]. *Commun. Partial Differ. Equ.* **9** (13), 1265–1298.

Gill, A.F. (1982). *Atmosphere-Ocean Dynamics* (Academic Press, New York).

Grisvard, P. (1985). *Elliptic Problems in Non-Smooth Domains (Monographs Studies in Mathematics, vol. 24)* (Pitman, London).

Guillén-González, F., Masmoudi, N., Rodríguez-Bellido, M.A. (2001). Anisotropic estimates and strong solutions of the primitive equations. *Differ. Integral Equ.* **14** (11), 1381–1408.

Haltiner, G., Williams, R. (1980). *Numerical Weather Prediction and Dynamic Meteorology, second ed.* (Wiley, New York).

Hale, J.K. (2001). Attractors and dynamics in partial differential equations. In: *From Finite to Infinite Dimensional Dynamical Systems (Cambridge, 1995), of NATO Sci. Ser. II Math. Phys. Chem., vol. 19* (Kluwer Acad. Publ., Dordrecht), pp. 85–112.

HANEY, R. (1971). Surface thermal boundary condition for ocean circulation models. *J. Phys. Oceanogr.* **1**, 241–248.

HU, C., TEMAM, R., ZIANE, M. (2002). Regularity results for linear elliptic problems related to the primitive equations. *Chin. Ann. Math. Ser. B.* **23** (2), 277–292.

HU, C., TEMAM, R., ZIANE, M. (2003). The primitive equations of the large scale ocean under the small depth hypothesis. *Discrete Cont. Dyn. Syst.* **9** (1), 97–131.

JU, N. (2007). The global attractor for the solutions to the 3D viscous primitive equations. *Discrete Contin. Dyn. Syst.* **17** (1), 159–179.

KOBELKOV, G.M. (2006). Existence of a solution 'in the large' for the 3D large-scale ocean dynamics equations. *C. R. Math. Acad. Sci. Paris.* **343** (4), 283–286.

KOZLOV, V.A., MAZYA, V.G., ROSSMANN, J. (1997). *Elliptic Boundary Value Problems in Domains with Point Singularities* Math. Surveys Monogr., vol. 52, (Amer. Math. Soc., Providence, RI).

KUKAVICA, I., ZIANE, M. (2007a). On the regularity of the primitive equations of the ocean. *C. R. Math. Acad. Sci. Paris.* **345** (5), 257–260.

KUKAVICA, I., ZIANE, M. (2007b). On the regularity of the primitive equations. *Nonlinearity*. **20**, 2739–2753.

LAPRISE, R. (1992). The Euler equation of motion with hydrostatic pressure as independent coordinate. *Mon. Wea. Rev.* **120** (1), 197–207.

LIONS, J.L. (1969). *Quelques Méthodes De Résolution Des Problèmes Aux Limites Non Linéaires* (Dunod, Paris); re-edited (2002).

LIONS, J.L., MAGENES, E. (1972). *Nonhomogeneous Boundary Value Problems and Applications, vol. 1* (Springer-Verlag, New York).

LIONS, J.L., MALGRANGE, B. (1960). Sur l'unicité rétrograde dans les problèmes mixtes paraboliques. *Math. Scand.* **8**, 277–286.

LIONS, J.L., TEMAM, R.,WANG, S. (1992a). New formulations of the primitive equations of the atmosphere and applications. *Nonlinearity*, **5**, 237–288.

LIONS, J.L., TEMAM, R., WANG, S. (1992b). On the equations of the large-scale ocean. *Nonlinearity*, **5**, 1007–1053.

LIONS, J.L., TEMAM, R., WANG, S. (1993a). Models of the coupled atmosphere and ocean (CAO I). *Comput. Mech. Adv.* **1**, 5–54.

LIONS, J.L., TEMAM, R., WANG, S. (1993b). Numerical analysis of the coupled models of atmosphere and ocean (CAO II). *Comput. Mech. Adv.* **1**, 55–119.

LIONS, J.L., TEMAM, R., WANG, S. (1995). Mathematical study of the coupled models of atmosphere and ocean (CAO III). *J. Math. Pures Appl.* **74**, 105–163.

MA, T., WANG, S. (2005a). Geometric theory of incompressible flows with applications to fluid dynamics. *Mathematical Surveys and Monographs, 119* (American Mathematical Society, Providence, RI), pp. x+234.

MA, T.,WANG, S. (2005b). Bifurcation theory and applications. *World Scientific Series on Nonlinear Science. Series A: Monographs and Treatises, vol. 53* (World Scientific Publishing Co. Pvt. Ltd., Hackensack, NJ), pp. xiv+375.

MAZYA,V.G., ROSSMANN, J. (1994). On the behaviour of solutions to the Dirichlet problem for second order elliptic equations near edges and polyhedral vertices with critical angles. *Z. Anal. Anwendungen*, **13** (1), 19–47.

MOISE, I., SIMONNET, E.,TEMAM, R., ZIANE, M. (1998). Numerical simulation of differential systems displaying rapidly oscillating solutions. *J. Eng. Math.* **34**, 201–214.

MOISE, I., TEMAM, R., ZIANE, M. (1997). Asymptotic analysis of the Navier-Stokes equations in thin domains (volume dedicated to O.A. Ladyzhenskaya). *Topol. Methods Nonlinear Anal.* **10** (2), 249–282.

ORENGA, P. (1995). Un théorème d'existence de solutions d'un problème de shallow water (in French) [A theorem on the existence of solutions of a shallow–water problem]. *Arch. Ration. Mech. Anal.* **130** (2), 183–204.

PEDLOSKY, J. (1987). *Geophysical Fluid Dynamics*, second ed. (Springer-Verlag, NewYork).

PETCU, M. (2006). On the three dimensional primitive equations. *Adv. Differ. Equ.* **11**, 1201–1226.

PETCU, M. (2007). On the backward uniqueness of the primitive equations. *J. Math. Pure Appl.* **87**, 275–289.

PETCU, M. (2004). Gevrey class regularity for the primitive equations in space dimension 2. *Asymp. Anal.* **39**, 1–13.

PETCU, M., TEMAM, R., WIROSOETISNO, D. (2004a). Existence and Regularity for the primitive equations in space dimension 2. *Comm. Pure Appl. Anal.* **3** (1), 115–131.

PETCU, M., TEMAM, R., WIROSOETISNO, D. (2004b). Renormalization group method applied to the primitive equation. *J. Differ. Equ.* **208** (1), 215–257.

PETCU, M., WIROSOETISNO, D. (2005). Sobolev and Gevrey regularity for the primitive equations in a space dimension 3. *Appl. Anal.* **84**, 769–788.

SCHLICHTING, H. (1979). *Boundary Layer Theory*, seventh ed. (McGraw-Hill, New York).

SIMONNET, E., DIJKSTRA, H.A., GHIL, M. (2008). Bifurcation analysis of ocean, atmosphere and climate models. In this volume.

SMAGORINSKY, J. (1963). General circulation experiments with the primitive equations. I. The basic experiment. *Mon. Wea. Rev.* **91**, 98–164.

SMOLARKIEWICZ, P.K., MARGOLIN, L.G., WYSZOGRODZKI, A.A. (2001). A class of nonhydrostatic global models. *J. Atmos. Sci.* **58** (4), 349–364.

TEMAM, R. (1997). *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*, second ed. (Springer-Verlag, New York).

TEMAM, R. (1977). *Navier–Stokes Equations, Studies in Mathematics Application* (North-Holland, Amsterdam). [reedited, AMS-Chelsea Ser., Amer. Math. Soc. (Providence, RI) (2001)].

TEMAM, R., WIROSOETISNO, D. (2007). Exponential Approximations for the Primitive Equations of the ocean, in *Discrete and Continuous Dynamical Sytems-Series B*, **7** (2), 425–440.

TEMAM, R., WIROSOETISNO, D. (2008). Stability of the slow manifold in the primitive equations arXiv:0808.2878v1.

TEMAM, R., ZIANE, M. (1996). Navier-Stokes equations in three-dimensional thin domains with various boundary conditions. *Adv. Differ. Equ.* **1** (4), 499–546.

TEMAM, R., ZIANE, M. (1997). Navier–Stokes equations in thin spherical domains. *Contemp. Math.* **209**, 281–314.

TEMAM, R., ZIANE, M. (2004). Some mathematical problems in geophysical fluid dynamics. In: Friedlander, S., Serre, D. (eds.), *Handbook of Mathematical Fluid Dynamics*, **3**, (Elsevier).

TRENBERTH, K.E. (1992). *Climate System Modeling* (Cambridge Univ. Press, Cambridge).

VON NEUMANN, J. (1963). Collected works. In: Taub, A.H. (ed.), *Vol. VI: Theory of Games, Astrophysics, Hydrodynamics and Meteorology* (A Pergamon Press Book The Macmillan Co., New York), pp. 1963 x+538.

WANG, S. (1992a). On the 2D model of large-scale atmospheric motion: well-posedness and attractors. *Nonlinear Anal.* **18** (1), 17–60.

WANG, S. (1992b). Attractors for the 3D baroclinic quasi-geostrophic equations of largescale atmosphere. *J. Math. Anal. Appl.* **165** (1), 266–283.

WASHINGTON, W.M., PARKINSON, C.L. (1986). *An Introduction to Three Dimensional Climate Modeling* (Oxford Univ. Press, Oxford).

ZENG, Q.C. (1979). *Mathematical and Physical Foundations of NumericalWeather Prediction* (Science Press, Beijing, China).

ZIANE, M. (1995). Regularity results for Stokes type systems. *Appl. Anal.* **58**, 263–292.

ZIANE, M. (1997). Regularity results for stationary primitive equations of atmosphere and the ocean. *Nonlinear Anal.* **28**, 289–313.

ZIANE, M. (2000). unpublished manuscript.

Index

AB2–AM3 (Adams-Bashforth with Adams-Moulton) algorithm, 139–143
AB3–AM4 algorithm, 143–146, 162, 163
absolute angular momentum conservation law, 77
accuracy, 3, 5, 11, 12, 13, 14, 15, 18, 22, 23, 26, 27, 34, 36, 47, 48, 59, 60, 62, 101, 107, 113, 114, 118
accuracy, transport scheme, 13–15
adaptive mesh refinement (AMR), 341, 370–372
advection, 14, 51, 127
advection equation, 128, 147, 453
advection test case, 14, 15, 17, 22, 24, 28, 29, 34
Alfvenization, 360, 364
Alfvén's theorem, 356, 364
amplification factor, 38–41, 44
analytic wind field, 34
angular momentum, 4, 6, 62, 76, 77, 78, 87, 119
Antarctic circumpolar vortex, 322
anticyclonic gyres, 209, 210
anticyclonic rotating mixing layer, 326
anticyclonic vortices, 321, 330
approximate functions, 603–604
ARPEGE model, 389
arrival cell, 4
asymptotic instability, 130
atmosphere, bifurcation analysis of, 187–225
atmosphere, primitive equations of, 63–66, 606
atmospheric fluid flow, 211
atmospheric scales, large, 321
augmented pressure, 587

back-and-forth nudging (BFN) algorithm, 412–413
backward trajectories, 26–27
backward uniqueness, of PEs, 695–706
baroclinic instability, 238, 330, 336
baroclinic jet, 331–336
baroclinic PGF, 126, 127
baroclinic PGF, compressibility and, 169–171
baroclinic PGF, in Jacobian scheme, 165
baroclinic shear flows, steady parallel, 238
baroclinic time-stepping, 126, 149, 157
baroclinic wave-mean oscillation, 239, 241
baroclinic wave test, 102, 107, 113, 116
barotropic aliasing, 157
barotropic averaging, 126
barotropic mode, 126, 153, 483, 506, 507
barotropic mode-splitting, and compressibility, 171–173
barotropic PGF, 125
barotropic QG model, 213
barotropic streamfunction, 222, 223
barotropic time filters, 157
barotropic time-stepping, 126, 157
beta-plane approximation, 591, 609
BFN algorithm, 412–413
bifurcation analysis, 187–225
bifurcation diagrams, 206, 207, 222
bifurcation diagrams, of double-gyre circulation, 211–212
bifurcations, 189, 191, 192, 197, 205
bifurcations, flip or period-doubling, 196–198, 214
bifurcations, gluing, 201
bifurcations, homoclinic, 190, 198, 199, 214, 216
bifurcations, in primitive-equation models, 220–222
bifurcations, Neimark–Sacker or torus, 197
bifurcations, symmetry-breaking, 214
bifurcations, transcritical, 193, 195, 196, 205, 206
bifurcation theory, 190–201, 215, 222
block Gauss–Seidel preconditioner, 209
boundary conditions, 10, 100, 113, 524, 525, 590–591, 608–609, 613–614, 643
boundary conditions, modal equations and, 491–493
boundary-layer, turbulent, 318

boundary value problem, 207, 714–731
Boussinesq approximation, 169, 330, 585–587
Boussinesq approximation, accuracy of, 176–178
Boussinesq approximation, EOS and, 174–176
Boussinesq equations, 483
Boussinesq model, 153
Box–Müller technique, 298
branch switching, 205–207
Brownian motion, 281
Brunt-Väisälä (buoyancy) frequency, 331, 484, 547
B-spline basis function, 255
Burgers number, 665

Camassa-Holm (CH) equation, 249, 250, 253
CAO, 581, 613–616
cascade DCISL scheme, 30–34
cascade interpolation method, 30, 31
catastrophic loss, 194
cell-integrated methods. *See* finite-volume (FV) methods
cell-integrated semi-Lagrangian (CISL) schemes, 12
central limit theorem, 280–285
centroidal Voronoi tessellations (CVT), 419
CFC gases, 322
chaotic attractor, 198, 235, 242, 244
chaotic invariant set, periodic trajectories of, 235
chaotic Lorenz system, stochastic version of, 302, 303
chemical transport model, 109
CH equation, 249, 250, 253
chloro-fluoro-carbon (CFC) gases, 322
CH solitons, 249, 250, 268, 269, 271, 272
circulation integral, 276
circulation loop figures, 273, 274
CISL scheme, 12, 27, 36, 107
classical baroclinic instability theory, 241
classical fluid dynamics, time-periodic flows in, 231–244
Clebsch variational approach, 258
climate models, bifurcation analysis of, 187–225
climate system, 187, 188
climatic behavior, 190
collision, overtaking, 251, 271
collocation techniques, 207
commutative noise, 295
compatibility property, 18, 19
complete set of conservation laws (CSCL) model, 62–79
compressible flows, weakly, 315–319
computational costs of FV, 16–17, 103–104, 106, 113
computational fluid dynamics (CFD), 122
conservation laws, 4, 62–79, 83, 84, 102, 270. *See also* complete set of conservation laws (CSCL) model
conservation of angular momentum, 76–77
conservation of entropy, 76
conservation of linear correlations, 55
conservation of mass, 4–5
conservation of mass, of moist air, 5, 66–72
conservation of mass, of passive tracers, 73
conservation of mass, of water vapor, 73–74
conservation of total energy, 74–75, 450–454
conservation property, 18
conservative semi-Lagrangian schemes based on rational functions (CSLR) schemes, 59
conservative spatial discretization schemes, 4
consistency property, 18
constancy preservation, 13, 19, 51–55, 126, 148–152
constant density circulation, bifurcation diagram of, 222
continuation methods, 201–209
continuity equation (moist air), 4–6, 8, 10, 11, 12, 13, 18, 47, 55, 63, 64, 67, 73, 75–80, 84, 87–91, 93, 95, 97, 99, 106, 108, 109, 111
continuity equation, discretized, 257
continuity equation, for tracers, 11, 18, 73
continuity equation, for water vapor, 64, 75
continuous primitive equations, 63–65
control parameter, bifurcation diagram of, 222
convergence tests, 268, 269
Coriolis accelerations, 320
Coriolis force, 123, 129, 144, 145, 211, 321, 323, 584
corrector algorithm, 203
cost function, 385, 395, 398, 399, 401, 416
cotangent lift, 264
coupled atmosphere–ocean (CAO), 581, 613–616
coupled system, 743–747
Courant-Fredrichs-Levy (CFL) time-step restriction, 11, 12
Courant number, 46, 61, 123, 129
covariance evolution equation, 460
covariance localization, 428
Craik-Leibovich equations, 344, 346, 347, 355
curvelet-based multiscale methods, 433
cycle expansions, 234–238
cyclones, tropical, 321
cyclonic gyres, 209
cyclonic perturbation, 333
cyclonic vortices, 321

damped-advective modes, 241
data assimilation, 387, 420–428
DCISL scheme, 12, 13, 27, 29, 30, 34, 38, 39, 40, 47, 49, 57, 69
DCISL scheme, extension to spherical geometry, 41–46
DCISL scheme, fully 2D, 27–30
DCISL scheme, hybrid trajectories, 67–73
DCISL scheme *versus* flux-based schemes, 49, 57, 61–62, 113
deformational test cases, 14
degree of local mass conservation, 5, 34–37, 101, 113
degrees of freedom, 208, 209, 241, 250
density Jacobian schemes, 127, 156, 165, 177
density-weighted Favre-filtering, 317
departure cell, 4–13, 19, 20, 26–30, 34–36, 40, 46, 47, 57, 59, 61, 67–71, 79, 101, 112
departure cell-integrated semi-implicit semi-Lagrangian dynamical core, 79
departure cell-integrated semi-Lagrangian scheme. *See* DCISL scheme
deterministic initial value problem, 447–450
diagnostic variables, 588, 591–592, 609–610, 680
diffeomorphisms, 250, 252–254, 258, 259
diffusion processes, 103, 106
direct numerical simulations. *See* DNS
Dirichlet boundary condition, 711, 713, 714
Dirichlet problem, regularity of solutions to, 707, 708
Dirichlet–Robin mixed boundary value problem, 714–721
discrete EP equation, for VPM, 261–263
discrete method, constructing fully, 260, 261
discrete relabeling, 251, 266, 276
DNS, 324–327, 331, 341, 342, 355
DNS, limits of turbulence, 310
double-gyre circulation, bifurcation diagram of, 211–212
double-gyre quasi-geostrophic model, 210–218
double-gyre system, 209
downstream cell-integrated scheme, 29
dynamical system, 196, 198, 202, 224
dynamical systems theory, 189, 191, 192, 208, 196, 222

Eady model, 330, 331
earth atmosphere, 321
Earth-System Modeling Framework, 123
ECMWF. *See* European Center for Medium Range Weather Forecasting
eddy coefficients, 312, 315
eddy-damped quasi-normal Markovian (EDQNM) theory, 312, 313, 359
eddy diffusivity, 312, 313
eddy parameterization, 124
eddy-viscosity model, spectral, 312, 315
EDQNM theory, 312, 313, 359
effective advection speed, 48, 49
eigenmodes, spectral behavior of, 213
eigenvalue problem, 191, 205, 241, 415
eigenvalues, 192, 193, 196, 197, 199, 205, 213, 214, 236
eigenvectors, 197, 205
eigenvectors, Floquet, 240, 241
EKF. *See* extended Kalman filter
Ekman damping coefficient, 239, 241
Ekman dissipation, 239
Ekman-friction damping rate, 241
elliptic boundary value problems, in cylinder-type domains, 707–714
elliptic linear problems, in geophysical fluid dynamics, 706
El-Niño ocean–atmosphere system, 224
El Niño-Southern Oscillation (ENSO), 239, 241
empirical orthogonal function (EOF), 241, 413, 426
EnKF. *See* ensemble Kalman filtering methods
ensemble assimilation methods, 428
ensemble filters, deterministic update, 426, 427
ensemble generation theory, 238
ensemble Kalman filtering (EnKF) methods, 388, 420–428
ensemble square-root filter (EnSRF), 427
entropy conservation law, 76
EOS. *See* equation of state
EPDiff, 249, 250, 252, 253
EPDiff, constrained action principle for semidiscrete, 258–260
EPDiff, discretising, 254
EPDiff, Kelvin's circulation theorem for discrete, 266–268
EPDiff, Lagrangian for semidiscrete, 257, 258
EPDiff, variational principle for discrete, 258–261
equation of state (EOS), 127
equation of state (EOS), and Boussinesq approximation, 174–176
equation of state (EOS), compressibility of seawater in, 168
equivalent-barotropic model. *See* 1.5-layer model
Euler equations, 252, 579
Eulerian divergence, 91–92
Eulerian grid cells, 4, 6, 11, 12, 48, 58, 97, 111, 113, 258, 261

Eulerian latitudes, 31, 32
Eulerian longitudes, 30–32
Eulerian schemes, compatibility property for, 18, 19
Eulerian velocity, 258
Euler-Poincaré (EP) equations, 250
Euler scheme, 293, 299
European Center for Medium Range Weather Forecasting (ECMWF), 63, 388, 390–393
Ewald–Témam scheme, 292, 293, 297, 298
exact cell-integrated continuity equation, 5, 6–11
exact departure cell, 35
explicit energy conversion term, 93
explicit HIRLAM-DCISL dynamical core, 80–81
explicit integration procedure, 78
explicit semi-Lagrangian momentum equation, 88
explicit semi-Lagrangian prediction equations, 81
explicit time-stepping scheme, 12, 79, 84
extended Kalman filter (EKF), 415, 423, 424

fast Fourier transform (FFT), 350
fast-time averaging filter, design of, 156–158
fast-time averaging filter, step-multipliers for, 158, 159
FB (forward-backward) schemes, 128, 139, 140, 142, 162
Ferrel cells, 321
finite amplitude instability, 243
finite difference system (FDS) model, 418
finite-dimensional attractors, 189
finite dimensional subspace, 254
finite-volume (FV) methods, 3–5, 107. *See also* DCISL scheme; flux-based schemes
finite-volume (FV) methods, computational cost of, 16–17, 25, 103–104
finite-volume (FV) methods, mass conservation law for, 4, 6–11, 62
first-order symplectic Euler-A method, 261
flip doubling bifurcation, 196, 197
floating Lagrangian control volume vertical coordinate, 67
floating-point operations per second (FLOPS), 16
Floquet eigenvectors, 240, 241
Floquet multipliers, 196, 235
Floquet theory, 240
fluid exchange, 234
fluid motions, 231, 232
fluid trajectories, 233, 234
fluid turbulence, limits of direct-numerical simulation in, 310
fluid velocity, 232
flux-based/flux-form schemes, 5, 11, 12, 26, 47, 60, 61–63
flux-based schemes, based on mixing ratio, 54
flux-based schemes, based on operator splitting, 48–56
flux-based schemes *versus* DCISL schemes, 49, 57, 61–62, 113
flux corrected transport (FCT) filters, 60
flux-form Eulerian scheme, 12
flux limiters, 55, 56, 60, 61
Fokker–Planck equation, 283, 301
four-dimensional variatonal data assimilation. *See* 4D-VAR
4D-VAR, 388, 413. *See also* POD 4D-VAR
Fourier modes, 312
free-shear layers, rotating, 326, 327
free-shear layers, universality of, 327–330
frictional dissipation, 241
Froude number, 331, 332
full semi-implicit CSCL system, 86
fully 2D DCISL scheme, 27–30
fundamental matrix, 446, 449, 463, 474
FV dynamical core, 6, 62. *See also* HIRLAM-DCISL dynamical core; NCAR-FFSL dynamical core
FV Lagrangian pressure gradient force, 101
FV methods. *See* finite-volume (FV) methods

Galerkin approximation, 695
Galilean transformation, 331
Gaussian Rossby distribution, 327
Gauss's divergence theorem, 8
GCM. *See* general circulation model
general circulation model (GCM), 4, 63, 108, 111, 112, 188, 386, 387, 578
generalized Lagrangian mean (GLM) method, 343
geodesic motion, 252, 253
geophysical dynamics, time-periodic flows in, 231–244
geophysical flows, 190
geophysical flows, data for, 390–394
geophysical flows, models for, 388–390
geophysical flows, OBL role in, 343
geophysical flows, specificity of, 386–388
geophysical fluid dynamics. *See* GFD
geophysical processes, 386
geophysical turbulence, 340–343
geostationary satellites, 392, 393
Gevrey regularity, 678, 691
GFD, 319–336, 482, 578–579, 706
GFD, applications, 330
GFD, LES for, 309–336
GFD–Stokes problem, 731–743
Gibb's phenomena, 18
global mass conservation, 5, 10, 11, 29

gluing bifurcation, 201
Green-Nagdhi equation, 253
Green's function, 252, 253
Green's theorem, 164
grid momentum, 258, 259, 261
grid-point semi-Lagrangian continuity equation, 10
grid-point semi-Lagrangian schemes, 11, 38
grid resolution, 270, 271
grid velocity, 255, 259, 274, 276
Gronwall's lemma, 622, 623, 633, 684
Gulf stream region, 209, 210, 323
gyre mode, 214, 215, 217, 224

Hadley cells, 321
Hamiltonian particle-mesh (HPM) method, 253
Hamiltonian structure, 260
Hamilton's principle, 252
Hamming window, 158, 159
Helmholtz equation, 82, 86
Helmholtz operator, 253, 254, 269, 365
Hessian, second-order methods, 402–407
Hessian-free truncated-Newton (HFTN) minimization algorithm, 419
Heun scheme, 293
high-resolution limited area model. *See* HIRLAM
Hilbert scalar products, 656
Hilbert space, 691
Hille–Phillips–Yoshida theorem, 554, 557
Hille–Yosida theorem, 496, 503, 560, 566, 570
HIRLAM, 34
HIRLAM-DCISL dynamical core, conservation properties, 100
HIRLAM-DCISL dynamical core, explicit, 80
HIRLAM-DCISL dynamical core, idealized test of, 105–107
HIRLAM-DCISL dynamical core, resolution, 106
HIRLAM-DCISL dynamical core, semi-implicit version of, 87–95
homoclinic bifurcations, 198, 199, 214, 216
homoclinic explosion, 200
homoclinic orbits, 198, 200, 201, 214, 224
homogeneous boundary condition, 567
homotopy methods, 205–207
homotopy parameter, perturbation, 206
Hopf bifurcation, 194, 200, 207, 211, 212, 214, 217
Hopf bifurcation, supercritical, 195
HPM method, 253
H^2 regularity, 708, 736–743, 745–747
hybrid trajectory scheme, 69
hydrodynamic turbulence, 357, 361
hydrostatic approximation, 587–589, 606–607
hydrostatic equation, 64, 78, 483, 484
hydrostatic inconsistency, 163
hyperbolic fixed point, 192. *See also* saddle-node fixed point
hyperviscosity, 315, 332

icosahedral-hexagonal grid, 16
ideal fluid, 252
ideal gas equation, 64
idealized baroclinic wave test, 102
images, assimilation of, 428–433
Imperial College Ocean Model (ICOM), 419
implicit free-surface model, 159
incompressible Navier-Stokes equations, 577, 579
infinite-dimensional space, 254
infinitesimal transformation, 265
integrated kernel design, 121–128
intermediate grid, 30, 46
intermittency, LAMHD equations, 366–367
internal energy, 74
interpolation methods, 387
"invariance principle," 280
invariant Cantor set, 200
inviscid PEs, 483–486
iso-amplification rates, 326
isotropic turbulence, decaying, 313, 314
Ito calculus, 283, 285, 293–295
Ito integrals, 283
Ito SDE, 284, 285
Ito systems, 283, 285, 292

Jacobian matrix, 203, 204, 208, 444, 445, 459, 460
Jacobian matrix, genuine nonlinearity and essential linearity, 468–472
Jacobian matrix, symmetric part of, 465–468
Jacobian scheme, 156, 165
Jacobi–Davidson QZ method, 205
JW06a tests, 102, 104, 105

Kalman filter (KF), computational cost of, 423
Kalman filter (KF), reduced rank, 426
Kalman filter (KF) methods, 388, 421, 422
Karhunen-Loeve expansion, 414
Kármán-Howarth theorem, 366
KdV equation, 249
Kelvin-Helmholtz instability, 326
Kelvin-Helmholtz vortices, 326
Kelvin-Noether theorem, 254
Kelvin's circulation theorem, 356, 364
Kelvin's circulation theorem, for discrete EPDiff, 266–268
kernel code, in ROMS, 124
KF. *See* Kalman filter
kinematic model, 233

kinetic energy, 75
Kolmogorov equation, 282, 283
Korteweg-deVries (KdV) equation, 249
Krylov subspace, 205
Kuramoto-Sivashinsky equation, 237
Kuroshio extension, 209

Lagrange function, 410, 411
Lagrangian-average MHD (LAMHD) equations, 357, 363, 365, 366
Lagrangian cell. *See* departure cell
Lagrangian divergence, 89, 90, 91, 92
Lagrangian fluid particles. *See* Lagrangian particles
Lagrangian grid, 26, 30, 31, 38, 46, 58
Lagrangian latitude, 26, 31, 32, 46
Lagrangian longitude, 26, 31, 33
Lagrangian model layer, 7, 10
Lagrangian motion, 233–234
Lagrangian particles, 255, 268
Lagrangian particle trajectory, 362
Lagrangian pole point, 45, 46
Lagrangian schemes, 11, 12, 13, 18, 22, 26, 38, 56, 58, 113, 114
Lagrangian's principle, 253
Lagrangian vertical coordinate, 6, 97
LAM, 481, 482
laminar flow, steady, 242–244
Laplace–Beltrami operators, 587
large eddy simulations. *See* LES
Lax–Milgram theorem, 619, 732, 745
1.5-layer model, 219
LES, 310–319, 324, 326, 327, 331, 341, 342, 346–350, 355
LES, for MHD, 357
LES, of low Mach number flows, 317
LES formalism, incompressible, 311, 312
LF (leapfrog) time-stepping, 122
LF–AM3 scheme, 160–162
LF–TR/LF–AM3 with FB feedback, 138, 139
Lie-algebra commutator, 252
limit cycles, local codimension-1 bifurcations of, 195–198
limited area models (LAM), 481, 482
Lin constraints, 258
linear disturbance growth, 238–242
linear eigenmodes, 239
linear stability analysis, 191, 202, 206, 223, 243, 331
linear stability problem, 204, 205
linear system, well-posedness of, 525–541
linear time-periodic propagating-wave flows, 232
linear 2D PEs, 490, 491
Lions' theorem, 387
Lipschitz convergence criterion, 15
Lipschitz domain, 631
local codimension-1 bifurcations, of limit cycles, 195–198
local codimension-1 bifurcations, of steady states, 193–195
locally mass conserving semi-Lagrangian scheme, 58
local mass conservation, 5, 11, 13, 17, 24, 26, 34–37, 58, 59, 91, 111
loop integrals, 274, 276
Lorentz force, 364
Lorenz bifurcation, 200
Lorenz system, 200, 208, 302, 303
low-frequency dynamics, quantization of, 215
low-frequency variability, 224, 225
low-frequency variability, of double-gyre wind-driven circulation, 209
Lyapunov characteristic vectors, 242
Lyapunov–Schmidt reduction, 192

Mach number flows, 317
macro-pressure, 311
macro-temperature, 317
magnetic Prandtl number, 357
magnetic Reynolds number, 357
magnetohydrodynamic. *See* MHD
mass-fixer, 11
mass-points, 32
mass-wind inconsistency, 6, 18, 55, 107–111
Maxwell's equations, 356
mean field theory (MFT), 358
mean-flow advection, 241
meridional Courant number, 46
message passing interface (MPI), 350
meteorological modeling, FV methods in. *See* finite-volume (FV) methods
meteorology, mesoscale and small-scale, 322
MFT, 358
MHD, fluids to, 355–357
MHD, LES for, 357
MHD equations, 324
MHD framework, 341
MHD turbulence, 357, 359
MHD turbulence, central problem in DNS of, 361
MHD turbulence, paradigmatic problems, 366
midlatitude atmosphere–ocean system, 225
midlatitude oceans, 209
Milsteyn scheme, 289, 290
Milsteyn scheme, explicit version, 294–295
Milsteyn scheme, implicit version, 296
mixing layer, anticyclonic rotating, 327

mixing layer, 2D, 326
mixing ratios, 19, 55, 56
Möbius band, 197
modal equations, two-dimensional $x - z$ case, 491–493
modal equations and boundary conditions, 491–493
model equations with wave effects, 347–349
Moderate Resolution Imaging Spectroradiometer (MODIS), 428
Modular Ocean Model (MOM), code organizational structure in, 123
moist air, conservation of mass of, 5–11, 66–72, 78–80, 85, 89–90, 93–94
moist air, continuity equation of, 5–11, 66–72, 78–80, 85, 89–90, 93–94
moist air, explicit continuity equation for, 66–72, 78–80, 93–94
moist air, semi-implicit continuity equation for, 78–80, 85, 89–90
moisture equation, 64
momentum filaments, 251
momentum maps, 264, 272
momentum maps, for lattice EPDiff, 247–276
momentum maps, left-action, 259, 264
momentum maps, right-action, 265, 266, 268, 270, 272
monodromy matrix, 196
monophasic Newtonian flow, 310
monotonic filter, 22–25
monotonicity-preserving advection schemes, 147
multichannel singular spectrum analysis (M-SSA), 217

Navier-Stokes-alpha model, Lagrangian-averaged, 253
Navier–Stokes equations, 311, 320, 336, 341, 345, 358, 390, 485, 553, 577, 579, 632, 734
Navier–Stokes equations, 2D, 210
NCAR-FFSL dynamical core, conservation properties, 101
NCAR-FFSL dynamical core, 3D transport scheme, 95–98
NCAR-FFSL dynamical core, governing equations, 97, 98
NCAR-FFSL dynamical core, idealized test of, 103–105
NCAR-FFSL dynamical core, resolution, 106
NCAR-FFSL dynamical core, time-stepping, 98–100
Neimark–Sacker bifurcation, 197, 198
Neumann boundary condition, 712
Neumann problem, 707, 708
Neumann–Robin boundary value problem, 721–731
Newton iteration, 240
Newton-Picard method, 240
Newton process, 206
Newton–Raphson iterations, 203, 208, 209
NLEV layers, 68
non-Boussinesq models, 173, 177
noncatastrophic loss, 194
nonhomogeneous boundary conditions, 537–540, 570–572
nonhydrostatic models, 111, 112
nonlinear Boussinesq approximation, weakly, 249
nonlinear coupling term, 445, 446
nonlinear 3D PEs, 547
nonlinear internal waves, interactions of, 247–249
nonlinear 2D PEs, 490, 491
nonstochastic filters, 426
normal form theory, 192
normal mode analysis, 238, 242
normal mode expansion, 486–489, 547–550, 571
normal mode expansion, full 3D linear case, 548–550
normal mode expansion, space dimension 2.5, 522–524
normal modes, 240, 242
nudging data assimilation (NDA) method, 408
nudging method, 388, 408–413
numerical weather prediction (NWP) models, 425

OBL. *See* oceanic boundary layer
ocean, bifurcation analysis of, 187–225
ocean, Boussinesq approximation, 585–587
ocean, hydrostatic approximation, 587–589
ocean, PEs of, 584, 593
ocean circulation, wind-driven, 188–190
oceanic basins, 211, 215
oceanic boundary layer (OBL), 340, 341
oceanic boundary layer (OBL), in geophysical flows, 343
oceanic boundary layer (OBL), waves, currents, and turbulence in, 343–355
oceanic circulation, 322, 323
Oceanic General Circulation Models (OGCMs), 121–123, 148, 176, 177, 578
ocean models, zero-dimensional, 188
oceanography, mesoscale and small-scale, 323
ODEs, 188, 190, 191, 285
offline coupling, 109
OGCMs, 121–123, 148, 176, 177, 578
1D DCISL scheme, 38, 47–48
1D EPDiff equation, numerical solution to, 269

1D semi-implicit CSCL shallow water models, 81–87
1D subgrid-cell reconstructions, 20–24
1D energy balance models, 208
online coupling, 108, 114
operator-split schemes, 49–56, 61
optimality system, 396, 397
optimal control, 395
optimal nudging coefficients, parameter estimation of, 410–412
ordinary differential equations (ODEs), 188, 190, 191, 285
Ornstein–Uhlenbeck (OU) process, 299, 302
ozone hole, 322

parabolic spline method (PSM), 22, 120
parametrization, of subgrid effects, 387
partial differential equation (PDE) models, 188, 189, 201, 202, 208, 223
particle-mesh calculus, 254–258
passive tracers, conservation of mass, 72, 73
passive tracers, explicit continuity equations for, 73, 91
peakons. *See* CH solitons
Pedley's analysis, 326
period-doubling bifurcation, 196–198, 214
periodic boundary conditions, 511, 512–514
periodic orbit analysis, 237
periodic orbits, 197–201, 207–209, 217, 235, 237
periodic orbit theory, 235
periodic wave-mean oscillation, 240
Perron-Frobenius operator, 236
perturbation homotopy parameter, 206
perturbations, 191, 194–196
PEs, 63, 66, 75, 76, 79, 87, 102, 117, 119, 188, 482–486, 491, 492, 520–525, 547, 577, 579, 580, 647–664, 678, 695
PEs, bifurcations in, 220, 224
PEs, existence and uniqueness, 641, 668
PEs, full 3D linear case, 546–572
PEs, inviscid, 483–486
PEs, linear 2D, 490, 491
PEs, mathematical background, 579–581
PEs, of large-scale ocean, 588
PEs, physical background of, 578–579
PEs, strong solutions, 616, 641
PEs, weak formulation, 592, 610, 614–616, 643–645
phase space trajectories, 195, 199
physical space eddy coefficients, 315
piecewise constant function, 20
piecewise constant subgrid-cell reconstructions, 41
piecewise cubic method, 22, 33
piecewise linear representation, 20
piecewise parabolic method, 21, 22, 39
pitchfork bifurcation, 193, 194, 196, 197, 200, 205, 206, 211, 212, 216, 217, 224
pitchfork bifurcation, perturbed, 195
pitchfork bifurcation, supercritical and subcritical, 218
plane channel, nonrotating, 314–316
plateau-peak behavior, 313
Platen method, explicit version, 296
POD 4D-VAR, adaptive, 415–417
POD model, 388, 419
POD model, discrete case, 414, 415
POD model, goal-oriented model-based reduction, 417, 418
POD model, reduction methods application to geosciences, 413–420
Poincaré constant, 657
Poincaré map, 196, 197, 200, 207–208
Poincaré map, boundary value problem, 207
Poincaré map, fixed points of, 207–208
Poiseuille flow, 314
polar caps, 44, 46
polar-orbiting satellites, 392, 393
posterior energy fixer, 103
posterior mass fixer, 103, 109
potential energy, 74, 100, 101
PoU (partition-of-unity) property, 254, 255
PoU vector field, 255, 262, 264, 265
Prandtl number, 313
Prandtl number, low magnetic, 367–370, 373
predictor-corrector algorithm, 79, 90, 95, 133, 138
pressure fixer, 110
pressure gradient force, 82, 87, 88, 100, 101
pressure-gradient force (PGF) error, 152–156, 163–168. *See also* baroclinic PGF
primitive equations. *See* PEs
principal component analysis (PCA), 413
priori estimates, 627–640, 696
prognostic equation, 4, 10, 19, 47, 68, 73, 75, 77, 96, 111
prognostic variables, 4, 19, 56, 58, 86, 106, 107, 111, 588, 609–610, 680
proper orthogonal decomposition. *See* POD
pseudo-arclength methods, 202–204, 208
pseudogeometrical domain, 608, 613
pseudo-spectral methods, 313, 314, 365

QG model. *See* quasi-geostrophic model
quadratic invariants, MHD equations, 356
quasi-bicubic interpolation, 46
quasi-biparabolic subgrid-cell representation, 25

quasi-DNS, 342
quasi-geostrophic (QG) equations, 210, 223
quasi-geostrophic (QG) model, 387, 388, 400–402
quasi-geostrophic (QG) model, double-gyre, 210–218
quasi-geostrophic theory, Rossby waves of, 232
quasi-horizontal Lagrangian trajectories, 67
quasi-horizontal momentum equation, 63
quasi-hydrostatic atmosphere, CSCL for, 63–79
quasi-hydrostatic FV methods, 107

radiosonde measurements, 390, 391
random number generators, 298
Rayleigh-Bénard convection, 200, 242
Regional Oceanic Modeling System. *See* ROMS
regional problems, at midlatitudes, 609
relaxation oscillations, 215, 217
Reynolds-averaged-Navier-Stokes (RANS), 322, 342, 348
Reynolds numbers, 242, 243, 317, 318, 320, 326, 327, 355, 665
Riemann sum, 261, 267, 283
RK2 algorithm with FB feedback, 130–138
ROMS, code architecture, 127
ROMS, free-surface model, 125
ROMS, kernel code, 124
ROMS, main time-stepping procedure of, 125
Rossby basin mode, 216
Rossby number, 320–321, 323, 324, 326, 328, 331, 332, 665
Rossby number, local vorticity-based, 324
Rossby waves, of quasi-geostrophic theory, 232
Rossby waves propagation, 211, 223
rotating channel, 324, 325
rotating mixing layer, 326
roundoff error, numerical, 273, 274, 276
Runge–Kutta scheme, 261, 292–294, 297, 303, 368

saddle-node bifurcation, 193, 194, 197, 204, 206, 212, 216, 243
saddle-node fixed point, 198, 199, 201, 204
Saint-Venant's equations, 389, 398–400
satellites, geostationary, 392, 393
satellites, polar-orbiting, 392, 393
scalar-advection tests, 15
SDEs. *See* stochastic differential equations
sea surface temperature (SST), 428
second-moment closure equations, 445, 459–465
second-moment closure equations, energetic consistency of, 464–465
second-order accurate control volume discretization method, 221
second-order adjoint model, 402, 419
second-order implicit midpoint method, 26
second-order methods, 402–408
SEEK filter. *See* singular evolutive extended Kalman filter
semidiscrete EPDiff, constrained action principle for, 258–260
semi-implicit continuity equations, 79, 89, 90, 91
semi-implicit energy conversion term, 93–95
semi-implicit equation, 84, 85
semi-implicit scheme, 81, 82, 84, 103, 116, 117
semi-implicit semi-Lagrangian model, 11
semi-implicit time-stepping, 11, 63, 79
semi-implicit tracer continuity equation, 93
semi-Lagrangian inherently conserving and efficient scheme. *See* SLICE scheme
semi-Lagrangian scheme, grid point, 11, 12, 22, 26, 38, 58
semi-Lagrangian scheme, FV/cell integrated, 12, 13, 18, 26, 58, 114, 119
semi-Lagrangian scheme, flux-based, 56
semi-monotonic filter, 22, 23
shallow water dynamics, 253
shallow water equations (SWEs), 133, 482, 579
shallow-water model, 218–220, 224
shape-preservation property, 18
shared-memory computers, parallelization on, 123
sharp catastrophic loss. *See* catastrophic loss
shear-Coriolis instability, 326, 327
shear flows, 324
Shilnikov chaos, 216
Shilnikov phenomenon, 200, 201
Shilnikov wiggles, 201, 217
simultaneous iteration technique (SIT), 205
singular belt, 44–45
singular evolutive extended Kalman (SEEK) filter, 420, 426
singular value decomposition (SVD), 414
SLICE scheme, 32, 35, 46
SLICE scheme, extension to spherical geometry, 46
Smagorinsky's model, 312, 315
Sobolev regularity, 678
soft catastrophic loss. *See* noncatastrophic loss
South China Sea (SCS), 247, 248
space shuttle, 247, 248
specific concentration, 5, 18, 56, 73, 78, 108, 110
split explicit time-stepping, 12, 78, 79
stationary equations, 550–552
steady shear flows, 241
steady states, 201–202, 205, 220
steady states, local codimension-1 bifurcations of, 193–195

stochastic differential equations (SDEs), 280, 283–285, 292, 303, 304
stochastic initial value problem, 454–459
stochastic methods, 388
stochastic numerical techniques, 279–304
stochastic perturbation equation, 462, 463
stochastic Taylor series, application to numerical schemes, 288, 289
stochastic Taylor series, basic structure of, 285–288
stochastic Taylor series, order of approximation, 290–292
Stokes drift, 345, 347
Stokes formula, 558
storm formation, 330
strange attractor, 198
strange attractor, aperiodic and unstable, 234
strange attractor, chaotic motion on, 235
Stratonovich calculus, 283, 285, 294–296
Stratonovich integrals, 283, 287, 290
Stratonovich SDE, 283, 284, 292, 303
Stratonovich system, 285, 292
streamfunctions, 211–213, 223, 233
subcritical modes, full 3D linear case, 554–560
subgrid-cell distributions, 20–25, 38
subgrid effects, parametrization of, 387
subgrid models, 315
subgrid-scale fluxes, 347, 348
subgrid-scale TKE equation, 348
subgrid-stress (SGS) tensors, 358, 359
subpolar gyre, 211, 212
subtropical gyre, 211, 212
supercritical modes, full 3D linear case, 560–566
SWEs. *See* Shallow-Water equations
swirling shear flow test, 15
symmetry-breaking bifurcations, 214
synthetic aperture radar (SAR) images, 247, 248

thermal-wind balance, 321
thermobaric effect, 173, 176
thermodynamic equation, 63, 80, 81, 87, 93, 98, 103, 106, 113
thermohaline circulation, 188
theta-method, 159
3D DCISL scheme, 13
3D PEs, 520, 547, 691, 704–706
3D trajectories, 67–69, 96, 97, 108, 111, 113
tidal, residual, intertidal mudflat (TRIM) model, 132
tilde filtering. *See* density-weighted Favre-filtering
time-dependent flows, 238
time-independent bounds, 471
time-independent flow, 275
time-periodic flows, in classical fluid dynamics, 231–244
time-periodic flows, in geophysical dynamics, 231–244
time-periodic flows, nonlinear, 232
time-split schemes. *See* operator-split schemes
time-stepping algorithms, 128
time-stepping algorithms, FB schemes, 128
time-stepping algorithms, nonlinear system, 160
TKE, 347, 352
torus bifurcation, 197, 198
total energy, 75
total energy conservation law, 74–75, 87
"total potential" energy, 74
tracer conservation, 126, 148–152
tracers, continuity equation for, 13
tracers, mixing ratio for, 19, 56, 58
trace theorem, 550, 552, 553, 734
trade winds, 321, 322
trajectory algorithm, 15, 26, 35, 69
transcritical bifurcation, 193, 196, 205, 206
transcritical bifurcation, perturbed, 195
translational passive advection tests, 14
transparent boundary conditions, 511
transport schemes, computational efficiency of, 16, 17
transport schemes, desirable properties, 13–20
transport schemes, in one and two dimensions, 12
transverse internal wave interactions, 247–250
tropical cyclones, 321
turbulence, decaying isotropic Euler, 313, 314
turbulence, on Jupiter, 310
turbulence, transition to, 242–244
turbulent behavior, 187
turbulent behavior, for boundary layers, 318
turbulent Ekman layer, 322, 323
turbulent flows, 124
turbulent kinetic energy (TKE), 347, 352
2D fast Fourier transform (FFT), 350
2D flows, 269, 270
2D flux-based/flux-form schemes, 5, 11, 12, 26, 47, 49, 55, 57, 61, 62
2D meandering jet, 233
2D models, eddy-resolving, 188
2D Navier-Stokes equations, 210
2D oceanic flows, 209
2D PEs, 642, 643, 647–654, 664, 695, 699–704
2D PEs, linear and nonlinear, 490, 491
2D PEs, strong solutions, 641, 668–678
2D PEs, uniqueness of z-weak solutions, 654–664
2D PEs, weak and functional formulations, 643–645
2D PEs, weak solutions, 665–667
2D subgrid-cell reconstructions, 24, 25

2D (x, z) model, finite differences in time and space, 508–510
2D (x, z) model, numerical results, 510, 511
2D (x, z) model, numerical simulations, 507–510
2D (x, z) model, theoretical framework, 494, 495
2D (x, z) model, vertical decomposition, 507, 508
2D (x, z) model, well-posedness results, 493–507
2.5D, impenetrable boundaries, 541–546
2.5D, well-posedness of linear system, 525–541
2.5D PEs, 520–525

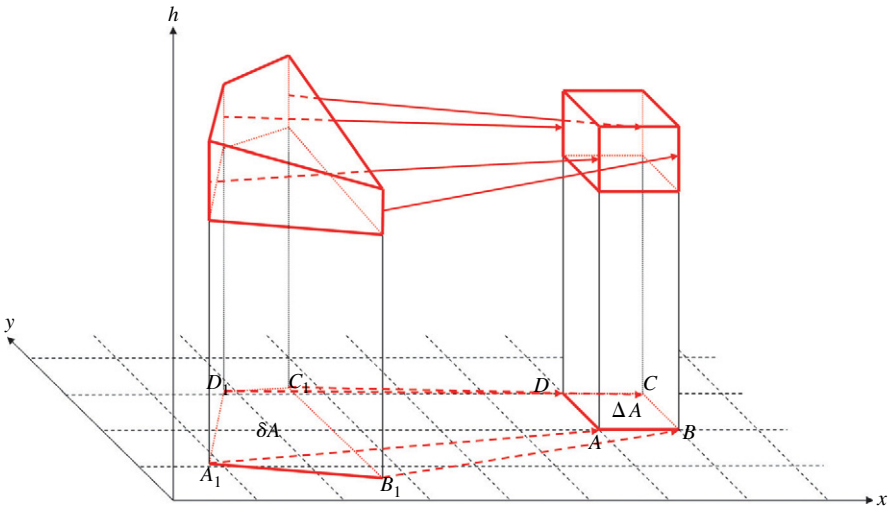
upstream integration, 70, 80, 91

variational methods, 388, 394–412
variational methods, ingredients of, 395, 396
variational methods, optimality system, 396, 397
variational particle mesh (VPM) method, 250, 253, 254, 269
variational particle mesh (VPM) method, discrete EP equation for, 261–263
variational principle, for discrete EPDiff, 258–261
velocity field, 269
velocity magnitude, surfaces of, 272
vertical averaging, 2D PEs, 645–647
vertical discretization, 62, 65, 97
vertical mode-splitting, 147–156
viscous dissipation, 349
von Neuman stability analysis, 37, 39, 41, 129
vortex forces, 344, 347
vortices, 328, 330, 336
vortices, anticyclonic and cyclonic, 321, 322, 331
vortices, boundary-layer, 318
vortices, coherent, 313, 317
vortices, oceanic, 323
vortices, quasi 2D cyclonic, 331

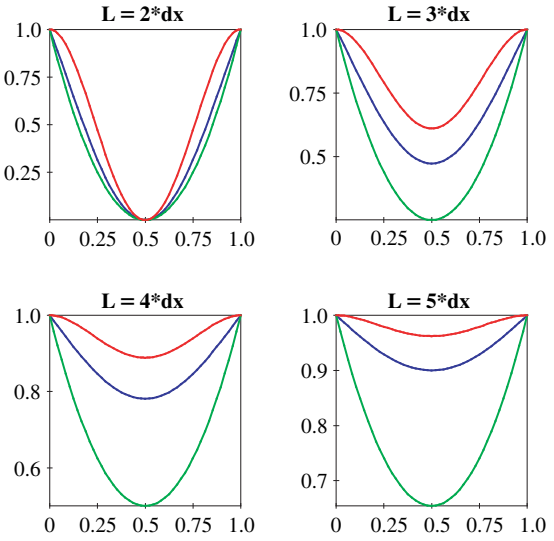
wake, rotating, 327
wall clock time (WT), 103
water vapor, conservation of mass of, 73
wave-current interactions, 344–346
wave-dynamical modes, 241
wave effects, 347
wave influences, simulations with, 350–355
wave-mean oscillation, 239–241
waves propagation, 248
wave system, 128, 146
wave trains, great lines on sea, 247
wave trains, nonlinear internal, 247
weakly compressible flows, 315–319
well-posedness of linear system, 525–541
Wiener process, 281, 286, 293, 299
wiggling behavior, 214
wind-driven double-gyre circulation, 209
wind-driven ocean circulation, 188–190, 209–222, 224
wind stress intensity, 218

zero mode, full 3D linear case, 553, 554
z-strong solutions, 696
z-weak solutions, 695, 699

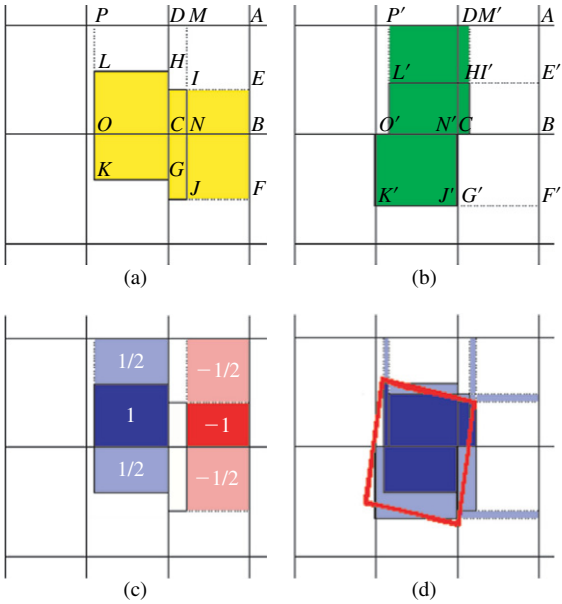
CHAPTER I



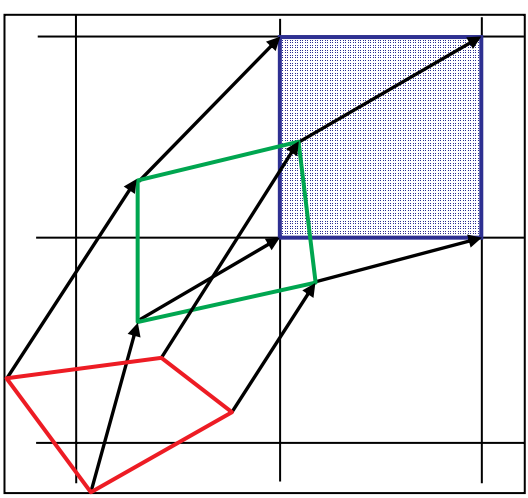
COLOR PLATE 1 (FIG. 1.1) Conceptual sketch showing a cell that is moving with the flow in a Lagrangian model layer during a time-step Δt. To the left is shown the cell at time t (the so-called departure cell). The horizontal velocity $\vec{V}$ within the model layer is assumed independent of height so that the cell walls, which initially at time t are vertical, remain vertical. The cell ends up at time $t + \Delta t$ as the horizontally regular Eulerian grid cell (the so-called arrival cell) shown in the vertical column to the right. Just four trajectories are shown. The projections on a horizontal plane are shown in more detail in Fig. 1.2.



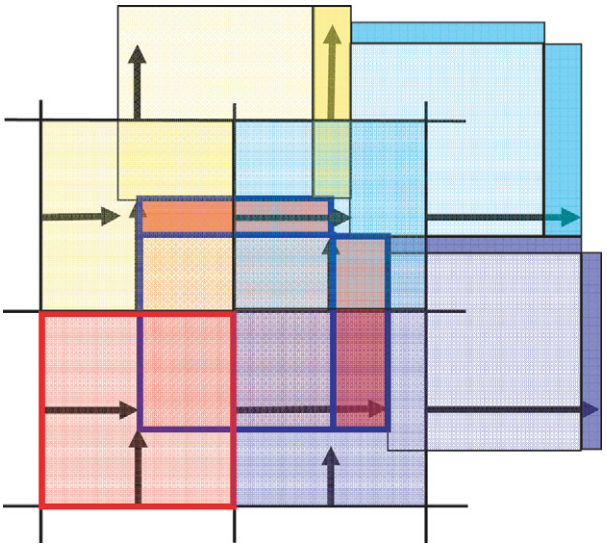
COLOR PLATE 2 (FIG. 2.12) Squared modulus of the amplification factor as a function of α for the (a) $2\Delta x$, (b) $3\Delta x$, (c) $4\Delta x$, and (d) $5\Delta x$ waves. Red and green lines are for the DCISL scheme using PPM2 and piecewise constant subgrid-cell representation, respectively. For comparison, the squared modulus of the amplification factor for the traditional semi-Lagrangian scheme based on cubic Lagrange interpolation (blue line) is shown as well.



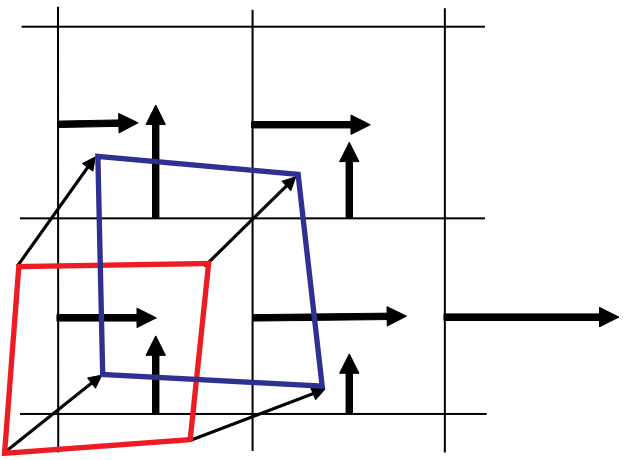
COLOR PLATE 3 (FIG. 2.11) A graphical illustration of the LIN and ROOD [1996] scheme for the idealized test case for assessing the degree of local mass conservation. The arrival cell is the north-eastern most regular grid cell in all plots. The capital letters on (a) and (b) refer to the vertices located south-west of the letter in question except for J′ and N′ that refer to the vertice to the south-east of the letter in question. The notation $\overline{ABCD}$ will refer to the average value in the cell with vertices at A, B, C, and D. (a) and (c) illustrate $X_C\left[\tfrac{1}{2}\left(\overline{\psi}^n + \overline{\psi}_{AY}\right)\right]$, where $\overline{\psi}_{AY}$ (yellow area) is computed using an advective operator. (a) Following the conceptual illustration of LEONARD, LOCK and MACVEAN [1996], $X_C\left[\tfrac{1}{2}\left(\overline{\psi}^n + \overline{\psi}_{AY}\right)\right]$ is given by $\tfrac{1}{2}\left(\overline{DCOP} + \overline{HGKL}\right) - \tfrac{1}{2}\left(\overline{ABNM} + \overline{EFJI}\right)$. (c) shows the cell averages with weight one (dark blue), half (light blue), minus one (red), and minus half (light red), for the contribution from X_C. (b) Similarly for Y_C, we get that $Y_C\left[\tfrac{1}{2}\left(\overline{\psi}^n + \overline{\psi}_{AX}\right)\right] = \tfrac{1}{2}\left(\overline{BF'G'C} + \overline{N'J'K'O'}\right) - \tfrac{1}{2}\left(\overline{AE'H'D} + \overline{M'I'L'P'}\right)$ and the green area is $\overline{\psi}_{AX}$. (d) shows the final forecast with the same coloring as in (c). The red rectangle is the exact departure area.



COLOR PLATE 4 (FIG. 3.3) Illustrating the different areas in Eqs. (3.113) and (3.114): δA_k^n(*red*), the departure area at time $n\Delta t$, $\delta A_k^{n+1/2}$(*green*), the "mid-way" area at time $(n + 1/2)\Delta t$, and δA_k^{n+1}(*blue*), the arrival area at time $(n + 1)\Delta t$.

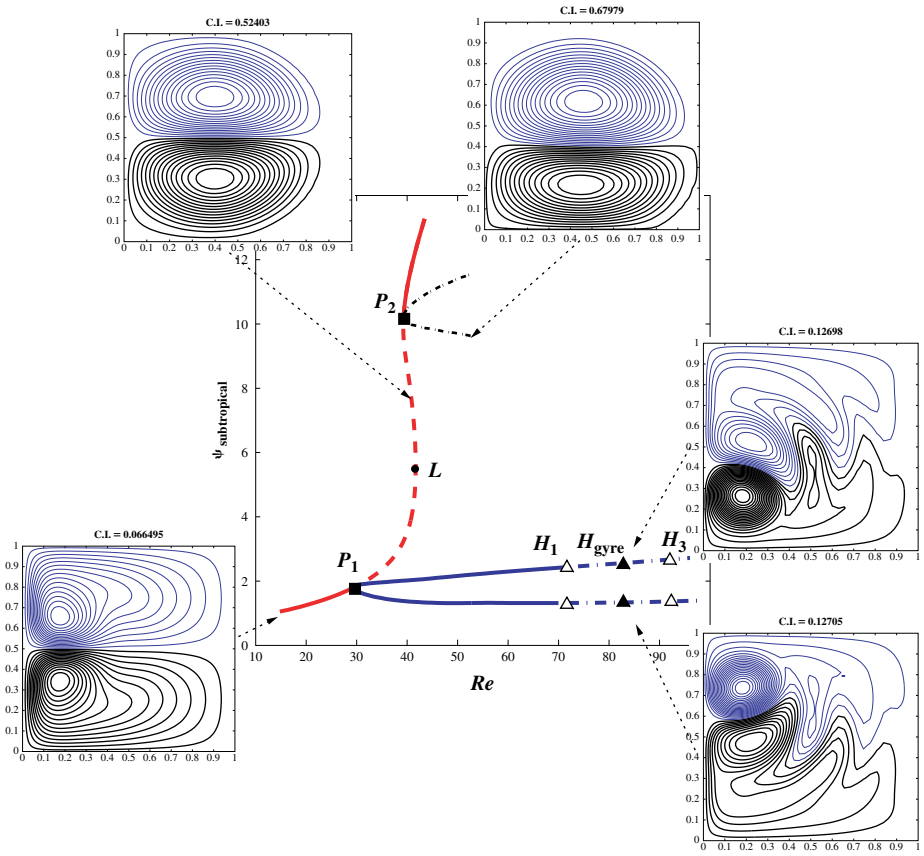


(A)

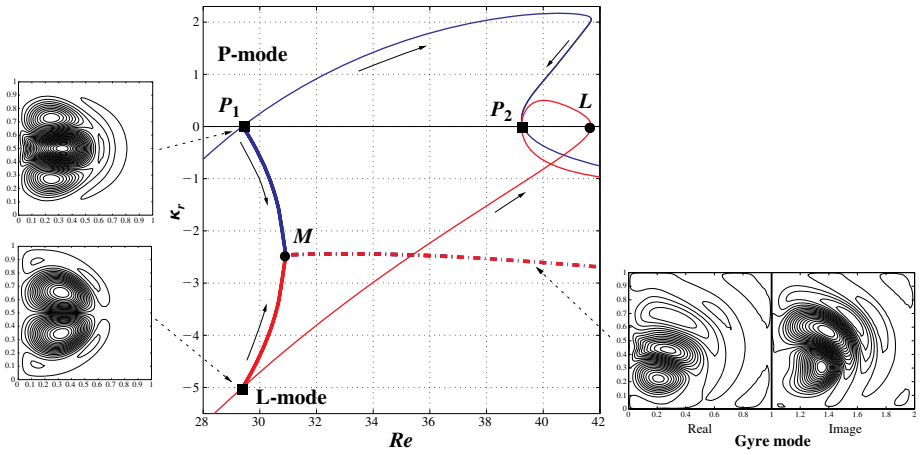


(B)

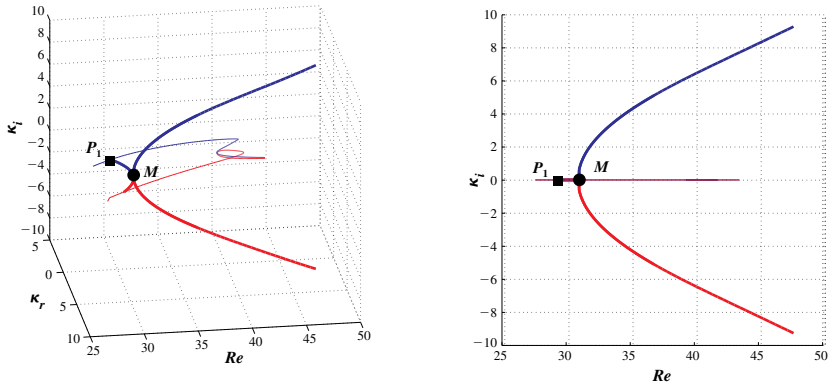
COLOR PLATE 5 (FIG. 3.4) Panel A: Illustrating the Lagrangian divergence $D_k^n\left(\vec{V}_k^n\right) = \frac{1}{\Delta A}\frac{\delta A_k^{n+1} - \Delta A}{\Delta t}$, which corresponds to the Eulerian divergence (Eq. (3.116)). The periphery of a regular departure area, ΔA, is marked red and the periphery around its arrival area, δA_k^{n+1}, is marked blue. Additional departure and arrival areas for three neighbor cells are shown. Note that the areas δA_k^{n+1} do not cover the whole domain; there are cracks between them. Panel B: Illustrating the Lagrangian divergence $D_k^n\left(\vec{V}_k^n\right) = \frac{1}{\Delta A}\frac{\delta A_k^{n+1/2} - \delta A_k^{n-1/2}}{\Delta t}$, where the departure area $\delta A_k^{n-1/2}$ is marked red and the arrival area $\delta A_k^{n+1/2}$ is marked blue. Obviously, $D_k^n\left(\vec{V}_k^n\right)$ is generally different from $D_k^n\left(\vec{V}_k^n\right)$ in A. In both panels: Black arrows are velocity components in the C-grid.



COLOR PLATE 6 (FIG. 4.2) Bifurcation diagram for the QG model (Eq. (4.1)). Here, the maximum value of the streamfunction of the subtropical gyre ($\Psi_{\text{subtropical}}$) is plotted versus the Reynold number $Re = UL_x/\nu$, where U is a typical characteristic horizontal velocity. The bifurcations P_1, P_2 correspond to the pitchfork bifurcations (squares), whereas the triangles indicate the location of the Hopf bifurcations. The saddle-node bifurcation L is indicated by a dot. The panels correspond to the upper layer streamfunctions, positive contours are represented by thick lines, negative contours by thin lines together with contour intervals (from SIMONNET and DIJKSTRA [2002]).



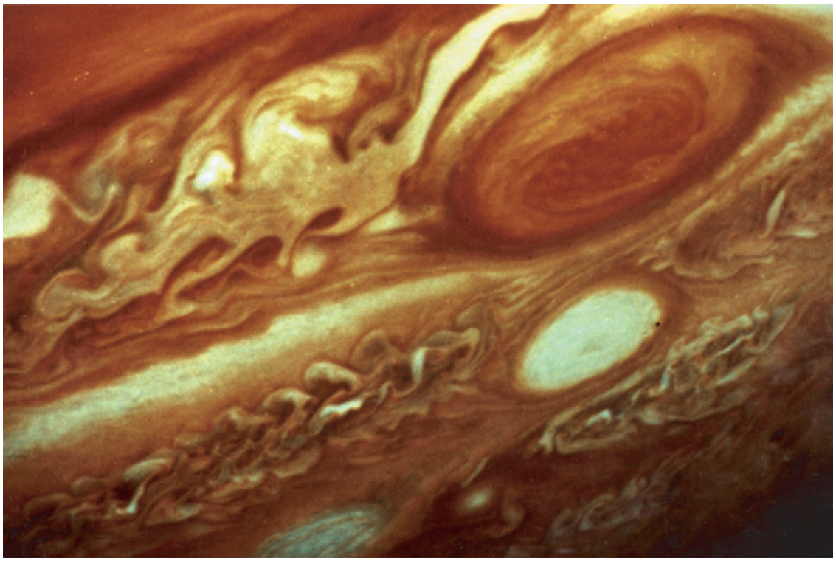
(a)



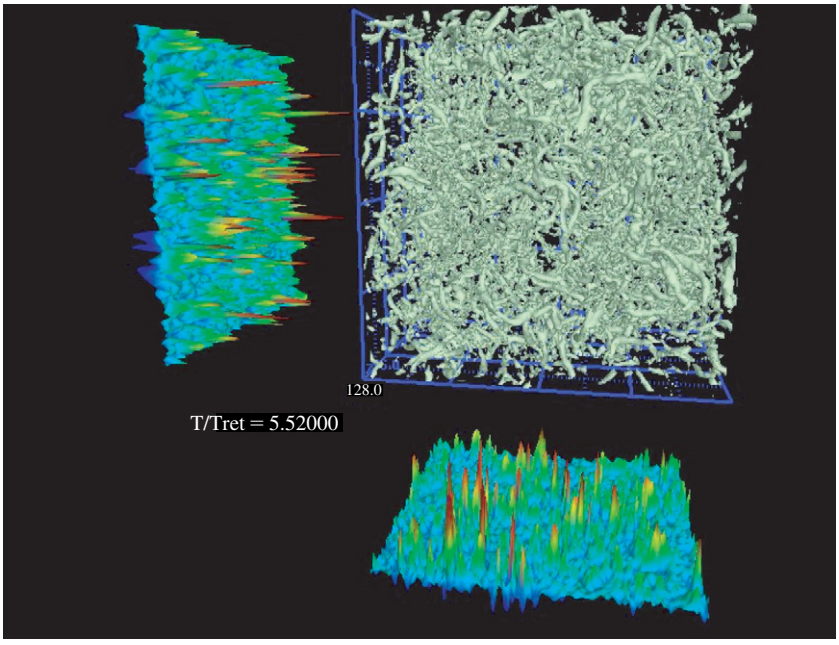
(b)

COLOR PLATE 7 (FIG. 4.3) Spectral behavior of the eigenmodes involved into the various bifurcations off the antisymmetric branch for the barotropic QG model shown in red and blue in Fig. 4.2. The real part, panel (a), and imaginary part, panel (b), of the P-mode and L-mode eigenvalues are plotted versus Re. Thick lines, beginning at P_1 in Fig. 4.2 and ending at M, correspond to the path of these two modes on the asymmetric branch, thin lines to the path on the antisymmetric branch. The dash-dot thick line indicates a non-zero imaginary part. At M, the merging between the P-mode and L-mode occurs. The various panels show the streamfunction patterns at the locations indicated by the arrows (from SIMONNET and DIJKSTRA [2002]).

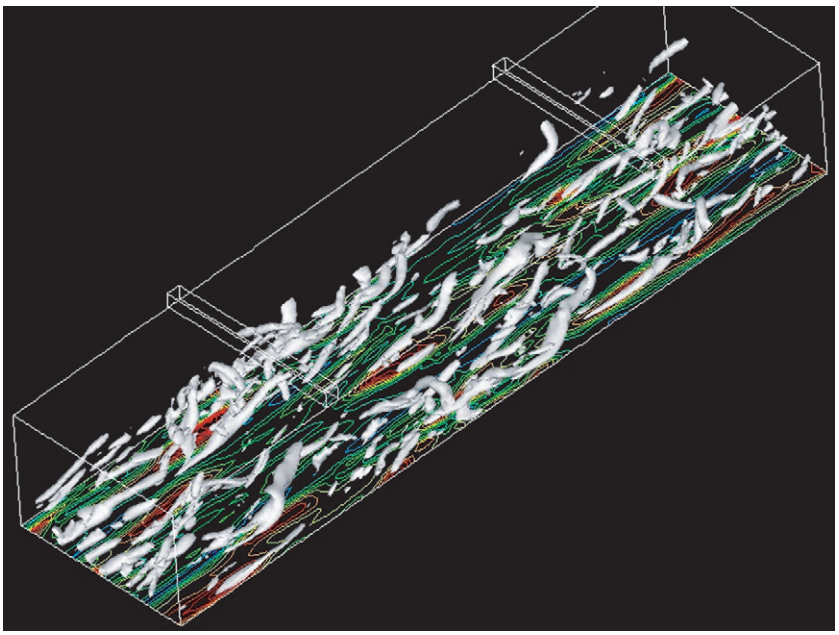
CHAPTER VII



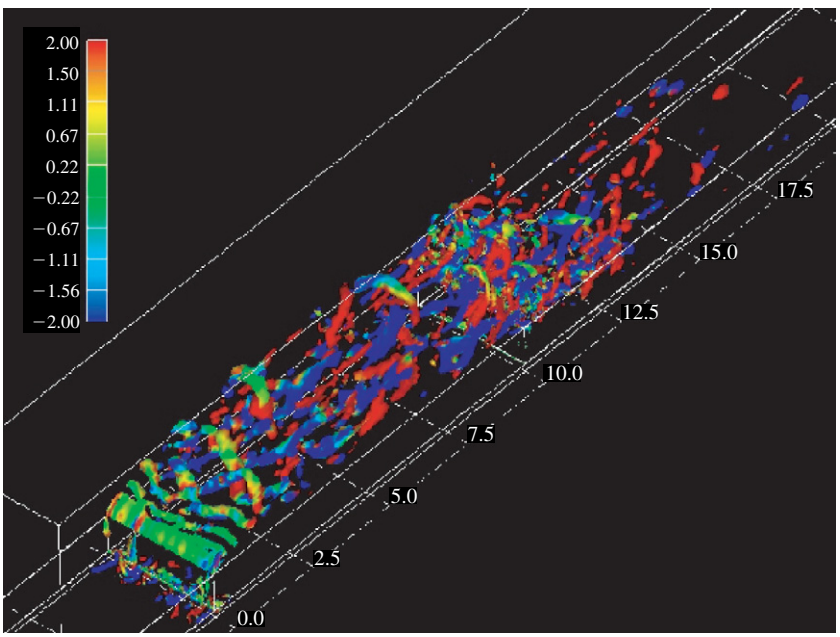
COLOR PLATE 8 (FIG. 1.1) Turbulence on Jupiter around the great red spot (picture courtesy of Jet Propulsion Laboratory, Pasadena).



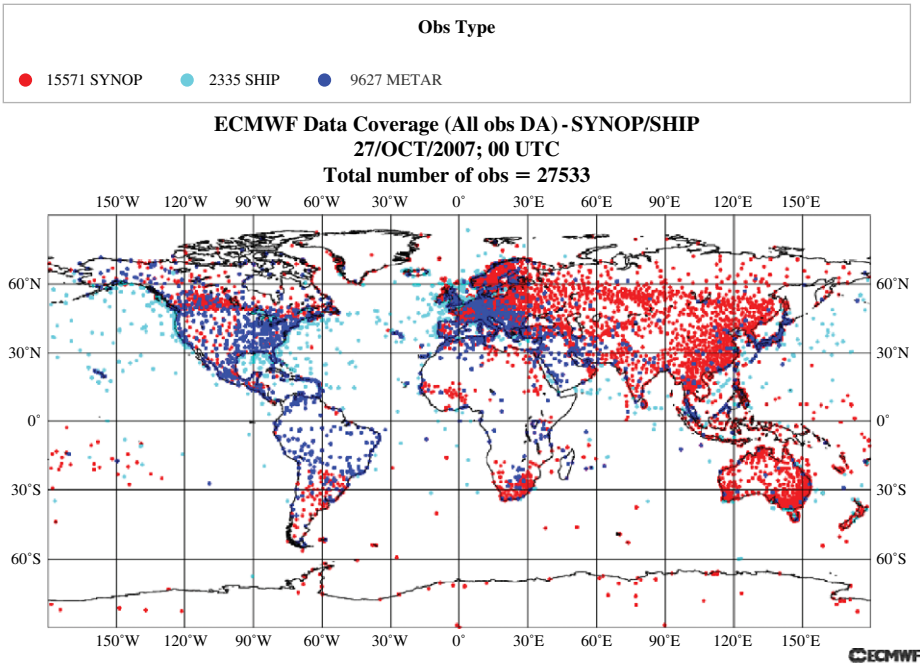
COLOR PLATE 9 (FIG. 1.2) Isotropic decaying turbulence. View of positive-Q isosurfaces (from LESIEUR, MÉTAIS and COMTE [2005], courtesy of Cambridge University Press).



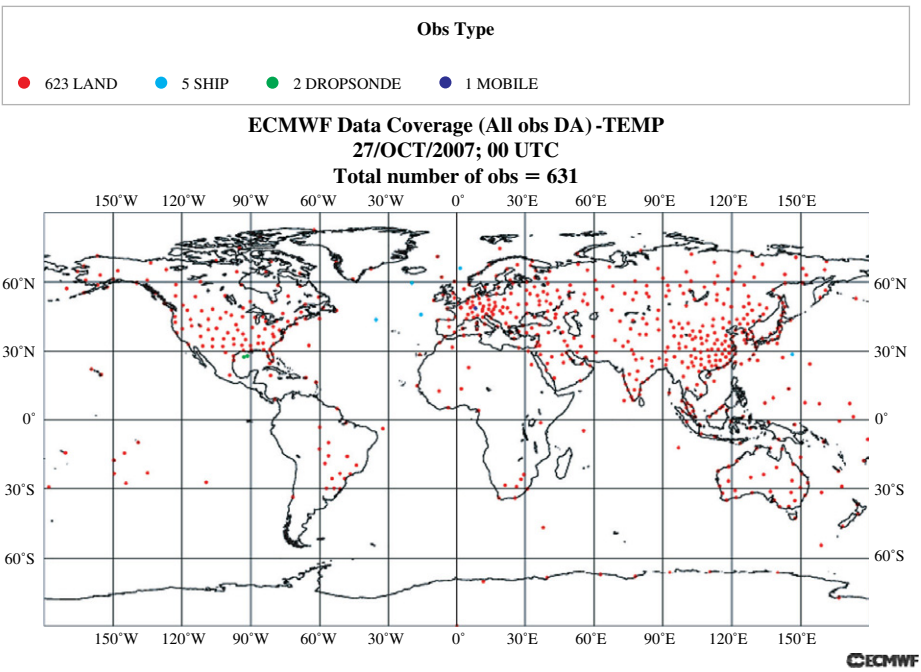
COLOR PLATE 10 (FIG. 1.4) $Q > 0$ isosurfaces in a weakly compressible channel (picture courtesy of Y. DUBIEF, from DUBIEF and DELCAYRE [2000]).



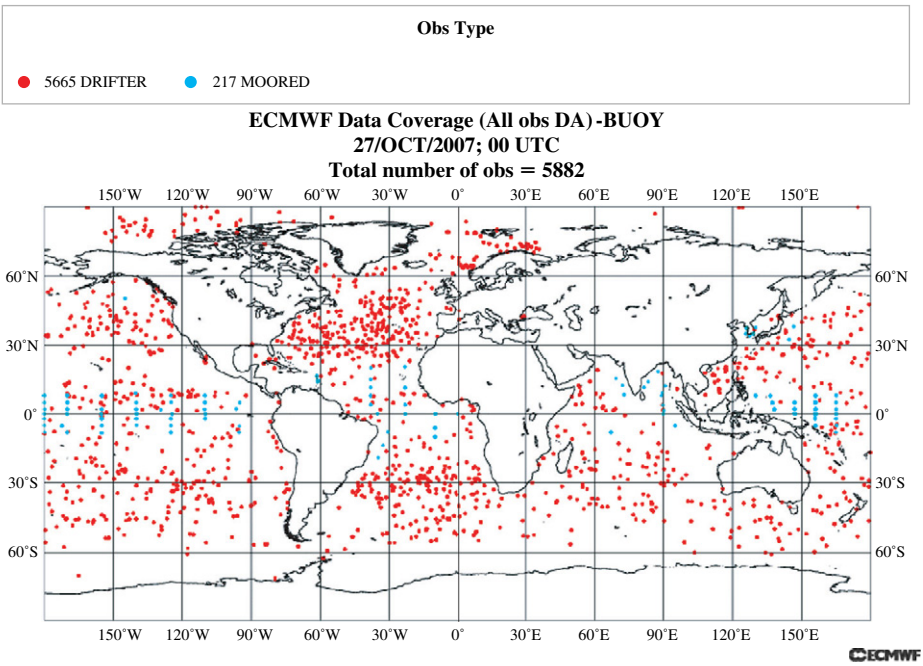
COLOR PLATE 11 (FIG. 1.6) Obstacle with wall effect. Positive Q isosurfaces with positive and negative longitudinal vorticity (from LESIEUR, MÉTAIS and COMTE [2005], courtesy of Cambridge University Press).



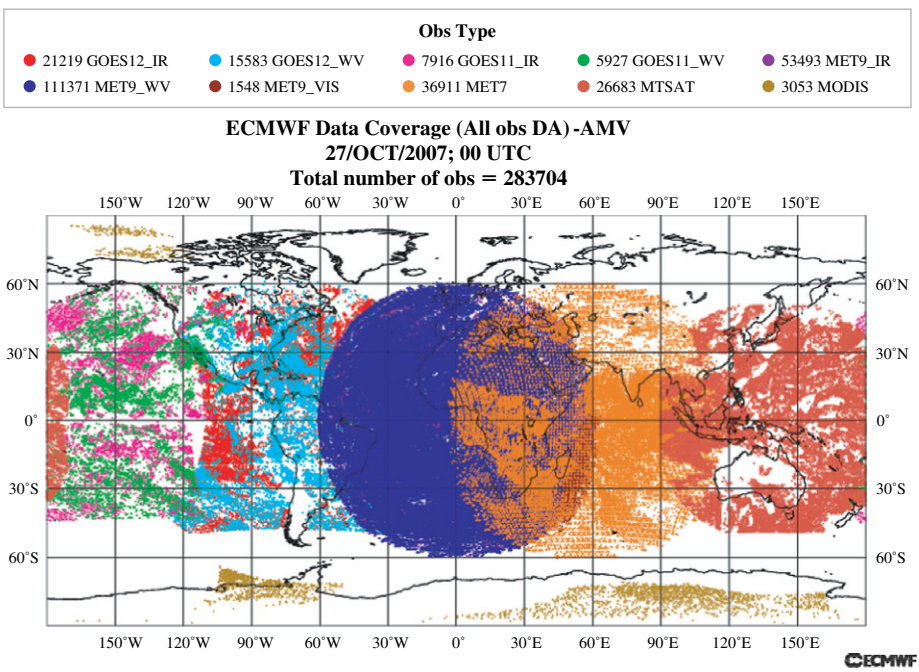
COLOR PLATE 12 (FIG. 2.1) SYNOP/SHIP data: synoptic networks in red, airport data in blue, and ship data in green.



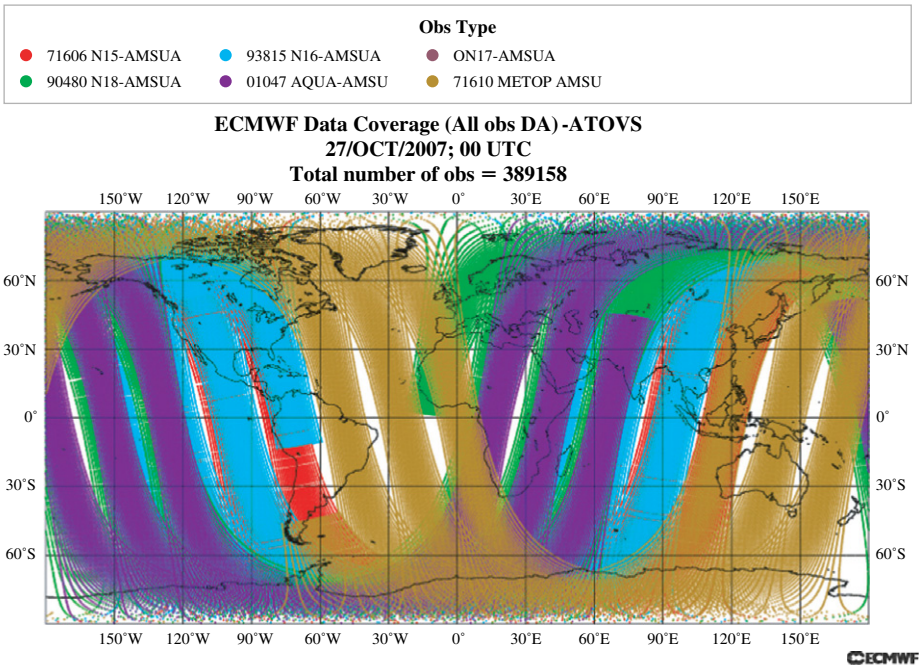
COLOR PLATE 13 (FIG. 2.2) Radiosonde measurements.



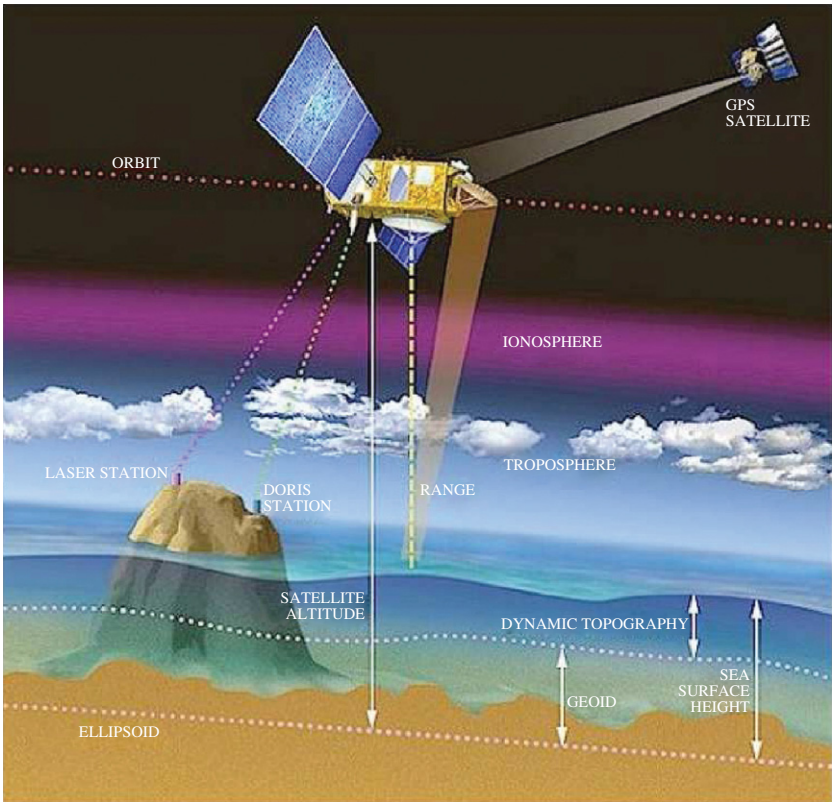
COLOR PLATE 14 (FIG. 2.3) Drifting and moored buoys.



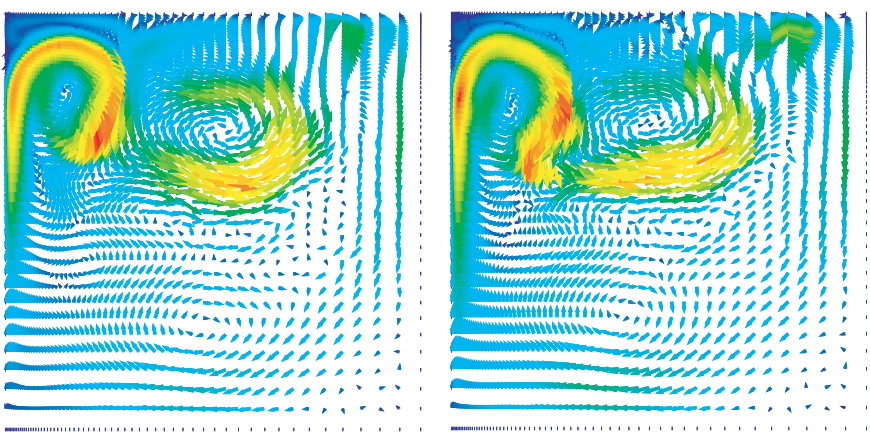
COLOR PLATE 15 (FIG. 2.4) Observations from ten geostationary satellites.



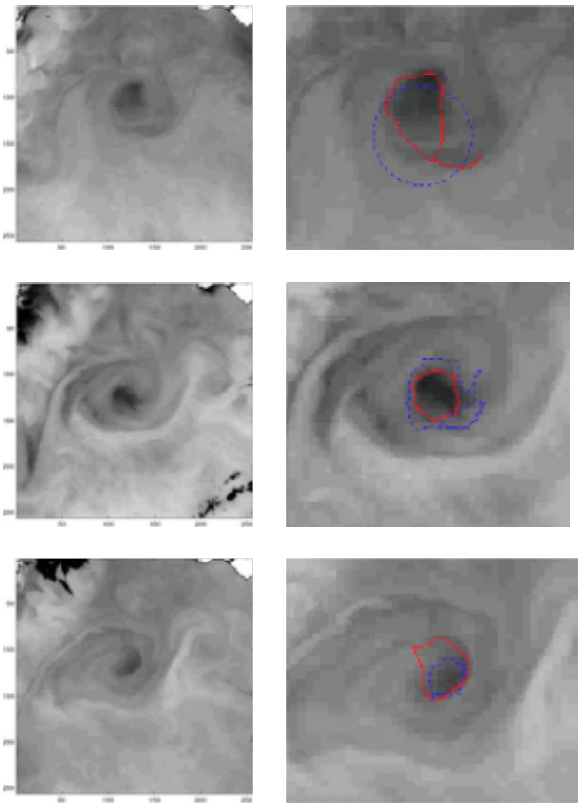
COLOR PLATE 16 (FIG. 2.5) Trajectories of six polar-orbiting satellites.



COLOR PLATE 17 (FIG. 2.6) Satellite altimetry (CNES figure from Aviso web site http://www.aviso.oceanobs.com).



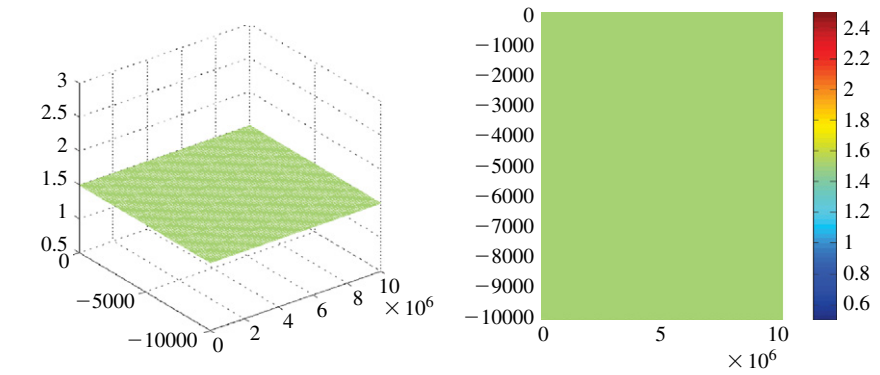
COLOR PLATE 18 (FIG. 6.1) Application of POD model reduction method to Imperial College Ocean Model (ICOM) adaptive finite element ocean Model. Full model and reduced-order model based on first 30 base functions.



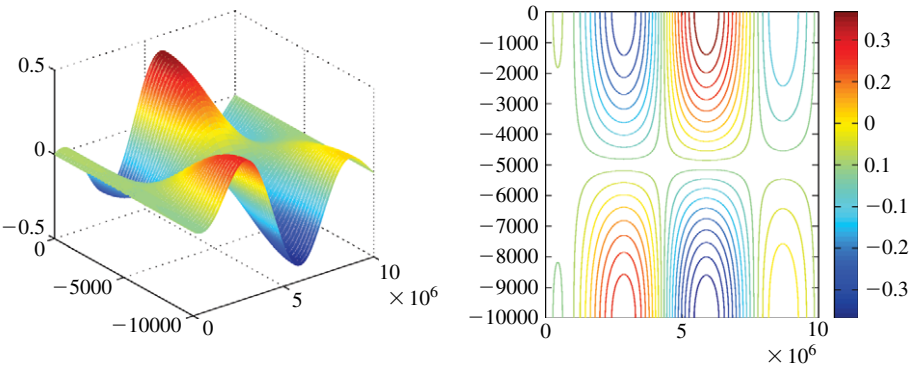
Evolution of a vortex in SST

Tracking in triplet successive frames using the curvelets based method. The initial and final snakes are displayed.

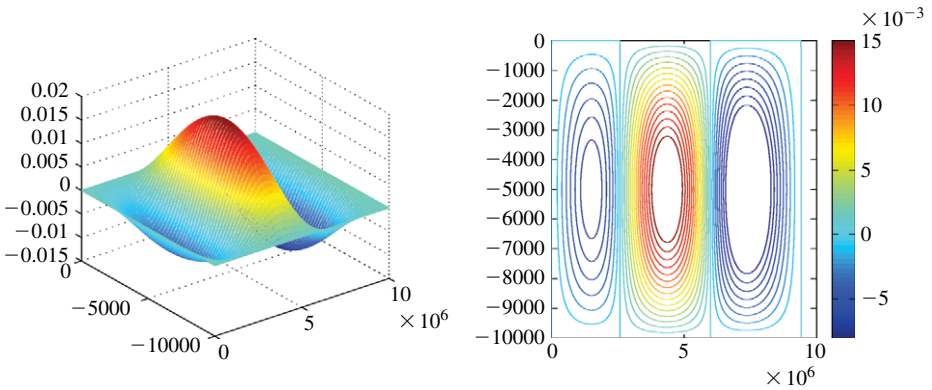
COLOR PLATE 19 (FIG. 8.3) Direct assimilation of images.



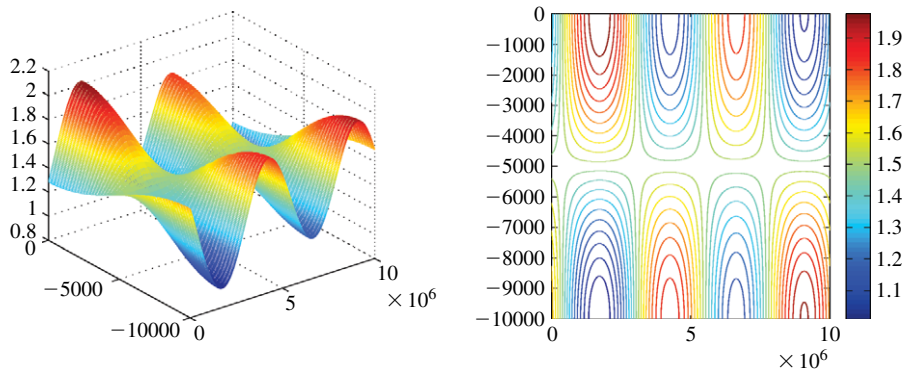
COLOR PLATE 20 (FIG. 2.2) Periodic boundary condition. Initial data u_0.



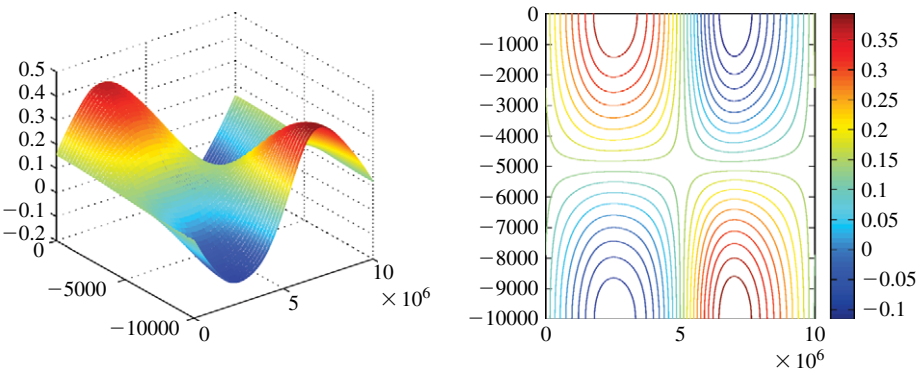
COLOR PLATE 21 (FIG. 2.3) Periodic boundary condition. Initial data v_0.



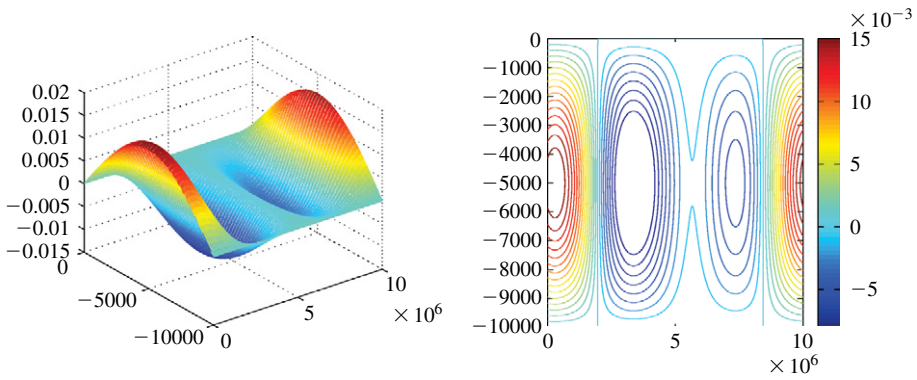
COLOR PLATE 22 (FIG. 2.4) Periodic boundary condition. Initial data ψ_0.



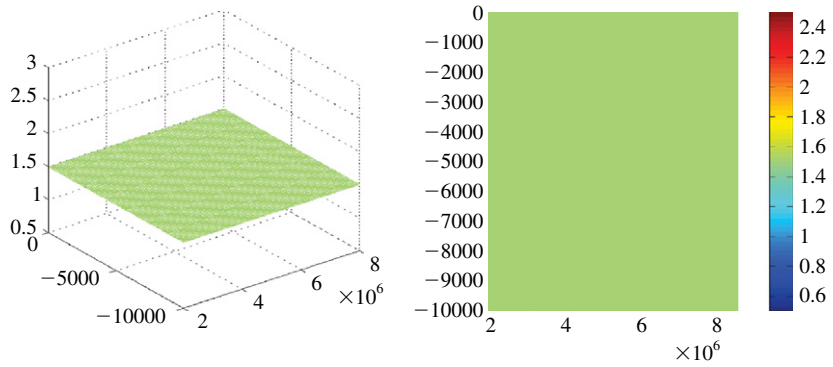
COLOR PLATE 23 (FIG. 2.5) Periodic boundary condition. Values of u at $t = t_1$.



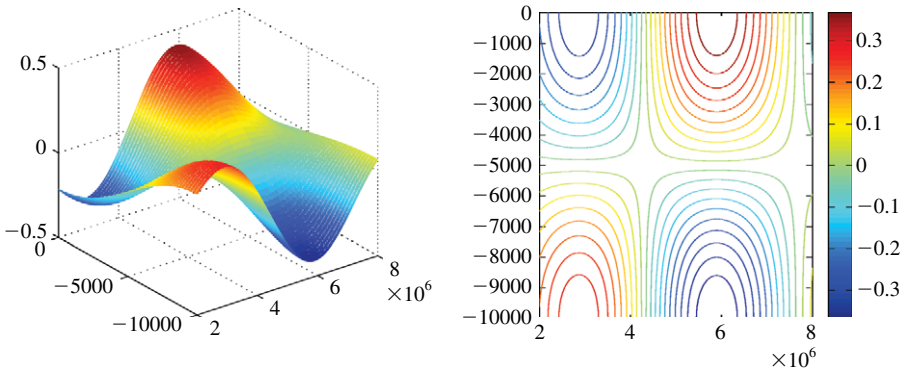
COLOR PLATE 24 (FIG. 2.6) Periodic boundary condition. Values of v at $t = t_1$.



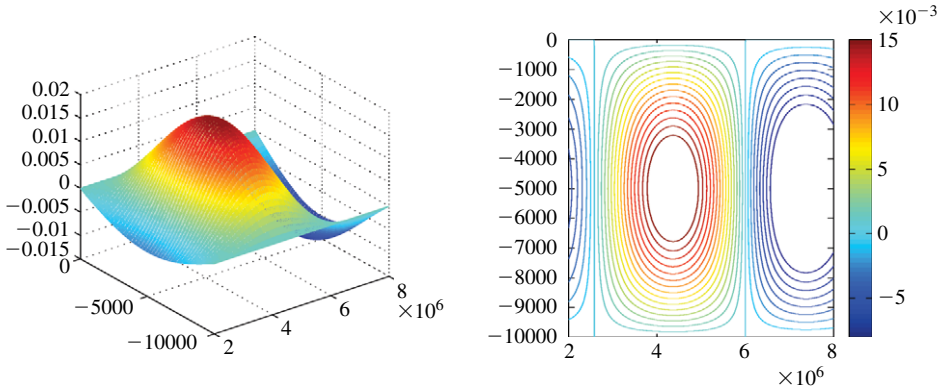
COLOR PLATE 25 (FIG. 2.7) Periodic boundary condition. Values of ψ at $t = t_1$.



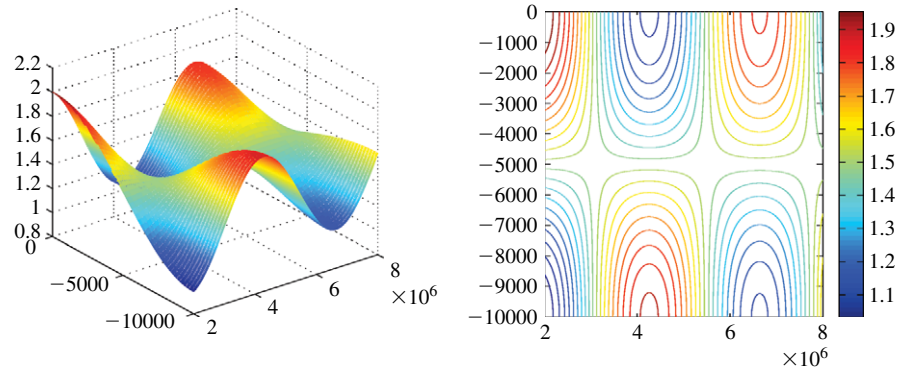
COLOR PLATE 26 (FIG. 2.9) Transparent boundary condition. Initial data u_0.



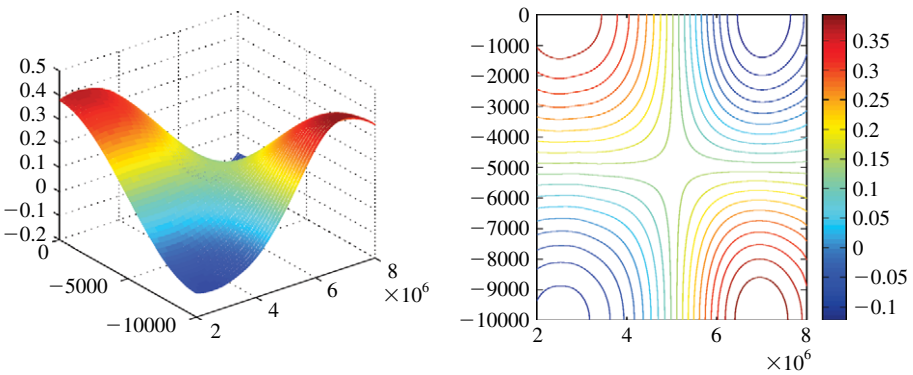
COLOR PLATE 27 (FIG. 2.10) Transparent boundary condition. Initial data v_0.



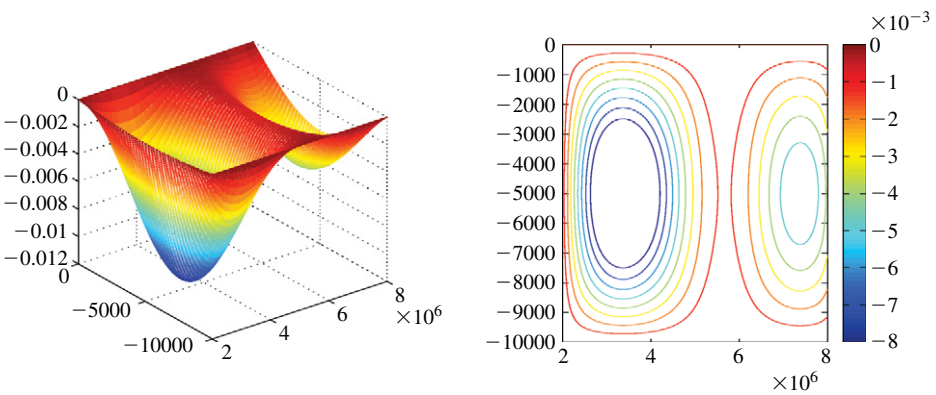
COLOR PLATE 28 (FIG. 2.11) Transparent boundary condition. Initial data ψ_0.



COLOR PLATE 29 (FIG. 2.12) Transparent boundary condition. Values of u at $t = t_1$.



COLOR PLATE 30 (FIG. 2.13) Transparent boundary condition. Values of v at $t = t_1$.



COLOR PLATE 31 (FIG. 2.14) Transparent boundary condition. Values of ψ at $t = t_1$.