HAL
open science

New numerical methods for shallow water flows

Abdelaziz Beljadid

To cite this version:

Abdelaziz Beljadid. New numerical methods for shallow water flows. Numerical Analysis [math.NA]. Université Pierre et Marie Curie - Paris VI; Université d'Ottawa, 2015. English. NNT : 2015PA066355 . tel-01266679

HAL Id: tel-01266679

https://theses.hal.science/tel-01266679v1

Submitted on 3 Feb 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Université Pierre et Marie Curie - École Doctorale de Sciences Mathématiques de Paris Centre et Université d'Ottawa - Département de Génie Civil

THÈSE EN COTUTELLE

pour obtenir le grade de

Docteur en Mathématiques appliquées de l'Université Pierre et Marie Curie
Docteur en Génie civil de l'Université d'Ottawa

Présentée par

Abdelaziz Beljadid

Nouvelles méthodes numériques pour les écoulements en eaux peu profondes

---

Directeurs de thèse : Philippe G. LeFloch et Abdolmajid Mohammadian

Rapporteurs :

Alina Chertouk	North Carolina State University, USA
Daniel Y. Le Roux	Université Claude Bernard Lyon 1, France

Soutenue le 9 juillet 2015 devant le jury composé de :

Alina Chertouk	North Carolina State University, USA
Amir Hakami	Carleton University, Canada
Philippe G. LeFloch	Université Pierre et Marie Curie, France
Ousmane Seidou	Université d'Ottawa, Canada
Stéphane Zaleski	Université Pierre et Marie Curie, France

À ma femme Mounia et mes fils Hamza, Adam et Ali

À la mémoire de mon père, à ma mère, mes sœurs et frères

REMERCIEMENTS

Je voudrais tout d'abord remercier mes directeurs de thèse, en mathématiques Mr Philippe G. LeFloch et en génie civil Mr Abdolmajid Mohammadian pour m'avoir donné la possibilité de travailler sur un sujet de recherche très intéressant. Leurs manières d'encadrement et leurs discussions scientifiques très stimulantes et enrichissantes ont constituées pour moi des éléments moteurs pour la bonne réussite de la thèse.

Je remercie très chaleureusement mes deux examinateurs de thèse Mme Alina Chertouk et Mr Daniel Y. Le Roux. Je remercie sincèrement Mr Amir Hakami, Mr Ousmane Seidou et Mr Stéphane Zaleski pour m'avoir fait l'honneur d'être membre de Jury.

Je suis extrêmement reconnaissant à Mme Johanne Bruyère pour son travail efficace et ses efforts considérables pour l'établissement et le suivi de la cotutelle. Je remercie également Mme Patricia Zizzo pour sa patience et ses efforts remarquables qu'elle n'a cessé de déployer pour l'élaboration de la convention de cotutelle.

Enfin une personne d'une importance colossale dans ma vie, qui m'accompagne et qui me soutient quotidiennement et surtout dans les périodes difficiles. Il s'agit de ma femme, dont je ne pourrai quantifier l'apport et elle a été l'autre moi. Je te remercie infiniment Mounia.

RÉSUMÉ

Dans ce projet de recherche, on s'intéresse au développement et à l'évaluation de nouvelles méthodes numériques pour les écoulements peu profonds. De nouvelles techniques de discrétisation spatiales et temporelles des équations sont proposées. Une partie de la thèse est dédiée au développement d'une méthode des volumes finis explicite d'ordre élevé et d'une famille de schémas semi-implicites qui sont efficaces pour la modélisation des processus lents et rapides dans les écoulements océaniques et atmosphériques. La deuxième partie du projet de recherche concerne la construction d'un schéma numérique efficace sans solveur de Riemann pour les écoulements peu profonds avec une topographie variable sur un maillage non structuré. Dans cette partie de la thèse, une nouvelle approche est proposée pour l'analyse de stabilité des schémas numériques non structurés pour les équations en eaux peu profondes. Dans la troisième partie de la thèse, deux schémas de volumes finis sont développés pour les lois de conservation sur des surfaces courbes qui ont un large potentiel d'être appliqués aux écoulements peu profonds sur la sphère. Dans ces cas, les schémas numériques sont développés en adoptant la démarche suivie par Stanley Osher. Cette démarche consiste à utiliser des systèmes hyperboliques simples qui génèrent des phénomènes d'ondes complexes et des solutions qui ont différentes structures. Ces solutions sont très efficaces pour tester les méthodes numériques. Dans notre cas, nous avons utilisé les équations de Burgers qui ont joué un rôle très important dans le développement des schémas numériques à capture de chocs en mécanique des fluides.

Dans le premier article, une nouvelle méthode des volumes finis décentrée explicite est proposée pour le système de Saint-Venant avec un terme source qui comprend le paramètre de Coriolis en utilisant un maillage non structuré. La plupart des schémas numériques décentrés, efficaces pour les ondes rapides (ondes de gravité), conduisent à un niveau d'amortissement élevé pour les ondes lentes (ondes de Rossby). La méthode proposée donne de bons résultats à la fois pour les ondes de gravité et les ondes de Rossby. Les techniques proposées sont suffisantes pour supprimer le bruit numérique des ondes courtes sans amortissement des ondes longues, telles que les ondes de Rossby qui sont essentielles dans le transport de l'énergie dans les océans et l'atmosphère.

Dans le cas où le système comprend une large gamme de fréquences des ondes, ce qui est le cas des écoulements atmosphériques, il est important d'utiliser des méthodes semi-implicites afin d'opter pour un pas de temps optimal. La méthode semi-implicite semi-lagrangienne à deux niveaux (SETTLS) proposée par Hortal (2002) a une région de stabilité absolue indépendante du nombre de Courant-Friedrichs-Lewy (CFL). La plupart des modèles de prévision numérique atmosphérique utilisent cette méthode comme schéma temporel. Cependant, la méthode SETTLS peut générer des oscillations pour le traitement du terme non linéaire surtout pour le cas des solutions qui ont un caractère

oscillatoire. Pour remédier à ce problème, dans le deuxième article, nous avons proposé une nouvelle classe de schémas semi-implicites semi-lagrangiens potentiellement applicables aux modèles atmosphériques. Cette classe de schémas numériques présente plusieurs avantages de stabilité, de précision et de convergence. De bons résultats sont obtenus en comparaison à d'autres schémas semi-implicites semi-lagrangiens et méthodes semi-implicites de type prédicteur-correcteur.

Dans le troisième article, un nouveau schéma équilibre partiellement centré est développé pour la résolution numérique des équations de Saint-Venant avec une topographie variable sur un maillage non structuré. Cette méthode est stable et simple puisqu'elle ne fait pas appel à la résolution du problème de Riemann. La méthode proposée est précise pour le cas des solutions discontinues et peut être appliquée aux écoulements peu profonds avec une topographie variable et une géométrie complexe où l'utilisation des maillages non structurés est avantageuse.

Motivé par de nombreuses applications en dynamique des fluides, dans le projet de thèse on s'intéresse également au développement de méthodes numériques dans le cas des surfaces courbes. L'objectif est de concevoir des méthodes numériques robustes et efficaces pour le cas des solutions discontinues et qui préservent la structure fondamentale des équations, notamment les propriétés liées à la géométrie. Pour développer ces méthodes, l'approche suivie par Stanley Osher est adoptée et les équations de Burgers sont utilisées vu leur importance pour le développement des schémas numériques à capture de chocs.

Dans le quatrième article, une méthode des volumes finis satisfaisant la compatibilité géométrique est développée pour les lois de conservation sur la sphère. Cette méthode est basée sur la résolution du problème de Riemann généralisé et l'approche du «splitting» directionnel en latitude et en longitude sur la sphère. Les dimensions géométriques sont considérées de manière analytique et la forme discrète du schéma numérique proposé respecte la propriété de compatibilité géométrique. La méthode proposée est stable et précise pour le cas des solutions discontinues de grands chocs et amplitudes en comparaison avec des schémas numériques très connus. Une nouvelle classification des flux est proposée en introduisant les notions de flux feuilletés et de flux génériques. Le comportement asymptotique des solutions est étudié en fonction de la nature du flux et les propriétés des solutions discontinues sont analysées. Les résultats démontrent la capacité et le potentiel de la méthode proposée pour la résolution des lois de conservation sur la sphère dans le cas des solutions discontinues. Ce schéma numérique pourrait être étendu au cas des équations de Saint-Venant sur la sphère.

Dans le cinquième article, on propose un schéma numérique efficace respectant la propriété de compatibilité géométrique pour les lois de conservation sur la sphère. La méthode proposée présente plusieurs avantages, notamment de bons résultats dans le cas des solutions discontinues avec des chocs d'amplitudes moyennes, une faible dis-

sipation numérique et une simplicité puisqu'elle ne fait pas appel à la résolution du problème de Riemann. Cette méthode pourrait être étendue au cas des équations de Saint-Venant sur la sphère.

Dans le sixième article, une nouvelle approche est proposée pour analyser la stabilité des schémas numériques appliqués aux écoulements peu profonds. Cette méthode utilise la notion du pseudo spectre des matrices. La méthode proposée est efficace en comparaison avec les méthodes couramment utilisées telles que la stabilité asymptotique et la stabilité de Lax-Richtmyer. Cette approche est utile pour le choix du type de maillage, des emplacements appropriés des variables primitives (hauteur et vitesses), et de la méthode de discrétisation la plus stable.

ABSTRACT

This research project focuses on the development and evaluation of numerical methods for shallow flows by proposing new spatial and temporal discretization techniques. First, a new high-order explicit finite volume method and a class of semi-implicit schemes are introduced which are effective for modelling fast and slow waves in oceanic and atmospheric flows. In the second part of the research project, a central-upwind scheme is proposed for shallow water flows on variable topography using unstructured grids. In this part of the project, a new approach is proposed for the stability analysis of unstructured numerical schemes for shallow water equations. In the third part of the thesis, two finite volume methods are developed for the conservation laws on curved geometries which are potentially applicable to shallow flows on a sphere. For such cases, numerical schemes are developed by using the approach followed by Stanley Osher. This approach employs simple hyperbolic systems which generate complex wave phenomena, and solutions that are effective for assessing numerical methods. In our case, Burgers' equations are used since they have played an important role in the development of shock-capturing schemes in fluid mechanics.

In the first paper, a new explicit upwind finite volume method is proposed for the Saint-Venant system using unstructured grids, with a source term which is assumed to include the Coriolis Effect. Most upwind schemes that perform well for fast gravity waves lead to a high level of damping for slow modes such as Rossby waves. The developed method leads to accurate results for both gravity and Rossby waves. The proposed techniques are enough to suppress the short-wave numerical noise without damping long waves essential in the transport of energy in the ocean and atmosphere, such as Rossby waves. However, it is useful to use semi-implicit methods for systems which include different scales of wave speeds, such as atmospheric flows, in order to use practical time steps. Hortal (2002) proposed a two-time-level semi-Lagrangian method called the Stable Extrapolation Two-Time-Level Scheme (SETTLS), which has a region with absolute stability independent of the Courant-Friedrichs-Lewy number (CFL). Most weather-prediction models use this method as a temporal scheme. However, the SETTLS method can generate high noise for a purely oscillatory nonlinear term, depending on the size of the time step. The goal of the second paper is to deal with the issues of instability associated with the treatment of the non-linear part of the forcing term. A class of semi-implicit semi-Lagrangian schemes is developed which is potentially applicable to atmospheric models. The proposed class of schemes performs well in terms of stability, accuracy, convergence, and efficiency in comparison with other previously known semi-implicit semi-Lagrangian schemes and semi-implicit predictor corrector methods.

In the third paper, a well-balanced positivity-preserving cell-vertex central-upwind

scheme is proposed for the Saint-Venant system with variable bottom topography. The advantages are that the proposed method is Riemann-problem-solver-free, stable, and accurate for discontinuous solutions. This scheme can be applied to problems with complex geometries, where the use of unstructured grids is advantageous.

Motivated by numerous applications in fluid dynamics, we are also interested in the development of numerical methods for conservation laws on curved geometries, with the objective being to design robust and efficient numerical approximation methods. These schemes allow the computation of discontinuous solutions and preservation of the fundamental structure of the equations, especially geometry-related properties. In order to develop these methods, the approach followed by Stanley Osher is adopted by using Burgers' equations.

In the fourth paper, a geometry-preserving finite volume method is developed for conservation laws on a sphere. The proposed method is based on a generalized Riemann solver and an operator-splitting approach using latitude and longitude on a sphere. The geometric dimensions are considered in an analytical way, which leads to a discrete form of the scheme that respects the geometric compatibility property. The proposed method performs well in terms of stability and accuracy for discontinuous solutions with large amplitudes and shocks compared to some well-known schemes. A new classification of the flux vector is introduced in which the foliated and generic fluxes are distinguished. The properties of the solutions are investigated and their late-time asymptotic behavior is studied using the properties of the flux vector field. The results demonstrate the proposed scheme's potential and ability to resolve discontinuous solutions with large amplitudes and shocks for conservation laws on a sphere. This method could be extended to shallow water models on a sphere.

An efficient finite volume method is introduced in the fifth paper for conservation laws on a sphere. The main advantages of the designed scheme are its low numerical dissipation and simplicity, since no Riemann solvers are used. The proposed method has good resolution of conservation laws for discontinuous solutions with shocks of average amplitude. The scheme respects the geometric compatibility property, and it is efficient for nonlinear hyperbolic conservation laws on a sphere. This method could be extended to the shallow water systems on a sphere.

In the sixth paper, a new method using pseudospectra is proposed for stability analysis of unstructured finite volume methods for shallow water equations. The proposed approach is effective compared to the commonly used methods such as asymptotic stability and Lax-Richtmyer stability. The proposed approach can be helpful for choosing the type of mesh, the appropriate placements of the primitive variables on the grids, and a suitable discretization method which is stable for a wide range of modes.

TABLE DES MATIÈRES

REMERCIEMENTS iii

RÉSUMÉ iv

ABSTRACT vii

TABLE DES MATIÈRES ix

LISTE DES TABLEAUX xiii

LISTE DES FIGURES xiv

CHAPITRE 1 Introduction 1
- 1.1 Objectifs 1
- 1.2 Revue de littérature 4
  - 1.2.1 Méthodes des volumes finis décentrées pour les écoulements peu profonds 5
  - 1.2.2 Schémas semi-implicites semi-lagrangiens et leurs applications aux écoulements atmosphériques 7
  - 1.2.3 Méthodes des volumes finis partiellement centrées pour les écoulements peu profonds 8
  - 1.2.4 Méthodes des volumes finis pour les lois de conservation sur des surfaces courbes 9
- 1.3 Méthodologie et présentation du travail de recherche 9
  - 1.3.1 Une méthode des volumes finis non structurée pour les écoulements à grande échelle 10
  - 1.3.2 Une classe de schémas semi-implicites semi-lagrangiens potentiellement applicables aux modèles atmosphériques 12
  - 1.3.3 Un schéma équilibre partiellement centré de type «Cell-Vertex» préservant la positivité pour les écoulements en eaux peu profondes 14
  - 1.3.4 Analyse des méthodes des volumes finis non structurées pour les écoulements en eaux peu profondes en utilisant le pseudo spectre 15
  - 1.3.5 Une méthode des volumes finis respectant la condition de compatibilité géométrique pour les lois de conservation sur des surfaces courbes 16
  - 1.3.6 Un schéma équilibre partiellement centré pour les lois de conservation sur des surfaces courbes 17

CHAPITRE 2 Une méthode des volumes finis non structurée pour les écoulements

à grande échelle utilisant le schéma d'Adams du quatrième ordre 19
2.1 Introduction 20
2.2 Shallow water equations 22
2.2.1 Linear SWEs 22
2.2.2 Nonlinear SWEs 23
2.3 Finite volume method 24
2.3.1 Unstructured grid implementation 24
2.3.2 The proposed high-order upwind scheme 25
2.3.3 Temporal integration method 27
2.4 Numerical experiments for symmetric linear equatorial Rossby waves . 31
2.4.1 Symmetric Equatorial Rossby waves of Index 1 31
2.4.2 Performance of the proposed method 32
2.5 Numerical experiments for nonlinear cases 34
2.5.1 Nonlinear Gravity waves 34
2.5.2 Parabolic flood waves 34
2.5.3 Nonlinear Rossby soliton waves 35
2.5.4 Effect of Grid Structure 36
2.6 Conclusions 36

CHAPITRE 3 Analyse théorique et numérique d'une classe de schémas semi-implicites semi-lagrangiens potentiellement applicables aux modèles atmosphériques 48
3.1 Introduction 50
3.2 Problem and notation 51
3.3 Explicit scheme for the nonlinear term 52
3.3.1 The proposed explicit predictor corrector scheme 52
3.3.2 Stability analysis of the proposed scheme 54
3.3.3 Accuracy of oscillatory solutions 55
3.4 The proposed scheme for the linear term 56
3.4.1 Treatment of the linear term 56
3.4.2 Stability analysis and choice of parameter μ 56
3.5 Numerical experiments of the semi-Lagrangian semi-implicit combination 58
3.5.1 The full semi-Lagrangian semi-implicit scheme 58
3.5.2 Linear stability and error analysis of the two-frequency system 58
3.5.3 Accuracy, efficiency, and convergence of the proposed class of schemes 60
3.6 Conclusions 64

CHAPITRE 4 Un schéma équilibre partiellement centré de type «Cell-Vertex» préservant la positivité pour les écoulements en eaux peu profondes 79
4.1 Introduction 80
4.2 The Cell-Vertex Central-Upwind Scheme 82

4.2.1 Cell-Vertex Grid and Notations 82
4.2.2 The Semi-Discrete Form of the Scheme 84
4.2.3 Continuous Piecewise Linear Approximation of the Bottom . . . 85
4.2.4 Piecewise Linear Reconstruction 85
4.3 Positivity Preserving Reconstruction for Water Surface Elevation 87
4.4 Well-Balanced Discretization of the Source Term 89
4.5 Positivity Preserving Property of the Scheme 90
4.6 Numerical Examples 91
4.7 Conclusions 99

CHAPITRE 5 Analyse des méthodes de volumes finis non structurées pour les écoulements en eaux peu profondes en utilisant le pseudo spectre 101
5.1 Introduction 101
5.2 Discretization of Shallow Water Equations 103
5.2.1 Shallow water equations 103
5.2.2 Finite volume schemes 103
5.3 Discrete Operators of the Schemes and Analysis of Stability 106
5.3.1 Temporal discretization 106
5.3.2 Stability analysis techniques 106
5.4 Numerical Tests of Stability 108
5.4.1 Dimensionless form of SWEs 108
5.4.2 Test cases using the Crank−Nicolson method 108
5.5 Numerical tests for Kelvin waves 111
5.6 Conclusions 113

CHAPITRE 6 Une méthode des volumes finis respectant la condition de compatibilité géométrique pour les lois de conservation sur des surfaces courbes . . 115
6.1 Introduction 116
6.2 Geometric Burgers models on the sphere 119
6.2.1 Geometric hyperbolic conservation laws on manifolds 119
6.2.2 The models of interest in this paper 121
6.3 Classes of flux vector fields 122
6.3.1 Foliated flux vector fields 122
6.3.2 Notion of independent domains 124
6.3.3 Genuine nonlinearity and late-time asymptotics 124
6.4 Special classes of solutions 125
6.4.1 Wave structure 125
6.4.2 Spherical coordinates 126
6.4.3 Solutions for linear foliated flux 126
6.4.4 Non-trivial steady state solutions 127
6.5 Geometric finite volume method on the sphere 128

6.5.1 Discrete form of the divergence operator 128
6.5.2 Equations for the splitting approach 130
6.5.3 Second-order approximations based on generalized Riemann problems 132
6.5.4 The proposed piecewise linear reconstruction 134
6.6 Analysis of the spatial and temporal orders of accuracy of the scheme . 135
6.6.1 Spatial order of the scheme 135
6.6.2 Impact of the splitting approach on the temporal order of the scheme 136
6.7 Numerical experiments 137
6.7.1 First test case with linear foliated flux 137
6.7.2 Second test case with nonlinear foliated fluxes 138
6.7.3 Third test case with nonlinear foliated fluxes –an alternative form 142
6.7.4 Fourth test case with fully coupled flux vector fields 145
6.7.5 Fifth test case: revisiting the asymptotic convergence property . 147
6.8 Concluding remarks 149

CHAPITRE 7 Un schéma numérique efficace respectant la condition de compatibilité géométrique pour les lois de conservation sur la sphère 153
7.1 Introduction 154
7.2 Governing equations 156
7.3 Derivation of the proposed scheme 157
7.3.1 Notations 157
7.3.2 Discretization of the divergence operator 158
7.3.3 Reconstruction 159
7.3.4 Evolution and projection 161
7.3.5 Geometry-compatible condition 165
7.4 The proposed scheme using the latitude-longitude grid on the sphere . 166
7.4.1 Computational grid on the sphere 166
7.4.2 Non-oscillatory piecewise linear reconstruction 167
7.5 Geometry-compatible flux vectors and particular solutions of interest . 168
7.6 Numerical experiments 170
7.7 Concluding remarks 174

CHAPITRE 8 Conclusions et recommandations 176

RÉFÉRENCES 180

LISTE DES TABLEAUX

Table 2.1 Temporal rate of convergence of the proposed method at $t = 250$ 40

Table 2.2 L^2 error for various grids 40

Table 3.1 Value of $(\omega\Delta t)_{max}$ of the segment of imaginary axis $[-(\omega\Delta t)_{max}, (\omega\Delta t)_{max}]$ included in the region of absolute stability depending on the values of the approximation parameter α and the decentring parameter θ. 69

Table 3.2 Errors of the amplification factors using the proposed schemes and the method studied by Cullen (2001). 70

Table 6.1 Temporal rate of convergence of the proposed scheme at $T = 5$. 137

LISTE DES FIGURES

Figure 2.1 Parameters at the right (R) and left (L) sides of the interface of computational cells used in the κ scheme 41
Figure 2.2 Cells used in the proposed method 41
Figure 2.3 Water surface elevation for Rossby waves : the analytical form at time $t = 500$ 42
Figure 2.4 Water surface elevation for Rossby waves using the fourth-order Adams method and Quick scheme at time $t = 250$ with $CFL = 0.1$ 42
Figure 2.5 Water surface elevation using the proposed method for Rossby waves at time $t = 500$ with $CFL = 0.1$ 43
Figure 2.6 Comparison of numerical solutions using the proposed method, κ schemes, and the analytical solution of water surface elevation for Rossby waves at time $t = 500$ with $CFL = 0.1$ 43
Figure 2.7 Evolution of L^2 error in log-log scale for Rossby waves at time $t = 250$, where $x_m = \sqrt{\Omega_m}$ 44
Figure 2.8 Change in kinetic energy for Rossby waves using the proposed method and κ schemes until time $t = 500$ with $CFL = 0.1$ . . . 44
Figure 2.9 Comparison of the solution for the gravity waves using the proposed method and the reference solution at time $t = 50$ with $CFL = 0.4$ 45
Figure 2.10 A three-dimensional view of water surface elevation using the proposed method at time $t = 50$ with $CFL = 0.4$ 45
Figure 2.11 Comparison of the solution for a parabolic flood waves using the proposed method and the analytical solution with $CFL = 0.5$. 46
Figure 2.12 Free surface elevation for nonlinear Rossby soliton waves using the proposed method at time $t = 40$ with $CFL = 0.3$ 46
Figure 2.13 Evolution of total energy for nonlinear Rossby soliton waves using the proposed method until time $t = 40$ with $CFL = 0.3$ 47
Figure 2.14 Triangular meshes used in the analysis of grid effects 47
Figure 3.1 Region of absolute stability for six values of the parameter α plotted for fifteen CFL Number values. The central white region corresponds to the absolutely stable region independent of the CFL Number for a decentring parameter $\theta = 0.7$. 69
Figure 3.2 Region of absolute stability of the proposed schemes ($\alpha = 1/4$) for four values of the decentring parameter θ plotted for fifteen CFL number values. The central white region corresponds to the absolutely stable region independent of the CFL number. 70

Figure 3.3 Region of absolute stability of the scheme without corrector step and using the value $N_\alpha^{n+1/2}$ in the step using BDF2 method, for six values of the parameter α plotted for fifteen CFL number values. 71
Figure 3.4 Solutions of oscillatory equation using the proposed schemes for the nonlinear term. (a): Module of the errors of the amplification factors $|A_k - A_{an}|$. (b): Damping of the solution. (c): Relative phase change. (d): Module of the computational mode. 72
Figure 3.5 Treatment of the linear term. (a): Module of the amplification factor, for oscillation equation (3.9) using $\theta = 0.5$, plotted as function of μ and $\omega\Delta t$. (b): Module of the amplification factor for Equation (3.5) using $\mu = 0.5$ and $\theta = 0.5$ plotted on the $\lambda\Delta t$ -$\omega\Delta t$ plane. 73
Figure 3.6 Treatment of the linear term. Module of the amplification factor for Equation (3.5), plotted on the $\lambda\Delta t$-$\omega\Delta t$ plane. (a): Example of stable case using $\mu = 0.5$ and $\theta = 0.55$. (b): Example of unstable case using $\mu = 0.5$ and $\theta = 0.45$. 73
Figure 3.7 The proposed full semi-lagrangian semi-implicit schemes using $\mu = 0.5$ and $\theta = 0.7$. (a): Module of the amplification factors for five CFL number values. (b): Module of the errors of the amplification factors $|A_k - A_{an}|$ for $ks = 2\pi/3$. (c): Module of the errors for $ks = \pi$. (d): Module of the errors for $ks = 4\pi/3$. . 74
Figure 3.8 Module of the errors of the amplification factors $|A_k - A_{an}|$ for the schemes proposed by Clancy and Pudykiewicz (2013). (a): Method AM2-LFT. (b): Method AM2-ABM. (c): Method T-ABT. (d): Method AM2-ABT. 75
Figure 3.9 The scheme studied by Cullen (2001). (a):Module of the errors of the amplification factors $|A_k - A_{an}|$. (b): Relative phase change. 76
Figure 3.10 Module of the computational mode for the proposed schemes ($\theta = 0.7$). 76
Figure 3.11 Module of the errors of the amplification factors $|A_k - A_{an}|$ for the proposed schemes ($\theta = 0.7$). 77
Figure 3.12 Evolution of the errors until time $T = 50$ using the proposed schemes and the method studied by Cullen (2001) for an initial condition which combined two modes ($\nu_R = 1$, $\nu_I = -1$) with $\Delta t = 0.03$. (a): Errors of the parameter D. (b): Errors of the parameter ϕ. 77

Figure 3.13 Evolution of the error of the solution $(D, \phi)^T$ until time $T = 50$ using the proposed method with $\Delta t = 0.030$ and the method studied by Cullen (2001) with $\Delta t = 0.024$ for an initial condition which combined two modes. (a) : case $\nu_R = 1$ and $\nu_I = 1$ and (b): case $\nu_R = 1$ and $\nu_I = 4$ 78

Figure 3.14 (a): The error function for the method studied by Cullen (2001) using the time step Δt and the error function for the proposed method using the time step $1.25\Delta t$. (b): The error function for the proposed method using the time step Δt 78

Figure 4.1 Sample of the cell-vertex. Solid lines represent the primary triangular grids and the dashed lines show the computational polygonal cells. 83

Figure 4.2 Example 1: Solution (w) computed by the well-balanced cell-vertex central-upwind scheme for $\varepsilon = 10^{-2}$ (left) and $\varepsilon = 10^{-3}$ (right). 93

Figure 4.3 Example 1: Top (left) and three-dimensional (3-D) (right) views of the solution (w) computed by the well-balanced cell-vertex central-upwind scheme for $\varepsilon = 10^{-4}$. 94

Figure 4.4 Example 1: Solutions (w) computed by the non well-balanced cell-vertex central-upwind scheme using the fine (left) and coarse (middle) grids and by the well-balanced cell-vertex central-upwind scheme using the coarse grid (right). Here, $\varepsilon = 10^{-4}$. 95

Figure 4.5 Examples 2 and 3: One-dimensional slices of the bottom topographies (4.24), left, and (4.25), right. These plots are not to scale. 96

Figure 4.6 Example 2: Solution (w) computed by the well-balanced cell-vertex central-upwind scheme. 97

Figure 4.7 Example 2: Solution (w) computed by the non well-balanced cell-vertex central-upwind scheme. 97

Figure 4.8 Example 3: Solution (w) computed by the well-balanced cell-vertex central-upwind scheme. The circle in the center of the computational domain represents the part of the bottom that is above the water surface. 98

Figure 4.9 The same as Figure 4.8, but the solution (w) is computed by the non well-balanced cell-vertex central-upwind scheme. 98

Figure 4.10 Example 4: Solution at time $t = 15$ computed by the proposed cell-vertex central-upwind scheme. In the left and middle graphs, a 1-D slice of the solution along the line $y = 0$ is shown. There, w_2 and w_1 are the water surface elevations computed using average cell areas $|M_j| = 0.50$ and $|M_j| = 2.30$, respectively, w_0 is the initial condition. The bottom topography B is plotted at the left. The 3-D view of the computed water surface is on the right. 100

Figure 4.11 The same as Figure 4.10, but at a later time $t = 55$. 100
Figure 5.1 Sample of the unstructured grids used in the analysis. Grid i for method i-CN, where $i = 1, 2, 3, 4, 5$. 104
Figure 5.2 Left: Evolution of the parameter $\|\mathbf{C}^n\|$ using method 1-CN (test a) with periodic boundary conditions in the x- and y-directions. Right: Eigenvalue spectrum and pseudospectra of the same method with the same boundary conditions 109
Figure 5.3 Left : Eigenvalue spectrum and pseudospectra of method 1-CN (test b) using periodic boundary condition in the x-direction and wall boundary condition in the y-direction. Right : Eigenvalue spectrum and pseudospectra of method 2-CN using periodic boundary conditions in the x- and y-directions. 109
Figure 5.4 Left : Eigenvalue spectrum and pseudospectra of method 3-CN (test a) using uniform unstructured grid and periodic boundary conditions. Right : Eigenvalue spectrum and pseudospectra of method 3-CN (test b) using general unstructured grid and periodic boundary conditions. 110
Figure 5.5 Left: Eigenvalue spectrum and pseudospectra of method 4-CN using unstructured grid and periodic boundary conditions. Right : Eigenvalue spectrum and pseudospectra of method 5-CN (test a) using uniform unstructured grid and periodic boundary conditions. 111
Figure 5.6 Change in total energy for Kelvin waves using the schemes i-CN, i=1, 2, 3, 4, and 5 for two time periods with $CFL = 0.5$ 112
Figure 5.7 The isolines of the schemes 3-CN (left) and 1-CN (right) at one time period 113
Figure 5.8 The isolines of the schemes 2-CN (left) and 5-CN (right) at one time period 113
Figure 5.9 Three dimensional view of the water surface elevation using the schemes 3-CN (left) and 4-CN (right) at one time period 113
Figure 6.1 Types of grids used on the sphere 130
Figure 6.2 Evolution of L^1-error in log-log scale until time $T = 5$ ($dx = \log(\Delta\lambda)$) for the nonlinear foliated flux defined on the basis of the scalar potential $h(x, u) = (x.a)f(u)$. Left: For the nonlinear foliated flux defined by using $a = i_1$. Right: For the nonlinear foliated flux defined by using $a = i_1 + i_2 + i_3$. 136
Figure 6.3 Solutions of Test 1-a (left) and Test 1-b (right) at time $t = 50$ with $\Delta t = 0.01$, $\Delta\lambda = \pi/96$, and $\Delta\phi = \pi/96$. 138

Figure 6.4 Solution of Test 2-a on the entire sphere (left) at time $t = 100$. Evolution of L^1-error of the solution until time $t = 30$ for Test 2-b (right) using the proposed scheme and Central-upwind scheme with $\gamma = 0.1$. The grid with $\Delta\lambda = \pi/96$ and $\Delta\phi = \pi/96$ is used for the two tests cases. 139

Figure 6.5 Solutions of Test 2.b at time $t = 5$ for the cases $\gamma = 0.1$ (left) and $\gamma = 0.3$ (right) using the proposed method with $\Delta t = 0.03$, $\Delta\lambda = \pi/96$, and $\Delta\phi = \pi/96$. 140

Figure 6.6 Convergence for Test 2-b using $\gamma = 1$. Left: Evolution of the parameters $E_1(u) = \|u(t) - \overline{u}_I\|_{L^1}$ in domain 1 and $E_2(u) = \|u(t) - \overline{u}_{II}\|_{L^1}$ in domain 2 with $\Delta t = 0.05$, $\Delta\lambda = \pi/96$ and $\Delta\phi = \pi/96$. Right: Two-dimensional view of the solution at time $t = 50000$. 141

Figure 6.7 Time-variation diminishing property (6.4) for the first order of the scheme (left) and for the second order of the scheme (right) until time $t = 50$ with $\Delta t = 0.01$, $\Delta\lambda = \pi/96$ and $\Delta\phi = \pi/96$. . 142

Figure 6.8 Contraction property (6.5) with $\Delta t = 0.01$, $\Delta\lambda = \pi/96$ and $\Delta\phi = \pi/96$. Left : for the case of the first and second order schemes. Right: for the case of the first-order scheme using several functions. 143

Figure 6.9 Left: Evolution of L^1-error for Test 3-a until time $t = 30$ with $\Delta t = 0.02$, $\Delta\lambda = \pi/96$ and $\Delta\phi = \pi/96$ for the three cases $k = 0.1$, $k = 0.2$ and $k = 0.3$. Right: Evolution of $\|u_1(t) - \overline{u}_1\|_{L^1}$ for large simulation time with $\Delta t = 0.05$, $\Delta\lambda = \pi/96$ and $\Delta\phi = \pi/96$ 144

Figure 6.10 Stability property for Test 3-a (left) and property (6.4) for the first-order scheme until time $t = 50$ (right) with $k = 1$, $\Delta t = 0.01$, $\Delta\lambda = \pi/96$ and $\Delta\phi = \pi/96$. 145

Figure 6.11 Property (6.4) for the second-order scheme until time $t = 50$ (left) and contraction property (6.5) for the first-order scheme (right) with $\Delta t = 0.01$, $\Delta\lambda = \pi/96$ and $\Delta\phi = \pi/96$. 145

Figure 6.12 Convergence and stability for Test 4 with $\Delta\lambda = \pi/96$ and $\Delta\phi = \pi/96$. Left: Evolution of the parameter $E(u) = \|u_2(t) - \overline{u}_2\|_{L^1}$ for generic flux with different values of the parameter s with $\Delta t = 0.05$. Right: Entropy stability property (6.3) for the generic flux with $\Delta t = 0.01$. 146

Figure 6.13 Time-variation diminishing property (6.4) (left) and contraction property (6.5) for the first-order scheme (right) for the generic flux with $\Delta t = 0.01$, $\Delta\lambda = \pi/96$ and $\Delta\phi = \pi/96$. 147

Figure 6.14 Initial condition u_1 (left) and solution u_1 at time $t = 50000$ (right) with $\Delta t = 0.05$, $\Delta\lambda = \pi/96$ and $\Delta\phi = \pi/96$. 148

Figure 6.15 Initial condition u_2 (left) and solution u_2 at time $t = 50000$ (right) with $\Delta t = 0.05$, $\Delta\lambda = \pi/96$ and $\Delta\phi = \pi/96$. 149
Figure 7.1 Schematic view of the decomposition of the control volume . . . 160
Figure 7.2 Type of grid used on the sphere 167
Figure 7.3 (a)-(b)-(c): Types of grids used on the sphere. (d): The domain $D_{jk} = D_{jk}^{+} \cup D_{jk}^{-}$. 169
Figure 7.4 Solutions on the entire sphere at time $t = 5$ for Test 1 (left) and Test 2 (right) 172
Figure 7.5 Solutions on the entire sphere at time $t = 5$ for Test 3 (left) and Test 4 (right) 172
Figure 7.6 Solutions on the entire sphere at time $t = 5$ for Test 5 (left) and Test 6 (right) 173

CHAPITRE 1 Introduction

1.1 Objectifs

Le projet de recherche de la thèse concerne le développement et l'évaluation de nouvelles méthodes numériques pour les écoulements peu profonds. Dans cette section, on présente les objectifs et on introduit le cadre général des différentes thématiques qui seront détaillées dans les chapitres qui vont suivre.

De nouvelles techniques de discrétisations spatiales et temporelles des équations sont proposées. L'objectif est de développer des schémas numériques pour les écoulements peu profonds qui sont stables, non dispersifs et non diffusifs. Ces méthodes doivent présenter un très bon compromis entre la précision et le coût de calcul et doivent être faciles à implémenter dans le cas des maillages non structurés et pour le cas des géométries courbes.

Dans notre démarche, les quatre parties suivantes sont traitées dans la thèse:

- Développement de méthodes numériques efficaces pour la modélisation des processus lents dans les écoulements océaniques et atmosphériques,
- Construction de schémas numériques efficaces sans solveurs de Riemann pour les écoulements peu profonds,
- Introduction d'une nouvelle approche pour l'analyse de stabilité des méthodes numériques pour les écoulements en eaux peu profondes,
- Développement de schémas numériques pour les lois de conservation sur des surfaces courbes, qui ont un large potentiel d'être appliqués pour les écoulements peu profonds sur la sphère.

Dans la première partie de la thèse, on s'intéresse à la modélisation des processus lents dans les écoulements océaniques et atmosphériques. Ces écoulements sont forcés à de très grandes échelles spatiales et temporelles, en particulier par les effets qui dépendent de la rotation terrestre. La variation du paramètre de Coriolis avec la latitude génère les ondes de Rossby, qui sont aussi appelées les ondes planétaires. Ces ondes ont des fréquences très faibles et elles se propagent lentement.

Dans les écoulements océaniques et atmosphériques, il est nécessaire de calculer avec une bonne précision à la fois les modes lents et les modes rapides. Dans un premier temps, un schéma de volume finis décentré est proposé pour le cas des équations de Saint-Venant avec le terme source qui comprend le paramètre de Coriolis. Le système de Saint-Venant, en dépit de sa simplicité, comprend tous les aspects de la dynamique

de l'atmosphère et de l'océan à grande échelle. Les équations de Saint-Venant permettent la modélisation des écoulements de fluides en milieux peu profonds. Le modèle basé sur ces équations est validé expérimentalement et il est largement utilisé pour simuler de nombreux phénomènes naturels dans l'atmosphère, les océans, les rivières et les estuaires. Les équations de Saint-Venant sont dérivées à partir des équations de Navier-Stokes (Vreugdenhil, 1994) après intégration suivant la verticale en adoptant l'hypothèse de pression hydrostatique et en négligeant l'accélération verticale et l'effet de la viscosité du fluide.

L'objectif est de proposer une méthode des volumes finis explicite sur un maillage non structuré pour les écoulements à grande échelle. Cette méthode doit être précise pour ces types d'écoulements en présence des ondes lentes (ondes de Rossby) et des ondes rapides (ondes de gravité). La méthode doit être capable de supprimer le bruit numérique des ondes courtes sans amortissement des ondes longues qui ont une importance significative dans la dynamique de l'océan et de l'atmosphère.

Le schéma numérique développé est explicite. Néanmoins, en présence des termes sources dans le système qui génèrent des ondes de grandes vitesses de propagation, il est nécessaire d'utiliser des méthodes implicites pour l'intégration temporelle de ces termes. Aussi, dans le cas où le système comprend une large gamme de fréquences des ondes, ce qui est le cas des écoulements atmosphériques, il est important d'utiliser des méthodes semi-implicites afin d'opter pour un pas de temps optimal.

La plupart des centres météorologiques utilisent des méthodes semi-implicites semi-lagrangiennes dans leurs modèles de prévision numérique atmosphérique. Environnement Canada utilise la méthode semi-implicite semi-lagrangienne nommée SETTLS, établie par Hortal (2002), comme schéma numérique temporel dans ses modèles atmosphériques. Néanmoins, cette méthode nécessite parfois des itérations supplémentaires pour éviter les oscillations numériques. Cette méthode peut générer des oscillations pour le traitement du terme non linéaire pour le cas des solutions qui ont un caractère quasi-oscillatoire. Des travaux de recherche sont effectués dans la thèse pour remédier à ce problème. Ces travaux sont menés en collaboration avec l'équipe de recherche en prévision numérique atmosphérique d'Environnement Canada. L'objectif de cette partie du projet de recherche est de remédier aux problèmes d'instabilité associés au traitement de la partie non linéaire du terme source. La méthode à proposer doit améliorer le modèle météorologique d'Environnement Canada de point de vue stabilité sans avoir d'impact sur la précision des résultats et avec un impact négligeable en temps de calcul.

Dans le cadre des travaux de recherche de la thèse, un autre schéma numérique efficace et sans solveur de Riemann est développé pour le cas des équations de Saint-Venant avec une topographie variable. Les applications visées concernent la simulation de différents types d'écoulement environnementaux. À titre d'exemple, ce système peut être

appliqué pour étudier les écoulements à surface libre, les aménagements hydrauliques et le dimensionnement des ouvrages hydrauliques, l'étude des crues et la conception des ouvrages de protection contre les inondations, et la prévision par modélisation numérique de la zone de risque suite à une rupture de barrage. Ce système peut être éventuellement couplé avec les équations de transport de sédiments et ce pour étudier la dynamique sédimentaire dans les rivières, les estuaires et les zones côtières. L'objectif est de développer des méthodes des volumes finis du type Godunov qui ont pour avantage principal d'éviter la résolution du problème de Riemann aux interfaces des volumes de contrôle.

Les solutions du système des équations de Saint-Venant peuvent développer des discontinuités en temps finis. Les méthodes des volumes finis centrées ne sont pas recommandées pour ces types de solutions et elles sont généralement instables. Les méthodes numériques couramment utilisées sont décentrées. Ces méthodes utilisent une résolution exacte ou des approximations du problème de Riemann au niveau des interfaces des cellules de calcul. Ceci rend leur résolution numérique plus chère. Le schéma de volumes finis à développer est partiellement centré. Dans la formulation de ce schéma, aucune résolution du problème de Riemann n'est effectuée. Le schéma doit assurer de bons résultats dans le cas des solutions discontinues. L'équilibre entre les termes convectifs et le terme de topographie doit être préservé pour les solutions stationnaires. Le schéma numérique doit avoir de bonnes qualités liées à la stabilité et la convergence et en particulier ce schéma doit avoir de bonnes habilités de résolution des faibles perturbations des états d'équilibre. En plus, la méthode des volumes finis à proposer doit assurer la positivité de la hauteur d'eau au cours du temps.

Dans le projet de recherche, on propose une nouvelle approche pour l'analyse de stabilité des méthodes des volumes finis appliquées aux écoulements peu profonds. Cette approche utilise le pseudo spectre de l'opérateur discret du schéma numérique. On montre l'utilité de la méthode proposée par rapport à la stabilité asymptotique et à la stabilité au sens de Lax-Richtmyer.

Dans la thèse, on s'intéresse aussi au développement des schémas numériques dans le cas des surfaces courbes. L'objectif final est de proposer des méthodes des volumes finis qui sont potentiellement applicables aux écoulements peu profonds sur la sphère, qui décrivent des écoulements sur la surface de la terre. En plus des difficultés rencontrées par les schémas numériques dans le cas des systèmes d'écoulement en coordonnées cartésiennes, une attention particulière doit être accordée à l'impact de la géométrie dans le cas des surfaces courbes. Les dimensions géométriques ne doivent pas influencer la condition de compatibilité géométrique du schéma numérique sous sa forme semi-discrète. En d'autres termes, l'équilibre entre les termes convectifs et les termes résultants de la variation de la géométrie (et éventuellement le terme de topographie s'il est considéré dans le terme source) doit être préservé pour les solutions stationnaires. Dans cette partie du projet de recherche, nous adoptons la démarche suivie par Stanley

Osher pour l'établissement des schémas numériques robustes. Cette approche consiste à considérer des systèmes hyperboliques simples mais qui peuvent générer des solutions qui ont différentes structures d'ondes et qui constituent un moyen efficace pour tester les méthodes numériques à développer. Dans notre cas, nous considérons les équations de Burgers sur la sphère qui sont largement utilisées pour le développement des schémas numériques à capture de chocs en mécanique des fluides.

Deux schémas de volumes finis sont développés pour les systèmes hyperboliques de lois de conservation sur la sphère. Le premier schéma est plus approprié pour ces systèmes en présence de solutions discontinues de grands chocs et amplitudes. Une résolution du problème de Riemann généralisé est considérée dans la formulation de la méthode développée. Les dimensions géométriques de la surface courbe sont considérées de manière analytique pour aboutir à une forme semi-discrète qui satisfait d'une manière exacte la condition de compatibilité géométrique. Dans la formulation de la méthode proposée, une nouvelle reconstruction est conçue afin d'assurer la stabilité et de bonnes précisions pour les solutions discontinues de grands chocs et amplitudes. Des analyses approfondies sont effectuées pour comprendre l'évolution des solutions et leurs comportements asymptotiques qu'on ne peut pas connaitre analytiquement. Ces analyses sont effectuées en fonction des classes de flux et de leur caractère de linéarité.

Le deuxième schéma numérique proposé est simple puisqu'il n'utilise aucun solveur de Riemann. Ce schéma est plus adapté aux systèmes hyperboliques de lois de conservation sur la sphère dans le cas des solutions d'amplitudes et chocs moyens. Le schéma est non diffusif et sa forme semi-discrète respecte la condition de compatibilité géométrique.

Les deux schémas de volumes finis développés pour les lois de conservation scalaires hyperboliques non linéaires sur la sphère pourraient être étendus aux systèmes hyperboliques multidimensionnels et aux modèles d'écoulements peu profonds sur la sphère.

1.2 Revue de littérature

L'objet de cette section est de présenter une synthèse bibliographique sur l'ensemble des parties développées dans la thèse. La revue de littérature est décrite en quatre parties :

- Méthodes des volumes finis décentrées pour les écoulements peu profonds,
- Schémas semi-implicites semi-lagrangiens et leurs applications aux écoulements atmosphériques,
- Méthodes des volumes finis partiellement centrées pour les écoulements peu profonds,

- Méthodes des volumes finis pour les lois de conservation sur des surfaces courbes.

Un état des connaissances pour chaque partie du projet sera présenté. On trouve également dans chaque partie, une formulation claire et précise des problématiques à résoudre et les objectifs à atteindre dans le cadre du projet de recherche de la thèse.

1.2.1 Méthodes des volumes finis décentrées pour les écoulements peu profonds

La méthode des volumes finis (VF) est largement utilisée pour la modélisation numérique des écoulements peu profonds en raison de son avantage de conservation de masse et de quantité de mouvement. Les schémas VF décentrés sont devenus très populaires pour la résolution numérique des équations de Saint-Venant dans le cas des solutions discontinues. Ces schémas assurent un niveau faible de diffusion numérique en utilisant un nombre acceptable de cellules de maillage.

Les schémas VF décentrés nécessitent la résolution du problème de Riemann pour le calcul du flux au niveau de chaque interface des cellules de calcul. Plusieurs méthodes de résolution du problème de Riemann exacte ou approchée ont été développées. L'algorithme le plus populaire est la méthode de Godunov (Godunov, 1959; Godunov et al., 1961). Ce schéma numérique est du premier ordre en précision. Les données initiales dans les cellules sont représentées par des constantes par morceaux avec des discontinuités au niveau des interfaces de cellules. La solution exacte du problème de Riemann au niveau de chaque interface des cellules est utilisée pour le calcul du flux à cette interface. L'extension de la méthode de Godunov au deuxième ordre ou à un ordre élevé est possible en utilisant d'autres formes de reconstruction des variables primitives du système au niveau des cellules de calcul (approximations linéaire, parabolique, etc.). La résolution exacte du problème de Riemann coûte très cher numériquement, ce qui a conduit au développement de méthodes d'approximation qui sont relativement moins coûteuses en termes de temps de calcul. La méthode de Roe (1981) est l'une des méthodes d'approximation qui sont largement appliquées en dynamique des fluides. Cette méthode est basée sur la linéarisation du système d'équation et elle donne de bonnes approximations des solutions discontinues. En présence du terme source, les méthodes VF peuvent engendrer des oscillations numériques causées par le déséquilibre entre le terme de flux et le terme source. Plusieurs techniques ont été développées pour assurer l'équilibre entre ces termes (exemple : Vázquez-Cendón, 1999; Gallouet et al., 2003; Mohammadian et al., 2005; Mohammadian et Le-Roux, 2006; Stewart et al., 2011). Parmi les termes sources considérés, il y a ceux qui sont dûs à la topographie variable, au frottement et à l'effet de Coriolis. Le terme de Coriolis est considéré surtout dans le cas des écoulements à grande échelle. Plusieurs travaux ont été effectués pour développer ou évaluer les schémas numériques appliqués à ces types d'écoulement (exemple: Walters et Carey, 1983; Foreman, 1984; Hanert et al., 2004, 2005; Le Roux et Pouliot,

2008; Hanert et al., 2009; Walters et al., 2009; Le Roux et al., 2011). Cependant, peu de travaux ont été réalisés pour le cas des méthodes VF appliquées à ces types d'écoulement (exemple : Lin et al., 2003; Mohammadian et Le Roux, 2008; Castro et al., 2008; Beljadid et al., 2012).

Les problèmes liés à la stabilité des méthodes numériques sont encore amplifiés dans le cas de discrétisation des équations en eaux peu profondes sur des maillages non structurés. Ceci est principalement dû aux modes numériques qui sont générés par les effets liés à la structure du maillage. On obtient des matrices non normales pour les opérateurs linéaires des schémas numériques, ce qui représente une source d'amplification des solutions en temps finis et de problèmes de stabilité. Néanmoins, les maillages non structurés sont nécessaires dans plusieurs contextes et ils offrent une bonne flexibilité pour la discrétisation des domaines complexes et des conditions aux frontières en utilisant différentes tailles des cellules de calcul. Les maillages non structurés nécessitent l'utilisation d'une méthode de discrétisation des équations qui assure une bonne précision avec moins de sensibilité à la structure du maillage. La méthode doit être stable et les amplifications numériques doivent être faibles pour une large gamme de modes.

Les méthodes qui utilisent des directions alternées (ou l'approche du «splitting» directionnel) sont largement utilisées dans le cas des maillages structurés. Dans le cas des maillages non structurés, cette approche peut souffrir des effets d'orientation de la grille et des difficultés rencontrées pour préserver le caractère multidimensionnel du système lors de sa discrétisation.

Les schémas d'intégration temporelle jouent également un rôle important dans la performance globale d'une méthode numérique. Même pour des conditions initiales lisses, les solutions obtenues peuvent développer des discontinuités en temps finis. Les méthodes d'intégration temporelle pour ces cas peuvent engendrer des oscillations numériques. Des méthodes d'intégration temporelle TVD ont été développées et utilisées avec succès dans de nombreux problèmes en raison de leur capacité à éviter la naissance des oscillations, à assurer la stabilité, et à améliorer la précision des résultats. Une classe de schémas numériques d'ordres élevés a été développée par Shu et Osher (1988). Cette classe de méthodes a été étudiée de façon approfondie par Shu (2002) et, Spiteri et Ruuth (2002).

La plupart des méthodes numériques décentrées, avec un bon choix du schéma temporel, aboutissent à de bons résultats pour les ondes de gravité mais ces méthodes échouent à calculer les ondes lentes (ondes de Rossby). Les schémas centrés qui sont efficaces pour le calcul des ondes de Rossby rencontrent plusieurs difficultés (ils sont généralement instables) pour simuler les ondes de gravité.

Le développement de méthodes numériques efficaces à la fois pour le cas des ondes rapides (ondes de gravité) et des ondes lentes (ondes de Rossby) nécessite de nouvelles techniques pour assurer l'équilibre entre le terme de flux et le terme de Coriolis. L'étude

de ce problème fait partie du projet de thèse.

1.2.2 Schémas semi-implicites semi-lagrangiens et leurs applications aux écoulements atmosphériques

La méthode semi-implicite semi-lagrangienne (SL) a été proposée pour la première fois par Robert (1981) pour l'intégration temporelle des modèles numériques atmosphériques. Un état de l'art sur l'application des méthodes SL en prévision numérique atmosphérique est donné par Staniforth et Côté (1991) et d'autres études ont été menées pour examiner les méthodes SL (exemple: Bonaventura, 2000; White–III et Dongarra, 2011). En 1991, la méthode SL avec trois niveaux de calcul en temps a été mise en place et utilisée par le centre européen de prévision météorologique à moyen terme. Cette méthode est décrite en détail par Ritchie et al. (1995). Tanguay et al. (1990) ont généralisé l'utilisation de la méthode semi-implicite pour intégrer les équations pour un fluide compressible dans le cas non hydrostatique. Le développement des méthodes SL à deux étapes par McDonald et Bates (1987) et Temperton et Staniforth (1987) a motivé l'utilisation de ce type de schéma SL vu son avantage concernant le temps de calcul et la réduction du nombre et de l'impact des modes d'origine non-physique (McDonald et Haugen, 1992; Temperton et al., 2001; Hortal, 2002). En effet, l'utilisation de plus de deux niveaux de temps de calcul conduit en général à des modes numériques ayant des amplitudes comparables à celles des modes physiques. Ceci peut influencer la précision de la solution numérique si aucune technique efficace n'est utilisée pour amortir les modes numériques. Dans ce sens, Hortal (2002) a proposé une méthode SL à deux niveaux (SETTLS) qui présente une zone de stabilité absolue indépendante de la condition de stabilité de Courant-Friedrichs-Lewy.

D'après Hortal (2002), dans la formulation de la méthode SETTLS, l'équation de la trajectoire est obtenue en adoptant une approximation explicite de la moyenne de l'accélération entre le point de départ et le point d'arrivée. Ce schéma peut être considéré comme un cas particulier, et il démontre le plus de stabilité, de la famille des schémas numériques de second ordre proposée par Gospodinov et al. (2001). Ces schémas sont paramétrés par un nombre indéterminé α. Durran et Reinecke (2004) ont montré que la taille de la région de stabilité absolue de la famille des schémas proposée par Gospodinov et al. (2001) varie considérablement en fonction de ce paramètre et que la région optimale est obtenue pour la valeur $\alpha = 1/4$, ce qui correspond à la méthode SETTLS. Pourtant, SETTLS n'est pas un choix idéal vu que les points sur l'axe imaginaire du plan complexe sont en dehors du domaine de stabilité absolue, sauf le point d'origine. Par conséquent, la méthode SETTLS peut générer des oscillations pour le traitement du terme non linéaire surtout pour le cas des solutions qui ont un caractère quasi-oscillatoire. Pour remédier à ce problème, une classe de schémas semi-implicites semi-lagrangiens est développée dans ce projet de recherche.

1.2.3 Méthodes des volumes finis partiellement centrées pour les écoulements peu profonds

Les solutions du système de Saint-Venant peuvent développer des discontinuités en temps finis. Les méthodes couramment utilisées pour la résolution de ce système sont les schémas de volumes finis décentrés. Tel qu'indiqué auparavant, ce type de schéma nécessite la résolution du problème de Riemann au niveau des interfaces des cellules de calcul. Un aperçu de l'état de l'art concernant ces schémas est donné dans la section 1.2.1. Sommairement, la différence principale entre les schémas décentrés et les schémas centrés est que les schémas décentrés utilisent de manière intensive les informations sur les caractéristiques de propagation des ondes, tandis que les schémas centrés sont principalement basés sur les moyennes de flux sans utiliser ces informations. Les schémas centrés ont attiré beaucoup d'attention après le travail important de Nessyahu et Tadmor (1990), où un schéma centré à capture de chocs est proposé. Un état de l'art à propos des schémas centrés et leurs extensions est donné par Russo (2002). Les schémas centrés peuvent être améliorés de façon considérable en utilisant quelques informations sur les vitesses de propagation des ondes. Ceci a conduit au développement d'une classe de schémas partiellement centrés qui sont proposés par Kurganov et co-auteurs (Kurganov et al., 2001; Kurganov et Petrova, 2001; Kurganov et Tadmor, 2000a). Ces schémas sont simples vu qu'ils ne font pas appel à la résolution du problème de Riemann au niveau des interfaces des cellules de calcul. Ils présentent une bonne résolution par rapport aux schémas centrés pour les solutions discontinues. Ces schémas sont appliqués avec succès à plusieurs problèmes d'écoulements peu profonds et pour les systèmes hyperboliques (Bollermann, 2013; Kurganov et Levy, 2002; Kurganov et Petrova, 2005, 2007).

Bryson et al. (2011) ont proposé un schéma partiellement centré pour les équations de Saint-Venant avec une topographie variable en utilisant un maillage triangulaire. Le développement et l'extension des schémas partiellement centrés pour plusieurs autres contextes (type de maillage, autres termes sources,...) sont nécessaires pour élargir leurs applications. Deux propriétés sont nécessaires pour le développement de ces schémas. La première concerne l'équilibre entre le terme source et le terme de flux, en particulier le schéma doit préserver les solutions statiques. La deuxième propriété est celle relative à la positivité : la méthode doit garantir des valeurs positives de la hauteur d'eau au cours du temps. Dans le projet de thèse, une méthode des volumes finis partiellement centrée est développée pour le système de Saint-Venant avec une topographie variable. Un maillage non structuré est utilisé sous forme de polygones construits à partir d'un maillage triangulaire initial.

1.2.4 Méthodes des volumes finis pour les lois de conservation sur des surfaces courbes

L'étude des systèmes hyperboliques sur des surfaces courbes a connu un intérêt accru vu les nombreuses applications dans la dynamique des fluides, en particulier pour le cas des lois de conservation scalaires comme les équations de Burgers sur la sphère. Malgré l'apparence simple de ce modèle, il génère des phénomènes d'ondes complexes qui ne sont pas observés en l'absence de l'effet de la géométrie. Les équations de Burgers ont une grande utilité dans le développement des schémas numériques à capture de chocs en mécanique des fluides. Ces équations devraient permettre de fournir un moyen efficace pour construire des schémas numériques pour le cas des systèmes hyperboliques et le système de Saint-Venant sur des surfaces courbes. Les propriétés mathématiques des solutions aux lois de conservation sur des variétés sont largement étudiées par LeFloch et co-auteurs (exemple : Amorim et al., 2005, 2008; Ben-Artzi et LeFloch, 2007; Ben-Artzi et al., 2009; LeFloch, 2011; LeFloch et Okutmustur, 2008). Les lois de conservation sur des surfaces en évolution sont aussi étudiées par Dziuk, Kroöner et Müller, et Giesselman (2009). Ben-Artzi, Falcovitz, et LeFloch (2009) ont proposé un schéma de volumes finis pour les lois de conservation sur la sphère avec une approximation du second ordre basée sur la résolution du problème de Riemann généralisé. Dans cette méthode, un «splitting» directionnel en latitude et longitude sur la sphère est utilisé pour simplifier la résolution du problème de Riemann. La forme discrète de la méthode des volumes finis satisfait la condition de « divergence nulle ». Cette méthode permet de préserver les solutions stationnaires non triviales avec une bonne précision. Une extension de ce schéma pour un ordre temporel élevé nécessite l'analyse de l'impact du «splitting» directionnel adopté sur l'ordre du schéma temporal à utiliser. D'autres aspects liés à la classification des flux et à l'analyse de l'évolution des solutions sont nécessaires pour étudier l'impact de la nature du maillage sur la précision du schéma et pour prédire la convergence asymptotique des solutions. Ces éléments sont traités dans la thèse et un nouveau schéma de volumes finis est proposé.

On note que les schémas partiellement centrés démontrent une bonne résolution des solutions discontinues sur une géométrie plane pour le cas des systèmes hyperboliques et le système de Saint-Venant. Aucune extension de ces schémas n'a été faite pour le cas des surfaces courbes. Dans ce projet de recherche, un nouveau schéma partiellement centré est proposé pour les systèmes hyperboliques sur la sphère.

1.3 Méthodologie et présentation du travail de recherche

La méthodologie suivie dans ce projet de recherche peut être scindée en trois phases principales. Une première phase porte sur la définition précise du problème et des objectifs. La deuxième phase concerne les développements théoriques et l'établissement de nouvelles techniques pour la formulation de nouveaux schémas numériques. La

dernière phase est consacrée aux études analytiques et numériques pour évaluer les performances des méthodes proposées et leurs comparaisons avec celles des méthodes existantes. Dans ce qui suit, pour les six contributions scientifiques, on présente le modèle des équations qui a fait l'objet d'étude, la problématique à résoudre et l'objectif, et un sommaire des résultats obtenus.

1.3.1 Une méthode des volumes finis non structurée pour les écoulements à grande échelle

La force de Coriolis a un effet important sur la dynamique des écoulements géophysiques. Dans les travaux décrits lors du deuxième chapitre, on s'intéresse à la modélisation des processus lents et rapides dans l'océan et l'atmosphère. Un calcul précis des deux processus nécessite des méthodes numériques efficaces. Les équations de Saint-Venant sont considérées avec un terme source qui comprend le paramètre de Coriolis. L'application des schémas de volumes finis pour ce système pose problème à cause du déséquilibre rencontré entre le terme de flux et le terme source. Les méthodes actuellement disponibles sont en général incapables de garantir une bonne précision à la fois pour les ondes lentes (ondes de Rossby) et les ondes rapides (ondes de gravité). La plupart des schémas numériques décentrés donnent de bons résultats pour les ondes rapides mais ils échouent dans la résolution des ondes lentes. Les schémas centrés efficaces pour les ondes de Rossby rencontrent des problèmes pour le cas des ondes de gravité. Les schémas de volumes finis décentrés en combinaison avec la méthode TVD Runge-Kutta d'ordre 3 pour l'intégration temporelle donnent des résultats précis pour le cas des ondes de Kelvin et les ondes de gravité. Les résultats obtenus pour le cas des ondes de Rossby en utilisant ces schémas décentrés ne sont pas satisfaisants.

Les équations des écoulements peu profonds considérées sont décrites par :

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial \mathbf{E}}{\partial x} + \frac{\partial \mathbf{G}}{\partial y} = \mathbf{S}, \tag{1.1}$$

Les équations de type linéaire et non linéaire sont définies en fonction des termes du vecteur des variables primitives $\mathbf{U}$ et des termes des flux $\mathbf{E}$ et $\mathbf{G}$.

Pour les équations linéaires ces termes sont définis par:

$$\mathbf{U} = \begin{bmatrix} \eta \\ u \\ v \end{bmatrix} \quad , \quad \mathbf{E} = \begin{bmatrix} Hu \\ g\eta \\ 0 \end{bmatrix} \quad , \quad \mathbf{G} = \begin{bmatrix} Hv \\ 0 \\ g\eta \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} 0 \\ fv \\ -fu \end{bmatrix} \quad , \tag{1.2}$$

où η est l'élévation de la surface au-dessus de l'élévation moyenne H, u et v sont les composantes moyennes de la vitesse d'écoulement respectivement selon les directions x et y. Le terme de Coriolis est noté par f et l'accélération dûe à la gravité est notée par g.

Pour les équations non linéaires, les vecteurs **U**, **E**, et **G** sont donnés par :

$$\mathbf{U} = \begin{bmatrix} h \\ hu \\ hv \end{bmatrix} \quad , \quad \mathbf{E} = \begin{bmatrix} hu \\ hu^2 + 0.5gh^2 \\ huv \end{bmatrix} \quad , \quad \mathbf{G} = \begin{bmatrix} hv \\ huv \\ hv^2 + 0.5gh^2 \end{bmatrix} \tag{1.3}$$

Le terme source **S**, qui inclut le paramètre de Coriolis, est donné par

$$\mathbf{S} \ = (0, fhv, -fhu)^T, \tag{1.4}$$

où h est la hauteur totale d'eau.

Dans le cas d'une approximation $\beta-$plan, le terme de Coriolis peut s'exprimer sous la forme $f = f_0 + \beta y$ où on considère $f_0 = 0$ étant donné qu'on s'intéresse au cas des ondes équatoriales. Le paramètre β est donné par:

$$\beta = 2\mathbf{\Omega}/\mathbf{R} = 2.29 \times 10^{-11}\, m^{-1}s^{-1}, \tag{1.5}$$

où $\mathbf{\Omega}$ et $\mathbf{R}$ sont respectivement la vitesse angulaire et la valeur moyenne du rayon de la terre ($\mathbf{\Omega} = 7.29 \times 10^{-5}\, rad\, s^{-1}, \mathbf{R} = 6371\, km$).

Une nouvelle méthode des volumes finis décentrée sur un maillage non structuré est proposée pour les écoulements à grande échelle. Cette méthode utilise la méthode d'Adams du quatrième ordre combinée avec le «splitting» directionnel pour l'intégration temporelle. Le terme de Coriolis est intégré analytiquement avant et après l'intégration numérique du terme de flux. Une nouvelle reconstruction est proposée et elle consiste à utiliser une approximation de troisième ordre orthogonalement à l'interface des cellules de calcul. Le long de l'axe de l'interface de chaque cellule, on utilise une approximation du premier ordre. Les analyses montrent que le «splitting» directionnel et la structure du maillage ont un léger impact sur l'ordre de la méthode d'Adams utilisée pour l'intégration temporelle. Les techniques proposées sont suffisantes pour supprimer les oscillations numériques liées aux ondes courtes sans amortissement des ondes longues. Les analyses confirment que dans la nouvelle reconstruction, le premier ordre utilisé dans la direction de chaque interface des cellules de calcul a moins d'impact sur l'ordre spatial du schéma proposé. L'équilibre entre le terme de flux et le terme de Coriolis est préservé par la méthode proposée. Cette méthode donne de bons résultats à la fois pour le cas des ondes de gravité et des ondes de Rossby.

1.3.2 Une classe de schémas semi-implicites semi-lagrangiens potentiellement applicables aux modèles atmosphériques

Les modèles généralement appliqués en prévision numérique atmosphérique s'écrivent sous la forme semi-lagrangienne suivante :

$$\frac{d\psi}{dt} = N + L, \tag{1.6}$$

où le terme source à droite de l'équation (1.6) est décomposé en un terme linéaire noté L qui est responsable des ondes rapides et le reste correspond au terme non linéaire noté par N qui génère les ondes lentes. Dans l'équation (1.6), les deux termes sources linéaire et non linéaire dépendent de la fonction $\psi(x,t)$, de la position x et du temps.

En général, pour résoudre l'équation (1.6) des méthodes explicites sont utilisées pour l'intégration temporelle du terme non linéaire N, et le terme linéaire L est intégré de manière implicite. Dans ce qui suit, le pas du temps est noté par Δt et $t_p = p\Delta t$ désigne le temps à l'étape p. Les notations "+", "0" and "−" sont utilisées pour désigner respectivement les variables aux étapes $t+\Delta t$, t et $t-\Delta t$. Les notations A et D sont utilisées respectivement pour le point de départ $x_D = x(t)$ et le point d'arrivée $x_A = x(t+\Delta t)$ de la trajectoire semi-lagrangienne. Pour tout point x_j de la grille de discrétisation, on désigne par $\tilde{x}_j^n$ l'estimation de la position du point de départ de la parcelle du fluide au temps t_n qui arrive à la position d'arrivée (x_j, t_{n+1}). À l'étape p, la valeur du terme non linéaire est notée par $N^p = N(\tilde{x}_j^p, t_p)$ et la valeur du terme linéaire est notée par $L^p = L(\tilde{x}_j^p, t_p)$ au point de départ $(\tilde{x}_j^p, t_p)$ de la trajectoire qui arrive à la position (x_j, t_{n+1}).

Dans notre méthodologie pour construire la nouvelle famille de schémas semi-implicites semi-lagrangiens, on considère les approximations de second ordre suivantes établies par Gospodinov et al. (2001) pour l'équation (1.6)

$$\frac{\psi_A^+ - \psi_D^0}{\Delta t} = (\frac{3}{4}-\alpha)N_A^0 + (\frac{3}{4}+\alpha)N_D^0 - (\frac{1}{4}-\alpha)N_A^- - (\frac{1}{4}+\alpha)N_D^- + \frac{1}{2}(L_A^+ + L_D^-), \tag{1.7}$$

où α est un paramètre arbitraire.

La méthode nommée SETTLS, établie par Hortal (2002), utilise une approximation du type Gospodinov et al. (2001) donnée par (1.7) pour le cas $\alpha = 1/4$.

La plupart des centres météorologiques utilisent SETTLS comme schéma numérique pour l'intégration temporelle de leur modèle atmosphérique décrit par une équation de type (1.6). Cette méthode peut générer des oscillations pour le cas des solutions qui ont un caractère quasi-oscillatoire. Dans le modèle d'Environnement Canada, la méthode SETTLS nécessite parfois des itérations supplémentaires pour éviter les oscillations numériques.

Une analyse détaillée des conditions de stabilité des schémas semi-implicites semi-lagrangiens basés sur les approximations (1.7) de Gospodinov et al. (2001), est effectuée par Durran et Reinecke (2004). En particulier, les auteurs ont étudié les problèmes de stabilités liés au traitement du terme non linéaire dans le cas de la méthode SETTLS. Dans leur analyse, les auteurs ont remarqué que la zone de stabilité de la méthode SETTLS se réduit à un seul point sur l'axe imaginaire du plan complexe, ce qui constitue un inconvénient pour le cas des solutions à caractère oscillatoire.

Dans le cadre de la thèse, des analyses théoriques et numériques des propriétés de certaines méthodes semi-lagrangiennes complexes sont effectuées pour faire face aux problèmes d'instabilité associés au traitement de la partie non linéaire du terme source. L'objectif est de développer un schéma numérique qui possède une zone de stabilité plus large que celle de la méthode SETTLS, en particulier sur l'axe imaginaire du plan complexe. Les techniques à développer pour améliorer la stabilité ne doivent pas être coûteuses de point de vue temps de calcul et doivent garantir une bonne précision.

La classe des schémas numériques développée utilise une version modifiée de la méthode TR-BDF2 qui est une combinaison de la règle du trapèze (TR) et de la formule de différenciation en arrière d'ordre deux (BDF2). Les analyses menées par Dharmaraja (2007) ont montré que la méthode TR-BDF2 présente de bonnes qualités de stabilité pour la résolution des équations différentielles. Une bonne précision est obtenue pour les systèmes d'équations différentielles raides en utilisant un pas de temps acceptable.

La classe de schémas semi-implicites semi-lagrangiens proposée utilise trois étapes. Le terme non linéaire est traité de manière explicite dans les deux premières étapes comme prédicteur et correcteur en utilisant la méthode du trapèze et dans la troisième étape, on applique la méthode BDF2 explicite. Pour le terme linéaire, la méthode du trapèze implicite est utilisée dans la première étape, la méthode du trapèze explicite est utilisée dans la deuxième étape et la méthode BDF2 est appliquée de manière implicite dans la troisième étape.

L'approximation du type Gospodinov et al. (2001) est considérée dans l'évaluation du prédicteur pour le terme non linéaire, dans laquelle on utilise la valeur intermédiaire du terme non linéaire donnée par :

$$N_\alpha^{n+1/2} = (\frac{3}{4}-\alpha)N(x_j,t_n)+(\frac{3}{4}+\alpha)N(\tilde{x}_j^n,t_n)-(\frac{1}{4}-\alpha)N(x_j,t_{n-1})-(\frac{1}{4}+\alpha)N(\tilde{x}_j^n,t_{n-1}), \tag{1.8}$$

où $N(x,t_n) := N(\psi(x,t_n),x,t_n)$.

Les analyses ont montré que la valeur la plus adéquate en termes de stabilité et précision pour le paramètre arbitraire est $\alpha = 1/4$.

Les schémas utilisés pour le traitement du terme linaire L^p et non linéaire N^p interagissent d'une manière complexe. D'après les analyses effectuées, il est recommandé d'utiliser le même décentrement pour les deux termes dans la deuxième étape pour ne

pas influencer la précision de la méthode. Les schémas semi-implicites semi-lagrangiens proposés sont définis par le paramètre arbitraire θ qui correspond au décentrement au niveau de la deuxième étape relative au correcteur. Afin de garantir une bonne précision, le décentrement θ doit être dans l'intervalle $[0.5, 0.75]$.

Les nouvelles techniques développées ont permis de construire de nouveaux schémas semi-implicites semi-lagrangiens qui présentent des domaines de stabilité absolue très larges. Cette classe de schémas présente de bonnes qualités de stabilité pour le cas des solutions à caractère oscillatoire. Les tests numériques ont confirmé que cette classe de schémas présente plusieurs avantages de stabilité, de précision et de convergence.

Les schémas proposés présentent l'avantage d'utiliser la même hypothèse utilisée dans la méthode SETTLS, relative à l'approximation de l'accélération entre le point de départ et le point d'arrivée sur la trajectoire semi-lagrangienne. Cette valeur est utilisée dans les méthodes proposées pour obtenir la position du point milieu de la trajectoire. Les performances de la classe de schémas proposée en termes de précision ont un grand avantage pour l'utilisation des pas de temps qui sont larges et qui satisfont le critère de Lipschitz. Ce critère constitue une condition suffisante pour éviter l'intersection des trajectoires calculées. Ceci permet de réduire le coût de calcul de la classe des schémas numériques proposée. L'utilisation de la méthode explicite pour le terme linéaire dans la deuxième étape permet de réduire davantage le temps de calcul. Ces schémas sont concurrentiels en termes de temps de calcul en comparaison avec des schémas numériques à deux étapes très connus. La classe des schémas proposée présente de bonnes qualités en termes de stabilité, de précision et de convergence par rapport à des méthodes semi-implicites semi-lagrangiennes et des schémas prédicteurs-correcteurs très connus.

1.3.3 Un schéma équilibre partiellement centré de type «Cell-Vertex» préservant la positivité pour les écoulements en eaux peu profondes

On considère les équations de Saint-Venant en deux dimensions avec une topographie variable

$$\begin{cases} h_t + (hu)_x + (hv)_y = 0, \\ (hu)_t + \left(hu^2 + \dfrac{g}{2}h^2\right)_x + (huv)_y = -ghB_x, \\ (hv)_t + (huv)_x + \left(hv^2 + \dfrac{g}{2}h^2\right)_y = -ghB_y, \end{cases} \tag{1.9}$$

où h désigne la profondeur d'eau, $(u, v)^T$ est le vecteur vitesse, $B(x, y)$ est la topographie et g est l'accélération de la pesanteur.

Pour le développement des schémas partiellement centrés il est recommandé (Bryson et al., 2011) d'utiliser les variables $(w = B+h, hu, hv)$ au lieu des variables conservatives

(h, hu, hv). Ainsi, le système (1.9) peut s'écrire sous la forme suivante :

$$\mathbf{U}_t + \mathbf{F}(\mathbf{U}, B)_x + \mathbf{G}(\mathbf{U}, B)_y = \mathbf{S}(\mathbf{U}, B), \tag{1.10}$$

avec

$$\begin{aligned}
\mathbf{F}(\mathbf{U}, B) &= \Big(p,\ \frac{p^2}{w-B} + \frac{g}{2}(w-B)^2,\ \frac{pq}{w-B}\Big)^T, \\
\mathbf{G}(\mathbf{U}, B) &= \Big(q,\ \frac{pq}{w-B},\ \frac{q^2}{w-B} + \frac{g}{2}(w-B)^2\Big)^T, \\
\mathbf{S}(\mathbf{U}, B) &= \Big(0,\ -g(w-B)B_x,\ -g(w-B)B_y\Big)^T.
\end{aligned} \tag{1.11}$$

La solution du système (1.9) peut développer des discontinuités en temps finis. Les méthodes couramment utilisées pour ces cas sont des schémas de volumes finis décentrés. Ce type de schéma nécessite la résolution du problème de Riemann au niveau des interfaces des cellules de calcul. Un nouveau schéma partiellement centré est développé en utilisant un maillage non structuré sous forme de polygones construits à partir d'un maillage triangulaire initial. Ce maillage présente des cellules de calcul avec une uniformité spatiale meilleure que le maillage triangulaire initial. Delis et al. (2011) ont montré l'intérêt et les avantages de ce type de maillage par rapport au maillage basé sur des cellules centrées.

De nouvelles reconstructions sont proposées pour la topographie et l'élévation de la surface libre. Les techniques développées dans ces reconstructions permettent d'assurer la stabilité de la méthode et la positivité de la hauteur d'eau au cours du temps. La nouvelle discrétisation du terme source dû à la topographie permet la préservation d'une manière exacte des états d'équilibre du système. Les performances de la méthode des volumes finis proposée sont testées par des exemples numériques. Les résultats confirment que cette méthode assure l'équilibre entre le terme de flux et le terme source et préserve la positivité de la hauteur d'eau au cours du temps. Cette méthode est efficace et peut être appliquée au système de Saint-Venant lorsque la topographie est discontinue ou hautement variable et sur des domaines complexes qui nécessitent l'utilisation des maillages non structurés.

1.3.4 Analyse des méthodes des volumes finis non structurées pour les écoulements en eaux peu profondes en utilisant le pseudo spectre

Dans cette partie du projet, on propose une nouvelle méthode pour l'analyse de stabilité des méthodes des volumes finis. On considère les équations des écoulements en eaux peu profondes avec l'effet de Coriolis sur des maillages non structurés. Dans l'analyse, on considère cinq schémas de volumes finis qui utilisent différents types de discrétisations spatiales et la méthode de Crank-Nicolson comme schéma temporel.

La discrétisation des équations en eaux peu profondes dans le cas des maillages non

structurés entraine des problèmes de stabilité à cause des modes numériques. Un maillage non structuré conduit en général à un opérateur discret non normal de la méthode numérique. Les vecteurs propres de l'opérateur discret peuvent être loin de l'orthogonalité, ce qui peut causer des amplifications des solutions en temps finis. Les conditions de stabilité asymptotique ne permettent pas de fournir assez d'information sur le comportement des solutions en temps finis. Des amplifications des solutions peuvent être observées même si le schéma numérique est stable au sens de Lax–Richtmyer. La nouvelle approche de stabilité des schémas numériques est basée sur la notion du pseudo spectre. Cette nouvelle approche est utile pour le choix du maillage, des emplacements les plus convenables des variables primitives au niveau des cellules du maillage et de la méthode de discrétisation la plus stable.

1.3.5 Une méthode des volumes finis respectant la condition de compatibilité géométrique pour les lois de conservation sur des surfaces courbes

Dans cette partie du projet, on considère les systèmes hyperboliques sur des surfaces courbes. L'objectif est de développer une méthode des volumes finis d'ordre élevé qui respecte la condition de compatibilité géométrique et qui est efficace pour le calcul des solutions discontinues pour ces systèmes. L'équation de base suivante est considérée sur la sphère S^2

$$\partial_t u + \nabla \cdot F(\cdot, u) = 0, \qquad u = u(t) : S^2 \to \mathbb{R}, \tag{1.12}$$

où $u(t, x)$ est la fonction scalaire représentant l'inconnu du problème sous la condition initiale

$$u(0, x) = u_0(x), \tag{1.13}$$

et le flux est donné par l'équation

$$F(x, \bar{u}) = n(x) \wedge \Phi(x, \bar{u}), \qquad x \in S^2,\ \bar{u} \in \mathbb{R}, \tag{1.14}$$

où $\Phi = \Phi(x, \bar{u})$ est un champ de vecteur dans l'espace $\mathbb{R}^3$ et $n(x)$ désigne le vecteur unitaire normal à la sphère.

On s'intéresse aux vecteurs flux qui satisfont la condition de compatibilité géométrique suivante:

$$\nabla \cdot (F(\cdot, \bar{u})) = 0, \tag{1.15}$$

où $\bar{u}$ est une constante arbitraire. Plus particulièrement, on s'intéresse à la classe de flux définie par:

$$\Phi(x, \bar{u}) = \nabla h(x, \bar{u}), \qquad x \in S^2, \quad \bar{u} \in \mathbb{R}, \tag{1.16}$$

où $h = h(x, \bar{u})$ est une fonction lisse au voisinage de la sphère S^2 et ∇ est l'opérateur du gradient dans $\mathbb{R}^3$.

Cette partie du projet de recherche se base sur les travaux de Ben Artzi et LeFloch (2007) et Ben-Artzi, Falcovitz, et LeFloch (2009). Un schéma de volumes finis est proposé en se basant sur la résolution du problème de Riemann généralisé, la méthode du «splitting» directionnel en latitude et longitude, et la méthode Runge-Kutta d'ordre trois (TVDRK3). Une nouvelle reconstruction linéaire est utilisée en prenant en considération les valeurs de la solution aux centres des cellules de calcul et les valeurs des solutions du problème de Riemann aux interfaces des cellules. Ces dernières sont obtenues en utilisant les approximations de second ordre basées sur la résolution du problème de Riemann généralisé. La méthode proposée est utilisée pour étudier numériquement les propriétés des solutions discontinues pour les deux versions du schéma du premier et deuxième ordre.

La méthode des volumes finis proposée présente un lien très fort entre sa forme semi-discrète et l'équation du système (1.12). La forme semi-discrète du schéma numérique respecte d'une manière exacte la condition de compatibilité géométrique. La reconstruction linéaire proposée a permis d'améliorer nettement la qualité des résultats des solutions discontinues du système. La méthode est d'ordre deux dans l'espace et l'approche du «splitting» directionnel combinée avec la méthode TVDRK3 constituent une méthode efficace pour l'intégration temporelle. L'ordre trois de la méthode TVDRK3 est moins influencé par la méthode du «splitting» directionnel utilisée. Le schéma proposé est efficace dans le cas des solutions discontinues de grands chocs et amplitudes en comparaison avec d'autres schémas numériques très connus.

La méthode proposée est utilisée pour étudier le comportement asymptotique des solutions. Une classification des flux est proposée où les notions de flux feuilletés et flux génériques sont introduites. Cette classification et la linéarité des flux constituent un concept très important et suffisant pour prédire le comportement asymptotique des solutions du système (1.12). Les résultats obtenus pour le cas des flux feuilletés non linéaires présentent un intérêt particulier pour construire des solutions stationnaires non triviales. Pour ces flux, les solutions qui sont constantes le long des lignes de niveau sont des solutions stationnaires non triviales du système (1.12). Ces solutions sont utilisées dans les tests numériques pour évaluer les performances de la méthode proposée. Cette méthode peut être étendue au cas des équations de Saint-Venant sur la sphère.

1.3.6 Un schéma équilibre partiellement centré pour les lois de conservation sur des surfaces courbes

Dans cette partie du projet, on considère le système hyperbolique non linéaire sur des surfaces courbes donné par l'équation (1.12). L'objectif est de développer une méthode des volumes finis partiellement centrée pour le calcul des solutions discontinues de ce système. Ceci constitue le premier schéma numérique de type partiellement centré

pour les lois de conservation sur des surfaces courbes. Comme mentionné auparavant, la démarche suivie par Stanley Osher est adoptée pour l'établissement des schémas numériques robustes. Les équations de Burgers sont utilisées pour développer et valider la nouvelle méthode proposée.

On note que la méthode décrite en 1.3.5 pour la résolution du système (1.12) utilise l'approche du «splitting» directionnel pour simplifier la résolution du problème de Riemann généralisé. Ceci présente un coût supplémentaire en termes de temps de calcul. L'objectif est de développer un schéma de volumes finis sans solveur de Riemann, non diffusif et qui ne fait pas appel à l'approche du «splitting» directionnel. Une discrétisation est proposée pour l'opérateur de divergence qui satisfait la condition de compatibilité géométrique en utilisant la fonction h. La forme semi-discrète du nouveau schéma numérique est construite en suivant les étapes de reconstruction, évolution et projection. Dans notre démarche, on suppose que les dérivées spatiales de la fonction u sont bornées indépendamment du pas du temps considéré. Cette hypothèse est suffisante pour obtenir une forme simplifiée du schéma sans résoudre le problème de Riemann au niveau des interfaces des cellules de calcul. D'après les analyses effectuées, l'hypothèse adoptée est plus convenable pour le cas des solutions discontinues avec des amplitudes et chocs moyens.

Les valeurs extrêmes des vitesses locales de propagation des ondes au niveau des interfaces des cellules sont utilisées pour la construction du nouveau schéma numérique. La méthode développée respecte la condition de compatibilité géométrique et sa forme semi-discrète est fortement liée aux propriétés analytiques de l'équation (1.12) et à la géométrie de la sphère.

Une reconstruction non oscillatoire est proposée dans laquelle le gradient de chaque variable est calculé en utilisant une fonction Minmod pour garantir la stabilité de la méthode. Le schéma de volumes finis proposé est simple et moins coûteux de point de vue temps de calcul puisque l'approche du «splitting» directionnel et les solveurs de Riemann sont évités. Les solutions stationnaires non triviales obtenues sur la base des flux feuilletés non linéaires sont utilisées pour tester la nouvelle méthode proposée. Cette méthode est efficace pour le cas des solutions discontinues d'amplitudes et chocs moyens pour les lois de conservation scalaires hyperboliques non linéaires sur la sphère. Le schéma de volumes finis proposé pour les lois de conservation sur la sphère peut être étendu au cas du système de Saint-Venant sur la sphère.

CHAPITRE 2 Une méthode des volumes finis non structurée pour les écoulements à grande échelle utilisant le schéma d'Adams du quatrième ordre

An unstructured finite volume method for large-scale shallow flows using the fourth-order Adams scheme [1]

Résumé

Dans ce chapitre, on s'intéresse à la modélisation des processus lents et rapides dans l'océan et l'atmosphère. Un calcul précis des deux processus nécessite des méthodes numériques efficaces qui permettent de calculer les ondes relativement rapides sans amortissement des ondes longues. Une nouvelle méthode des volumes finis décentrée est proposée sur un maillage non structuré pour les écoulements peu profonds à grande échelle. Le système de Saint-Venant est considéré avec un terme source qui inclut le paramètre de Coriolis. Ce système est un prototype important pour étudier plusieurs aspects de la dynamique à grande échelle de l'atmosphère et l'océan.

La plupart des méthodes numériques décentrées, qui aboutissent à de bons résultats pour les ondes rapides (ondes de gravité), conduisent à un niveau d'amortissement élevé pour les modes qui sont lents tels que les ondes de Rossby. Les schémas numériques centrés qui sont efficaces pour les ondes de Rossby rencontrent des problèmes pour simuler la propagation des ondes de gravité.

Dans la méthode des volumes finis proposée, on utilise une nouvelle approche dans laquelle la méthode d'Adams du quatrième ordre est combinée avec le «splitting» directionnel pour l'intégration temporelle des équations. Une nouvelle technique est proposée pour le traitement du terme de Coriolis dans laquelle ce terme est intégré analytiquement avant et après la résolution numérique de l'équation relative au terme de flux. Le traitement des flux convectifs est effectué en utilisant la méthode de Roe (1981). Une nouvelle reconstruction des variables primitives est proposée et elle consiste à utiliser une approximation de troisième ordre orthogonalement à chaque interface des cellules de calcul. Le long de l'axe de chaque interface des cellules, on utilise une approximation du premier ordre. Les techniques proposées et la méthode d'Adams d'ordre quatre, utilisée pour l'intégration temporelle sans aucune itération sur le correcteur, sont suffisantes pour supprimer le bruit numérique des ondes courtes sans amortissement des ondes longues.

Les performances de la méthode proposée ont été validées par des tests numériques

[1] Cet article est réalisé en collaboration avec A. Mohammadian et H. Qiblawey, et publié sous la forme: A. Beljadid, A. Mohammadian, H. Qiblawey, 2013, An unstructured finite volume method for large-scale shallow flows using the fourth-order Adams scheme. Computers & Fluids (Elsevier), 88, 579-589

pour les modèles d'équations de types linaire et non linéaire. Les résultats des tests numériques montrent que l'approche du «splitting» directionnel a une légère influence sur l'ordre de la méthode d'Adams utilisée pour l'intégration temporelle. Les analyses confirment que dans la nouvelle reconstruction, le premier ordre utilisé dans la direction des interfaces des cellules de calcul a un impact négligeable sur l'ordre spatial du schéma numérique. La structure des cellules du maillage a un effet négligeable sur la qualité des résultats. La méthode proposée préserve l'équilibre entre le terme de flux et le terme de Coriolis. De bons résultats sont obtenus à la fois pour les ondes de gravité et les ondes de Rossby, qui ont une importance cruciale dans la simulation des écoulements peu profonds à grande échelle.

La conservation de l'énergie du système est considérablement améliorée par rapport à d'autres schémas décentrés qui sont largement utilisés pour la simulation des ondes de gravité comme le schéma décentré de troisième ordre. Ces schémas conduisent à un niveau élevé de dissipation d'énergie et d'oscillation numérique causées par le déséquilibre entre le terme de flux et le terme source.

Les tests numériques montrent qu'il n'est pas avantageux de considérer d'itération ou de modificateur supplémentaire à l'étape de correction. Une seule itération au niveau du correcteur est suffisante pour obtenir de bons résultats. Ceci rend la méthode proposée moins chère que la méthode Runge-Kutta d'ordre quatre, dans laquelle le flux doit être intégré quatre fois pour chaque étape.

2.1 Introduction

Shallow water equations (SWEs) are used to describe many physical phenomena in oceans, rivers, the atmosphere,etc. These equations are applicable when the vertical velocity component is negligible compared to the horizontal components, and are obtained by assuming hydrostatic pressure distribution (e.g. Vreugdenhil, 1994). The three-dimensional incompressible Navier-Stokes equations are averaged over the depth to obtain the SWEs. In the absence of viscous terms, SWEs can be considered a hyperbolic system. The finite volume (FV) methods are most convenient for modeling these systems since they have a conservative form. Upwind finite volume (UFV) methods can numerically solve these systems with good accuracy and an acceptable computational cost.

UFV schemes use exact or approximate methods to solve the Riemann problem at the interface of computational cells. Godunov's method (Godunov, 1959; Godunov et al., 1961) is the most popular scheme using the exact solution of the Riemann problem. Its extension to second-order and to high-order schemes is given by Van Leer (1979) and Colella and Woodward (1984), respectively. The exact algorithms are computationally expensive compared to the approximate methods. Roe's method (1981), which is applied in this work, is the most popular approximate method. It requires

an accurate estimation of parameter values near the interface on both sides of the computational cell. In the presence of source terms in the SWEs, the UFV schemes may lead to numerical oscillations due to the imbalance between the source and flux terms. To overcome this problem, some special treatments can be applied for balancing the source and flux terms. A large number of studies have been conducted in this direction, such as Vázquez-Cendón (1999), Gallouet et al. (2003), Mohammadian et al. (2005), Mohammadian and Le-Roux (2006), and Stewart et al. (2011). Other studies have been conducted to evaluate the performance of various schemes for large-scale shallow flows (e.g. Walters and Carey, 1983; Foreman, 1984; Hanert et al., 2004, 2005; Le-Roux and Pouliot, 2008; Hanert et al., 2009; Walters et al., 2009; Le-Roux et al., 2011). Nevertheless, UFV methods are considered in a limited number of studies (e.g. Lin et al., 2003; Mohammadian and Le-Roux, 2008; Castro et al., 2008; Beljadid et al., 2012).

The performance of numerical methods is greatly influenced by the temporal schemes used. Total Variation Diminution (TVD) temporal integration methods, developed by Shu and Osher (1988), are among the most popular temporal integration schemes. They are widely used for their ability to avoid oscillations and to maintain stability. Furthermore, some higher-order TVD schemes are insensitive to the values of Courant-Friedrichs-Lewy (CFL) numbers and present highly accurate results over a wide range of CFL numbers.

Beljadid et al. (2013b) studied the performance of UFV schemes and examined several aspects, including mass and energy conservation, numerical diffusion, and numerical oscillations for certain waves. The accuracy of various schemes was analyzed for different types of waves. Through numerical experiments, it was demonstrated that UFV schemes provide accurate results for various waves. However, these schemes fail in the modeling of Rossby waves, which have a particular behavior and are difficult to capture by several well-known upwind schemes. In this paper we propose a new upwind finite volume method which presents a good improvement for the modeling of Rossby waves. A high-order spatial scheme based on polynomial fitting is proposed. Operator splitting and the fourth-order Adams method are used for temporal integration.

The paper is organized as follows: SWEs are presented in Section 2.2. In Section 2.3, the proposed finite volume method is described. Section 2.4 presents some numerical experiments for equatorial Rossby waves. In Section 2.5, some numerical experiments are performed using the proposed method for nonlinear SWEs. Some concluding remarks complete the study.

2.2 Shallow water equations

In this section, linear and nonlinear shallow water equations are presented. The conservative form of the 2D shallow water equations is written as (Vreugdenhil, 1994):

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial \mathbf{E}}{\partial x} + \frac{\partial \mathbf{G}}{\partial y} = \mathbf{S}, \tag{2.1}$$

The linear and nonlinear equations are defined in terms of parameters **U**, **E**, **G**, and **S**.

2.2.1 Linear SWEs

For linear shallow water equations, the parameters **U**, **E**, **G**, and **S** are defined as:

$$\mathbf{U} = \begin{bmatrix} \eta \\ u \\ v \end{bmatrix} \quad , \quad \mathbf{E} = \begin{bmatrix} Hu \\ g\eta \\ 0 \end{bmatrix} \quad , \quad \mathbf{G} = \begin{bmatrix} Hv \\ 0 \\ g\eta \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} 0 \\ fv \\ -fu \end{bmatrix} \quad , \tag{2.2}$$

where η represents the water surface elevation, u and v are the depth-averaged velocity components in the x-and y-directions, respectively, f is the Coriolis parameter, g is the gravity acceleration, H is the average value of the water depth, and $(H + \eta)$ is the total water depth.

The term **S** may include various source terms such as bed friction, bed topography, and wind stress. Since this paper concentrates on Rossby waves, the source term **S** is assumed to include the Coriolis parameter.

The beta-plane approximation to the Coriolis parameter is considered ($f = \beta y$), where β is the linear coefficient of variation of f with respect to y. The variable y is considered as the meridional distance from the equator (positive northward). The parameter β is given as:

$$\beta = 2\mathbf{\Omega}/\mathbf{R} = 2.29 \times 10^{-11}\, m^{-1} s^{-1}, \tag{2.3}$$

where $\mathbf{\Omega}$ and **R** are the angular speed of the Earth's rotation and the mean radius of the Earth, respectively ($\mathbf{\Omega} = 7.29 \times 10^{-5}\, rad\, s^{-1}, \mathbf{R} = 6371\, km$).

The dimensionless form of SWEs is used in this paper. The model equations (2.1) and (2.2) are converted into dimensionless form on an equatorial beta-plane using the variables $\tilde{x} = x/L^*$, $\tilde{y} = y/L^*$, $\tilde{\eta} = \eta/H^*$, $\tilde{u} = u/U^*$ and $\tilde{v} = v/U^*$. The reference values of the depth (H^*), time (T^*), length (L^*) and velocity (U^*) scales are expressed

as:

$$
\begin{aligned}
H^* &= H \\
T^* &= \beta^{-1/2}(gH)^{-1/4} \\
L^* &= \frac{1}{\beta T^*} \\
U^* &= V^* = \frac{L^*}{T^*}
\end{aligned}
\tag{2.4}
$$

The resulting system, the Jacobian matrix, and the corresponding eigenvalues and eigenvectors are given in Appendix I.

2.2.2 Nonlinear SWEs

For nonlinear shallow water equations, the parameters **U**, **E**, **G**, and **S** are defined as:

$$
\mathbf{U} = \begin{bmatrix} h \\ hu \\ hv \end{bmatrix} \quad , \quad \mathbf{E} = \begin{bmatrix} hu \\ hu^2 + 0.5gh^2 \\ huv \end{bmatrix} \quad , \quad \mathbf{G} = \begin{bmatrix} hv \\ huv \\ hv^2 + 0.5gh^2 \end{bmatrix}
\tag{2.5}
$$

The source term **S** is assumed to include the Coriolis effect

$$
\mathbf{S} \ = (0, fhv, -fhu)^T,
\tag{2.6}
$$

where h is the total fluid depth.

In the presence of the Coriolis effect, the nonlinear SWEs are converted into a dimensionless form on an equatorial beta-plane using the variables $\tilde{x} = x/L^*$, $\tilde{y} = y/L^*$, $\tilde{h} = h/H^*$, $\tilde{u} = u/U^*$ and $\tilde{v} = v/U^*$. The characteristic time (T^*), length (L^*) and velocity (U^*) scales are expressed in terms of the parameter β in the same way using equations (2.4), where the parameter H^* is the mean water depth.

When the Coriolis force is absent, the following reference parameters are used to convert the nonlinear SWEs to a dimensionless form:

$$
\begin{aligned}
T^* &= L^*/\sqrt{gH^*} \\
U^* &= V^* = \frac{L^*}{T^*},
\end{aligned}
\tag{2.7}
$$

where the characteristic length L^* can be arbitrarily chosen and the parameter H^* can be chosen with the same order as the mean water depth h in the system.

2.3 Finite volume method

An upwind finite volume method on an unstructured grid is employed in this paper. The variables are located at the geometric centers of the computational grids. Each triangle represents a control volume. The SWEs are integrated over every control volume as:

$$\int_{\Omega} \left(\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial \mathbf{E}}{\partial x} + \frac{\partial \mathbf{G}}{\partial y} - \mathbf{S} \right) d\Omega = 0, \tag{2.8}$$

where Γ and Ω denote the boundary and the area of the domain, respectively.

By using the divergence theorem, the flux integral is transformed into a boundary integral:

$$\int_{\Omega} \left(\frac{\partial \mathbf{E}}{\partial x} + \frac{\partial \mathbf{G}}{\partial y} \right) d\Omega = \int_{\Gamma} \mathbf{F} \cdot \mathbf{n} \, d\Gamma, \tag{2.9}$$

where $\mathbf{F} = (\mathbf{E}, \mathbf{G})^t$ is the flux vector and $\mathbf{n}$ is the unit outward normal vector to the boundary Γ. Then, (2.8) leads to

$$\frac{d}{dt} \int_{\Omega} \mathbf{U} d\Omega + \int_{\Gamma} \mathbf{F} \cdot \mathbf{n} \, d\Gamma = \int_{\Omega} \mathbf{S} d\Omega \tag{2.10}$$

2.3.1 Unstructured grid implementation

For an unstructured triangular grid, the boundary integral $\int_{\Gamma} \mathbf{F} \cdot \mathbf{n} \, d\Gamma$ in (2.10) may be approximated by a summation over the triangle edges as:

$$\int_{\Gamma} \mathbf{F} \cdot \mathbf{n} \, d\Gamma = \sum_{k=1}^{3} \int_{\Gamma_k} \mathbf{F} \cdot \mathbf{n} \, d\Gamma_k = \sum_{k=1}^{3} \left(\mathbf{F}_k \cdot \mathbf{n}_k \right) \; l_k, \tag{2.11}$$

where Γ_k, $\mathbf{F}_k$, $\mathbf{n}_k$, and l_k, $k = 1, 2, 3$, are respectively the triangle edges, the outward fluxes, the unit outward normal vectors, and the lengths corresponding to the edges of a triangular cell.

The convective flux $\mathbf{F}$ can be calculated by various schemes. Most schemes may be written in a general form as:

$$\mathbf{F} = 0.5 \left(\mathbf{F}_R + \mathbf{F}_L - \Delta \mathbf{F}^* \right), \tag{2.12}$$

where $\mathbf{F}_L = \mathbf{F}\left(\mathbf{U}_L\right)$ and $\mathbf{F}_R = \mathbf{F}\left(\mathbf{U}_R\right)$ are the left and right flux vectors.

The flux difference $\Delta \mathbf{F}^*$, which plays the role of stabilization, is computed based on Roe's linearization:

$$\Delta \mathbf{F}^* = \sum_{k=1}^{3} \tilde{\alpha_k} \; \left| \tilde{a_k} \right| \tilde{\mathbf{e}}_k, \tag{2.13}$$

where $\tilde{a_k}$, $\tilde{\mathbf{e}}_k$ and $\tilde{\alpha_k}$ are, respectively, the eigenvalues and the eigenvectors of the

approximate Jacobian $\widetilde{\mathbf{J}}$ and the coefficients of decomposition of $\Delta\mathbf{U} = \mathbf{U}_R - \mathbf{U}_L$ on the basis of the eigenvectors, as explained in Appendix I.

In the κ scheme, $\mathbf{U}_L$, and $\mathbf{U}_R$ are calculated at the interface as:

$$\begin{aligned} \mathbf{U}_L &= \mathbf{U}_w + \frac{s}{4}\left[(1-\kappa s)(\mathbf{U}_w - \mathbf{U}_{ww}) + \delta(1+\kappa s)(\mathbf{U}_e - \mathbf{U}_w)\right] \\ \mathbf{U}_R &= \mathbf{U}_e + \frac{s}{4}\left[(1-\kappa s)(\mathbf{U}_e - \mathbf{U}_{ee}) + \delta(1+\kappa s)(\mathbf{U}_w - \mathbf{U}_e)\right], \end{aligned} \tag{2.14}$$

with $\delta = \frac{2L_w}{L_w+L_e}$, where L_w and L_e are defined in Figure 2.1. The slope limiter s is calculated in this paper using:

$$s = \frac{2\Delta_-\Delta_+}{\Delta_+^2 + \Delta_-^2 + \varepsilon}, \quad \varepsilon > 0, \tag{2.15}$$

with $\Delta_+ = \mathbf{U}_w - \mathbf{U}_{ww}$ and $\Delta_- = \mathbf{U}_e - \mathbf{U}_w$.

The parameter ε is a small positive number chosen according to the order of the scheme. It should be limited in order to not affect the order of accuracy of the numerical scheme. In the numerical experiments presented in this paper, for a numerical scheme of order r, the parameter ε is chosen as $0 < \varepsilon < \Omega_m^{r+1}$, where Ω_m is the area of the smallest cell in the entire computational domain.

Depending on κ, equations (2.14) lead to the following schemes:

$$\kappa = \begin{cases} 0, & simplified\ Fromm\ scheme, \\ 1/6, & cell-based\ third\ order\ upwind\ method, \\ 1/3, & third-order\ upwind\ method, \\ 1/2, & Quick\ scheme, \\ 1, & upwind-centered\ scheme \end{cases}$$

The case $\kappa = -1$ corresponds to the second-order upwind scheme. This scheme is not considered in this paper since it leads to inaccurate results for Kelvin, Yanai, and Poincaré waves.

2.3.2 The proposed high-order upwind scheme

The method introduced in this paper includes a high-order upwind interpolation scheme. Upwind methods can be improved if the values of the parameters on both sides of the interface are estimated with more accuracy. The proposed method uses polynomials with two variables. In this paper, we use the third-order Lagrange polynomials in σ, and linear interpolation in τ, where the variable σ denotes the axis perpendicular to the interface and τ coincides with the interface as they are shown in Figure 2.2. The value of $\mathbf{U}_L$ is obtained by interpolation using three grid points upstream of the interface

and one grid point downstream of the interface. The parameter $\mathbf{U}(\sigma,\tau)$ is obtained by the following interpolation

$$\mathbf{U}(\sigma,\tau) = \sum_{i=1}^{i=4} L_i^{(l)}(\sigma)(\mathbf{U}_i - Q^{(l)}(\tau_i)) + Q^{(l)}(\tau), \tag{2.16}$$

where $Q^{(l)}(\tau)$ is a polynomial which depends on the variable τ and $L_i^{(l)}$ is the Lagrange polynomial associated with the cell i, obtained from the parameters of three cells on the left-hand side and one cell on the right-hand side of the interface.

$$L_i^{(l)}(\sigma) = \prod_{j=1, j\neq i}^{j=4} \frac{(\sigma - \sigma_j)}{(\sigma_i - \sigma_j)} \tag{2.17}$$

The parameter $\mathbf{U}_L$ is obtained by integration along the left hand-side of the interface as:

$$\mathbf{U}_L = \frac{1}{l}\int_{\frac{-l}{2}}^{\frac{l}{2}} \mathbf{U}(0^-,\tau)d\tau, \tag{2.18}$$

where l is the length of the interface of the computational cell. Since $\sum_{i=1}^{i=4} L_i^{(l)}(\sigma) = 1$, thus:

$$\mathbf{U}(\sigma,\tau) = \sum_{i=1}^{i=4} L_i^{(l)}(\sigma)(\mathbf{U}_i - Q^{(l)}(\tau_i) + Q^{(l)}(\tau)) \tag{2.19}$$

The form of the above equation allows defining the polynomial $Q^{(l)}$ without a constant. Then, for the linear case, the polynomial $Q^{(l)}$ is defined by $Q^{(l)}(\tau) = \mu\tau$. A similar approach is applied for the right-hand side value $\mathbf{U}_R$ by using Lagrange polynomials $L_i^{(r)}$ obtained from the parameters of three cells on the right-hand side and one cell on the left-hand side of the interface.

$$\mathbf{U}(\sigma,\tau) = \sum_{i=1}^{i=4} L_i^{(r)}(\sigma)(\mathbf{U}_i - Q^{(r)}(\tau_i)) + Q^{(r)}(\tau) \tag{2.20}$$

The parameter $\mathbf{U}_R$ is obtained by integration along the right-hand side of the interface as:

$$\begin{aligned} \mathbf{U}_R &= \frac{1}{l}\int_{\frac{-l}{2}}^{\frac{l}{2}} \mathbf{U}(0^+,\tau)d\tau \\ \mathbf{U}(\sigma,\tau) &= \sum_{i=1}^{i=4} L_i^{(r)}(\sigma)(\mathbf{U}_i - Q^{(r)}(\tau_i) + Q^{(r)}(\tau)) \end{aligned} \tag{2.21}$$

For the linear case in the τ direction, we consider $Q^{(l)}(\tau) = Q^{(r)}(\tau) = \mu\tau$.
In the numerical experiments, it was found that the slope μ greatly influences the results. For example, when the slope μ is calculated by using the grid point nearest to the computational cell, which differs from the four cells used for Lagrange polynomials of the variable σ, the results are not accurate. This is one of the reasons that the

κ scheme leads to inaccurate results, as will be shown in Section 2.4. Indeed, the κ scheme only performs an interpolation in the perpendicular direction and does not consider the coupling of the two directions.

In the proposed method, the first-order expansion of the parameter $\mathbf{U}$ is used to estimate the value of $\mathbf{U}_{P_2}$ and $\mathbf{U}_{P_1}$ at the extremities of the interface and the slope μ

$$\mu = \frac{\mathbf{U}_{P_2} - \mathbf{U}_{P_1}}{l} \tag{2.22}$$

and

$$d\mathbf{U} = \frac{\partial \mathbf{U}}{\partial x} dx + \frac{\partial \mathbf{U}}{\partial y} dy \tag{2.23}$$

Then,

$$\mathbf{U}_{P_{1i}} = \mathbf{U}_{P_1} + a_1 \Delta x_{1i} + b_1 \Delta y_{1i}, \tag{2.24}$$

where $a_1 = (\frac{\partial \mathbf{U}}{\partial x})_{P_1}$,$b_1 = (\frac{\partial \mathbf{U}}{\partial y})_{P_1}$, Δx_{1i}=$x_{P_{1i}}$-x_{P_1} and Δy_{1i}=$y_{P_{1i}}$-y_{P_1}. The values (x_{P_1}, y_{P_1}) are the coordinates of point P_1, and the values $(x_{P_{1i}}, y_{P_{1i}})$ are the coordinates of the points P_{1i} surrounding the point P_1.

The method of least squares is used to obtain the parameters $\mathbf{U}_{P_1}$, a_1, and b_1. The residual is given by:

$$R_1 = \sum_{i=1}^{i=N} (\mathbf{U} + a\Delta x_{1i} + b\Delta y_{1i} - \mathbf{U}_{P_{1i}})^2, \tag{2.25}$$

where N is the number of points surrounding point P_1. We search for the optimal values of $\mathbf{U}_{P_1}$, a_1, and b_1 for the variables $\mathbf{U}$, a, and b, respectively. Therefore,

$$(\mathbf{U}_{P_1}, a_1, b_1) = \underset{\mathbf{U},a,b}{\texttt{argmin}} \quad R_1 \tag{2.26}$$

$\mathbf{U}_{P_1}$, a_1, and b_1 can thus be obtained by solving the following system:

$$\begin{cases} \frac{\partial R_1}{\partial \mathbf{U}} = 0 \\ \frac{\partial R_1}{\partial a} = 0 \\ \frac{\partial R_1}{\partial b} = 0 \end{cases} \tag{2.27}$$

The parameters at the vertex P_2 are obtained in the same way as for P_1.

2.3.3 Temporal integration method

In most finite volume schemes, a Runge-Kutta method is used for temporal integration. A popular approach is the third-order TVD Runge-Kutta (TVDRK3) method, which is explained in Appendix II. The TVDRK3 method combined with the κ scheme leads to inaccurate results for Rossby waves due to an imbalance between the flux and Coriolis

terms. In this paper, the fourth-order Adams method is proposed for the temporal scheme, with operator splitting for the Coriolis and flux terms. The process includes three stages: in the first and third steps the Coriolis term is integrated analytically, and in the second step the flux term is integrated numerically. In the following, the temporal integration method is explained for linear SWEs. The method for nonlinear SWEs can be achieved by replacing u and v with hu and hv, respectively. Following Beljadid et al. (2012), first, the effect of the source term (the Coriolis effect) is considered:

$$\frac{\partial \mathbf{U}}{\partial t} = \mathbf{S} \tag{2.28}$$

Then, the other terms are added:

$$\frac{\partial \mathbf{U}}{\partial t} = \mathbf{F} \tag{2.29}$$

where

$$\mathbf{F} = -\left(\frac{\partial \mathbf{E}}{\partial x} + \frac{\partial \mathbf{G}}{\partial y}\right) \tag{2.30}$$

The system with Coriolis term only, i.e.,

$$\begin{aligned} \frac{\partial u}{\partial t} &= fv \\ \frac{\partial v}{\partial t} &= -fu \\ \frac{\partial \eta}{\partial t} &= 0 \end{aligned} \tag{2.31}$$

can be solved analytically as:

$$\begin{aligned} u &= u_0 \cos(ft) + v_0 \sin(ft) \\ v &= v_0 \cos(ft) - u_0 \sin(ft) \end{aligned} \tag{2.32}$$

In the following, Δt represents the time step size and η^n, u^n, and v^n are respectively the water surface elevation and the $x-$ and $y-$velocities at time $t^n = n\Delta t$.

First, (2.28) is integrated over half of the time step:

$$\begin{aligned} u^* &= u^n \cos(f\Delta t/2) + v^n \sin(f\Delta t/2) \\ v^* &= v^n \cos(f\Delta t/2) - u^n \sin(f\Delta t/2) \\ \eta^* &= \eta^n \end{aligned} \tag{2.33}$$

Then, (2.29) is integrated over the entire time step using the fourth-order Adams method, as explained later. The result after integration will be noted by $\mathbf{U}^{**}$. Finally,

(2.28) is integrated over the second half of the time step, i.e.,

$$\begin{aligned} u^{n+1} &= u^{**}\cos(f\Delta t/2) + v^{**}\sin(f\Delta t/2) \\ v^{n+1} &= v^{**}\cos(f\Delta t/2) - u^{**}\sin(f\Delta t/2) \\ \eta^{n+1} &= \eta^{**} \end{aligned} \tag{2.34}$$

The fourth-order Adams method (Appendix II) uses the fourth-order explicit Adams-Bashforth scheme as the predictor, and the fourth-order Adams-Moulton method for the corrector, as explained below. We denoted by $\mathbf{f}^n = \sum_{k=1}^{3} (\mathbf{F}_k \cdot \mathbf{n}_k)\ l_k$ the result of the integration of the flux on the control volume at time $t_n = n\Delta t$.

The result $\mathbf{U}^*$ of equations (2.33) is used to obtain the predictor value:

$$\mathbf{U}^{pred} = \mathbf{U}^* + \Delta t \left(\frac{55}{24}\mathbf{f}^* - \frac{59}{24}\mathbf{f}^{n-1} + \frac{37}{24}\mathbf{f}^{n-2} - \frac{9}{24}\mathbf{f}^{n-3} \right), \tag{2.35}$$

where $\mathbf{f}^*$ is the result of the integration of the flux using $\mathbf{U}^*$.

The corrector value is obtained as:

$$\mathbf{U}^{**} = \mathbf{U}^* + \Delta t \left(\frac{9}{24}\mathbf{f}^{pred} + \frac{19}{24}\mathbf{f}^* - \frac{5}{24}\mathbf{f}^{n-1} + \frac{1}{24}\mathbf{f}^{n-2} \right), \tag{2.36}$$

where $\mathbf{f}^{pred}$ is the result of the integration of the flux using the predictor value $\mathbf{U}^{pred}$.

In the numerical experiments presented below, we will show that it is not advantageous to consider any additional iteration or modifier at the corrector step, and that even with only one iteration, optimal results are obtained. Note that in the above algorithm, the integration of the flux term (the values $\mathbf{f}^*$ and $\mathbf{f}^{pred}$), which is the most computationally expensive part, is required only twice. Therefore, the above version of the fourth-order Adams method is less expensive than the fourth-order Runge-Kutta method (RK4), in which the flux must be integrated four times for each step. We used an Intel Core i7-2670QM in our calculations. For the non-linear gravity test, presented in Section 2.5.1 at time $t = 50$, the CPU time for the Adams method is 191 while the CPU time is 361 for the RK4 method. The computational cost for the Adams method is about 53 % of the cost of the fourth-order Runge-Kutta scheme. It should be mentioned that the fourth-order Adams method uses the values of the parameters and the derivative of the flux function from the earlier steps. Therefore, the first three time steps are required in order to begin the method. In this paper, the first three steps are calculated by using the RK4 method for temporal integration in order to remain consistent in the order of accuracy.

The proposed algorithm can be summarized as:

- **From step n=1 to step 3:**

- Integrate the source term over $\Delta t/2$ using equations (2.33) to obtain the parameter $\mathbf{U}^*$,
- Evaluate the four terms of flux, k_1, k_2, k_3 and k_4 using equations (2.61) of the RK4 method (Appendix II.2). Those terms are the result of integration of the function flux on the control volume using the proposed high-order upwind scheme. They are also the derivatives of the parameter $\mathbf{U}$, particularly $k_1 = (\frac{\partial U}{\partial t})_{t_n} = \mathbf{f}^n$, which is used after step 3,
- Integrate the numerical flux using equation (2.60) of RK4 (Appendix II.2) over Δt to obtain the parameter $\mathbf{U}^{**}$,
- Integrate the source term over $\Delta t/2$ using equations (2.34) to obtain the parameter $\mathbf{U}^{n+1}$

- **For step n $\geq$ 4 :**
  - Integrate the source term over $\Delta t/2$ using equations (2.33) to obtain the parameter $\mathbf{U}^*$,
  - Evaluate the flux $\mathbf{f}^*$ on the control volume using the proposed high-order upwind scheme,
  - Determine the predictor value $\mathbf{U}^{pred}$ using the Adams-Bashforth formula (2.35),
  - Evaluate the flux $\mathbf{f}^{pred}$ on the control volume using the proposed high-order upwind scheme,
  - Determine the corrector value $\mathbf{U}^{**}$ using the Adams-Moulton formula (2.36),
  - Integrate the source term over $\Delta t/2$ using equations (2.34) to obtain the parameter $\mathbf{U}^{n+1}$

Finally, in the numerical test cases presented below, the CFL number is defined as:

$$CFL = max\left(\frac{(\sqrt{u^2+v^2}+c)\Delta t}{d}\right), \tag{2.37}$$

where d is the distance between the geometric center of a triangle and its neighbor. The parameters u and v are the velocity components in the triangle, and c is the wave speed.

Note that the proposed method solves exactly stationary solutions corresponding to water at rest (h =Constant and $u = v = 0$) for systems (2.1) and (2.2). If this solution is considered as the initial condition, the operations of interpolation introduced in Section 2.3.2, for the spatial scheme of the proposed method, preserve the constant solutions and lead to flux equaling zero. The splitting before and after treatment of the flux does not affect the solution, since the condition $u = v = 0$ gives the same solution for

equations (2.33) and (2.34). Then the parameters $\mathbf{f}^n$, $\mathbf{f}^*$, and $\mathbf{f}^{pred}$ which are used in the fourth-order Adams-Bashforth predictor formula and the Adams-Moulton corrector formula are zero (equations (2.35) and (2.36)). Similarly, $k_i = 0$, $i = 1, 2, 3, 4$ are the coefficients used in RK4 (equations (2.60) and (2.61) in Appendix II.2), which is used to start the method in the first three steps. Then the constant solutions are preserved, which is also confirmed by the numerical results (not shown). In the numerical tests presented below, we also examine the well-balanced property for non-trivial steady state solutions. In particular, the steady state solutions in a moving frame, such as the exact solutions of the linear beta-plane equatorial shallow water equations, are considered.

2.4 Numerical experiments for symmetric linear equatorial Rossby waves

In this section, we perform a test for linear beta-plane equatorial shallow water equations by using large-scale equatorial waves, and test the ability of the proposed method to capture the slow waves. In particular, we consider the symmetric equatorial Rossby waves of Index 1, which are exact solutions of the linear beta-plane equatorial shallow water equations.

2.4.1 Symmetric Equatorial Rossby waves of Index 1

Equatorial Rossby waves, also called planetary waves, are found near the equator. They propagate westward and are slow (low-frequency) and long. These waves play an important role in the transfer of energy in the ocean and atmosphere. For the equatorial β-plane approximation $f = \beta y$, Rossby waves are exact solutions of linear SWEs. These solutions are in a steady state in a moving frame. When these solutions are considered as the initial condition, they must be preserved by an ideal numerical method. However, most well-known schemes fail in preserving these solutions. In this section, we give some details about the analytical solution of SWEs corresponding to the symmetric equatorial Rossby waves of Index 1, which will be used as the initial condition to test the proposed method.

We consider a Rossby wave with wavelength $L_w = 5500\,km$ and a domain $[0, 2L_w] \times [0, L_w]$. The mean depth and the reduced gravity are taken as $H = 300\,m$ and $g^{'} = 3 \times 10^{-2}\,ms^{-2}$, respectively. Using the characteristic depth, length, time, and velocity values given by equations (2.4) in Section 2.2.1, we obtain a dimensionless domain $[0, 2L] \times [0, L]$ with $L = 16$ and a dimensionless wavelength $X = 16$. In the following, for simplicity, we drop the ˜ sign for dimensionless parameters.

The analytical solution is given by:

$$\begin{cases} v(x,y,t) = -\bar{r}y\cos(kx-\omega t)e^{-y^2/2} \\ r^+(x,y,t) = \bar{r}\dfrac{2y^2-1}{k-\omega}\sin(kx-\omega t)e^{-y^2/2} \\ r^-(x,y,t) = \bar{r}\dfrac{-1}{k+\omega}\sin(kx-\omega t)e^{-y^2/2} \\ \eta(x,y,t) = \dfrac{r^-(x,y,t)-r^+(x,y,t)}{2} \\ u(x,y,t) = -\dfrac{r^-(x,y,t)+r^+(x,y,t)}{2} \end{cases} \tag{2.38}$$

where $r^+ = -u - \eta$ and $r^- = -u + \eta$ are the Riemann invariants, $k = 2\pi/X$ is the wavenumber, and ω is the smallest root in magnitude of the dispersion relation:

$$\omega^2 - k^2 - \frac{k}{\omega} = 3 \tag{2.39}$$

The dispersion relation (2.39) for $X = 16$ leads to $\omega = -0.12512089$. The dimensionless wave period is thus $\tilde{T} = 50$, which corresponds to a period of $70\,days$ in a real scale. The wave amplitude can be selected by setting the value of the constant $\bar{r}$. Figure 2.3 shows the analytical form of the Rossby waves ($\eta/\bar{r}$) at time $t = 500$ (≈ 23 months).

2.4.2 Performance of the proposed method

a. Analysis of the accuracy

In this section, the analytical solution of SWEs corresponding to the symmetric equatorial Rossby waves of Index 1 is used as the initial condition to test the proposed method. The results are compared with those of the κ schemes using a triangular mesh with a cell area of $1100\,km^2$. When κ schemes are combined with a fourth-order Adams method as a temporal integration scheme, the results are not accurate for Rossby waves, and a high level of numerical diffusion is observed. Figure 2.4 shows the water surface elevation using the Quick scheme, which is the best one among κ schemes, for Rossby waves at time $t = 250$ (five periods). A high level of damping is already observed at this simulation time, which shows that the κ schemes are not accurate in the modeling of Rossby waves.

Figures 2.5 and 2.6 show the water surface elevation using the proposed method for Rossby waves at time $t = 500$. A great improvement is observed compared with Figure 2.4, while the symmetric form of the flow is still well preserved. The analytical solutions of water surface elevation for the Rossby waves and water surface elevation using other upwind methods are also shown in Figure 2.6. As can be observed in those figures, while the κ schemes have highly damped the waves, the proposed method leads to

accurate results for Rossby waves, and the damping and phase errors are negligible.

We will use the algorithm used by Bona et al. (1995) to obtain the spatial and temporal orders of accuracy. To determine numerically the spatial convergence rate we use very small time step Δt in order to render the temporal errors negligible. Different sizes of the computational cells are used to obtain the numerical value of the spatial order of the proposed method. To measure the errors we will use the mesh-size weighted L^2-norm, denoted by $\|.\|_{L^2}$. Figure 2.7 shows the L^2 error in log-log scale for the test of Rossby waves at time $t = 250$, and we obtain an order of spatial accuracy of 2.74. The combined effects of spatial and temporal errors are in general difficult to distinguish. For a fixed size of mesh $\Delta x = \sqrt{\Omega_m}$, we consider a reference solution at time t which can be obtained by using a small time step Δt_{ref}. This reference solution will differs from the exact solution by an error that is almost purely from the spatial discretization. This solution is used to cancel the spatial errors. Then for a fixed spatial size Δx of the computational cell, we define a modified error at time t denoted by $E^*(t)$ associated to the values of Δt that are larger than Δt_{ref}, as follows:

$$E^*(t) = \|h^{(n)}(\Delta x, \Delta t) - h^{(n)}_{ref}(\Delta x, \Delta t_{ref})\|_{L^2} / \|h^{(a)}\|_{L^2},$$

where $h^{(n)}$ and $h^{(a)}$ are respectively the numerical and the analytical water depth at time t obtained by using the spatial step Δx and the temporal step Δt. The reference solution $h^{(n)}_{ref}$ is the numerical water depth at the same time t which is obtained by using the reference time step Δt_{ref} and the same spatial step Δx. For small values of Δt which are larger than Δt_{ref}, the temporal rate of convergence can be visible because when we subtract the reference solution $h^{(n)}_{ref}(\Delta x, \Delta t_{ref})$ from the approximate solution $h^{(n)}(\Delta x, \Delta t)$, the spatial errors are almost canceled. The temporal rate of convergence at time $t = 250$ is shown in Table 2.1 for $\Delta t_{ref} = 0.001$. The results confirm that the splitting method has less impact on the fourth order of the accuracy of the Adams method used as temporal scheme.

Symmetric Equatorial Rossby waves of Index 1, used in this section, are harder to capture by the numerical methods compared to Kelvin and Yanai waves commonly used in the numerical tests for well-balanced property (Khouider and Majda, 2005). Following the accuracy of the proposed method for this type of waves, we conclude that the balance between the flux and Coriolis terms is preserved.

b. Energy conservation

We now study the behavior of the proposed method and κ schemes with respect to numerical energy dissipation. Since the waves used in this section are steady state in a moving frame, they should be preserved by the numerical methods. In particular, their kinetic energy should be also preserved. Figure 2.8 shows the change in kinetic energy up to time $t = 500$ using the proposed method and κ schemes. The upwind-centered

scheme shows large oscillations in kinetic energy. For the other κ schemes the energy decreases rapidly. At time $t = 250$, the energy of the Fromm scheme is reduced to 50%, while for the proposed method at this time, the rate of energy dissipation is negligible (less than 1%). Therefore, the proposed method performs very well in the conservation of kinetic energy, even for long simulation times. This is of crucial importance in the simulation of large-scale oceanic and atmospheric flows, where Rossby waves play an essential role in the transfer of energy.

2.5 Numerical experiments for nonlinear cases

In this section, in order to validate the performance of the proposed method for nonlinear SWEs, some nonlinear test cases are presented. The tests are performed for nonlinear gravity waves, parabolic flood waves and nonlinear Rossby soliton waves. Finally, the effect of the mesh structure on the solution quality is analyzed using different grids.

2.5.1 Nonlinear Gravity waves

Here, nonlinear gravity waves are considered in order to test the proposed method. We consider a non-dimensional domain $[-L/2, L/2] \times [-L/2, L/2]$, with $L = 150$. A Gaussian distribution of water surface elevation is assumed as the initial condition for the dimensionless form of the nonlinear SWEs (2.1) and (2.5) without Coriolis effect.

$$\begin{aligned} \eta(x, y, 0) &= 0.05e^{-0.1(x^2+y^2)}, \\ u(x, y, 0) &= 0, \\ v(x, y, 0) &= 0 \end{aligned} \tag{2.40}$$

Figure 2.9 shows the solution for the proposed method at time $t = 50$ with $CFL = 0.4$ and cell area $\Omega_m = 2.25$. Since no explicit exact solution is available for this case, we use the third-order upwind scheme combined with the TVDRK3 method and a fine mesh with cell area $\Omega_m = 1$ to compute a reference solution. This scheme is chosen in order to obtain a reference solution for gravity waves only, since it performs well for this type of wave, but it leads to a high level of damping for Rossby waves, as shown in Section 2.4 for all κ schemes. As seen in Figures 2.9 and 2.10, the proposed method leads to good results; the solution remains symmetric, free of numerical oscillations, and accurate over the entire domain.

2.5.2 Parabolic flood waves

In this section, the proposed method is analyzed using an exact solution which corresponds to time-dependent flows for non linear SWEs. The analytical solution given in

Thacker (1981) for parabolic flood waves is employed. This solution is written as:

$$\begin{aligned} u(x,y,t) &= \frac{xt}{t^2+T^2}, \\ v(x,y,t) &= \frac{yt}{t^2+T^2}, \\ h(x,y,t) &= h_0[\frac{T^2}{t^2+T^2} - \frac{x^2+y^2}{R_0^2}(\frac{T^2}{t^2+T^2})^2] \end{aligned} \tag{2.41}$$

where h_0 is the initial height of the peak of the parabolic water surface and the parameters R_0 and T satisfy the following equation

$$T = R_0(2gh_0)^{-1/2} \tag{2.42}$$

In this test, a non-dimensional domain $[-6,6] \times [-6,6]$ is considered with $h_0 = 1$ and $R_0 = 14$. The initial condition is given as:

$$\begin{aligned} u(x,y,0) &= v(x,y,0) = 0, \\ h(x,y,0) &= h_0(1 - \frac{x^2+y^2}{R_0^2}) \end{aligned} \tag{2.43}$$

The exact solution is used to calculate the boundary conditions of the domain. Figure 2.11 shows the temporal evolution of the numerical solution using the proposed method and the analytical solution along section $y = 0$ using the cell area $\Omega_m = 0.01$. The $3D$ views of the solution (not shown here) show that the symmetric form of the flow is still well preserved. Following those tests, we conclude that the proposed method produces stable results while maintaining good accuracy for long time steps until the solution reaches its asymptotic convergence.

2.5.3 Nonlinear Rossby soliton waves

In this part, we present a test for an equatorial non-linear Rossby wave. This wave is driven by gravity and rotational forces. Asymptotic solutions of the system are approximated by using Boyd's previous work on equatorial Rossby soliton waves (Boyd, 1980, 1985). A non-dimensional domain $[-24,24] \times [-8,8]$ is considered in the present test. The boundary conditions used are non-flux along the walls. Equations (2.5) and (2.6) are considered by using the parameters $h = H + \eta$ with $H = 1$, $g = 1$, and $f(y) = y$. The approximate asymptotic solution is obtained with a first-order expansion of function with an order of 12 for Hermite polynomials, as explained in Boyd (1985). The simulations are performed up to $t = 40$ with $CFL = 0.3$ and using the cell area $\Omega_m = 0.04$. Even though the asymptotic analytical solution is only a first-order approximation, this solution can be used to check the phase speed of the wave and the conservation of total energy.

Figure 2.12 shows the free surface elevation $\tilde{\eta}$ of the nonlinear Rossby soliton wave in three dimensions for the proposed method at time $t = 40$. At this time, the analytical solution predicts that the peak of the wave will be $\tilde{\eta}_{max} = 0.156$ and for the proposed method the peak is $\tilde{\eta}_{max} = 0.154$. The proposed method preserves the shape of each part of the soliton and gives the correct position and phase speed of propagation of the wave compared to the asymptotic solution of Boyd (1985). For the first part of the soliton $\tilde{x}$-position $= -15.80$ and $\tilde{y}$-position=1.267, and for the second part of the soliton $\tilde{x}$-position $= -15.80$ and $\tilde{y}$-position=-1.267 .

Finally, Figure 2.13 presents the evolution of total energy and shows that the proposed method performs well for the nonlinear Rossby soliton wave in conservation of energy, which is a crucial requirement in atmospheric and oceanic simulations.

2.5.4 Effect of Grid Structure

In this section the effect of the mesh structure on the solution quality is analyzed. To characterize the mesh shapes, we choose a skewness parameter defined as:

$$Skewness = \frac{Optimal\,Cell\,Size - Cell\,Size}{Optimal\,Cell\,Size}, \tag{2.44}$$

where for a triangle grid the optimal cell size is the size of an equilateral triangle with the same circumscribed circle. The skewness parameter is widely used as an indicator of grid quality (e.g. Roldán et al., 2013). For an excellent mesh, the skewness parameter should be in the range values of $0-0.25$, and for good and acceptable grids respectively, it should be in the ranges of $0.25 - 0.50$ and $0.50 - 0.75$. In this section, we consider three triangular meshes, shown in Figure 2.14, with different grid qualities according to the skewness values. The first case (a) is for an acceptable triangular grid, the second case (b) represents a good mesh, and the third case is an excellent mesh, based on the skewness parameter. The tests are performed using parabolic flood waves until time $t = 100$ to verify the impact of the three meshes on the numerical results. The relative true errors of the water depth for the three cases at time $t = 100$ are presented in Table 2.2. The results demonstrate that for all three cases, the proposed scheme leads to a high level of accuracy and that the best result is obtained for case (c), which has a small skewness coefficient.

2.6 Conclusions

In this paper, a new numerical scheme using unstructured grids is developed for shallow water flows dominated by rotation effects. This method leads to accurate results for both gravity and Rossby waves, which is of crucial importance in the simulation of large-scale flows. A special treatment of the Coriolis effect was employed in which the

Coriolis term is integrated analytically before and after solving the conservation law. This method uses polynomial fitting with high accuracy on both sides of the interface of the computational cells. A fourth-order Adams method with an operator splitting scheme is used for temporal integration. This approach is enough to suppress the short-wave numerical noise without damping the long waves that are essential in the transport of energy Rossby waves in the ocean and atmosphere. The energy conservation of the scheme is considerably improved compared with other upwind schemes, such as the third-order scheme, the cell-based third-order scheme, the Fromm method, and the Quick scheme, which are widely used for simulation of gravity waves and lead to high levels of energy dissipation and numerical oscillation caused by the imbalance between the source and flux terms at the discrete level for Rossby waves. Finally, the employed fourth-order Adams method was found to be an accurate and efficient scheme for the time integration of large-scale shallow water equations, with no iteration on the corrector step being needed to stabilize the method, and optimal results are obtained with only one iteration.

Appendix I: The Jacobian matrix for SWEs

I.1 The Jacobian matrix for linear equations

The matrix $\widetilde{\mathbf{J}}$ satisfies $\Delta\mathbf{F} = \widetilde{\mathbf{J}}\ \Delta\mathbf{U}$ with:

$$\widetilde{\mathbf{J}} = \frac{\partial\left(\mathbf{F}\cdot\mathbf{n}\right)}{\partial\mathbf{U}} = \begin{pmatrix} 0 & Hn_x & Hn_y \\ gn_x & 0 & 0 \\ gn_y & 0 & 0 \end{pmatrix} \tag{2.45}$$

where $c = \sqrt{gH}$. The eigenvalues of $\widetilde{\mathbf{J}}$ are given by:

$$\tilde{a_1} = c, \quad \tilde{a_2} = 0, \quad \tilde{a_3} = -c, \tag{2.46}$$

with the corresponding eigenvectors

$$\tilde{\mathbf{e}}_1 = \begin{pmatrix} 1 \\ \lambda_0 n_x \\ \lambda_0 n_y \end{pmatrix} \quad , \quad \tilde{\mathbf{e}}_2 = \begin{pmatrix} 0 \\ -\lambda_0 n_y \\ \lambda_0 n_x \end{pmatrix} \quad , \quad \tilde{\mathbf{e}}_3 = \begin{pmatrix} 1 \\ -\lambda_0 n_x \\ -\lambda_0 n_y \end{pmatrix}, \tag{2.47}$$

where

$$\lambda_0 = \sqrt{\frac{g}{H}} \tag{2.48}$$

The coefficients $\tilde{\alpha_k}$, k=1,2,3, are computed as:

$$\begin{cases} \tilde{\alpha_1} = \dfrac{\Delta h}{2} + \dfrac{1}{2\lambda_0}\left[\Delta u\ n_x + \Delta v\ n_y\right] \\ \tilde{\alpha_2} = \dfrac{1}{\lambda_0}\left[\Delta v\ n_x - \Delta u\ n_y\right] \\ \tilde{\alpha_3} = \dfrac{\Delta h}{2} - \dfrac{1}{2\lambda_0}\left[\Delta u\ n_x + \Delta v\ n_y\right] \end{cases} \tag{2.49}$$

I.2 Non-dimensional linear system

The flux and source vectors in the dimensionless form become:

$$\mathbf{E} = \begin{bmatrix} u \\ \eta \\ 0 \end{bmatrix} \quad , \quad \mathbf{G} = \begin{bmatrix} v \\ 0 \\ \eta \end{bmatrix} \quad , \quad \mathbf{S} = \begin{bmatrix} 0 \\ yv \\ -yu \end{bmatrix} \tag{2.50}$$

The eigenvalues and eigenvectors of the nondimensional system can be calculated using those of the original system by simply setting $H = g = 1$.

I.3 The Jacobian matrix for nonlinear equations

The Jacobian matrix $\widetilde{\mathbf{J}}$ for the nonlinear case is given by:

$$\mathbf{J} = \frac{\partial\left(\mathbf{F}\cdot\mathbf{n}\right)}{\partial\mathbf{U}} = \begin{pmatrix} 0 & n_x & n_y \\ (\tilde{c}^2 - \tilde{u}^2)n_x - 2\tilde{u}\tilde{v}n_y & 2\tilde{u}n_x + \tilde{v}n_y & \tilde{u}n_y \\ -\tilde{u}\tilde{v}n_x + (\tilde{c}^2 - \tilde{v}^2)n_y & \tilde{v}n_x & \tilde{u}n_x + 2\tilde{v}n_y \end{pmatrix} \tag{2.51}$$

with the following eigenvalues and eigenvectors

$$\tilde{a}_1 = \tilde{u}n_x + \tilde{v}n_y + \tilde{c}, \quad \tilde{a}_2 = \tilde{u}n_x + \tilde{v}n_y, \quad \tilde{a}_3 = \tilde{u}n_x + \tilde{v}n_y - \tilde{c}, \tag{2.52}$$

$$\tilde{e}_1 = \begin{pmatrix} 1 \\ \tilde{u} + \tilde{c}n_x \\ \tilde{v} + \tilde{c}n_y \end{pmatrix} \quad , \quad \tilde{e}_2 = \begin{pmatrix} 0 \\ -\tilde{c}n_y \\ \tilde{c}n_x \end{pmatrix} \quad , \quad \tilde{e}_3 = \begin{pmatrix} 1 \\ \tilde{u} - \tilde{c}n_x \\ \tilde{v} - \tilde{c}n_y \end{pmatrix}, \tag{2.53}$$

where

$$\tilde{u} = \frac{u_R\sqrt{h_R} + u_L\sqrt{h_L}}{\sqrt{h_R} + \sqrt{h_L}}, \tilde{v} = \frac{v_R\sqrt{h_R} + v_L\sqrt{h_L}}{\sqrt{h_R} + \sqrt{h_L}}, \tilde{c} = \sqrt{g(h_R + h_L)/2}, \tag{2.54}$$

The coefficients $\tilde{\alpha_k}$ depend on the jumps $\Delta(.) = (.)_R - (.)_L$ as:

$$\begin{cases} \tilde{\alpha}_1 = \dfrac{\Delta h}{2} + \dfrac{1}{2\tilde{c}}(\Delta(hu)n_x + \Delta(hv)n_y - (\tilde{u}n_x + \tilde{v}n_y)\Delta h) \\ \tilde{\alpha}_2 = \dfrac{1}{\tilde{c}}((\Delta(hv) - \tilde{v}\Delta h)n_x - (\Delta(hu) - \tilde{u}\Delta h)n_y) \\ \tilde{\alpha}_3 = \dfrac{\Delta h}{2} - \dfrac{1}{2\tilde{c}}(\Delta(hu)n_x + \Delta(hv)n_y - (\tilde{u}n_x + \tilde{v}n_y)\Delta h) \end{cases} \tag{2.55}$$

Appendix II : Time integration methods

The following time integration methods are used to solve an ODE defined by:

$$\frac{d\mathbf{U}}{dt} = \mathbf{f}(\mathbf{U}, t),\ \mathbf{U}(t_0) = \mathbf{U}^0 \tag{2.56}$$

We use $\mathbf{U}^n$ to denote the computed approximation to the solution at time $t_n = n\Delta t$

II.1 The fourth-order Adams method

The fourth-order Adams method uses the fourth-order Adams-Bashforth scheme as predictor:

$$\mathbf{U}^{\Upsilon} = \mathbf{U}^n + \Delta t\left(\frac{55}{24}\mathbf{f}^n - \frac{59}{24}\mathbf{f}^{n-1} + \frac{37}{24}\mathbf{f}^{n-2} - \frac{9}{24}\mathbf{f}^{n-3}\right) \tag{2.57}$$

and the fourth-order Adams-Moulton scheme as corrector

$$\mathbf{U}^{n+1} = \mathbf{U}^n + \Delta t\left(\frac{9}{24}\mathbf{f}^{\Upsilon} + \frac{19}{24}\mathbf{f}^n - \frac{5}{24}\mathbf{f}^{n-1} + \frac{1}{24}\mathbf{f}^{n-2}\right) \tag{2.58}$$

where

$$\mathbf{f}^n = \mathbf{f}(\mathbf{U}^n, n\Delta t) \tag{2.59}$$

II.2 The fourth-order Runge-Kutta method

The value $\mathbf{U}^{n+1}$ is calculated using the following formula:

$$\mathbf{U}^{n+1} = \mathbf{U}^n + \frac{\Delta t}{6}(k_1 + 2k_2 + 2k_3 + k_4), \tag{2.60}$$

where

$$\begin{aligned} k_1 &= \mathbf{f}(\mathbf{U}^n, t_n) \\ k_2 &= \mathbf{f}(\mathbf{U}^n + k_1\Delta t/2, t_n + \Delta t/2) \\ k_3 &= \mathbf{f}(\mathbf{U}^n + k_2\Delta t/2, t_n + \Delta t/2) \\ k_4 &= \mathbf{f}(\mathbf{U}^n + k_3\Delta t, t_n + \Delta t) \end{aligned}, \tag{2.61}$$

II.3 The third-order TVD Runge-Kutta method

We consider the following ODE

$$\frac{d\mathbf{U}}{dt} = \mathbf{L}(\mathbf{U}),\ \mathbf{U}(t_0) = \mathbf{U}^0, \tag{2.62}$$

where $\mathbf{L}$ is a spatial operator. The TVDRK3 method (Shu and Osher, 1988) is performed via three stages to solve equation (2.62).

$$\begin{aligned}
\mathbf{U}^{(1)} &= \mathbf{U}^n + \Delta t \mathbf{L}(\mathbf{U}^n) \\
\mathbf{U}^{(2)} &= \frac{3}{4}\mathbf{U}^n + \frac{1}{4}\mathbf{U}^{(1)} + \frac{1}{4}\Delta t \mathbf{L}(\mathbf{U}^{(1)}) \\
\mathbf{U}^{n+1} &= \frac{1}{3}\mathbf{U}^n + \frac{2}{3}\mathbf{U}^{(2)} + \frac{2}{3}\Delta t \mathbf{L}(\mathbf{U}^{(2)})
\end{aligned} \tag{2.63}$$

Table 2.1 Temporal rate of convergence of the proposed method at $t = 250$

Δt	6E-03	7 E-03	8 E-03	9 E-03	1 E-02
Rate	4.00	4.00	4.00	3.98	3.95

Table 2.2 L^2 error for various grids

Meshes	Mesh (a)	Mesh (b)	Mesh (c)
Relative true errors	1.55 E-05	1.48 E-05	2.53 E-06

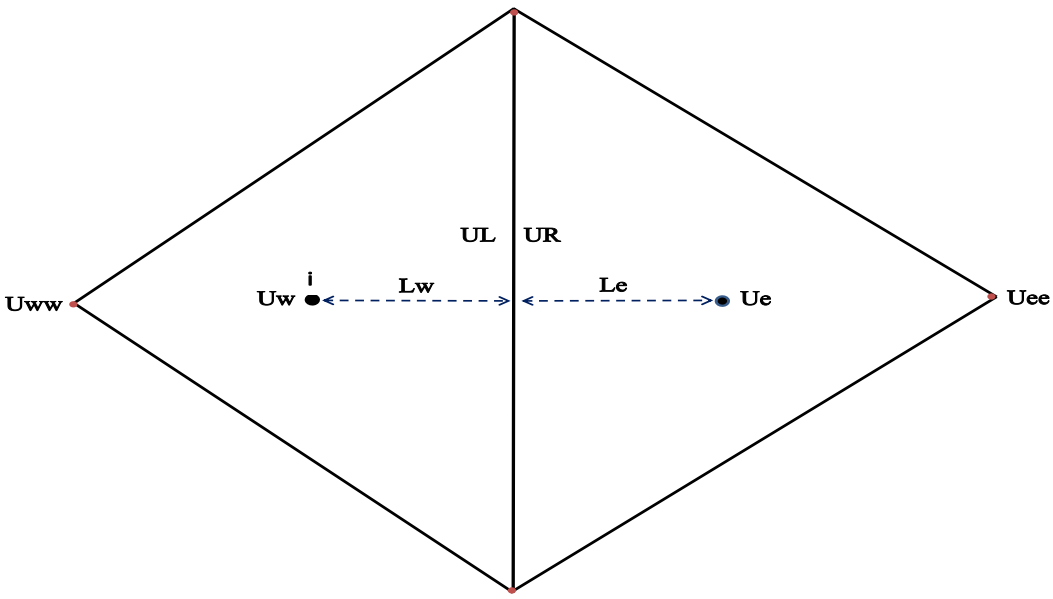

Figure 2.1 Parameters at the right (R) and left (L) sides of the interface of computational cells used in the κ scheme

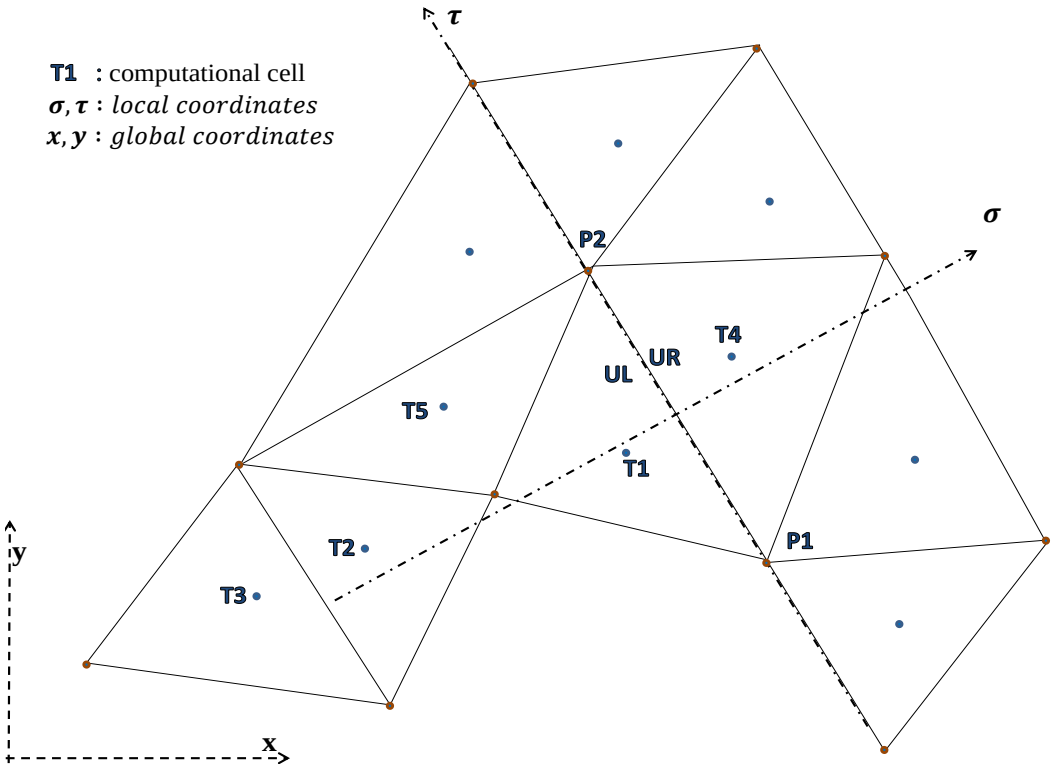

Figure 2.2 Cells used in the proposed method

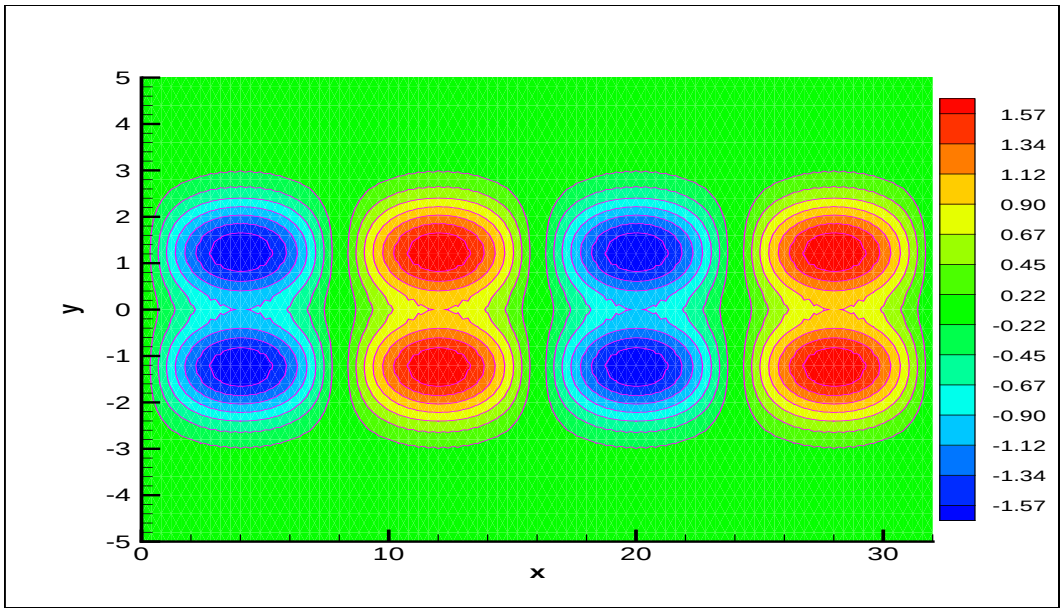

Figure 2.3 Water surface elevation for Rossby waves : the analytical form at time $t = 500$

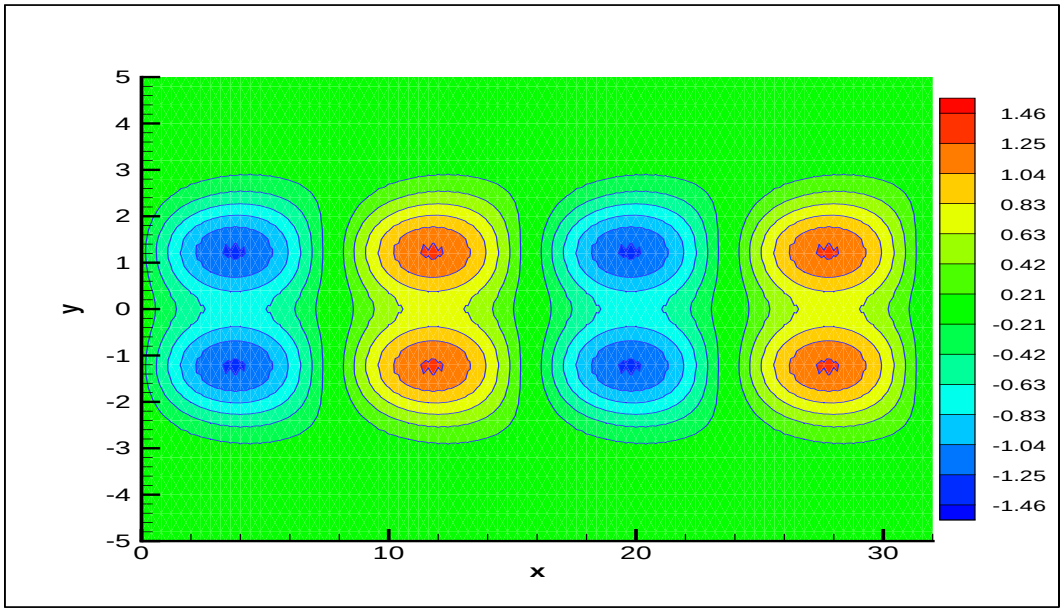

Figure 2.4 Water surface elevation for Rossby waves using the fourth-order Adams method and Quick scheme at time $t = 250$ with $CFL = 0.1$

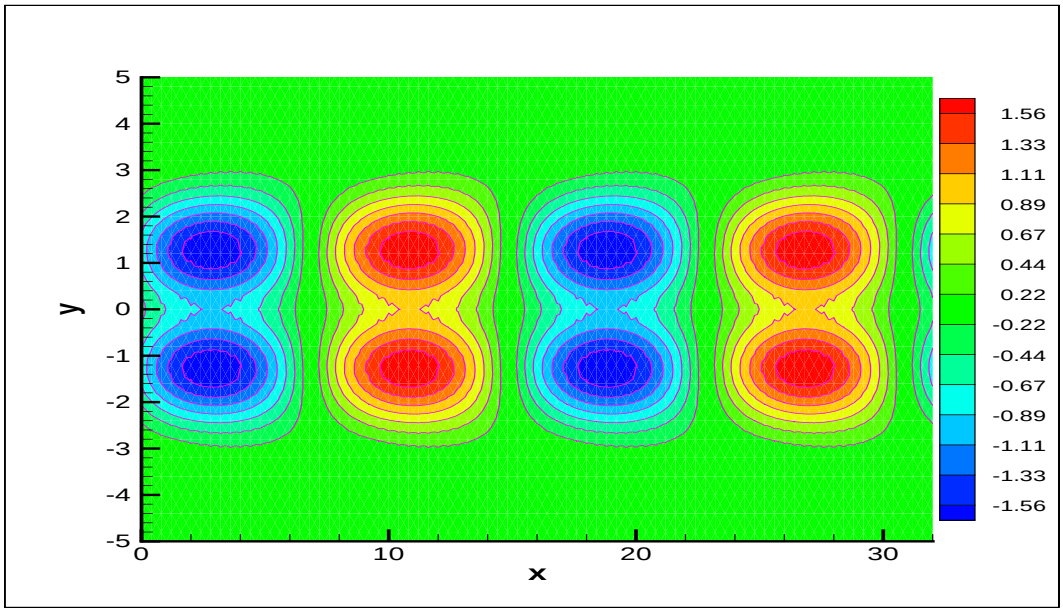

Figure 2.5 Water surface elevation using the proposed method for Rossby waves at time $t = 500$ with $CFL = 0.1$

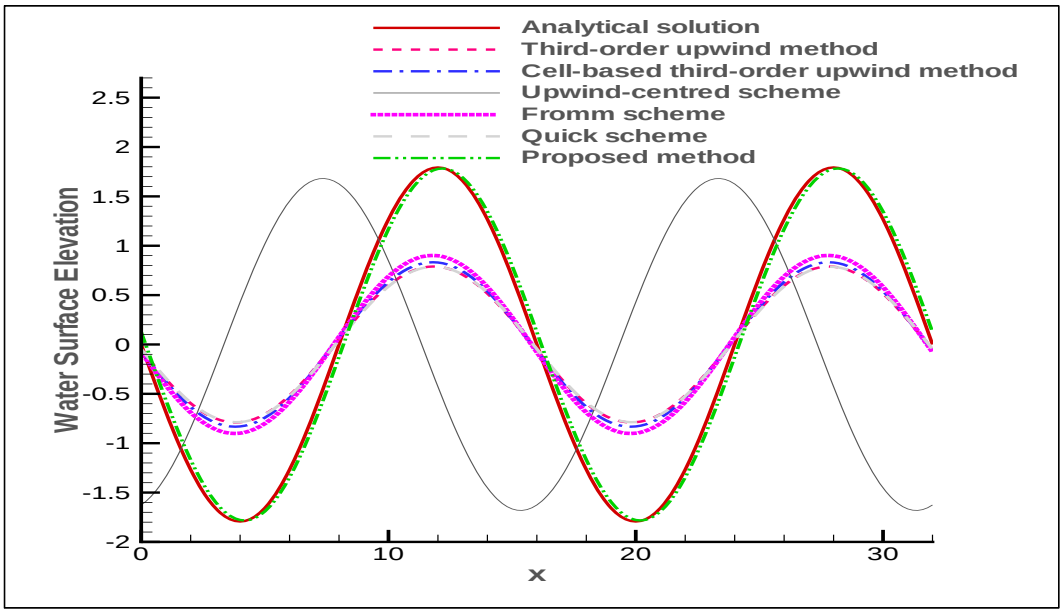

Figure 2.6 Comparison of numerical solutions using the proposed method, κ schemes, and the analytical solution of water surface elevation for Rossby waves at time $t = 500$ with $CFL = 0.1$

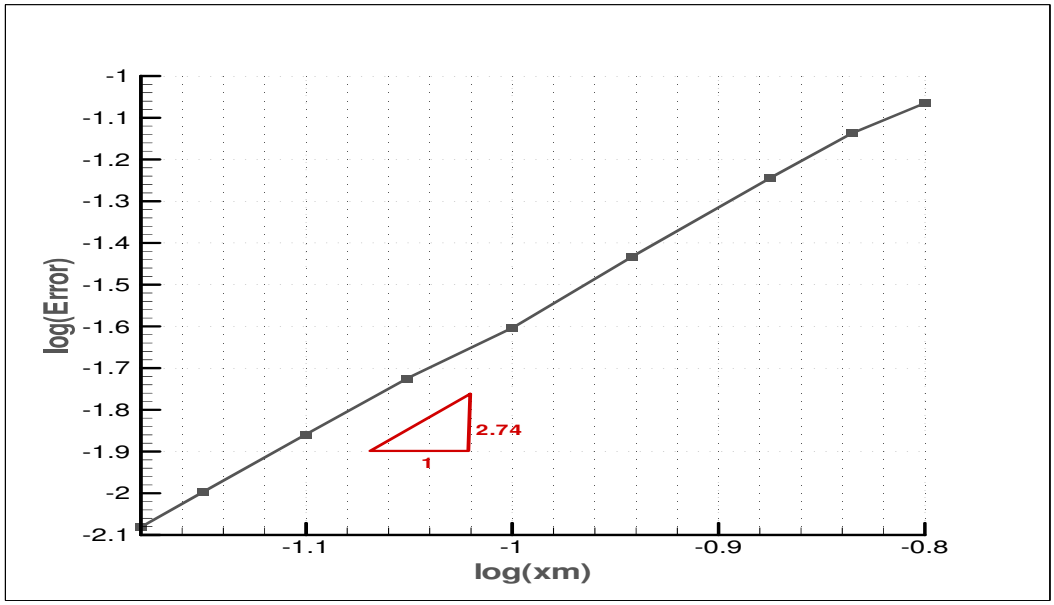

Figure 2.7 Evolution of L^2 error in log-log scale for Rossby waves at time $t = 250$, where $x_m = \sqrt{\Omega_m}$

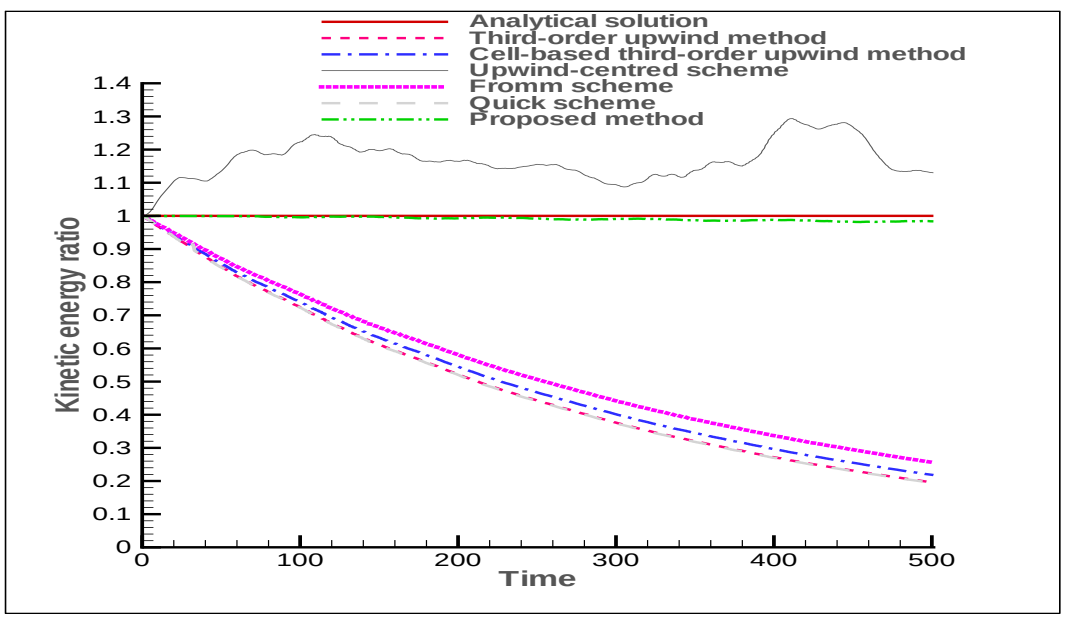

Figure 2.8 Change in kinetic energy for Rossby waves using the proposed method and κ schemes until time $t = 500$ with $CFL = 0.1$

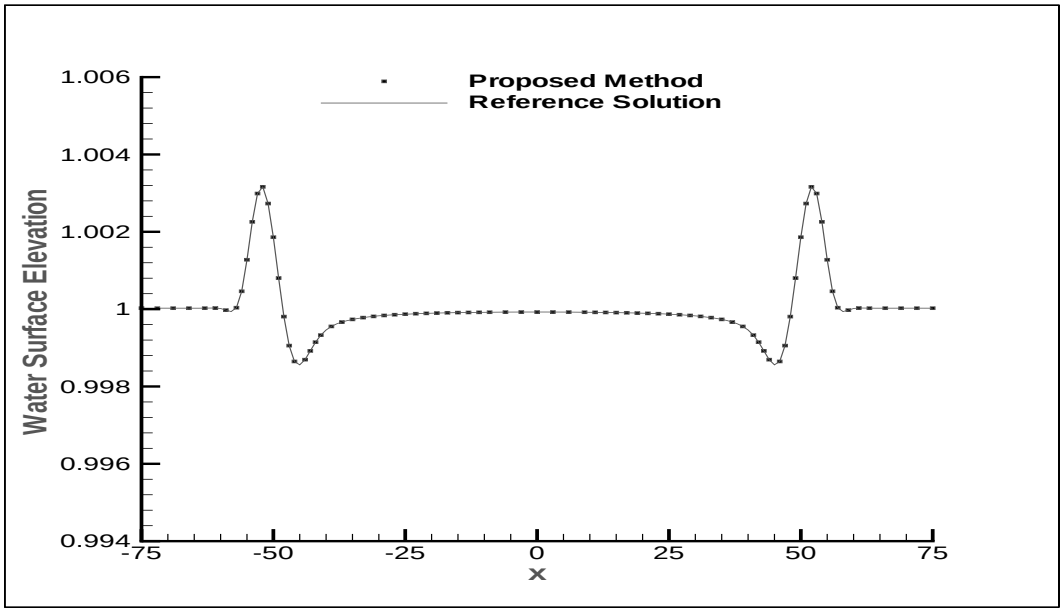

Figure 2.9 Comparison of the solution for the gravity waves using the proposed method and the reference solution at time $t = 50$ with $CFL = 0.4$

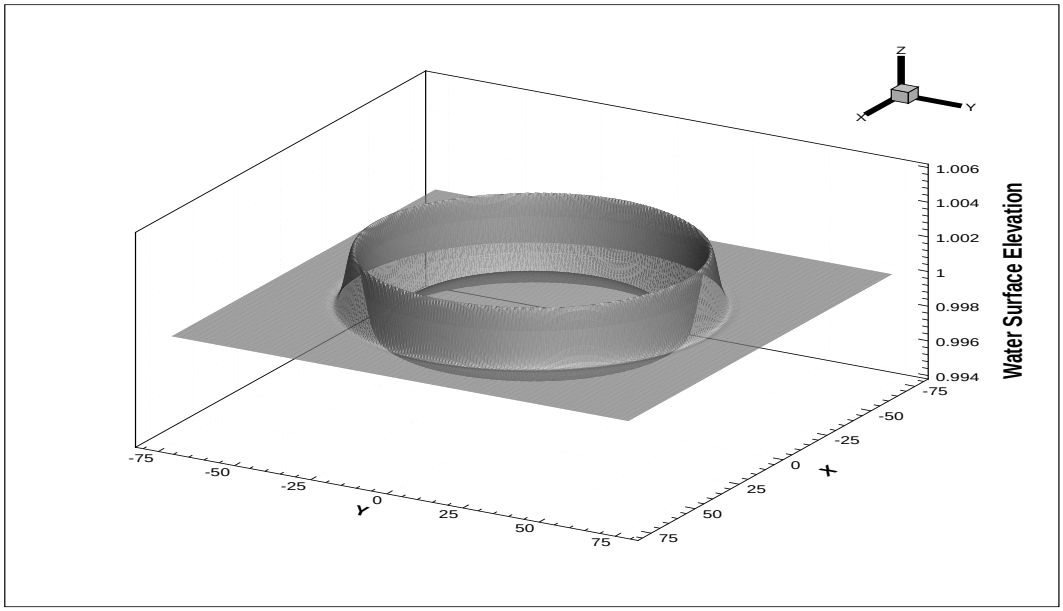

Figure 2.10 A three-dimensional view of water surface elevation using the proposed method at time $t = 50$ with $CFL = 0.4$

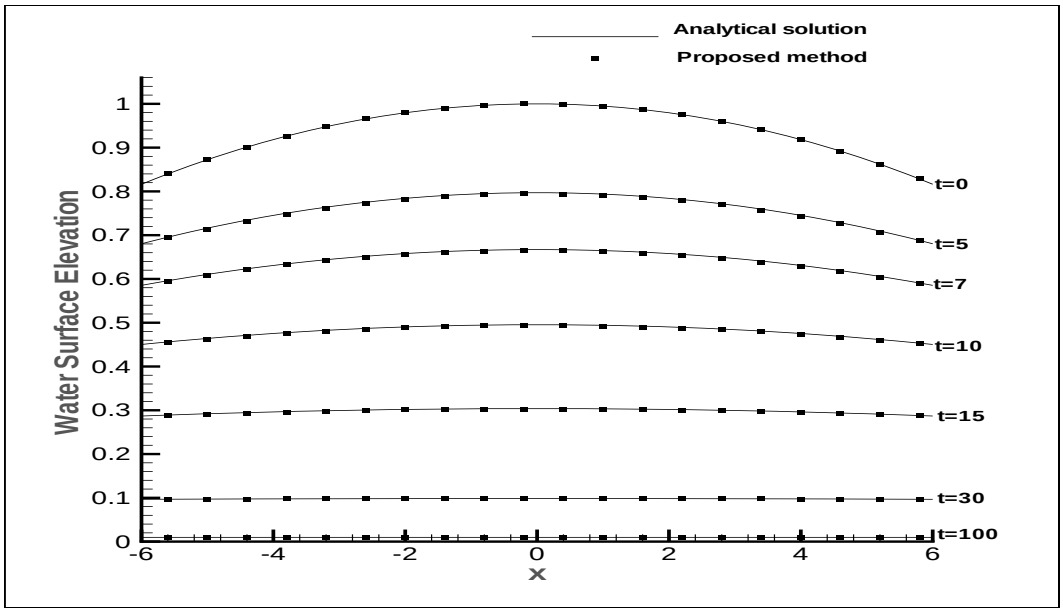

Figure 2.11 Comparison of the solution for a parabolic flood waves using the proposed method and the analytical solution with $CFL = 0.5$

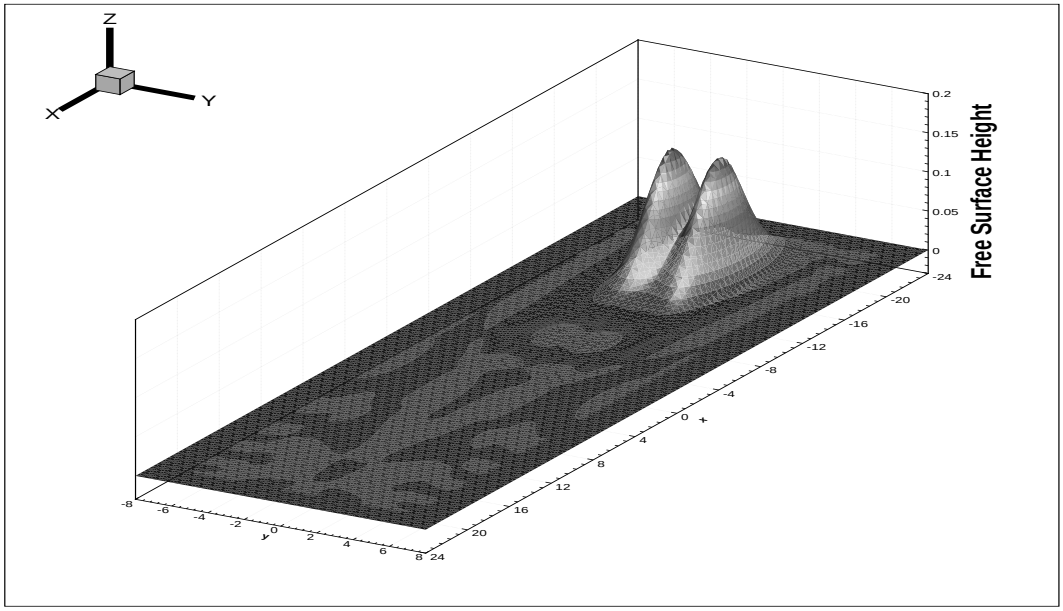

Figure 2.12 Free surface elevation for nonlinear Rossby soliton waves using the proposed method at time $t = 40$ with $CFL = 0.3$

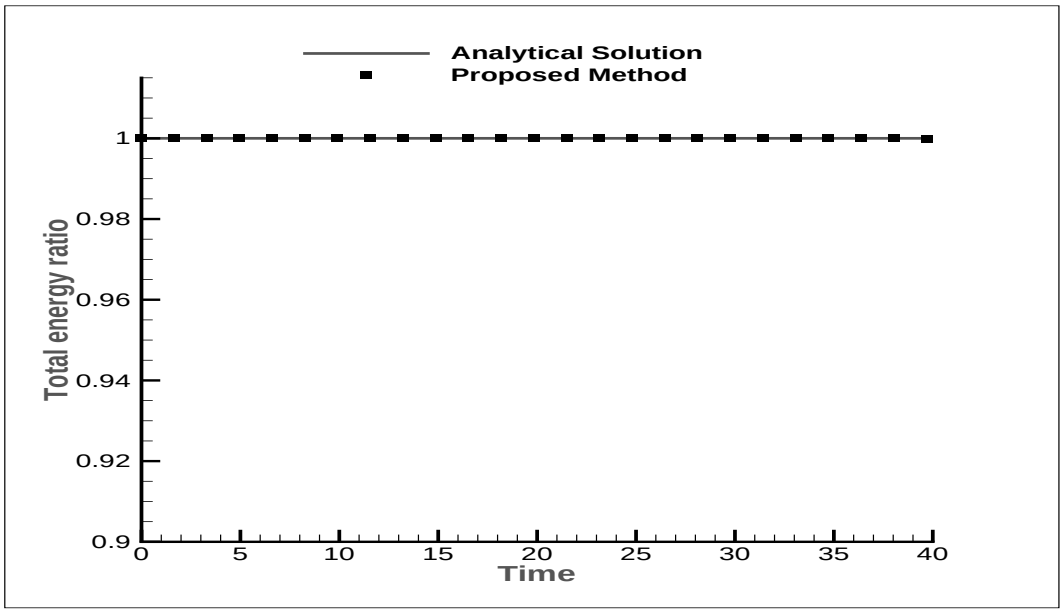

Figure 2.13 Evolution of total energy for nonlinear Rossby soliton waves using the proposed method until time $t = 40$ with $CFL = 0.3$

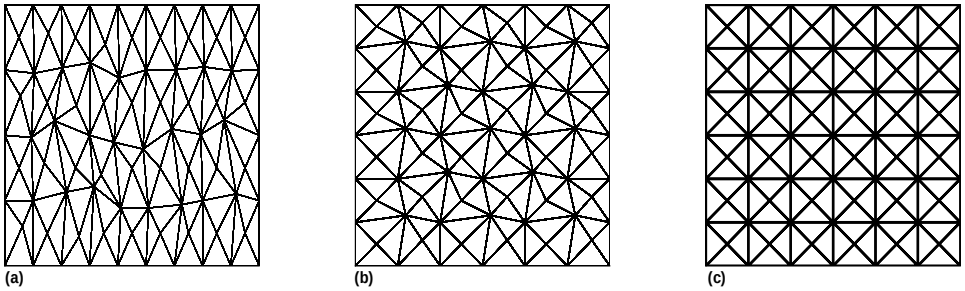

Figure 2.14 Triangular meshes used in the analysis of grid effects

CHAPITRE 3 Analyse théorique et numérique d'une classe de schémas semi-implicites semi-lagrangiens potentiellement applicables aux modèles atmosphériques

Theoretical and numerical analysis of a Class of Semi-implicit Semi-Lagrangian Schemes potentially applicable to Atmospheric Models [1]

Résumé

Le schéma explicite objet du chapitre 2 est efficace pour le cas du système de Saint-Venant avec le terme source qui inclut le paramètre de Coriolis. Néanmoins, dans le cas général où il y a des termes sources dans le système qui génèrent des ondes à hautes fréquences, il est utile d'utiliser des méthodes implicites dans l'intégration temporelle de ces termes afin d'opter pour des pas de temps plus pratiques. Par exemple, les schémas semi-implicites semi-lagrangiens sont largement utilisés pour les modèles atmosphériques où il y a différentes échelles de propagation des ondes. Ces schémas utilisent une décomposition du terme source en une partie linéaire et une partie non linéaire. La partie linéaire génère les ondes rapides dans le système et le reste non linéaire est responsable des ondes lentes. En général, la partie linéaire est traitée par des méthodes implicites tandis que la partie non linéaire est traitée en utilisant des schémas numériques explicites.

Une méthode semi-lagrangienne à deux niveau (SETTLS) ayant une région de stabilité absolue indépendante du nombre de Courant-Friedrichs-Lewy (CFL) a été proposée par Hortal (2002). La plupart des centres météorologiques utilisent cette méthode comme schéma temporel dans leur modèle de prévision numérique atmosphérique. Cependant, la méthode SETTLS peut générer des oscillations pour le traitement du terme non linéaire, surtout pour le cas des solutions qui ont un caractère quasi-oscillatoire. Environnement Canada utilise dans son modèle de prévision numérique atmosphérique la méthode SETTLS dans laquelle des itérations supplémentaires sont parfois ajoutées pour éviter les bruits numériques. Dans l'analyse de stabilité menée par Durran et Reinecke (2004), les auteurs ont remarqué que la zone de stabilité de la méthode SETTLS se réduit à un seul point sur l'axe imaginaire du plan complexe, ce qui constitue un inconvénient pour le cas des solutions à caractère oscillatoire.

Dans ce chapitre, on propose une classe de schémas semi-implicites semi-lagrangiens potentiellement applicables aux modèles atmosphériques. Des analyses théoriques et numériques des propriétés de certaines méthodes semi-lagrangiennes complexes sont

[1] Cet article est réalisé en collaboration avec A. Mohammadian, M. Charron et C. Girard, et publié sous la forme: A. Beljadid, A. Mohammadian, M. Charron, C. Girard, 2014, Theoretical and numerical analysis of a Class of Semi-implicit Semi-Lagrangian Schemes potentially applicable to Atmospheric Models. Monthly Weather Review - American Meteorological Society, vol 142, no. 12, 4458-4476.

menées pour faire face aux problèmes d'instabilité associés au traitement de la partie non linéaire du terme source. L'objectif est de développer un schéma numérique qui possède une zone de stabilité plus large que celle de la méthode SETTLS, en particulier sur l'axe imaginaire du plan complexe.

La classe de schémas semi-implicites semi-lagrangiens développée est basée sur une version modifiée de la méthode TR-BDF2, qui est une combinaison de la règle du trapèze (TR) et de la formule de différenciation en arrière d'ordre deux (BDF2). Dans cette version modifiée de la méthode TR-BDF2, on propose trois étapes. Le terme non linéaire est traité de manière explicite dans les trois étapes en utilisant la méthode du trapèze pour les deux premières étapes et la méthode BDF2 dans la troisième étape. Pour le terme linéaire, la méthode du trapèze implicite est appliquée dans la première étape, la méthode du trapèze explicite est utilisée dans la deuxième étape et la méthode BDF2 implicite est appliquée dans la troisième étape.

La combinaison des nouvelles techniques développées conduit à une famille de schémas numériques qui a une région de stabilité absolue relativement large en comparaison avec la méthode SETTLS. Cette classe de schémas numériques présente de bonnes qualités de stabilité pour le cas des solutions à caractère oscillatoire. Les résultats des tests numériques confirment les performances de cette classe de schémas en termes de stabilité, de précision et de convergence.

Dans tout modèle de circulation générale de l'atmosphère, la partie dynamique et la partie physique sont les plus coûteuses en temps de calcul. La partie dynamique est liée à la résolution implicite du terme linéaire et la partie physique est liée à la résolution explicite du terme non linéaire. Le coût de calcul de la résolution explicite du terme linéaire, ce qui est le cas de la deuxième étape dans les schémas proposés, est négligeable par rapport aux coûts de calcul des deux parties dynamique et physique. L'utilisation de la méthode explicite pour le traitement du terme linéaire dans la deuxième étape rend la famille des schémas numériques proposée moins coûteuse de point de vue temps de calcul. Ces schémas sont concurrentiels en termes de temps de calcul en comparaison avec des méthodes numériques à deux étapes très connues.

La classe de schémas proposée donne de bonnes approximations des solutions pour le cas des pas de temps qui sont larges et qui satisfont le critère de Lipschitz. Ce critère est une condition suffisante pour éviter l'intersection des trajectoires semi-lagrangiennes calculées. Ceci a permis de réduire davantage le coût de calcul de la classe de schémas numériques proposée. Cette classe de schémas présente plusieurs avantages de stabilité, de précision, et de convergence en comparaison à d'autres méthodes semi-implicites semi-lagrangiennes et des schémas semi-implicites de type prédicteur-correcteur très connus.

3.1 Introduction

Semi-Lagrangian (SL) semi-implicit integration methods, first proposed by Robert (1981), have been extensively studied and widely incorporated into atmospheric numerical models. A comprehensive review of the applications of SL methods in atmospheric problems was given by Staniforth and Côté (1991). Other extensive studies have also been conducted to examine SL methods (e.g. Bonaventura, 2000; White-III and Dongarra, 2011). A three-time-level SL method was implemented operationally at the European Center for Medium-Range Weather Forecasts in 1991, and was documented and described in Ritchie et al. (1995). Tanguay et al. (1990) generalized the use of a semi-implicit algorithm in order to integrate the fully compressible nonhydrostatic equations. Ritchie (1991), and Ritchie and Beaudoin (1994) applied the semi-Lagrangian semi-implicit method to a multilevel spectral primitive-equations model.

McDonald and Bates (1987) and Temperton and Staniforth (1987) proposed two-time-level SL methods. This motivated many authors to use two-time-level SL methods instead of three-time-level schemes, since they reduce the number and impact of computational modes (e.g. McDonald and Haugen, 1992; Temperton et al., 2001; Hortal, 2002). The use of more than two time-levels leads in general to computational modes having comparable amplitudes to those of physical modes, which can influence the accuracy of the solution if no efficient technique is used to damp those modes.

Hortal (2002) proposed a two-time-level SL method called the Stable Extrapolation Two-Time-Level Scheme (SETTLS), which has a region with absolute stability independently of the Courant number. Gospodinov et al. (2001) proposed a family of second-order two-time-level SL schemes that contain an undetermined parameter α. Durran and Reinecke (2004) showed that the size of the absolute stability region of the family of second-order schemes proposed by Gospodinov et al. (2001) varies dramatically according to the undetermined parameter, and that the optimal region is obtained for $\alpha = 1/4$, which corresponds to SETTLS. Still, SETTLS is not a perfect choice because the points on the imaginary axis are outside the domain of absolute stability, except for the origin point. Therefore, SETTLS can generate growth parasite for a purely oscillatory nonlinear term depending on the size of the time step.

Based on the analysis done by Dharmaraja (2007), the method TR-BDF2 performs well in terms of stability for solving partial differential equations. The method is able to compute the solutions using a reasonable step-size for stiff problems. This motivated us to use this method to develop the semi-Lagrangian schemes.

In this paper, theoretical and numerical analysis are performed to study the properties of more complex methods. The aim is to try to avoid the problems of instability associated with the treatment of the non-linear part of the forcing term. We propose a class of semi-Lagrangian semi-implicit schemes using a modified TR-BDF2 method. We use two stages as predictor and corrector in the trapezoidal method and one stage for

the BDF2 method. The family of second-order approximations proposed in Gospodinov et al. (2001) is used in the predictor of the TR method. For the linear term we use the implicit trapezoidal method in the first step, the explicit trapezoidal method in the second step, and the implicit BDF2 method in the third step. Following Hortal (2002), the equation of the semi-Lagrangian trajectory used in the SETTLS method is obtained using an average approximation of the acceleration between the departure point and the arrival point. Since the middle point is in the interval where the average approximation of the acceleration is considered, in the proposed method we use the same average of the acceleration to obtain the position of the middle point. An explicit equation is used for this point and the only iterative equation is the one used to obtain the departure point. The potential practical application of the proposed schemes to a weather-prediction model or any other atmospheric model is not analyzed in the current paper. The analysis of the application of the proposed class of schemes to these models will be considered in future studies.

The remainder of the paper is organized as follows. Section 3.2 introduces the notation and explicitly states the problem and expectations of the paper. Section 3.3 gives details on the proposed schemes for the nonlinear term as well as their linear stability. In Section 3.4, the proposed schemes for the linear term are presented as well as their stability analysis. In Section 3.5 the full semi-Lagrangian semi-implicit schemes are presented, and their accuracy, efficiency, and convergence are analyzed and compared to other existing schemes. Finally, Section 3.6 provides some concluding remarks.

3.2 Problem and notation

The forcing source term of the general equation in the framework of the SL method can be split into a fast linear term and a slower, residual nonlinear term as follows:

$$\frac{d\psi}{dt} = N + L. \tag{3.1}$$

We assume that the two terms depend on the field variable $\psi(x,t)$, x and time. At each time $t_n = n\Delta t$, where Δt is the time step, the values of the parameters ψ and the velocity $\mathbf{V}$ are knowns for all times $t_p = p\Delta t$ with $p = 0, 1, 2, ..., n$. The proposed method requires for each arrival point x_j the evaluation of the position at time t_n of the air parcel arriving at a point (x_j, t_{n+1}) and the position of the middle point on the semi-Lagrangian trajectory. The details of the computation of the departure point and the position of the middle point along the semi-Lagrangian trajectory are presented in Appendix II.

In general, for solving Equation (3.1), explicit methods are used for the nonlinear term N, and the linear term L is calculated using implicit methods. The superscripts "+", "0" and "−" are used to denote the variables at time levels $t + \Delta t$, t and $t -$

Δt, respectively. The subscripts A and D are used to denote the parameters at the departure point $x_D = x(t)$ and the arrival point $x_A = x(t + \Delta t)$, respectively. For the discrete form of the schemes, we define $\tilde{x}_j^n$ as an estimate of the departure point of the fluid parcel at time t_n that arrives at (x_j, t_{n+1}). To abbreviate the notation, for any step p, we use $N^p = N(\tilde{x}_j^p, t_p)$ and $L^p = L(\tilde{x}_j^p, t_p)$ as the values of the nonlinear term and the linear term, respectively, at the departure point of the trajectory originating at $(\tilde{x}_j^p, t_p)$ and arriving at (x_j, t_{n+1}). Following Theorem 1 in Gospodinov et al. (2001), the following discrete approximations of Equation (3.1)

$$\frac{\psi_A^+ - \psi_D^0}{\Delta t} = (\frac{3}{4} - \alpha)N_A^0 + (\frac{3}{4} + \alpha)N_D^0 - (\frac{1}{4} - \alpha)N_A^- - (\frac{1}{4} + \alpha)N_D^- + \frac{1}{2}(L_A^+ + L_D^-), \tag{3.2}$$

lead to second-order accuracy for any undetermined parameter α. We use the terminology *α-approximation* for these second-order approximations of the temporal derivative of the function ψ along the interval of endpoints A and D of the SL trajectory. Durran and Reinecke (2004) presented a stability analysis of a class of explicit semi-Lagrangian schemes using Equation (3.2), taking into consideration only the nonlinear term. They showed that the size of the region of absolute stability of the second-order schemes derived from (3.2) varies dramatically according to parameter α. SETTLS corresponds to the optimal region obtained by setting $\alpha = 1/4$. Following the stability analysis of this scheme using von Neumann's method, the imaginary points are outside the domain of absolute stability, except for the origin. In this study, we propose a family of semi-implicit semi-Lagrangian schemes which does not suffer from this disadvantage. We will use, as in the case of the SETTLS method, only one departure point, and we will examine several schemes by using the *α-approximation* for the treatment of the nonlinear term. Three steps are used, where the nonlinear term is treated explicitly in all steps, and the linear term is treated implicitly in the first and third steps and explicitly in the second step. For simplicity of terminology, in the second step we will use the term Trapezoidal method even if a decentring is applied. The proposed techniques for linear and nonlinear terms are combined to obtain the proposed class of semi-Lagrangian semi-implicit schemes, which provides interesting stability properties. Other advantages of this class of schemes will be demonstrated, such as the accuracy, convergence, and efficiency, which are compared to those obtained by using other existing semi-implicit predictor-corrector schemes and semi-Lagrangian semi-implicit methods.

3.3 Explicit scheme for the nonlinear term

3.3.1 The proposed explicit predictor corrector scheme

For the nonlinear term, we propose an explicit semi-Lagrangian scheme based on a modified TR-BDF2 method combining the trapezoidal rule and the second-order back-

ward differentiation formula. Some details about the initial BDF2 method are given in Appendix I. In the modified TR-BDF2 method, two stages are used as predictor and corrector in the trapezoidal method, and one stage for the BDF2 method.

$$\begin{aligned}
&\frac{\psi^* - \psi^n}{\Delta t/2} = \frac{1}{2}(N^n + N_\alpha^{n+1/2}),\\
&\frac{\psi^{n+1/2} - \psi^n}{\Delta t/2} = \theta N^* + (1-\theta)N^n,\\
&\psi^{n+1} - \frac{4}{3}\psi^{n+1/2} + \frac{1}{3}\psi^n = \frac{\Delta t}{3}(2N^{n+1/2} - N^n).
\end{aligned} \tag{3.3}$$

The terms in the right-hand side of each stage correspond to the explicit form for the nonlinear operator. In the first stage, we calculate the predictor value ψ^* by approximating the value of the nonlinear term $N^{n+1/4}$. This value is obtained by using a trapezoidal treatment of N^n and the *α-approximation* of the nonlinear term denoted by $N_\alpha^{n+1/2}$ at the midpoint of the SL trajectory originating at $(\tilde{x}_j^n, t_n)$ and arriving at (x_j, t_{n+1}). In the second stage, the corrector value of $\psi^{n+1/2}$ is calculated by using the trapezoidal method. We compute the value of N^* by using ψ^* obtained from the first stage, and apply a trapezoidal method using the two values N^n and N^* to obtain the source term for this corrector stage. A large domain of absolute stability of the scheme is obtained and, as will be demonstrated below, the decentring parameter θ has significant ability to improve the stability of this scheme for purely oscillatory solutions. The range values of θ will be chosen in order to ensure good accuracy of the proposed scheme.

In the third stage, the BDF2 method is applied to obtain the next value ψ^{n+1} by using an explicit approximation of the nonlinear term N^{n+1} in the right-hand side of the third equation in (3.3), where $N^{n+1/2}$ is computed by using the value of $\psi^{n+1/2}$ obtained from the second stage. The value $N_\alpha^{n+1/2}$ used in the predictor formula is estimated by using the *α-approximation* as

$$N_\alpha^{n+1/2} = (\frac{3}{4}-\alpha)N(x_j,t_n)+(\frac{3}{4}+\alpha)N(\tilde{x}_j^n,t_n)-(\frac{1}{4}-\alpha)N(x_j,t_{n-1})-(\frac{1}{4}+\alpha)N(\tilde{x}_j^n,t_{n-1}), \tag{3.4}$$

where e.g., $N(x,t_n) := N(\psi(x,t_n),x,t_n)$

A family of SL schemes is obtained based on the values of the undetermined parameter α and the decentring parameter θ. As will be confirmed in the next section, the use of the corrector in the second stage improves the stability of the proposed schemes. We will demonstrate that $\alpha = 1/4$ is the optimal choice with the decentring θ in the interval $[0.5, 0.75]$ to provide good accuracy.

3.3.2 Stability analysis of the proposed scheme

Following Durran (2010), the stability of the SL scheme defined by the undetermined parameter α and the decentring parameter θ will be analyzed by using the forced one-dimensional SL equation

$$\frac{d\psi}{dt} = \sigma\psi, \tag{3.5}$$

where $\sigma = \lambda + i\omega$, λ and ω are real constants. The stability will be analyzed using a constant advecting velocity U. Equation (3.5) becomes

$$\frac{d\psi}{dt} = [\frac{\partial}{\partial t} + U\frac{\partial}{\partial x}]\psi = \sigma\psi. \tag{3.6}$$

For any initial condition $\psi(x,0) = f(x)$, the analytical solution is

$$\psi(x,t) = f(x - Ut)e^{\sigma t}. \tag{3.7}$$

We assume a non-amplifying solution by imposing the condition $\lambda \leq 0$. The stability analysis of the SL schemes is examined in terms of the parameter α used in *α-approximation* of the predictor and the decentring parameter θ used in the second step. The analysis is performed using von Neumann's method, which is applied for a single Fourier mode of the form $\psi_j^n = A_k^n e^{i(kj\Delta x)}$ for any point $(j\Delta x, n\Delta t)$. When this mode is substituted in (3.5), letting $z = \lambda\Delta t + i\omega\Delta t$, one obtains after simplification the following equation for the amplification factor A_k

$$A_k^2 + [(\gamma_1 + \gamma_2(\frac{3}{4} + \alpha))e^{-iks} + \gamma_2(\frac{3}{4} - \alpha)]A_k - \gamma_2(\frac{1}{4} - \alpha) - \gamma_2(\frac{1}{4} + \alpha)e^{-iks} = 0, \tag{3.8}$$

where $\gamma_1 = -\frac{1}{24}(2\theta z^3 + (8 + 4\theta)z^2 + 24z + 24)$, $\gamma_2 = -\frac{\theta}{12}(z^3 + 2z^2)$ and $s = x_A - x_D$. The region of stability depends on the Courant number, which is proportional to the parameter s. The region on the $\lambda\Delta t - \omega\Delta t$ plane where the scheme, for the case $\theta = 0.7$, is absolutely stable is plotted in Figure 3.1 for several values of the parameter of approximation α. Since the solutions of (3.8) are periodic in ks, the curves plotted in Figure 3.1 are given for several discrete values of ks covering the whole periodic domain $[0, 2\pi]$. The regions of absolute stability are only slightly sensitive to the values of α, but the part of the imaginary axis inside the region of absolute stability is variable, and the most interesting is for $\alpha = 1/4$, which corresponds to $\omega\Delta t \in [-0.708, 0.708]$. This has a positive influence for stability, and especially for solutions which have a purely oscillatory character. The decentring acts directly on the size of the stable imaginary part as shown in Figure 3.2. Table 3.1 shows the values of $(\omega\Delta t)_{max}$ of the segments $[-(\omega\Delta t)_{max}, (\omega\Delta t)_{max}]$ of imaginary axis which are included in the region of absolute stability depending on the values of the approximation parameter α and the decentring parameter θ. To obtain the proposed class of SL schemes we set $\alpha = 1/4$,

which improves stability. As shown in Table 3.1, the use of this value ensures a large stability zone on the imaginary axis for all non high decentring parameter θ.

To conduct a comparison and in order to show the usefulness of using the trapezoidal corrector iteration, the regions of absolute stability are analyzed in the same way as the proposed class of schemes, but without using any corrector value for the variable $\psi^{n+1/2}$. We will use the two-step method and, first we consider the case in which the value $N_\alpha^{n+1/2}$ is considered in the BDF2 iteration an estimation of $N^{n+1/2}$. The same form as Equation (3.8) is obtained using the parameters $\gamma_1 = -1$ and $\gamma_2 = -z$. For this case, as can be seen in Figure 3.3, the results are similar to those obtained by Durran and Reinecke (2004) for the second-order two-time-level semi-Lagrangian schemes developed by Gospodinov et al. (2001) and in the particular case for the SETTLS method. The region of absolute stability is very sensitive and varies excessively depending on the value of parameter α. For the optimal region of absolute stability, the origin point is the only imaginary part.

In the case of the two-step method in which we consider the value of $N^{n+1/2}$ by using $\psi^{n+1/2}$ obtained from the first stage, a similar form of Equation (3.8) is obtained with $\gamma_1 = -\frac{1}{6}(z^3+4z+6)$ and $\gamma_2 = -\frac{1}{6}(z^3+2z)$. For this case, the region of absolute stability is still largely reduced compared to the region of absolute stability of the proposed class of schemes. This two-step method also presents the disadvantage of being stable on the imaginary axis of $\lambda\Delta t - \omega\Delta t$ plane at only a single origin point.

3.3.3 Accuracy of oscillatory solutions

In order to learn some basic information about the accuracy of the proposed class of schemes for oscillatory solutions, the following equation is analyzed in this section:

$$\frac{d\psi}{dt} = i\omega\psi, \tag{3.9}$$

where ω is a real number. In our analysis we will use the decentring parameter $\theta = 0.7$ in the second step of (3.3). Note that the analytical solution of oscillatory equation (3.9) corresponding to one mode can be obtained by setting $f(x) = e^{ikx}$ and $\sigma = i\omega$ in Equation (3.7), which leads to an analytical amplification factor $A_{an} = e^{i(\omega\Delta t - ks)}$. Therefore, for an accurate scheme, the module of the numerical amplification factor A_k and the relative phase change defined as $\phi_d = \frac{1}{\omega\Delta t}\arctan[\frac{\Im(A_k e^{iks})}{\Re(A_k e^{iks})}]$ should be close to unity. The polynomial equation for the numerical amplification factor can be obtained by setting $z = i\omega\Delta t$ in Equation (3.8). We use the $\omega\Delta t - ks$ plane to plot the parameters studied, where $\omega\Delta t \in [0, 0.7]$ and $ks \in [0, 2\pi]$. In Figure 3.4 we plot the module of the errors $|A_k - A_{an}|$ of the amplification factor, the damping, and the relative phase change. The errors are negligible, as shown in Figure 3.4-a, which is also confirmed in Figure 3.4-b and Figure 3.4-c, where we observe negligible damping and phase errors.

The proposed class of schemes presents the advantage of having a damped computational mode as shown in Figure 3.4-d compared to the computational modes of the class of semi-implicit predictor-corrector schemes developed by Clancy and Pudykiewicz (2013). Note that more comparisons in terms of accuracy between those schemes and the proposed full semi-implicit semi-Lagrangian schemes will be presented in detail in Section 3.5.

3.4 The proposed scheme for the linear term

3.4.1 Treatment of the linear term

We seek a numerical scheme for the linear term which is stable and accurate, in particular for the case of oscillatory solutions. This scheme should have some compatibility with the proposed scheme for the nonlinear term. Since three steps are used for the nonlinear term, in order to reduce the computational cost of the full scheme we try to use two implicit steps and one explicit step for the linear term. We will use the implicit trapezoidal method in the first step, the explicit trapezoidal method in the second step and the implicit BDF2 method in the third step

$$\begin{aligned}
&\frac{\psi^* - \psi^n}{\Delta t/2} = \frac{1}{2}(L^* + L^n),\\
&\frac{\psi^{n+1/2} - \psi^n}{\Delta t/2} = \theta L^* + (1-\theta)L^n,\\
&\psi^{n+1} - \frac{4}{3}\psi^{n+1/2} + \frac{1}{3}\psi^n = \frac{\Delta t}{3}((1+\mu)L^{n+1} - 2\mu L^{n+1/2} + \mu L^n).
\end{aligned} \tag{3.10}$$

The scheme for the linear term, which is the subject of this section, and the scheme for the nonlinear term given in the previous section are considered separately. It is well known that when the two schemes are combined, they interact in a complex way. Following our numerical tests, the use of the same decentring parameter θ in the second step for the schemes used for linear and nonlinear terms is required to avoid a large impact on the accuracy. In the right-hand side of the third equation of (3.10), a family of second-order approximations of the linear term L^{n+1} at time t_{n+1} is considered and the choice of the appropriate value of parameter μ is required.

3.4.2 Stability analysis and choice of parameter μ

First we will consider the case without decentring ($\theta = 0.5$) to determine the appropriate value of parameter μ for both stability and accuracy. In the same way as in the previous section, the stability analysis of the SL schemes (3.10) is examined using von Neumann's method, and a single Fourier mode of the form $\psi_j^n = A_k^n e^{i(kj\Delta x)}$ for any point $(j\Delta x, n\Delta t)$ is considered. Substituting this mode into Equation (3.5) and letting

$z = \lambda\Delta t + i\omega\Delta t$, one obtains the following equation for the numerical amplification factor A_k

$$A_k = e^{-iks}\frac{12-(4\mu-5)z-3\mu z^2}{12-(4\mu+7)z+(1+\mu)z^2}. \tag{3.11}$$

For the proposed schemes, we will try to choose the value of parameter μ which guarantees stability and accuracy, in particular for the purely oscillatory cases. Note that in general the amplification factor of the form

$$A = \pm e^{-iks}\frac{z+c}{z-c}, \tag{3.12}$$

which has the module

$$|A|^2 = \frac{(\lambda\Delta t + c)^2 + (\omega\Delta t)^2}{(\lambda\Delta t - c)^2 + (\omega\Delta t)^2}, \tag{3.13}$$

where c is a positive real number leads to stable results ($\lambda \leq 0$) and the scheme preserves perfectly the purely oscillatory solutions.

In our case we try to have the amplification factor A_k, obtained from (3.11), with the same form as (3.12). Therefore the two second-degree polynomials in the numerator and denominator of Equation (3.11) should have one common root and one root of the opposite sign. This condition leads to $\mu = 0.5$ and we obtain

$$A_k = -e^{-iks}\frac{z+2}{z-2}, \tag{3.14}$$

which leads to a stable scheme that perfectly preserves the oscillatory solutions.

In Figure 3.5-a we plot the module of the amplification factor, which is obtained by using (3.11) for the oscillatory equation (3.9) and without decentring ($\theta = 0.5$), as a function of the parameters μ and $\omega\Delta t$. In this case the schemes are stable for $\mu \leq 0.5$ and we do not observe any damping in the case of oscillatory solutions for $\mu = 0.5$. For general equation (3.5), in Figure 3.5-b we plot for the case $\theta = 0.5$ the module of the amplification factor $|A_k|$ of the solutions for the proposed schemes ($\mu = 0.5$). The schemes are stable and the oscillatory solutions are perfectly preserved. The analytical solutions which correspond to $|A_{an}| = e^{\lambda\Delta t}$ are represented in the same figure by the vertical lines.

Note that in the case in which decentring is considered, the value $\mu = 0.5$ leads to an amplification factor A_k closed to (3.12) with the remainder of order $|z|^2$, where $|z| << 1$. For this case, if we consider the purely oscillatory solutions, the module of the amplification factor can be approximated for $z = i\omega\Delta t$ and under the assumption $|z| << 1$ by

$$|A_k|^2 = 1 - \frac{8(\theta - 1/2)}{3}\frac{(\omega\Delta t)^2}{(\omega\Delta t)^2 + 4} + O((\omega\Delta t)^4) \quad . \tag{3.15}$$

In Figure 3.6-a we plot the module of the amplification factor for Equation (3.5) for $\theta = 0.55$, and the results confirm that the scheme is stable. In the same way, if we use $\theta = 0.45$ for Equation (3.5), the results are not stable, as shown in Figure 3.6-b. Finally, we use $\mu = 0.5$ for the full semi-implicit semi-Lagrangian combination, and θ is the only variable parameter, with $0.5 \leq \theta \leq 0.75$.

3.5 Numerical experiments of the semi-Lagrangian semi-implicit combination

3.5.1 The full semi-Lagrangian semi-implicit scheme

Following the previous sections, we will consider the following family of semi-Lagrangian semi-implicit schemes, which are obtained by considering $\alpha = 0.25$ and $\mu = 0.5$.

$$\begin{aligned}
&\frac{\psi^* - \psi^n}{\Delta t/2} = \frac{1}{2}(N^n + N_\alpha^{n+1/2}) + \frac{1}{2}(L^* + L^n),\\
&\frac{\psi^{n+1/2} - \psi^n}{\Delta t/2} = \theta N^* + (1-\theta)N^n + \theta L^* + (1-\theta)L^n,\\
&\psi^{n+1} - \frac{4}{3}\psi^{n+1/2} + \frac{1}{3}\psi^n = \frac{\Delta t}{3}(2N^{n+1/2} - N^n + \frac{3}{2}L^{n+1} - L^{n+1/2} + \frac{1}{2}L^n).
\end{aligned} \tag{3.16}$$

These schemes are examined, and we will demonstrate their advantages in terms of accuracy, efficiency, and convergence compared to other existing methods such as the semi-implicit predictor-corrector schemes proposed by Clancy and Pudykiewicz (2013) and the semi-Lagrangian semi-implicit scheme studied by Cullen (2001).

3.5.2 Linear stability and error analysis of the two-frequency system

We consider the following semi-Lagrangian equation

$$\frac{d\psi}{dt} = i\Omega\psi + i\omega\psi, \tag{3.17}$$

where Ω is the frequency of the fast waves, and ω is the frequency of the slow waves, which is the remaining nonlinear term ($|\Omega| > |\omega|$). If we consider the conjugate of the two members of Equation (3.17) we obtain a similar equation using the two frequencies $-\Omega$ and $-\omega$, and the conjugate function $\overline{\psi}$ is its solution. Therefore, in our analysis and without loss of generality, we will use the solutions of (3.17) under the assumption that Ω is positive and $\Omega > |\omega|$.

The analysis is performed using von Neumann's method for a single Fourier mode of the form $\psi_j^n = A_k^n e^{i(kj\Delta x)}$ for any point $(j\Delta x, n\Delta t)$. This mode is substituted into

(3.17) to obtain the following equation for the amplification factor A_k

$$A_k^2 + [(\gamma_1 + \gamma_2)e^{-iks} + \frac{\gamma_2}{2}]A_k - \frac{\gamma_2}{2}e^{-iks} = 0, \tag{3.18}$$

where

$$\gamma_1 = \frac{2 - z_1 + 2z_2}{6 - 3z_1} - \frac{\theta(4 - z_1 + 2z_2)(z_1 + z_2)(4 + z_1 + z_2)}{3(2 - z_1)(4 - z_1)} - \frac{(2 + (1 - \theta)(z_1 + z_2)(4 - z_1 + 2z_2))}{3(2 - z_1)},$$

and

$$\gamma_2 = \frac{\theta z_2(z_1 + z_2)(4 - z_1 + 2z_2)}{(2 - z_1)(4 - z_1)}.$$

The complex numbers z_1 and z_2 are related to the fast and slow waves, respectively, and are given by: $z_1 = i\Omega\Delta t$ and $z_2 = i\omega\Delta t$.

In our analysis we will consider a parameter of decentring $\theta = 0.7$ for the proposed schemes. In Figure 3.7-a we plot the module of the amplification factor A_k corresponding to the physical solution for five values of the parameter ks on the $\omega\Delta t - \Omega\Delta t$ plane. The dashed lines are the limit defined by $\Omega = |\omega|$ of the domain subject of our analysis. The results confirm the stability of the proposed schemes. In Figures 3.7-b, 3.7-c, and 3.7-d, we plot the module of the errors $|A_k - A_{an}|$ of the amplification factors, respectively, for the three values $ks = 2\pi/3$, $ks = \pi$, and $ks = 4\pi/3$. Figure 3.8 shows the module of the errors for the same system (3.17) for the four semi-implicit predictor-corrector schemes proposed in Clancy and Pudykiewicz (2013), which perform well for this test. The authors used the notations AM2*-LFT, AM2*-ABM, T-ABT, and AM2*-ABT for their schemes, where they used the Leapfrog trapezoidal (LFT) method, the Adams-Bashforth trapezoidal (ABT) scheme, and the Adams-Bashforth-Moulton (ABM) method as explicit predictor-corrector schemes. For the implicit schemes, the authors used the trapezoidal method (T) and a method denoted by AM2*. The schemes proposed in this paper perform very well in terms of accuracy compared to the three schemes AM2*-LFT, AM2*-ABM, and AM2*-ABT, as shown in Figures 3.7 and 3.8. Note that the best method in Clancy and Pudykiewicz (2013) for this test in terms of amplitude is the method T-ABT, but it presents some phase errors and it is less accurate than the proposed schemes, as shown in Figures 3.7 and 3.8.

In the same way, we will compare the errors of the amplification factor for the semi-Lagrangian semi-implicit method studied by Cullen (2001) and those obtained by using the proposed schemes. The amplification factor of the scheme studied by Cullen (2001) applied to the same equation (3.17) is as follows:

$$A_k = e^{-iks}[\frac{2+z_1+z_2}{2-z_1} + \frac{z_2(2+z_1+2z_2)}{(2-z_1)^2}],$$

where, as previously defined, $z_1 = i\Omega\Delta t$ and $z_2 = i\omega\Delta t$.

In Figure 3.9-a we plot the module of the errors $|A_k - A_{an}|$ of the amplification factor for the method studied by Cullen (2001). As can be seen in Figures 3.9 and 3.7, for positive and large values of $\omega\Delta t$ the proposed schemes are more accurate than the method studied by Cullen (2001). Note that the errors of this method are mainly due to the phase errors as shown in Figure 3.9-b.

3.5.3 Accuracy, efficiency, and convergence of the proposed class of schemes

In this section we will demonstrate the advantage of the proposed class of schemes in terms of accuracy, convergence, and efficiency. We will compare the results to those obtained by using the semi-Lagrangian semi-implicit scheme studied by Cullen (2001). We analyse the following system considered by Cullen (2001):

$$\begin{aligned} &\frac{\partial D}{\partial t} + \bigtriangledown^2\phi = 0, \\ &\frac{\partial \phi}{\partial t} + c^2 D = 0, \end{aligned} \tag{3.19}$$

where ϕ is the geopotential, D is the divergence term, and c is the gravity speed. We consider c_0 as the reference value of c which is used to split the source term c^2D into two parts. The first part, c_0^2D, based on the reference value c_0, is considered as the linear term which corresponds to the fast waves, and the second part is the residual term $(c^2 - c_0^2)D$, which is considered the nonlinear term. Since this second term is responsible for the slow waves, we will consider the condition $|c^2 - c_0^2| \leq c_0^2$, which leads to $\delta \geq 0.5$, where $\delta = (c_0/c)^2$. In the first equation of (3.19), the source term $\bigtriangledown^2\phi$ is considered as the linear term. The error analysis and the numerical tests will be performed using one mode as well as other solutions which are obtained by combination of two modes with different phases.

a. Error analysis using one mode

In our analysis we consider a single wave of the form $(D_j, \phi_j)_n^T = (D_0, \phi_0)^T A_k^n e^{ikj\Delta x}$ for any spatial temporal point $(j\Delta x, n\Delta t)$, where k is the wave number. We use the previous abbreviate notation for both linear and nonlinear terms in Equation (3.16), and we consider the parameters D^p and ϕ^p on the trajectory at each step p. The time discretization for the first step of the proposed scheme (3.16) applied to system (3.19)

can be written as:

$$\begin{cases} D^* = D^n + \frac{1}{4}k^2\Delta t(\phi^* + \phi^n), \\ \phi^* = \phi^n - \frac{1}{4}c_0^2\Delta t(D^* + D^n) - \frac{1}{4}(c^2 - c_0^2)\Delta t(\frac{5}{2}D^n - \frac{1}{2}D^{n-1}). \end{cases} \tag{3.20}$$

The time discretization for the second step is as follows:

$$\begin{cases} D^{n+1/2} = D^n + \frac{1}{2}k^2\Delta t(\theta\phi^* + (1-\theta)\phi^n), \\ \phi^{n+1/2} = \phi^n - \frac{1}{2}c_0^2\Delta t(\theta D^* + (1-\theta)D^n) - \frac{1}{2}(c^2 - c_0^2)\Delta t(\theta D^* + (1-\theta)D^n). \end{cases} \tag{3.21}$$

The third and last step discretization is:

$$\begin{cases} D^{n+1} - \frac{4}{3}D^{n+1/2} + \frac{1}{3}D^n = \frac{1}{3}k^2\Delta t(\frac{3}{2}\phi^{n+1} - \phi^{n+1/2} + \frac{1}{2}\phi^n), \\ \phi^{n+1} - \frac{4}{3}\phi^{n+1/2} + \frac{1}{3}\phi^n = \frac{\Delta t}{3}(-c_0^2(\frac{3}{2}D^{n+1} - D^{n+1/2} + \frac{1}{2}D^n) - (c^2 - c_0^2)(2D^{n+1/2} - D^n)). \end{cases} \tag{3.22}$$

Following the expressions of Equations (3.20), (3.21) and (3.22) respectively for the three steps, the terms $k^2\Delta t$, $c^2\Delta t$, and δ can be considered as dimensionless parameters. These parameters will be used to conduct a more general analysis of the characteristics of the proposed schemes and their comparison with those of the scheme studied by Cullen (2001). The three steps of the temporal discretization of the proposed schemes can be rewritten in the matrix form using the vector variable $\mathbf{U}^p = (D^p, \phi^p)^T$ at each step p. The vector value $\mathbf{U}^*$ of the first step is substituted in the second step, and the result obtained for $\mathbf{U}^{n+1/2}$ is substituted in the third step to obtain a fourth-degree polynomial equation for the amplification factor A_k. In Appendix II we give further explanations about the matrix form of the proposed schemes and the characteristics of the polynomial equation of the amplification factor. This polynomial has one root equal to zero. Therefore it can be reduced to a three-degree polynomial which has one real root corresponding to the computational mode. The two other roots are complex conjugate numbers and correspond to the inertio-gravity waves. The computational mode is largely damped, as shown in Figure 3.10 for the two cases $\delta = 0.75$ and $\delta = 1.25$. Note that one of the advantages of the proposed schemes is that they have a damped computational mode, as opposed to the method studied by Cullen (2001), which has one meteorological mode which is undamped (equal to unity).

In Figure 3.11, we plot the module of the errors $|A_k - A_{an}|$ of the amplification factor A_k in the $k^2\Delta t - c^2\Delta t$ plane for the cases $\delta = 0.5$ and $\delta = 1$. In order to compare between the accuracy of the proposed schemes and the method studied by Cullen (2001), the errors are integrated using

$$E^2 = \frac{1}{\mathcal{A}} \int\int |A_k - A_{an}|^2 d\sigma d\varrho, \tag{3.23}$$

where $\sigma = k^2\Delta t$, $\varrho = c^2\Delta t$, and $\mathcal{A}$ is the area of the domain used in the integration.

Table 3.2 shows the values of the errors E corresponding to the amplification factors which are obtained by using the method studied in Cullen (2001), and those obtained by using the proposed schemes with the decentring parameters $\theta = 0.7$ and $\theta = 0.75$ for five values of the dimensionless parameter δ. Following the results of this test we conclude that the proposed schemes provide more accurate results compared to those obtained by using the method studied by Cullen (2001). To further verify the resolution quality, the efficiency, and the convergence of the proposed schemes, in the following section severals tests are performed by using other solutions of Equation (3.19) which are the combination of two different modes.

b. Numerical test using two modes

In this section the numerical test is performed using an analytical solution which is the real part of the combination of two modes with different phases:

$$(D, \phi)^T = \Re((\nu Z_1 + \bar{\nu} Z_2)/2), \tag{3.24}$$

where

$$\begin{aligned} Z_1 &= (1, \frac{ic}{k})^T e^{ik(x+ct)}, \\ Z_2 &= (1, \frac{-ic}{k})^T e^{ik(x-ct)}, \end{aligned} \tag{3.25}$$

and $\nu = \nu_R + \nu_I i$ is a complex number with $\bar{\nu}$ as its conjugate. This analytical solution will be used as the initial condition to test the accuracy of the schemes. The numerical test is performed using the parameter $c = 0.01$ and the initial condition with $\nu_R = 1$, $\nu_I = -1$ and $k = 0.01$. Figure 3.12 shows the evolution of the errors for the parameters D and ϕ until time $T = 50$ using the proposed schemes and the scheme studied by Cullen (2001) with a time step $\Delta t = 0.03$. Note that for this test both methods lead to accurate results and the proposed schemes present a high accuracy compared to the method studied by Cullen (2001). In the following section we demonstrate that the high accuracy of the proposed schemes has a great advantage for using large time steps that meet the Lipschitz criterion in order to reduce the computational cost, which makes them competitive in terms of efficiency.

c. Efficiency of the proposed schemes

The most expensive parts in terms of computational cost are the dynamical part, which is related to the implicit resolution of the linear term, and the physical part, related to the explicit resolution of the nonlinear term. The computational cost of the explicit resolution of the linear term, which is the case in the second step in the proposed schemes, is negligible compared to the computational costs of the dynamical and physical parts.

For the proposed schemes we have one dynamical part and one physical part in the first step, one physical part in the second step, and one physical part and one dynamical part in the third step. The method studied by Cullen (2001) has one dynamical part and one physical part in each step. The efficiency of the proposed schemes and the efficiency of the method studied by Cullen (2001) will be compared by using the time steps satisfying the condition $\Delta t_m = 1.25\Delta t_c$, where Δt_m and Δt_c are the time steps used for the proposed schemes and the method studied by Cullen (2001), respectively. This condition is sufficient to give a reliable comparison of the efficiency, depending on the number of physical and dynamical parts in each method.

Equation (3.19) with the gravity speed $c = 0.01$ is solved, for two cases of analytic solutions of the form (3.24), using the proposed schemes and the method studied by Cullen (2001). We consider $k = 0.01$, and in the first case we use the initial condition with $\nu_R = 1$ and $\nu_I = 1$ and in the second case we consider $\nu_R = 1$ and $\nu_I = 4$. In Figure 3.13 we plot the evolution of the errors of the solutions $(D, \phi)^T$, which are obtained by using the proposed schemes with a time step $\Delta t = 0.030$ and the method studied by Cullen (2001) with a time step $\Delta t = 0.024$. The results of the proposed schemes remain more accurate in comparison to those obtained by using the method studied by Cullen (2001). Therefore we conclude that the proposed schemes perform well in terms of efficiency.

Our numerical tests show that the method used to compare the competitive efficiency by considering the time steps which satisfy the condition $\Delta t_m = 1.25\Delta t_c$ is valid for the large interval of time steps that meet the Lipschitz criterion. In the following we give another numerical test to study the efficiency of the proposed method using the two-frequency system (3.17). In this example we will consider the initial value of the trajectories $x(0) \in [0, 1]$ and the analytical solution $\psi(x, t) = (x + \rho x t)e^{(i\Omega + i\omega)t}$, where $\rho = 0.8$, $\omega = 1.0$, $\Omega = 1.8$ and the trajectories are given by $x(t) = x(0)/(1 + \rho t)$. We use the similar approach in Smolarkiewicz and Pudykiewicz (1992) to satisfy the Lipschitz criterion. The necessary condition the time step Δt should satisfy to avoid the intersection of the computed trajectories is:

$$\Delta t \| \frac{\partial \mathbf{V}}{\partial x} \| < 1, \tag{3.26}$$

which leads to the restrictive condition for the time step $\Delta t < 0.625$.

For each value of Δt we study the error defined by

$$E(\Delta t) = \max_{x \in I, t \in [0,1]} |\psi_n(x, t) - \psi_{exact}(x, t)|, \tag{3.27}$$

where $I = [\min_{t \in [0,1]} x(t), \max_{t \in [0,1]} x(t)]$, ψ_n is the numerical solution and ψ_{exact} is the exact solution.

Figure 3.14 (right) shows the error as a function of the variable time step Δt for the proposed method. The efficiency of the proposed method and the efficiency of the method studied by Cullen (2001) are compared by using the time steps that satisfy the condition $\Delta t_m = 1.25\Delta t_c$. The comparison is performed using the large interval of the time steps which meet the Lipschitz criterion. Figure 3.14 (left) shows the error for the method studied by Cullen (2001) using a time step Δt_c and the error for the proposed method using a time step $\Delta t_m = 1.25\Delta t_c$. The results of the proposed method remain accurate, which confirms that the proposed method performs well in terms of efficiency.

d. Convergence of the proposed schemes

In this section we will consider the variable $\mathbf{U} = (D, \phi)^T$ in our analysis. To study the convergence of the proposed schemes we use the following error for each global time T

$$E(T) = \max_{1 \leq n \leq N} \mid \mathbf{U}_n - \mathbf{U}(t_n) \mid, \tag{3.28}$$

where $\mathbf{U}_n$ and $\mathbf{U}(t_n)$ are the numerical solution and the analytical solution, respectively, at time $t_n = n\Delta t$. The time step is $\Delta t = T/N$, and N is the total number of steps necessary to reach the global time T. In the analysis, we seek for a small Δt an error model of the form $E(T) = C\Delta t^r$, where $E(T)$ is calculated using Equation (3.28). The least-squares method is used to find the values of the parameter C and the order of convergence r. The tests of convergence of the schemes are performed using the global time $T = 50$ for two cases. In the first case we use $c = 0.1$, and the initial condition is defined by setting $\nu_R = 1$, $\nu_I = 4$, and $k = 0.1$. For this case we obtain $C = 2.44 \times 10^{-11}$ and $r = 2.03$ for the proposed schemes, and $C = 2.46 \times 10^{-11}$ and $r = 2.02$ for the method studied by Cullen (2001). The proposed schemes are better in terms of convergence compared to the method studied by Cullen (2001). In some cases we observe the same order of convergence, such as the second case in which we use $c = 0.01$, $k = 0.01$, $\nu_R = 1$, and $\nu_I = 1$. For this case we obtain $C = 8.02 \times 10^{-6}$ and $r = 2$ for the proposed schemes and $C = 8.04 \times 10^{-6}$ and $r = 2$ for the method studied by Cullen (2001). For this case, in which we have the same order of convergence, the coefficient C for the proposed schemes is less than the coefficient obtained for the method studied by Cullen (2001). This is justified by the fact that the proposed schemes are more accurate as reported in the previous sections.

3.6 Conclusions

This paper presents theoretical and numerical analysis of the properties of more complex methods. We try to avoid the problems of instability associated with the treatment of the non-linear part of the forcing term. A class of semi-Lagrangian semi-implicit schemes is proposed which uses a modified TR-BDF2 method based on the Trapezoidal

Rule and the second-order Backward Differentiation Formula. For the nonlinear term we used two stages as predictor and corrector in the TR method and one stage for the BDF2 method. A family of second-order approximations derived by Gospodinov et al. (2001) is used in the first stage as the predictor of the TR method. For the linear term we used the implicit trapezoidal method in the first step, the explicit trapezoidal method in the second step and the implicit BDF2 method in the third step. The use of the corrector stage improves the stability of the proposed schemes, and a large stable imaginary part is obtained. Following the numerical tests, the second-order approximation using $\alpha = 1/4$, which corresponds to the approximation of SETTLS, is the appropriate choice which guarantees the large domain of absolute stability and the optimal intersection of the region of absolute stability with the imaginary axis. This intersection is improved by using a decentring in the second step to obtain the schemes which perform well for purely oscillatory solutions. The numerical analysis demonstrate that the proposed class of semi-Lagrangian semi-implicit schemes performs well in terms of stability, accuracy, convergence, and efficiency. The potential practical application of the proposed family of schemes to a weather-prediction model or any other atmospheric model is not analyzed and could be the subject of other forthcoming studies.

Appendix I: TR-BDF2 method

The composite second-order trapezoidal rule / backward differentiation method can be used to numerically solve the following ordinary differential equation:

$$\frac{d\mathbf{u}}{dt} = \mathbf{f}(\mathbf{u}),\ \mathbf{u}(t_0) = \mathbf{u}^0. \tag{3.29}$$

We use $\mathbf{u}^n$ to denote the computed approximation to the solution at time $t_n = n\Delta t$. The original TR-BDF2 scheme is performed in two steps. In the first step, we apply the trapezoidal rule (TR) to advance our solution from t_n to $t_{n+\gamma}$

$$\mathbf{u}^{n+\gamma} = \mathbf{u}^n + \frac{\gamma\Delta t}{2}(\mathbf{f}^n + \mathbf{f}^{n+\gamma}), \tag{3.30}$$

and then the second-order backward differentiation formula (BDF2) is used to advance the solution from $t_{n+\gamma}$ to t_{n+1}

$$\mathbf{u}^{n+1} = \frac{1}{\gamma(2-\gamma)}\mathbf{u}^{n+\gamma} - \frac{(1-\gamma)^2}{\gamma(2-\gamma)}\mathbf{u}^n + \frac{(1-\gamma)}{(2-\gamma)}\Delta t\mathbf{f}^{n+1}. \tag{3.31}$$

In our case we use $\gamma = 1/2$. The equations (3.30) and (3.31) can be solved by using implicit or explicit formulas for the source term.

Appendix II : Matrix form of the proposed scheme and trajectory evaluation

II-1: Matrix form of the proposed scheme

The three steps given in Section 3.5 by Equations (3.20), (3.21) and (3.22) for temporal discretization of the proposed schemes applied to solve Equation (3.19) can be rewritten in the following matrix form using the vector variable $\mathbf{U}^p = (D^p, \phi^p)^T$at each step p:

$$\begin{cases} \mathbf{M}_1\mathbf{U}^* = \mathbf{M}_2\mathbf{U}^n + \mathbf{M}_3\mathbf{U}^{n-1}, \\ \mathbf{U}^{n+1/2} = \mathbf{M}_4\mathbf{U}^n + \mathbf{M}_5\mathbf{U}^*, \\ \mathbf{M}_6\mathbf{U}^{n+1} = \mathbf{M}_7\mathbf{U}^{n+1/2} + \mathbf{M}_8\mathbf{U}^n. \end{cases} \tag{3.32}$$

The matrix $\mathbf{M}_j$, $j = 1, 2...8$ are given as:

$$\mathbf{M}_1 = \begin{pmatrix} 1 & -\frac{\Delta t k^2}{4} \\ \frac{\Delta t c^2}{4} & 1 \end{pmatrix}, \quad \mathbf{M}_2 = \begin{pmatrix} 1 & \frac{\Delta t k^2}{4} \\ \frac{3\Delta t c_0^2}{8} - \frac{5\Delta t c^2}{8} & 1 \end{pmatrix}, \quad \mathbf{M}_3 = \begin{pmatrix} 0 & 0 \\ \frac{\Delta t(c^2 - c_0^2)}{8} & 0 \end{pmatrix}, \tag{3.33}$$

$$\mathbf{M}_4 = \begin{pmatrix} 1 & \frac{\Delta t k^2(1-\theta)}{2} \\ \frac{\Delta t c^2(\theta-1)}{2} & 1 \end{pmatrix}, \quad \mathbf{M}_5 = \begin{pmatrix} 0 & \frac{\Delta t k^2\theta}{2} \\ \frac{\Delta t c^2\theta}{2} & 0 \end{pmatrix}, \quad \mathbf{M}_6 = \begin{pmatrix} 1 & -\frac{\Delta t k^2}{2} \\ \frac{\Delta t c_0^2}{2} & 1 \end{pmatrix}, \tag{3.34}$$

$$\mathbf{M}_7 = \begin{pmatrix} \frac{4}{3} & -\frac{\Delta t k^2}{3} \\ \Delta t c_0^2 - \frac{2\Delta t c^2}{3} & \frac{4}{3} \end{pmatrix}, \quad \mathbf{M}_8 = \begin{pmatrix} \frac{-1}{3} & \frac{\Delta t k^2}{6} \\ -\frac{\Delta t c_0^2}{2} + \frac{\Delta t c^2}{3} & \frac{-1}{3} \end{pmatrix}. \tag{3.35}$$

The system (3.32) can be rewritten in the following form:

$$\mathbf{U}^{n+1} = \mathbf{S}\mathbf{U}^n + \mathbf{T}\mathbf{U}^{n-1}, \tag{3.36}$$

where $\mathbf{S} = \mathbf{M}_6^{-1}[\mathbf{M}_7\mathbf{M}_4 + \mathbf{M}_7\mathbf{M}_5\mathbf{M}_1^{-1}\mathbf{M}_2 + \mathbf{M}_8]$ and $\mathbf{T} = \mathbf{M}_6^{-1}\mathbf{M}_7\mathbf{M}_5\mathbf{M}_1^{-1}\mathbf{M}_3$.

We obtain the equation of the amplification factor $\mathbf{P}(\tau) = 0$, where $\mathbf{P}$ is a fourth-degree polynomial given by:

$$\mathbf{P}(\tau) = \det(\tau^2\mathbf{I} - \tau\mathbf{S} - \mathbf{T}), \tag{3.37}$$

where $\mathbf{I}$ is the identity matrix 2×2. We have $\mathbf{P}(0) = \det(-\mathbf{T}) = 0$ since $\mathbf{M}_3$ is a singular matrix. If we exclude the zero root, we obtain a three-degree polynomial which has one real root, and the two other roots are complex conjugate numbers. The coefficients of matrices $\mathbf{M}_j$ can be expressed as a function of the parameters $k^2\Delta t$, $c^2\Delta t$, and $\delta = (c_0/c)^2$, which are used in Section 3.5 for the analysis.

II-2: Trajectory evaluation

We consider the following equation of the semi-Lagrangian trajectory

$$\frac{dx}{dt} = \mathbf{V}, \tag{3.38}$$

where x is the position vector of an air parcel and $\mathbf{V}$ is the corresponding three-dimensional velocity.

Following Hortal (2002), the approximation of the SETTLS method is obtained by using the Taylor expansion of the unknown position vector x around the departure point x_D^t of the semi-Lagrangian trajectory

$$x_A^{t+\Delta t} \approx x_D^t + \Delta t \cdot \left[\frac{dx}{dt}\right]_D^t + \frac{\Delta t^2}{2} \cdot \left[\frac{d^2x}{dt^2}\right]_{AV}, \tag{3.39}$$

where AV indicates the approximation of the average value along the semi-Lagrangian trajectory between the arrival point A at time $t + \Delta t$ and the corresponding departure point D at time t. The previous Equation (3.39) can be rewritten in the following form:

$$x_A^{t+\Delta t} \approx x_D^t + \Delta t \cdot \mathbf{V}_D^t + \frac{\Delta t^2}{2} \cdot \left[\frac{d\mathbf{V}}{dt}\right]_{AV}. \tag{3.40}$$

Hortal (2002) proposed the following approximation of the average of acceleration along the semi-Lagrangian trajectory between the departure point D at time t and the arrival point A at time $t + \Delta t$,

$$\left[\frac{d\mathbf{V}}{dt}\right]_{AV} \approx \frac{\mathbf{V}_A^t - \mathbf{V}_D^{t-\Delta t}}{\Delta t}, \tag{3.41}$$

which leads to the following equation of the semi-Lagrangian trajectory

$$x_A^{t+\Delta t} = x_D^t + \frac{\Delta t}{2}(\mathbf{V}_A^t + [2\mathbf{V}^t - \mathbf{V}^{t-\Delta t}]_D). \tag{3.42}$$

In the proposed method, for each arrival point x_j at time $t + \Delta t$ the departure point x_D^t is found by iteratively solving the previous equation.

Following Hortal (2002), SETTLS is based on the approximate value given by Equation (3.41) for the average of the acceleration $\left[\frac{d\mathbf{V}}{dt}\right]_{AV}$. The proposed method involves the evaluation of terms at the middle point of the semi-Lagrangian trajectory. The middle point is located in the part of the SL trajectory where the average of acceleration is estimated. This value is used in the proposed method to obtain the position of the middle point, denoted by $x^{t+\Delta t/2}$, of the SL trajectory. We will consider the Taylor expansion, using $\Delta t/2$, of the unknown position vector x around the departure point

x_D^t of the SL trajectory

$$x^{t+\Delta t/2} \approx x_D^t + \frac{\Delta t}{2} \cdot \mathbf{V}_D^t + \frac{\Delta t^2}{8} \cdot \left[\frac{d\mathbf{V}}{dt}\right]_{AV}, \tag{3.43}$$

where the average of the acceleration $\left[\frac{dV}{dt}\right]_{AV}$ is obtained using Equation (3.41). Therefore, the position of the middle point is obtained by using an explicit method, and the only iterative equation is the one used to obtain the departure point x_D^t.

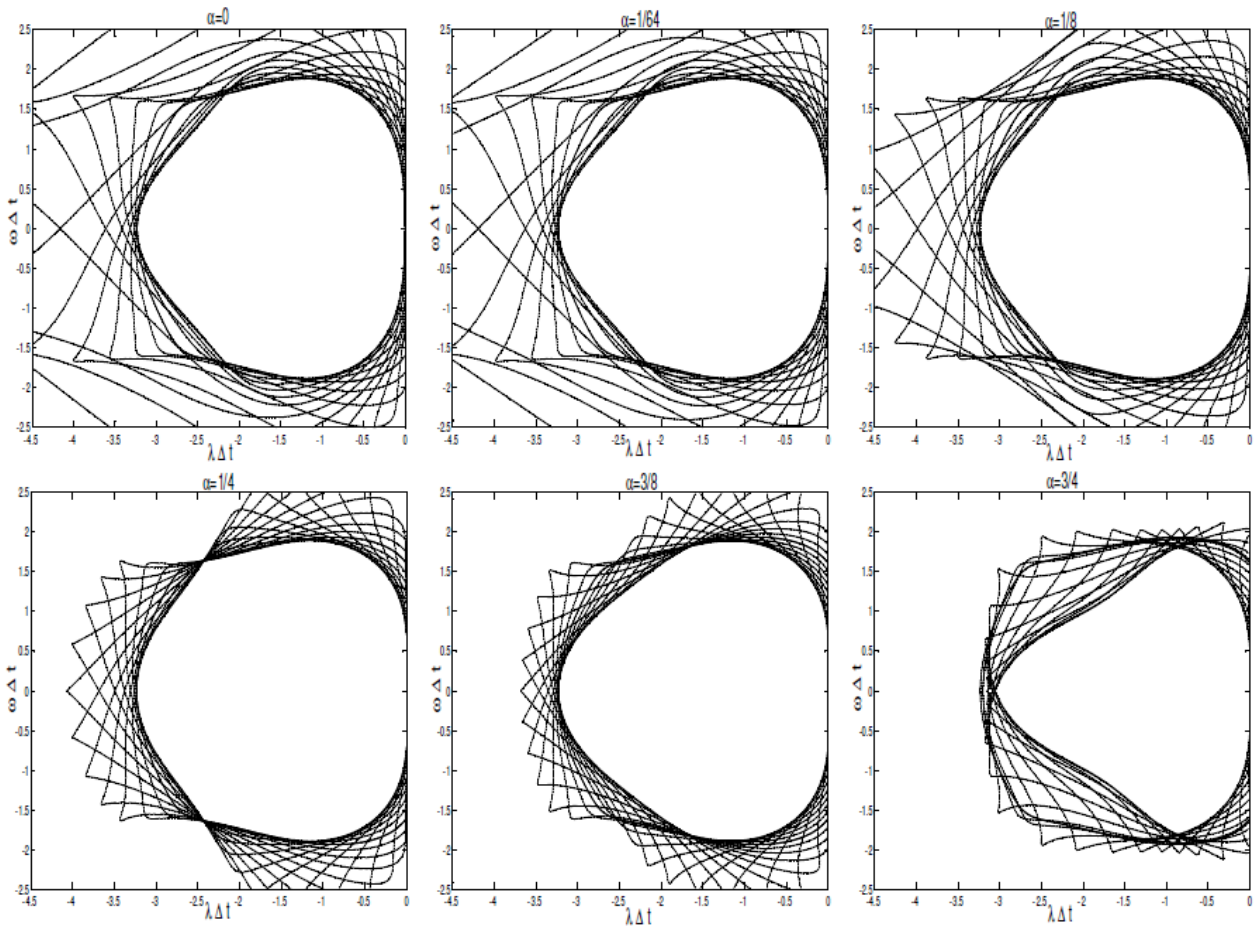

Figure 3.1 Region of absolute stability for six values of the parameter α plotted for fifteen CFL Number values. The central white region corresponds to the absolutely stable region independent of the CFL Number for a decentring parameter $\theta = 0.7$.

Table 3.1 Value of $(\omega\Delta t)_{max}$ of the segment of imaginary axis $[-(\omega\Delta t)_{max}, (\omega\Delta t)_{max}]$ included in the region of absolute stability depending on the values of the approximation parameter α and the decentring parameter θ.

	α						
θ	0	1/64	1/8	1/4	3/8	3/4	1
0.6	0.000	0.000	0.000	0.473	0.411	0.286	0.000
0.7	0.000	0.000	0.435	0.708	0.643	0.451	0.389
0.8	0.000	0.000	0.661	0.839	0.847	0.556	0.464
0.9	0.098	0.240	0.791	0.933	0.908	0.651	0.526
1.	0.343	0.449	0.876	1.000	0.954	0.745	0.581

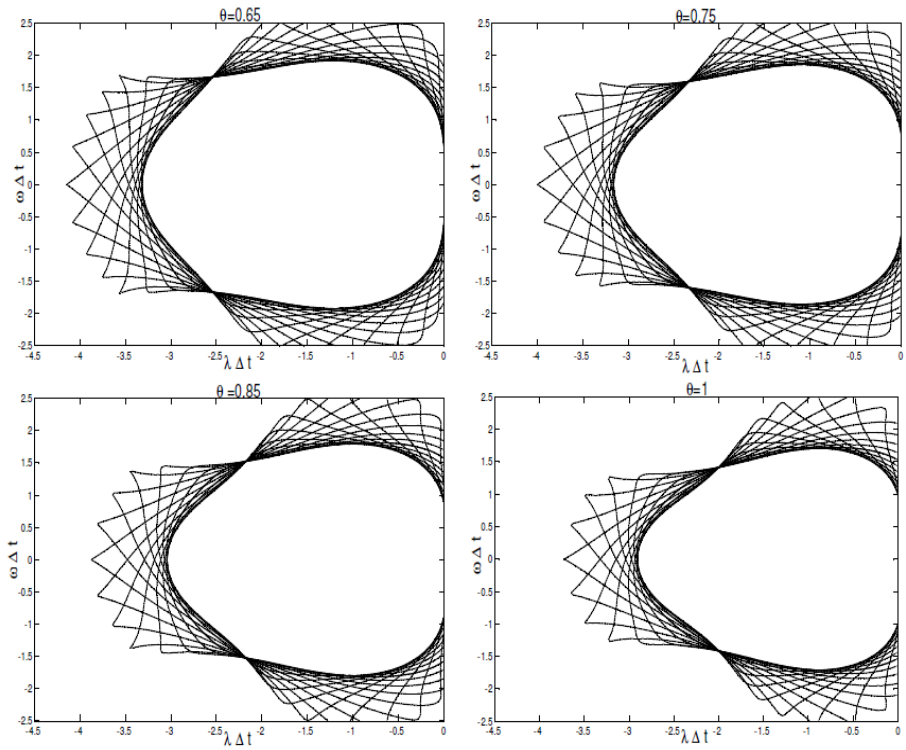

Figure 3.2 Region of absolute stability of the proposed schemes ($\alpha = 1/4$) for four values of the decentring parameter θ plotted for fifteen CFL number values. The central white region corresponds to the absolutely stable region independent of the CFL number.

Table 3.2 Errors of the amplification factors using the proposed schemes and the method studied by Cullen (2001).

δ	0.5	0.75	1	1.25	1.5
Proposed method $\theta = 0.70$	0.10	0.09	0.09	0.11	0.12
Proposed method $\theta = 0.75$	0.11	0.09	0.09	0.10	0.12
Cullen (2001)	0.12	0.12	0.12	0.12	0.12

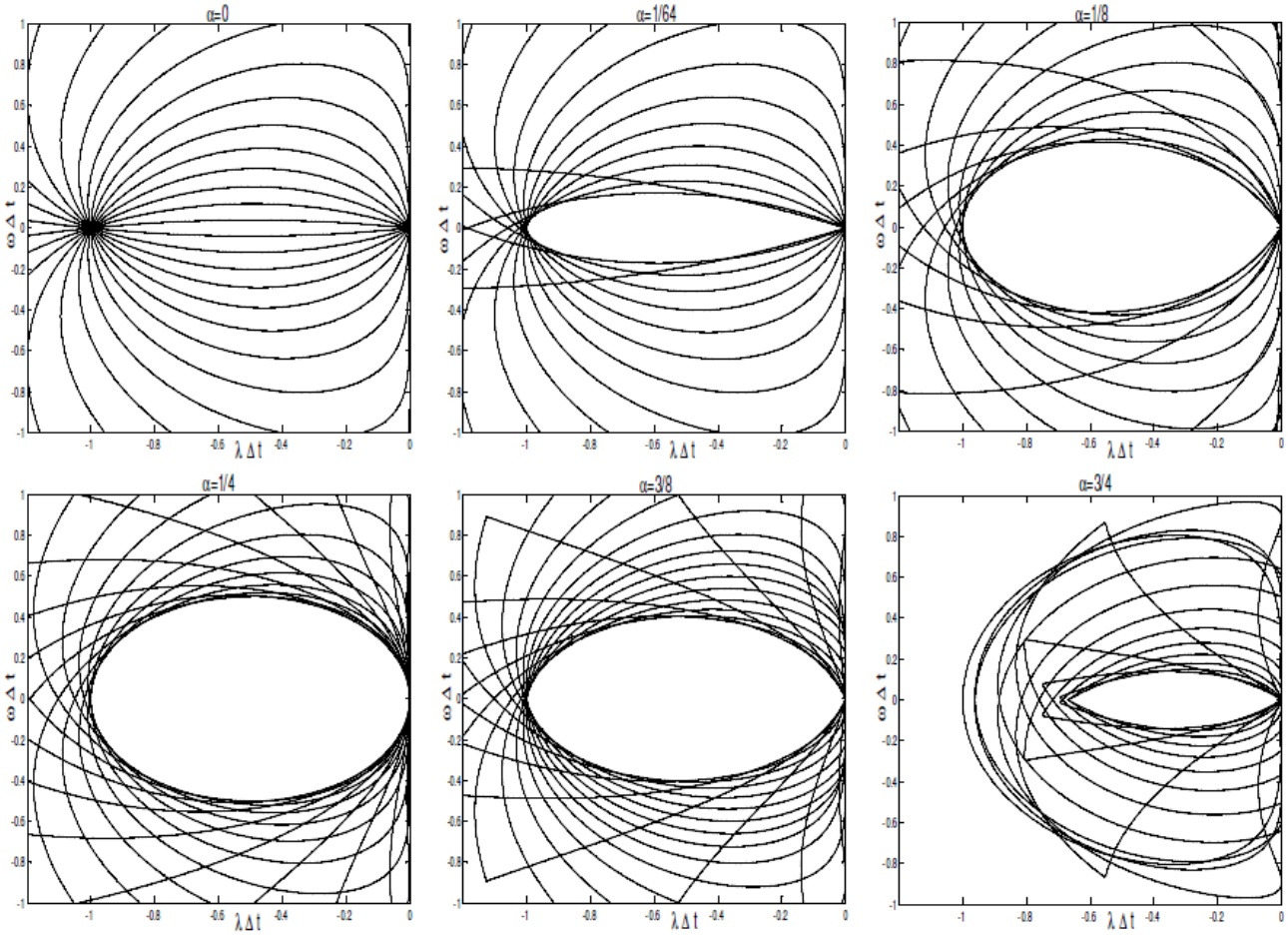

Figure 3.3 Region of absolute stability of the scheme without corrector step and using the value $N_\alpha^{n+1/2}$ in the step using BDF2 method, for six values of the parameter α plotted for fifteen CFL number values.

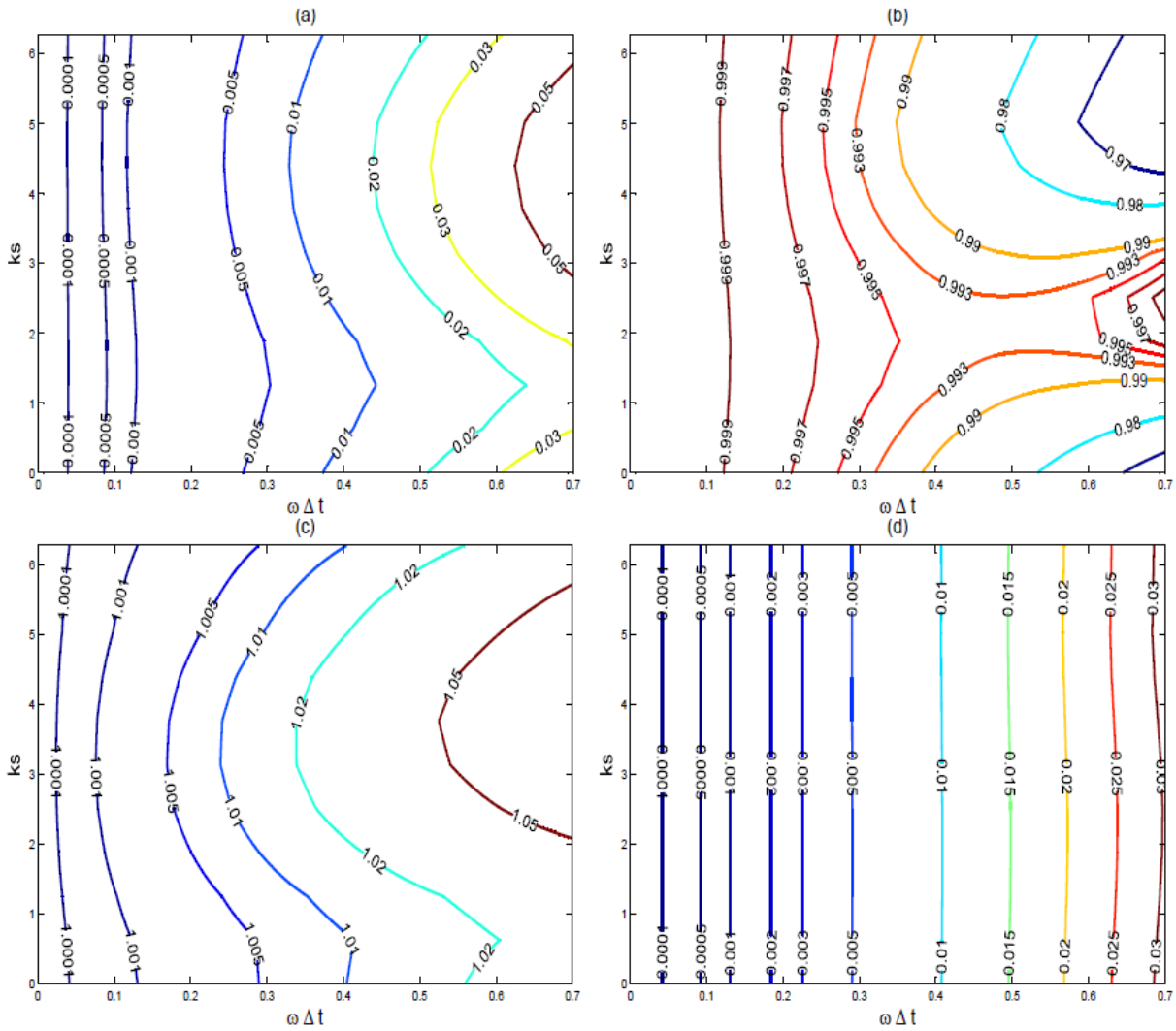

Figure 3.4 Solutions of oscillatory equation using the proposed schemes for the nonlinear term. (a): Module of the errors of the amplification factors $|A_k - A_{an}|$. (b): Damping of the solution. (c): Relative phase change. (d): Module of the computational mode.

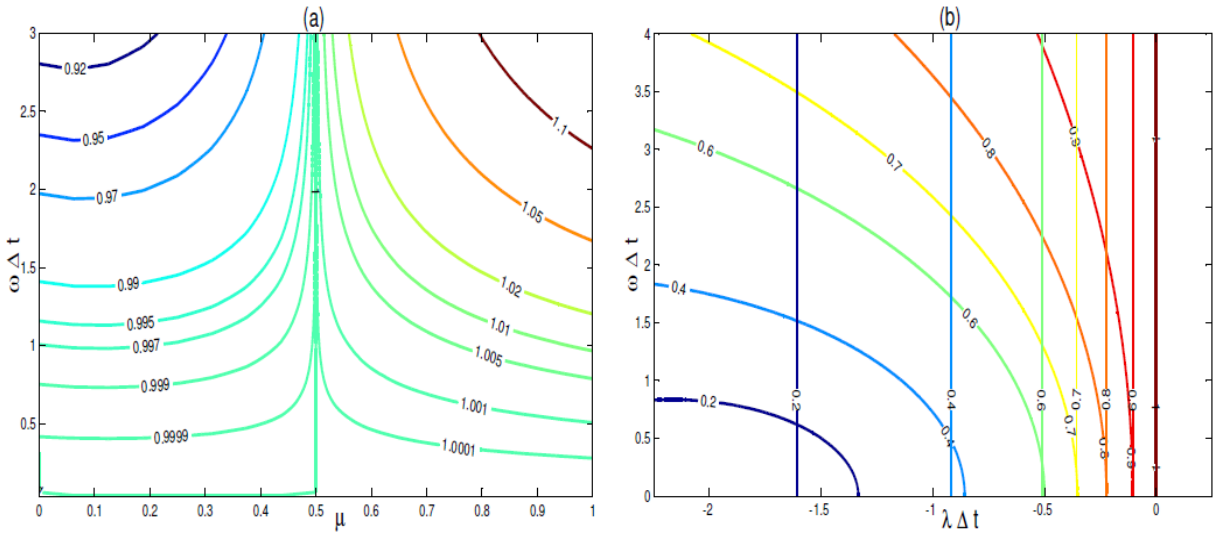

Figure 3.5 Treatment of the linear term. (a): Module of the amplification factor, for oscillation equation (3.9) using $\theta = 0.5$, plotted as function of μ and $\omega\Delta t$. (b): Module of the amplification factor for Equation (3.5) using $\mu = 0.5$ and $\theta = 0.5$ plotted on the $\lambda\Delta t$ -$\omega\Delta t$ plane.

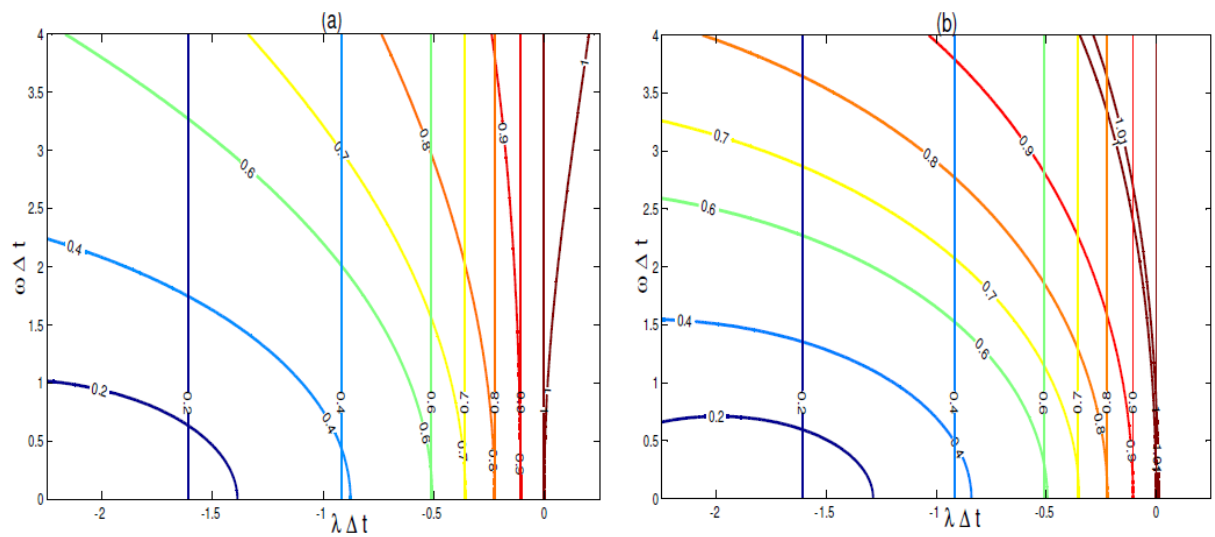

Figure 3.6 Treatment of the linear term. Module of the amplification factor for Equation (3.5), plotted on the $\lambda\Delta t$-$\omega\Delta t$ plane. (a): Example of stable case using $\mu = 0.5$ and $\theta = 0.55$. (b): Example of unstable case using $\mu = 0.5$ and $\theta = 0.45$.

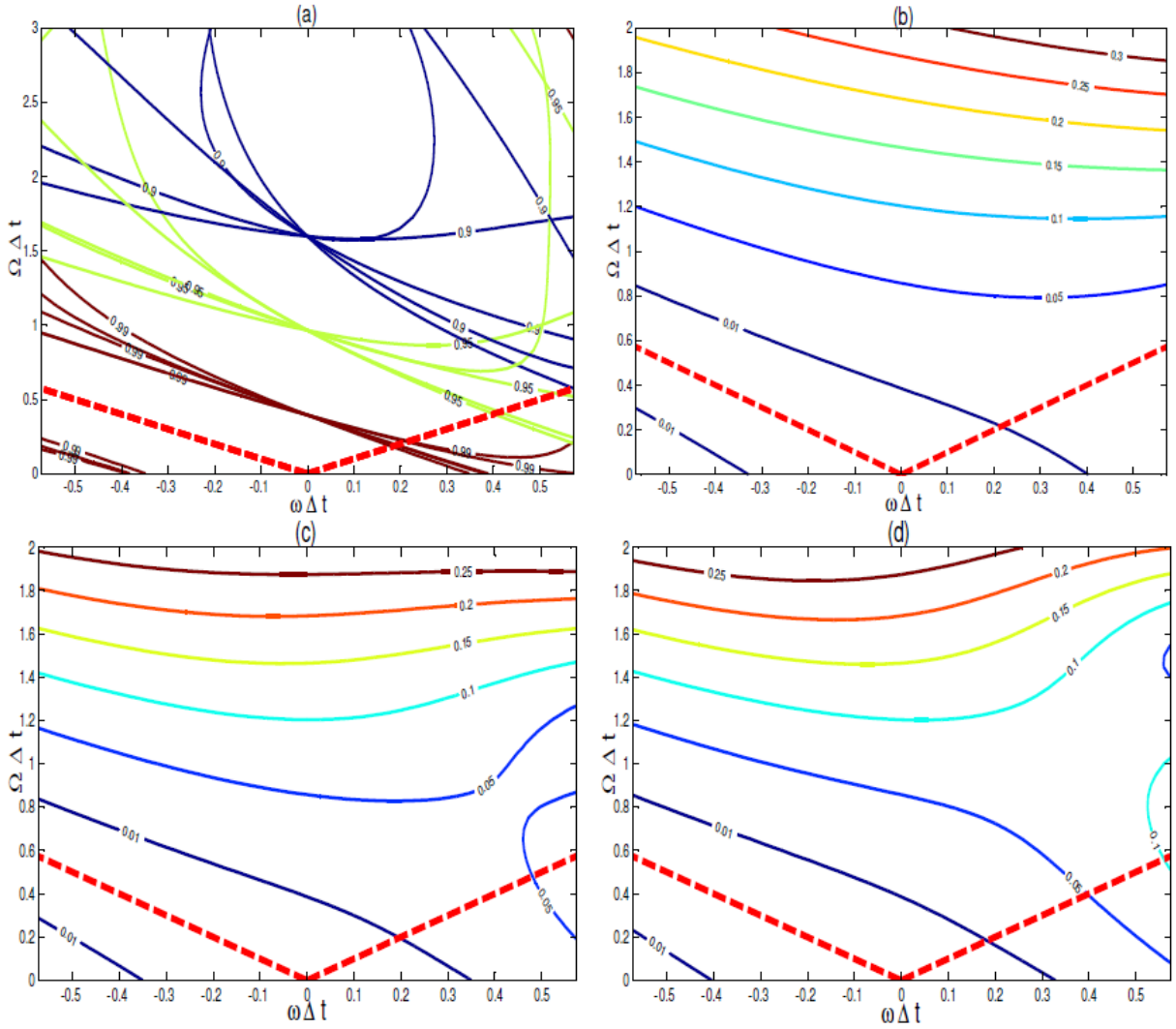

Figure 3.7 The proposed full semi-lagrangian semi-implicit schemes using $\mu = 0.5$ and $\theta = 0.7$. (a): Module of the amplification factors for five CFL number values. (b): Module of the errors of the amplification factors $|A_k - A_{an}|$ for $ks = 2\pi/3$. (c): Module of the errors for $ks = \pi$. (d): Module of the errors for $ks = 4\pi/3$.

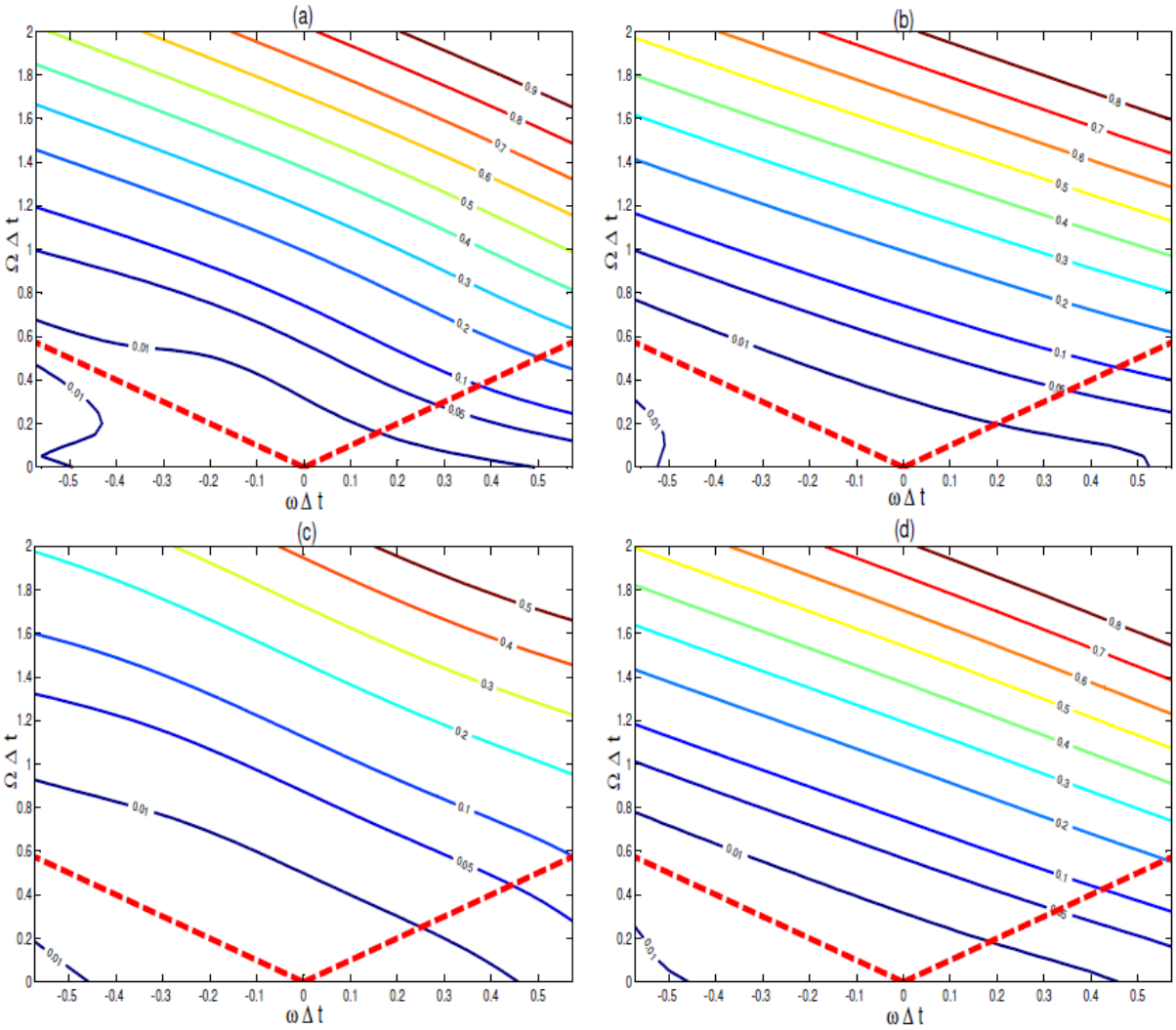

Figure 3.8 Module of the errors of the amplification factors $|A_k - A_{an}|$ for the schemes proposed by Clancy and Pudykiewicz (2013). (a): Method AM2-LFT. (b): Method AM2-ABM. (c): Method T-ABT. (d): Method AM2-ABT.

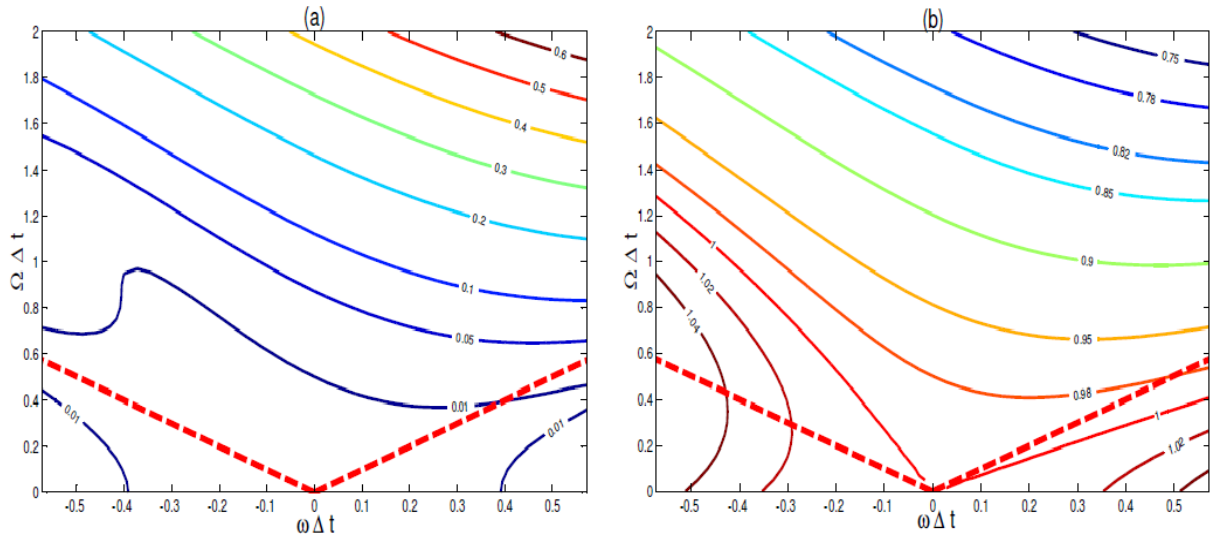

Figure 3.9 The scheme studied by Cullen (2001). (a):Module of the errors of the amplification factors $|A_k - A_{an}|$. (b): Relative phase change.

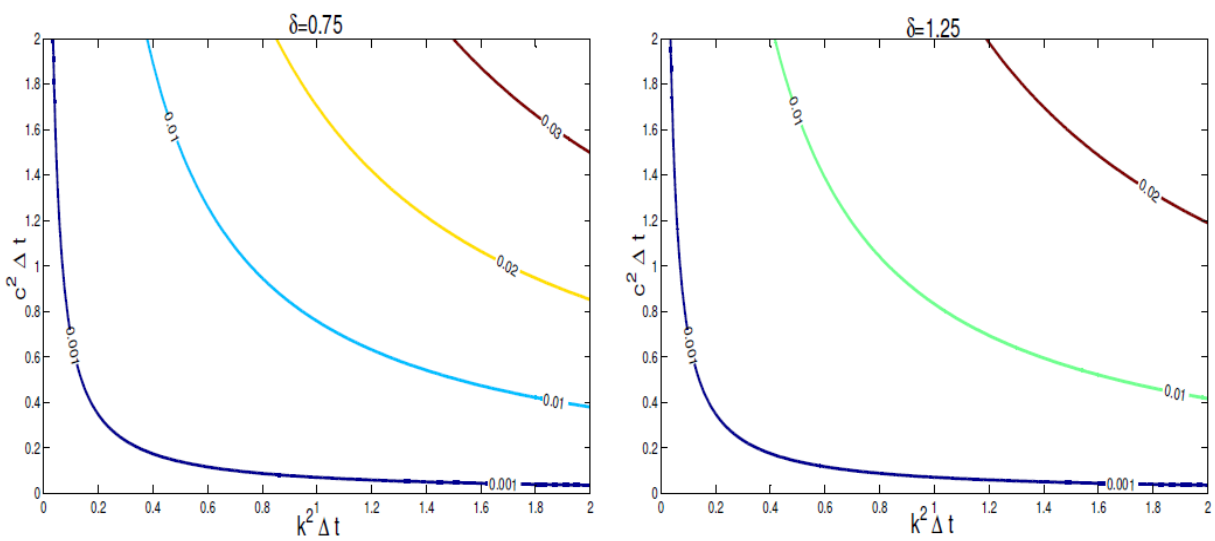

Figure 3.10 Module of the computational mode for the proposed schemes ($\theta = 0.7$).

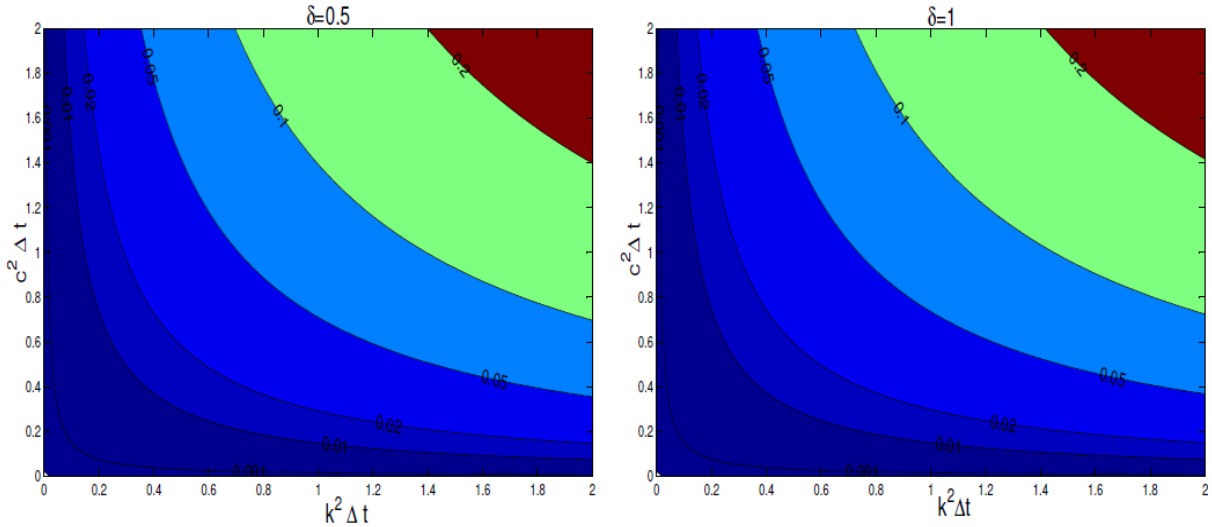

Figure 3.11 Module of the errors of the amplification factors $|A_k - A_{an}|$ for the proposed schemes ($\theta = 0.7$).

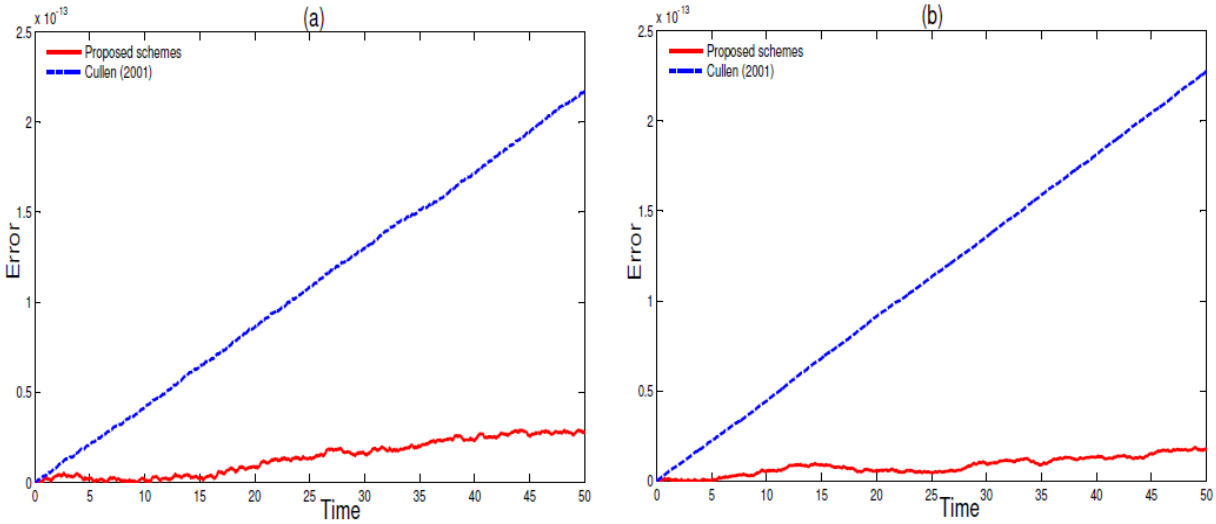

Figure 3.12 Evolution of the errors until time $T = 50$ using the proposed schemes and the method studied by Cullen (2001) for an initial condition which combined two modes ($\nu_R = 1$, $\nu_I = -1$) with $\Delta t = 0.03$. (a): Errors of the parameter D. (b): Errors of the parameter ϕ.

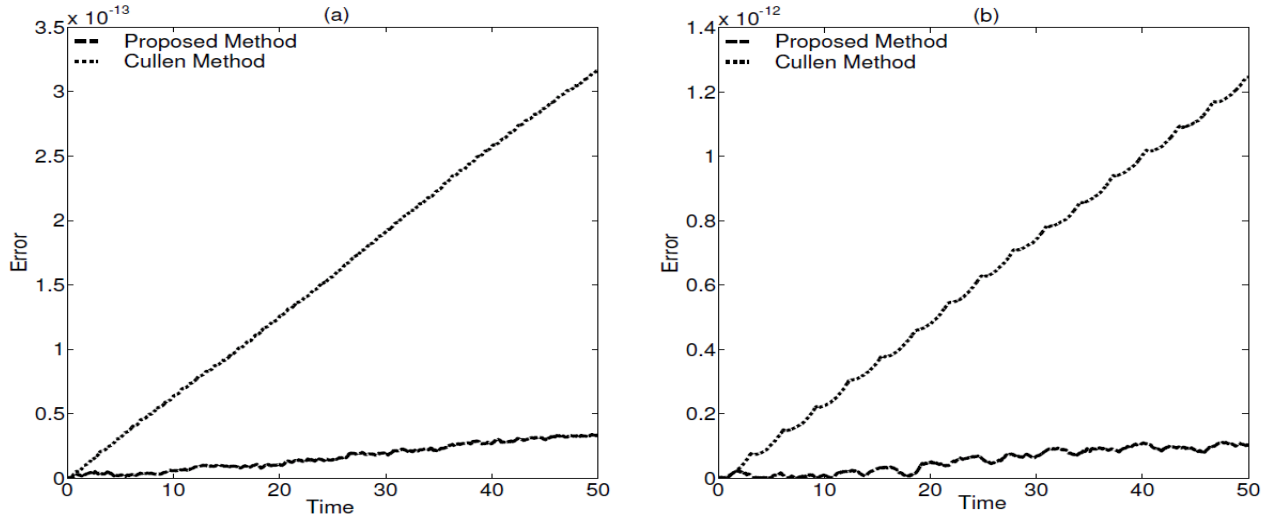

Figure 3.13 Evolution of the error of the solution $(D, \phi)^T$ until time $T = 50$ using the proposed method with $\Delta t = 0.030$ and the method studied by Cullen (2001) with $\Delta t = 0.024$ for an initial condition which combined two modes. (a) : case $\nu_R = 1$ and $\nu_I = 1$ and (b): case $\nu_R = 1$ and $\nu_I = 4$

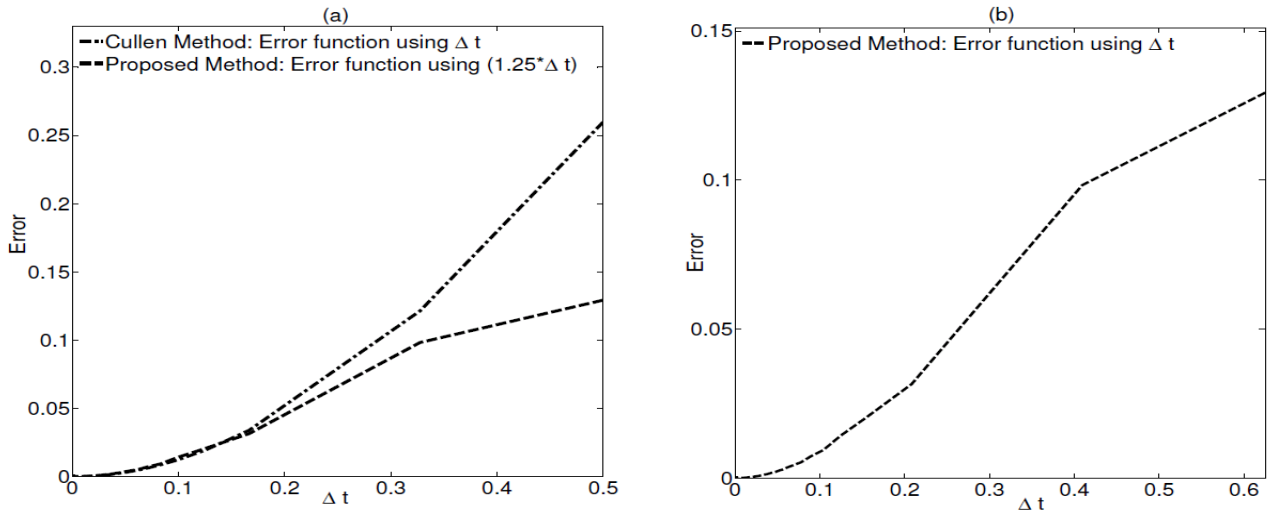

Figure 3.14 (a): The error function for the method studied by Cullen (2001) using the time step Δt and the error function for the proposed method using the time step $1.25\Delta t$. (b): The error function for the proposed method using the time step Δt

CHAPITRE 4 Un schéma équilibre partiellement centré de type «Cell-Vertex» préservant la positivité pour les écoulements en eaux peu profondes

Well-balanced positivity preserving Cell-Vertex Central-Upwind Scheme for Shallow Water Flows[1]

Résumé

Le système de Saint-Venant peut développer des discontinuités en temps finis. Les méthodes largement utilisées pour la résolution des équations de ce système sont des schémas de volumes finis décentrés. Ces méthodes nécessitent la résolution du problème de Riemann au niveau des interfaces des cellules de calcul.

Sommairement, la différence principale entre les schémas décentrés et les schémas centrés réside dans l'utilisation des informations sur les caractéristiques de propagation des ondes. Les schémas décentrés utilisent de façon intensive les informations sur les caractéristiques de propagation des ondes, alors que les schémas centrés sont principalement basés sur les valeurs moyennes des flux sans faire appel à ces informations. Les schémas centrés rencontrent plusieurs problèmes de stabilités surtout dans le cas des solutions discontinues. L'intérêt pour les schémas centrés s'est accru après le travail important de Nessyahu et Tadmor (1990), où un schéma centré à capture de chocs a été développé. Dans Russo (2002), une revue de littérature détaillée est donnée sur les schémas centrés et leurs extensions. Ces schémas peuvent être améliorés de façon considérable en utilisant quelques informations sur les vitesses de propagation des ondes. Ceci a conduit à la construction des schémas partiellement centrés qui sont développés par Kurganov et co-auteurs (Kurganov et al., 2001; Kurganov et Petrova, 2001; Kurganov et Tadmor, 2000a). Ces schémas sont simples vu qu'ils ne font pas appel à la résolution du problème de Riemann et ils ont une bonne résolution des lois de conservation dans le cas des solutions discontinues par rapport aux schémas centrés.

L'objectif de ce chapitre est de développer un schéma numérique sans solveur de Riemann pour les écoulements peu profonds. Un nouveau schéma équilibre partiellement centré est proposé pour la résolution numérique des équations de Saint-Venant avec une topographie variable sur un maillage non structuré. Le maillage utilisé est sous forme de polygones construits à partir d'un maillage triangulaire initial. Ce maillage présente des cellules de calcul avec une meilleure uniformité spatiale que le maillage triangulaire. L'utilité de ce type de maillage par rapport au maillage construit sur

[1]Cet article est réalisé en collaboration avec A. Mohammadian et A. Kurganov, et soumis pour publication sous la forme: A. Beljadid, A. Mohammadian, A. Kurganov, 2015, Well-balanced positivity preserving Cell-Vertex Central-Upwind Scheme for Shallow Water Flows. Computers & Fluids (Elsevier).

la base des cellules centrées est montrée par les analyses et comparaisons menées par Delis et al. (2011).

De nouvelles techniques sont proposées pour les reconstructions relatives à la topographie et aux variables primitives du système. Ces techniques permettent d'assurer la stabilité de la méthode et la positivité de la hauteur d'eau au cours du temps. On note que la condition de stabilité de Courant-Friedrichs-Lewy est moins restrictive que la condition établie pour garantir la positivité de la méthode proposée. Une nouvelle technique est proposée pour la discrétisation du terme source dû à la topographie pour préserver d'une manière exacte les états d'équilibre du système.

Les performances de la méthode proposée sont validées par des tests numériques. Les résultats obtenus montrent que la méthode proposée est stable et permet de résoudre les petites perturbations des états d'équilibre. Elle assure l'équilibre entre le terme de flux et le terme source et elle préserve la positivité de la hauteur d'eau au cours du temps. Cette méthode peut être appliquée au système de Saint-Venant lorsque la topographie est discontinue sur des domaines complexes qui nécessitent l'utilisation des maillages non structurés. Même si les grilles triangulaires ont été utilisées pour le maillage primaire, une extension de la méthode des volumes finis proposée pour le cas général des grilles primaires de type polygonal peut être menée de manière simple.

4.1 Introduction

This paper focuses on development of modern numerical methods for the two-dimensional (2-D) Saint-Venant system of shallow water equations (SWEs):

$$\begin{cases} h_t + (hu)_x + (hv)_y = 0, \\ (hu)_t + \left(hu^2 + \dfrac{g}{2}h^2\right)_x + (huv)_y = -ghB_x, \\ (hv)_t + (huv)_x + \left(hv^2 + \dfrac{g}{2}h^2\right)_y = -ghB_y. \end{cases} \tag{4.1}$$

Here, h is the water depth, $(u, v)^T$ is the velocity field, the function $B(x, y)$ represents the bottom elevation, and g is the acceleration due to gravity.

The system (4.1) is a hyperbolic system of conservation (if $B_x \equiv B_y \equiv 0$) or balance (if B is not a constant) laws. Many upwind and central schemes have been proposed for numerical solutions of hyperbolic (systems of) PDEs. Roughly speaking, the main difference between upwind and central schemes is that upwind schemes use characteristic information to determine nonlinear wave propagation, while central schemes are based on averaging over the waves without using their detailed structures. Even though upwind schemes may be highly accurate, they utilize (approximate) Riemann problem solvers and characteristics decomposition and thus they are typically computationally

expensive. Riemann-problem-solver-free central schemes have become very popular after the pioneer work of Nessyahu and Tadmor (1990), where a second-order, shock-capturing, finite volume central scheme on a staggered grid was proposed. Since 1990, several higher-order and multidimensional extensions and generalizations of staggered central schemes have been introduced (e.g. Russo, 2002, and references therein). Central schemes can be improved by using some characteristic information on local speeds of propagation. This leads to a class of central-upwind schemes developed by Kurganov and co-authors (e.g. Kurganov et al., 2001; Kurganov and Petrova, 2001; Kurganov and Tadmor, 2000a). The central-upwind schemes are simple Riemann-problem-solver-free methods that have been successfully applied to a variety of problems including gas dynamics (Karni et al., 2002; Kurganov et al., 2001; Kurganov and Petrova, 2001; Kurganov et al., 2007; Kurganov and Tadmor, 2000a, 2002) and several shallow water models (Bollermann et al., 2013; Chertock et al., 2014; Kurganov and Levy, 2002; Kurganov and Miller, 2014; Kurganov and Petrova, 2007, 2008, 2009). Kurganov and Petrova (2005) extended the central-upwind scheme to triangular grids for solving general 2-D systems of conservation laws.

SWEs and related models are of great interest for many atmospheric and oceanic applications as well as for modeling flows in the rivers and coastal areas. To be able to accurately model realistic situations, one has to develop numerical methods on unstructured grids due to their flexibility to represent irregular domains and convenience of local mesh refinement.

There are two main proprieties a good numerical method for SWEs should satisfy. The first one is called a *well-balanced property:* The scheme should exactly preserve "lake at rest" steady-state solutions. The second property is *positivity preserving:* The method should guarantee positivity of the computed values of the water depth in each point of the domain at all times.

The main widely used unstructured finite volume methods are *cell-centered* (CCFVM) and *cell-vertex* (CVFVM) ones. The cell-vertex methods are sometimes referred to as node-centered, mesh-vertex or vertex-centered methods. For the CCFVM, the cells are the triangles of the primary mesh. For CVFVM the cells are the dual of the primary mesh as explained in the next section. For a detailed discussion on the two methods we refer the reader to Blazek (2006); Mavriplis (2008); Morton and Sonar (2007).

Delis et al. (2011) have recently performed an extensive comparison between CCFVM and CVFVM for 2-D SWEs. They studied the performance, robustness and defectiveness of the two methods by comparing numerical results with both analytical solutions and experimental and field data. They found that CVFVM lead to identical convergence behavior for grids with various qualities (in terms of orientation and distortion) while in CCFVM, the results are influenced by the grid quality. The CVFVM produce smoother computational cells, even for highly distorted meshes (see Nikolos and Delis,

2009). The reason is that the cells in CVFVM are constructed in a way that leads to more spatial uniformity than CCFVM. This motivates the use of cell-vertex approach in this paper.

Bryson et al. (2011) have proposed a central-upwind scheme on triangular grids for the Saint-Venant system of SWEs with possibly discontinuous bottom topography. The authors have showed that their method is well-balanced and positivity preserving, and demonstrated the high resolution and robustness of the method. In this paper, we introduce a new well-balanced positivity preserving central-upwind scheme on *cell-vertex* grids (described in Section 4.2.1) for 2-D SWEs with variable topography.

The paper is organized as follows. In Section 4.2, we present the new cell-vertex semi-discrete central-upwind scheme for the SWEs (4.1). In Section 4.3, we propose a positivity preserving reconstruction for water surface elevation. The well-balanced discretization of the source term is developed in Section 4.4. The positivity preserving property of the proposed scheme is proved in Section 4.5. In Section 4.6, we demonstrate the high resolution and robustness of the proposed method by testing it on a variety of numerical experiments. The final Section 4.7 contains concluding remarks.

4.2 The Cell-Vertex Central-Upwind Scheme

In this section, we focus on the derivation of the proposed cell-vertex central-upwind scheme. First the cell-vertex unstructured grid and the notations used in this paper are described in Section 4.2.1. Then, we develop the central-upwind method over cell-vertex grids for the SWEs (4.1), which can be rewritten using the vector of variables $\boldsymbol{U} := (w, p, q)^T$ as

$$\boldsymbol{U}_t + \boldsymbol{F}(\boldsymbol{U}, B)_x + \boldsymbol{G}(\boldsymbol{U}, B)_y = \boldsymbol{S}(\boldsymbol{U}, B) \tag{4.2}$$

with

$$\begin{aligned}
\boldsymbol{F}(\boldsymbol{U}, B) &= \left(p,\ \frac{p^2}{w-B} + \frac{g}{2}(w-B)^2,\ \frac{pq}{w-B}\right)^T, \\
\boldsymbol{G}(\boldsymbol{U}, B) &= \left(q,\ \frac{pq}{w-B},\ \frac{q^2}{w-B} + \frac{g}{2}(w-B)^2\right)^T, \\
\boldsymbol{S}(\boldsymbol{U}, B) &= \Big(0,\ -g(w-B)B_x,\ -g(w-B)B_y\Big)^T,
\end{aligned} \tag{4.3}$$

where $w := h + B$ represents the water surface elevation and $p := hu$ and $q := hv$ denote the discharges in the x- and y-directions, respectively.

4.2.1 Cell-Vertex Grid and Notations

Unstructured cell-vertex grids are obtained using a triangular discretization of the global domain $\mathcal{D}$: The finite volume cells, denoted by M_j, are centered around the vertices as shown in Figure 4.1. There are various methods to define the dual grid. In this paper, the boundary ∂M_j of the cell M_j around each internal triangulation vertex

P_j is defined by connecting the centers of mass of the surrounding triangles that have P_j as a common vertex. The water surface elevation w and the discharges p and q are then represented by the corresponding cell averages over the cells M_j of size $|M_j|$ with the centers of mass denoted by $G_j \equiv (x_j, y_j)$.

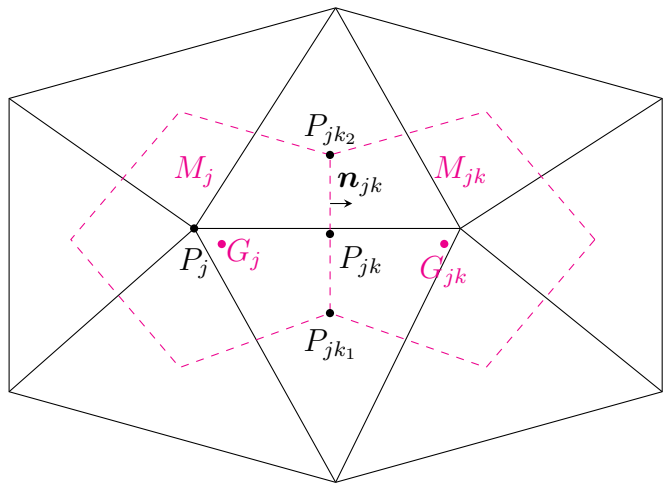

Figure 4.1 Sample of the cell-vertex. Solid lines represent the primary triangular grids and the dashed lines show the computational polygonal cells.

We assume that the discretization $\mathcal{D} = \bigcup_{j=1}^{N} M_j$ consists of N non-overlapping cells (N is equal to the number of nodes of the initial triangular gird). For each cell M_j we denote by m_j the number of its cell sides and by $M_{j1}, M_{j2}, \ldots, M_{jm_j}$ the neighboring cells that share with M_j a common side $(\partial M_j)_1$, $(\partial M_j)_2, \ldots, (\partial M_j)_{m_j}$, respectively. The side length and the outward unit normal vectors of the cell-interfaces are denoted by ℓ_{jk} and $n_{jk} := (\cos\theta_{jk}, \sin\theta_{jk})^T$, respectively, where θ_{jk} is the angle of the unit normal vector n_{jk} with the x-axis. The midpoint of the kth side of M_j is denoted by P_{jk} and the two nodes of this side are denoted by P_{jk_s}, where $s = 1, 2$. Note that the vertices of the cell M_j may be denoted by P_{jk_1}, which is the start point of the interface k of this cell. An anticlockwise orientation is used to define the start and end points of an interface. Similar indexing is used for other variables. For example, w_{jk_1}, h_{jk_1} and B_{jk_1} stand for the water surface elevation, water depth and the bottom elevation, respectively, at the vertex P_{jk_1}. Finally, for the sake of simplicity we use constant time step Δt and denote by $t^n := n\Delta t$ the nth time level.

Remark 1. It should be pointed out that although the initial grid is assumed to be triangular, the proposed cell-vertex method can be based on a quadrilateral or another polygonal unstructured initial grid. This leads to more flexibility of the cell-vertex methods compared to schemes that are restricted to triangular grids only.

4.2.2 The Semi-Discrete Form of the Scheme

The semi-discrete central-upwind scheme on cell-vertex grids can be derived following the procedure developed in Kurganov and Petrova (2005) and Bryson et al. (2011) for triangular grids. It can be shown that the resulting scheme is

$$\begin{aligned}\frac{d\overline{\boldsymbol{U}}_j}{dt} =& -\frac{1}{|M_j|}\sum_{k=1}^{m_j}\frac{\ell_{jk}\cos(\theta_{jk})}{a_{jk}^{\text{in}}+a_{jk}^{\text{out}}}\left[a_{jk}^{\text{in}}\boldsymbol{F}(\boldsymbol{U}_{jk}(P_{jk}),B_{jk})+a_{jk}^{\text{out}}\boldsymbol{F}(\boldsymbol{U}_j(P_{jk}),B_{jk})\right]\\ &-\frac{1}{|M_j|}\sum_{k=1}^{m_j}\frac{\ell_{jk}\sin(\theta_{jk})}{a_{jk}^{\text{in}}+a_{jk}^{\text{out}}}\left[a_{jk}^{\text{in}}\boldsymbol{G}(\boldsymbol{U}_{jk}(P_{jk}),B_{jk})+a_{jk}^{\text{out}}\boldsymbol{G}(\boldsymbol{U}_j(P_{jk}),B_{jk})\right]\\ &+\frac{1}{|M_j|}\sum_{k=1}^{m_j}\ell_{jk}\frac{a_{jk}^{\text{in}}a_{jk}^{\text{out}}}{a_{jk}^{\text{in}}+a_{jk}^{\text{out}}}\left[\boldsymbol{U}_{jk}(P_{jk})-\boldsymbol{U}_j(P_{jk})\right]+\overline{\boldsymbol{S}}_j,\end{aligned} \tag{4.4}$$

where $\overline{\boldsymbol{U}}_j \approx \frac{1}{|M_j|}\int_{M_j}\boldsymbol{U}(x,y,t)\,dxdy$ is the approximation of the cell averages of the solution at time t, and the quantity $\overline{\boldsymbol{S}}_j$ is a discretization of the cell averages of the source term, $\overline{\boldsymbol{S}}_j \approx \frac{1}{|M_j|}\int_{M_j}\boldsymbol{S}(\boldsymbol{U}(x,y,t),B(x,y))\,dxdy$, which will be discussed in Section 4.4.

The semi-discrete scheme (4.4) uses the bottom elevation $B_{jk} := B(P_{jk})$ at the midpoint of the kth cell interface, and the values $\boldsymbol{U}_j(P_{jk})$ and $\boldsymbol{U}_{jk}(P_{jk})$ at time t on the two sides of this interface (inside and outside of the cell M_j, respectively) are determined by using the positivity preserving piecewise linear reconstruction as explained in Sections 4.2.3, 4.2.4 and 4.3.

Finally, the one-sided local speeds of propagation at the kth interface of the cell M_j can be estimated using the smallest, $\lambda_1[V_{jk}]$, and largest, $\lambda_3[V_{jk}]$, eigenvalues of the Jacobian

$$V_{jk} = \cos(\theta_{jk})\frac{\partial\boldsymbol{F}}{\partial\boldsymbol{U}} + \sin(\theta_{jk})\frac{\partial\boldsymbol{G}}{\partial\boldsymbol{U}},$$

as follows:

$$\begin{aligned}a_{jk}^{\text{in}} &= -\min\left\{\lambda_1[V_{jk}(\boldsymbol{U}_j(P_{jk}))],\lambda_1[V_{jk}(\boldsymbol{U}_{jk}(P_{jk}))],0\right\},\\ a_{jk}^{\text{out}} &= \max\left\{\lambda_3[V_{jk}(\boldsymbol{U}_j(P_{jk}))],\lambda_3[V_{jk}(\boldsymbol{U}_{jk}(P_{jk}))],0\right\}.\end{aligned} \tag{4.5}$$

Remark 2. If the value of $a_{jk}^{\text{in}} + a_{jk}^{\text{out}}$ in equation (4.4) is zero or very close to zero (smaller than 10^{-10} in all of our numerical experiments), we avoid division by zero or by a very small number using the following approximations:

$$\begin{aligned}\frac{\left[a_{jk}^{\text{in}}\boldsymbol{F}(\boldsymbol{U}_{jk}(P_{jk}),B_{jk})+a_{jk}^{\text{out}}\boldsymbol{F}(\boldsymbol{U}_j(P_{jk}),B_{jk})\right]}{a_{jk}^{\text{in}}+a_{jk}^{\text{out}}} &\approx \frac{\left[\boldsymbol{F}(\boldsymbol{U}_{jk}(P_{jk}),B_{jk})+\boldsymbol{F}(\boldsymbol{U}_j(P_{jk}),B_{jk})\right]}{2},\\ \frac{\left[a_{jk}^{\text{in}}\boldsymbol{G}(\boldsymbol{U}_{jk}(P_{jk}),B_{jk})+a_{jk}^{\text{out}}\boldsymbol{G}(\boldsymbol{U}_j(P_{jk}),B_{jk})\right]}{a_{jk}^{\text{in}}+a_{jk}^{\text{out}}} &\approx \frac{\left[\boldsymbol{G}(\boldsymbol{U}_{jk}(P_{jk}),B_{jk})+\boldsymbol{G}(\boldsymbol{U}_j(P_{jk}),B_{jk})\right]}{2},\\ \frac{a_{jk}^{\text{in}}a_{jk}^{\text{out}}}{a_{jk}^{\text{in}}+a_{jk}^{\text{out}}}\left[\boldsymbol{U}_{jk}(P_{jk})-\boldsymbol{U}_j(P_{jk})\right] &\approx 0.\end{aligned}$$

Remark 3. The semi-discretization (4.4) is a system of ODEs, which has to be integrated in time using a sufficiently accurate and stable ODE solver. In all of the numerical experiments reported in Section 4.6, we have used the third-order strong stability preserving (SSP) Runge-Kutta method (see, e.g. Gottlieb et al., 2011, 2001).

4.2.3 Continuous Piecewise Linear Approximation of the Bottom

Let us assume that the bottom topography function B is a piecewise smooth function. In order to construct its continuous piecewise linear approximation, we first define the values of B at the vertices $P_{jk_i} = (x_{jk_i}, y_{jk_i})$ of the cell M_j. If the function B is continuous at P_{jk_i}, we simply take $B_{jk_i} := B(x_{jk_i}, y_{jk_i})$, otherwise we set

$$B_{jk_i} := \frac{1}{2}\left[\lim_{\varepsilon\to 0}\max_{\|\zeta\|=\varepsilon} B(x_{jk_i}+\zeta_x, y_{jk_i}+\zeta_y) + \lim_{\varepsilon\to 0}\min_{\|\zeta\|=\varepsilon} B(x_{jk_i}+\zeta_x, y_{jk_i}+\zeta_y)\right],$$

where $\zeta = (\zeta_x, \zeta_y)^T$.

We then obtain the approximate values of B at the midpoint of the interface connecting the points P_{jk_1} and P_{jk_2} using a linear approximation resulting in

$$B_{jk} = \frac{1}{2}\left(B_{jk_1} + B_{jk_2}\right).$$

Equipped with the approximate values of B at the midpoints of each cell interface, we approximate the value of B at the center of mass G_j by

$$B_j = \frac{1}{|M_j|}\int_{M_j} B(x,y)\,dxdy \approx \sum_{k=1}^{m_j} \mu_k B_{jk}, \tag{4.6}$$

where $\mu_k = \mathcal{A}_{jk}/|M_j|$ and $\mathcal{A}_{jk}$ is the area of the triangle $G_j P_{jk_1} P_{jk_2}$ (see Figure 4.1). Given the values at the vertices, B_j, B_{jk_1} and B_{jk_2}, we obtain a linear approximation of B over this triangle.

Finally, we obtain the continuous piecewise linear approximation of B at the cell M_j by taking the union of m_j planes over the corresponding triangles connecting the two neighboring vertices of M_j and its center of mass.

4.2.4 Piecewise Linear Reconstruction

In order to obtain a piecewise linear reconstruction of w, p and q, we need to approximate its gradients in each cell. The gradient of the ith component of $\boldsymbol{U}$ in the cell M_j,

denoted by $\nabla U_j^{(i)}$ $(i = 1, 2, 3)$, is computed using the Green-Gauss theorem as follows:

$$\nabla U_j^{(i)} = \frac{1}{|M_j|} \int\limits_{M_j} \nabla U_j^{(i)}(x, y)\, dxdy = \frac{1}{|M_j|} \sum_{k=1}^{m_j} \int\limits_{(\partial M_j)_k} \widetilde{U}_{jk}^{(i)} n_{jk}^{(i)}\, ds, \tag{4.7}$$

where $\widetilde{U}_{jk}^{(i)}$ is the estimated value of $U^{(i)}$ on the cell interface $(\partial M_j)_k$.

To prevent oscillations, we propose the following minmod-type reconstruction. To this end, we compute m_j gradients. The kth gradient $(k \in \{1, \ldots, m_j\})$ is calculated using equation (4.7) with the average values on each cell interface are obtained using the following two-step procedure:

- First, we take $\widetilde{U}_{js}^{(i)} = \left(\overline{U}_j^{(i)} + \overline{U}_{js}^{(i)}\right)/2$ for $i = 1, 2, 3$ and for all s except for $s = k$, for which we use an average of the values obtained at the interfaces attached to the kth interface.
- Then, the obtained values of $\widetilde{U}_{js}^{(1)} = \widetilde{w}_{js}$ are corrected: If $\widetilde{w}_{js} < B_{js}$, that is, if the estimated value of w is below the bottom elevation at the midpoint of the sth cell interface, we raise that value to $\widetilde{w}_{js} := B_{js}$.

Finally, for each variable w, hu and hv, out of the m_j gradient values we select the one that has the smallest magnitude and use the obtained numerical gradients $\nabla \boldsymbol{U}_j = ((\boldsymbol{U}_x)_j, (\boldsymbol{U}_y)_j)^T$ to build the corresponding linear pieces in the cell M_j:

$$\boldsymbol{U}_j(x, y) := \overline{\boldsymbol{U}}_j + (\boldsymbol{U}_x)_j(x - x_j) + (\boldsymbol{U}_y)_j(y - y_j). \tag{4.8}$$

The values $\boldsymbol{U}_j(P_{jk})$ required in (4.4), (4.5) are then obtained by substituting the coordinates of P_{jk} into (4.8).

Remark 4. Note that the reconstruction (4.8) satisfies the relationship similar to (4.6), established for the continuous piecewise linear reconstruction of the bottom topography in Section 4.2.3. In particular, for the water surface elevation w and the water depth $h := w - B$ we have

$$\sum_{k=1}^{m_j} \mu_k w(P_{jk}) = \overline{w}_j, \qquad \sum_{k=1}^{m_j} \mu_k h(P_{jk}) = \overline{h}_j, \tag{4.9}$$

which will be used in the proof of positivity preserving property of the scheme presented in Section 4.5.

We would like to point out that the piecewise linear reconstruction procedure for w presented in this section does not guarantee positivity of the reconstructed values of h. Therefore, this reconstruction has to be corrected to preserve the positivity of h.

4.3 Positivity Preserving Reconstruction for Water Surface Elevation

In this section, we propose an algorithm for the positivity preserving reconstruction of w. We say that the reconstruction is positivity preserving if it leads to nonnegative computed values of water depth at all of the cell vertices. The obtained reconstruction can be viewed as a correction of the basic piecewise linear reconstruction

$$w_j(x,y) := \overline{w}_j + (w_x)_j(x-x_j) + (w_y)_j(y-y_j) = \overline{w}_j + \nabla w_j \cdot (x-x_j, y-y_j)^T,$$

where the gradient $\nabla w_j = ((w_x)_j, (w_y)_j)^T$ is calculated using the modified minmod-type limiter described in Section 4.2.4.

We will distinguish between the three cases depending on the amount of water present in the cell M_j and on the local properties of the piecewise linear bottom approximation.

Case 1 (Wet Cells). We first consider the cells in which the water surface elevation $\overline{w}_j$ is greater than or equal to the bed elevation at all of the vertices of the cell M_j, that is, $\overline{w}_j \geq B_{jk_1}$ for all $k \in [1, m_j]$. In this case, it is possible to construct a single-plane reconstruction over the entire cell M_j. The reconstruction will take the form

$$w_j(x,y) := \overline{w}_j + \alpha \nabla w_j \cdot (x-x_j, y-y_j)^T, \tag{4.10}$$

where a proper selection of the parameter $\alpha \in [0,1]$ will help to respect positivity of the water depth.

To obtain the values of water surface elevation and water depth at the cell vertices (denoted by P_{jk_1}) we use

$$w_{jk_1} = \overline{w}_j + \alpha \nabla w_j \cdot \overrightarrow{G_j P_{jk_1}}, \tag{4.11}$$

and

$$h_{jk_1} = w_{jk_1} - B_{jk_1} = \overline{w}_j - B_{jk_1} + \alpha \nabla w_j \cdot \overrightarrow{G_j P_{jk_1}}. \tag{4.12}$$

The condition that the water surface elevation is greater than or equal to the bed elevation at all of the vertices of the cell M_j implies that the set of parameters $\alpha \in [0,1]$ which guarantee the positivity of h_{jk_1} at all of the cell vertices, is not empty since it contains $\alpha = 0$. We then consider the largest α in this set denoted by $\alpha_{\max}$ and we use the single-plane reconstruction based on the gradient $\alpha_{\max} \nabla w_j$. The parameter $\alpha_{\max}$ can be easily obtained by requiring $h_{jk_1} \geq 0$ in (4.12) for all k_1.

Case 2 (Partially Wet Cells with the Possibility of Single-Plane Reconstruction (4.10)*).* The second possible case corresponds to the situation, in which there are some cell vertices P_{jk_1} for which $\overline{w}_j < B_{jk_1}$. We split the vertices P_{jk_1}, $k = 1, \ldots, m_j$ into two separate sets: wet vertices where $\overline{w}_j \geq B_{jk_1}$, and dry vertices where $\overline{w}_j < B_{jk_1}$. Due to (4.9) and since $\overline{w}_j \geq B_j$, the set of wet vertices is not empty.

Similar to Case 1, we consider the parameter $\alpha_{\max}$ such that for all $\alpha \in [0, \alpha_{\max}]$ the values of the water depth obtained using equation (4.12) are nonnegative for all wet vertices. If h is also nonnegative for $\alpha = \alpha_{\max}$ at all of the dry vertices, we use $\alpha = \alpha_{\max}$ for the single-plane reconstruction (4.11), (4.12). Otherwise, no single-plane positivity preserving reconstruction is possible and we build a reconstruction consisting of m_j planes defined over the cell M_j.

Case 3 (Partially Wet Cells Not Included in Case 2). In this case, we propose a reconstruction with the minimal deviation from the direction of the initial gradient ∇w_j.

Case 3a: We first consider partially wet cells with only one dry vertex P_{jk_1} (that is, $\overline{w}_j < B_{jk_1}$). In the reconstruction, we set zero water depth at this point (that is, we set $w(P_{jk_1}) := B_{jk_1}$) and since the linear reconstruction should also satisfy $w(G_j) = \overline{w}_j$, we only need a third point to complete the reconstruction. To this end, we consider $m_j - 1$ planes passing through these two points and the point $(P_{jk_1'}, B_{jk_1'})$ for $k_1' \neq k_1$ and we compute their gradients $\{\nabla w_{k_1'}\}$. We then consider only those gradients that lead to positive reconstructions and out of them select the gradient which has the minimal deviation from the direction of the initial gradient ∇w_j by computing the angles between ∇w_j and $\nabla w_{k_1'}$. If none of the gradients $\nabla w_{k_1'}$ guarantees a positive reconstruction, we proceed with Case 3b.

Case 3b: Finally, we consider partially wet cells not covered by Case 3a. We now use a union of m_j planes defined over the cell M_j. First, we set zero depth at the cell vertices at which the condition $\overline{w}_j < B_{jk_1}$ is satisfied. There are many possibilities for the reconstruction, but in order to avoid oscillations we consider the constant depth denoted by $\widehat{h}_j$ at the other vertices where $\overline{w}_j \geq B_{jk_1}$. The value of $\widehat{h}_j$ can be obtained using the conservation requirement (4.9) as follows:

$$\widehat{h}_j = \frac{\overline{h}_j}{\sum_{k=1}^{m_j} \mu_k \varepsilon_k}, \quad \text{where} \quad \varepsilon_k = \begin{cases} 1, & \text{if } \overline{w}_j \geq B_{jk_1} \text{ and } \overline{w}_j \geq B_{jk_2}, \\ 0, & \text{if } \overline{w}_j < B_{jk_1} \text{ and } \overline{w}_j < B_{jk_2}, \\ 1/2, & \text{otherwise.} \end{cases}$$

Remark 5. In all of the cases considered above, the values of water surface elevation and water depth at the midpoints P_{jk} are obtained from the values at the cell vertices by

$$w(P_{jk}) = \frac{w(P_{jk_1}) + w(P_{jk_2})}{2}, \qquad h(P_{jk}) = \frac{h(P_{jk_1}) + h(P_{jk_2})}{2}.$$

Therefore, if the reconstructed water depth is nonnegative at all of the cell vertices, it will be also nonnegative over the entire cell, and in particular, at the midpoints of its interfaces. The positivity of the water depth at the midpoints of the interfaces will be crucial in the proof of the positivity preserving property of the scheme presented in Section 4.5.

4.4 Well-Balanced Discretization of the Source Term

The proposed semi-discrete central-upwind scheme (4.4) includes the cell average of the source term $\overline{\boldsymbol{S}}_j \equiv \left(0, \overline{S}_j^{(2)}, \overline{S}_j^{(3)}\right)^T$. To design a well-balanced scheme, that is, a scheme that exactly preserves "lake at rest" steady-state solutions satisfying $w \equiv C$, $u \equiv v \equiv 0$, where C is a constant, a special quadrature has to be designed.

Note that for a given "lake at rest" solution, $\boldsymbol{U}_j(P_{jk}) = \boldsymbol{U}_{jk}(P_{jk}) = (C, 0, 0)^T$, and the two momentum equations of the semi-discrete scheme (4.4) reduce to

$$\begin{aligned} -\frac{g}{2|M_j|}\sum_{k=1}^{m_j}\ell_{jk}\cos(\theta_{jk})(C-B_{jk})^2 + \overline{S}_j^{(2)} &= 0,\\ -\frac{g}{2|M_j|}\sum_{k=1}^{m_j}\ell_{jk}\sin(\theta_{jk})(C-B_{jk})^2 + \overline{S}_j^{(3)} &= 0. \end{aligned} \tag{4.13}$$

In the remaining part of the section, we derive a quadrature that satisfies the well-balancing conditions (4.13).

First, the source term $\overline{S}_j^{(2)}$ can be rewritten in the following form using the divergence theorem:

$$\begin{aligned} \overline{S}_j^{(2)} &= -\frac{g}{|M_j|}\int\limits_{M_j}(w-B)B_x\,dxdy = \frac{g}{2|M_j|}\int\limits_{M_j}((w-B)^2)_x\,dxdy - \frac{g}{|M_j|}\int\limits_{M_j}(w-B)w_x\,dxdy\\ &= \frac{g}{2|M_j|}\sum_{k=1}^{m_j}\int\limits_{(\partial M_j)_k}(w-B)^2\cos(\theta_{jk})\,ds - \frac{g}{|M_j|}\int\limits_{M_j}(w-B)w_x\,dxdy. \end{aligned} \tag{4.14}$$

We then apply the midpoint rule to approximate the integrals on the right-hand side (RHS) of (4.14) to obtain the well-balanced quadrature for $\overline{S}_j^{(2)}$:

$$\overline{S}_j^{(2)} = \frac{g}{2|M_j|}\sum_{k=1}^{m_j}\ell_{jk}(w_j(P_{jk})-B_{jk})^2\cos(\theta_{jk}) - g(w_x)_j(\overline{w}_j - B_j). \tag{4.15}$$

Similarly, the well-balanced quadrature for the source term $\overline{S}_j^{(3)}$ is

$$\overline{S}_j^{(3)} = \frac{g}{2|M_j|}\sum_{k=1}^{m_j}\ell_{jk}(w_j(P_{jk})-B_{jk})^2\sin(\theta_{jk}) - g(w_y)_j(\overline{w}_j - B_j). \tag{4.16}$$

Indeed, the quadratures (4.15) and (4.16) are well-balanced since the terms on the RHS of (4.15) and (4.16) containing the derivatives $(w_x)_j$ and $(w_y)_j$ vanish for the "lake at rest" solution $\boldsymbol{U} \equiv (C, 0, 0)^T$, and the well-balancing conditions (4.13) are satisfied.

4.5 Positivity Preserving Property of the Scheme

In this section, we prove the positivity preserving property of the proposed central-upwind scheme.

Theorem 1. Consider the semi-discrete central-upwind scheme (4.4) for the Saint-Venant system (4.2), (4.3). Let the ODE system (4.4) is integrated using the forward Euler method. We assume that at time $t = t^n$ the computed water depth is nonnegative, that is, $\overline{w}_j^n \geq B_j$ for all j and that the time step size is restricted by

$$\Delta t \leq \frac{1}{2a} \min_{j,k}\{d_{jk}\}, \tag{4.17}$$

where $a = \max\{a_{jk}^{\text{in}}, a_{jk}^{\text{out}}\}$ and d_{jk} is the distance between the center of mass G_j of the cell M_j and its kth interface $P_{jk_1}P_{jk_2}$.

Then $\overline{w}_j^{n+1} \geq B_j$ for all j at time $t = t^{n+1}$.

Proof. Applying the forward Euler temporal discretization to the first equation in (4.4) yields

$$\begin{aligned}
\overline{w}_j^{n+1} = \overline{w}_j^n &- \frac{\Delta t}{|M_j|} \sum_{k=1}^{m_j} \frac{\ell_{jk}\cos(\theta_{jk})}{a_{jk}^{\text{in}} + a_{jk}^{\text{out}}} \left[a_{jk}^{\text{in}}(hu)_{jk}(P_{jk}) + a_{jk}^{\text{out}}(hu)_j(P_{jk})\right] \\
&- \frac{\Delta t}{|M_j|} \sum_{k=1}^{m_j} \frac{\ell_{jk}\sin(\theta_{jk})}{a_{jk}^{\text{in}} + a_{jk}^{\text{out}}} \left[a_{jk}^{\text{in}}(hv)_{jk}(P_{jk}) + a_{jk}^{\text{out}}(hv)_j(P_{jk})\right] \\
&+ \frac{\Delta t}{|M_j|} \sum_{k=1}^{m_j} \ell_{jk} \frac{a_{jk}^{\text{in}} a_{jk}^{\text{out}}}{a_{jk}^{\text{in}} + a_{jk}^{\text{out}}} \left[w_{jk}(P_{jk}) - w_j(P_{jk})\right].
\end{aligned} \tag{4.18}$$

Recall that the reconstruction proposed in Sections 4.2.4 and 4.3 guarantees that the water surface elevation and the water depth satisfy the following equalities at time level $t = t^n$:

$$\overline{w}_j^n = \sum_{k=1}^{m_j} \mu_k w_j(P_{jk}), \qquad \overline{h}_j^n = \sum_{k=1}^{m_j} \mu_k h_j(P_{jk}), \tag{4.19}$$

and the inequalities $h_j(P_{jk}) \geq 0$ and $h_{jk}(P_{jk}) \geq 0$.

Since the piecewise linear reconstruction of the bottom topography is continuous we have $w_{jk}(P_{jk}) - w_j(P_{jk}) = h_{jk}(P_{jk}) - h_j(P_{jk})$ for each $k \in [0, m_j]$. We then use this equality together with (4.19) to rewrite equation (4.18) in the following form:

$$\begin{aligned}
\overline{h}_j^{n+1} = &\frac{\Delta t}{|M_j|} \sum_{k=1}^{m_j} h_{jk}(P_{jk}) \frac{\ell_{jk} a_{jk}^{\text{in}}}{a_{jk}^{\text{in}} + a_{jk}^{\text{out}}} \left[a_{jk}^{\text{out}} - u_{jk}^{\theta}(P_{jk})\right] \\
&+ \sum_{k=1}^{m_j} h_j(P_{jk}) \Big(\mu_k - \frac{\Delta t}{|M_j|} \cdot \frac{\ell_{jk} a_{jk}^{\text{out}}}{a_{jk}^{\text{in}} + a_{jk}^{\text{out}}} \left[a_{jk}^{\text{in}} + u_j^{\theta}(P_{jk})\right]\Big),
\end{aligned} \tag{4.20}$$

where

$$u^{\theta}_{jk}(P_{jk}) := \cos(\theta_{jk})u_{jk}(P_{jk})+\sin(\theta_{jk})v_{jk}(P_{jk}),\ \ u^{\theta}_{j}(P_{jk}) := \cos(\theta_{jk})u_{j}(P_{jk})+\sin(\theta_{jk})v_{j}(P_{jk}).$$

Since $a^{\text{out}}_{jk} \geq u^{\theta}_{jk}(P_{jk})$ and $h_{jk}(P_{jk}) \geq 0$, the first term on the RHS of (4.20) is nonnegative. Since

$$\frac{\Delta t}{|M_j|} \cdot \frac{\ell_{jk} a^{\text{out}}_{jk}}{a^{\text{in}}_{jk} + a^{\text{out}}_{jk}} \left[a^{\text{in}}_{jk} + u^{\theta}_{j}(P_{jk})\right] \leq \frac{\Delta t}{|M_j|}\, \ell_{jk} a^{\text{out}}_{jk} \ \text{ and } \ h_j(P_{jk}) \geq 0,$$

the positivity of the second term on the RHS of (4.20) can be ensured by enforcing

$$\Delta t \leq \frac{\mu_k |M_j|}{\ell_{jk} a^{\text{out}}_{jk}}. \tag{4.21}$$

Finally, since $\mu_k = \mathcal{A}_{jk}/|M_j|$ and $\mathcal{A}_{jk} = d_{jk}\ell_{jk}/2$, the condition (4.21) is followed from the time step restriction (4.17). □

Remark 6. The proof of Theorem 1 is still valid if the forward Euler temporal discretization is replaced with a high-order SSP ODE solver, since one step of any SSP method consists of a convex combination of several forward Euler steps.

Remark 7. We note that the condition (4.17) only ensures the positivity preserving property of the designed scheme, but does not a-priori guarantees its stability. Similar to the stability requirement of the central-upwind scheme on the triangular meshes (Bryson et al., 2011; Kurganov and Petrova, 2005), we can formulate the CFL condition for the proposed cell-vertex central-upwind scheme: No nonlinear (possibly discontinuous) waves generated at the cell interfaces should reach the center of mass of the computational cell over a time step Δt. This leads to the following time step restriction:

$$\Delta t < \frac{1}{a} \min_{j,k}\{d_{jk}\},$$

which is less restrictive than (4.17). Therefore, the condition (4.17) is expected to ensure both the stability and positivity.

4.6 Numerical Examples

In this section, we demonstrate the performance of the proposed central-upwind scheme on a variety of benchmarks. In all of the numerical experiments, we take $g = 1$ except for Example 4, where we set $g = 9.812$. In Examples 1–3, the proposed scheme is employed to compute small perturbations of the "lake at rest" steady states in different contexts. In Example 4, we simulate a rapidly varying flow arising in modeling dam

breaking over discontinuous bottom topography.

Example 1 : Small Perturbation over an Exponential Hump

In the first example, we consider the benchmark originally proposed in LeVeque (1998) and then widely used in the literature, as well as its more challenging version. We study the ability of the cell-vertex central-upwind scheme to accurately capture the propagation of a small perturbation of the "lake at rest" steady state over an exponential hump described by

$$B(x,y) = 0.8\exp(-5(x-0.9)^2 - 50y^2).$$

The computational domain is $[0,2]\times[-0.5,0.5]$, and the water surface is initially at rest everywhere except for the stripe $0.05 < x < 0.15$, where a small perturbation is initially located:

$$w(x,y,0) = \begin{cases} 1+\varepsilon, & 0.05 < x < 0.15, \\ 1, & \text{otherwise}, \end{cases} \qquad u(x,y,0) \equiv v(x,y,0) \equiv 0.$$

First, we take a relatively large perturbation $\varepsilon = 0.01$ and compute the solution using the grid with an average cell area $|M_j| = 2.24\cdot 10^{-5}$. The evolution (at times $t = 0.6$, 0.9, 1.2, 1.5 and 1.8) of the right-going portion of the water surface perturbation is shown in the left column of Figure 4.2. As one can see, the obtained solution is oscillation-free and the achieved resolution is comparable to the resolution achieved in Bryson et al. (2011); Kurganov and Levy (2002); LeVeque (1998). To further verify the robustness of the proposed method, we take smaller perturbation values $\epsilon = 10^{-3}$ and $\epsilon = 10^{-4}$ and compute the solution at the same time moments $t = 0.6$, 0.9, 1.2, 1.5 and 1.8 using the grids with average cell areas $|M_j| = 1.25\cdot 10^{-5}$ and $|M_j| = 7.78\cdot 10^{-6}$, respectively. The obtained results are shown in the right column of Figure 4.2 and in Figure 4.3. It should be observed that the computed solutions are still oscillation-free and highly resolved. This demonstrates the ability of the scheme to accurately capture quasi-steady states.

Next, we demonstrate the importance of the proposed well-balanced discretization of the source term. To this end, we design a non well-balanced cell-vertex central-upwind scheme by replacing the well-balanced quadratures (4.15) and (4.16) with the midpoint rule:

$$\overline{S}_j^{(2)} = -g(\overline{w}_j - B_j)(B_x)_j, \qquad \overline{S}_j^{(3)} = -g(\overline{w}_j - B_j)(B_y)_j, \tag{4.22}$$

where the components of ∇B are obtained using the divergence theorem:

$$(B_x)_j = \frac{1}{|M_j|}\sum_{k=1}^{m_j} \ell_{jk}B_{jk}\cos(\theta_{jk}), \qquad (B_y)_j = \frac{1}{|M_j|}\sum_{k=1}^{m_j} \ell_{jk}B_{jk}\sin(\theta_{jk}). \tag{4.23}$$

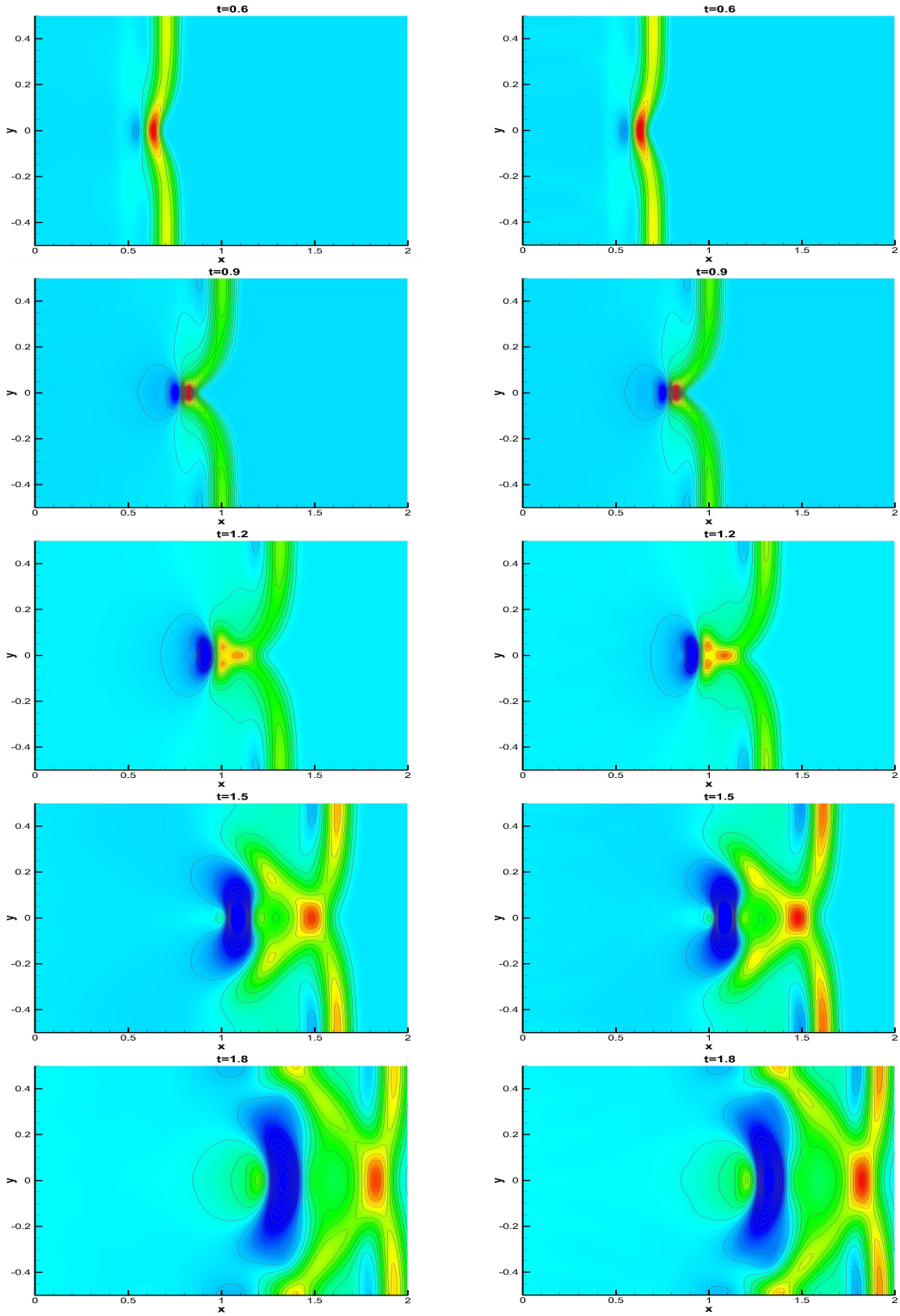

Figure 4.2 Example 1: Solution (w) computed by the well-balanced cell-vertex central-upwind scheme for $\varepsilon = 10^{-2}$ (left) and $\varepsilon = 10^{-3}$ (right).

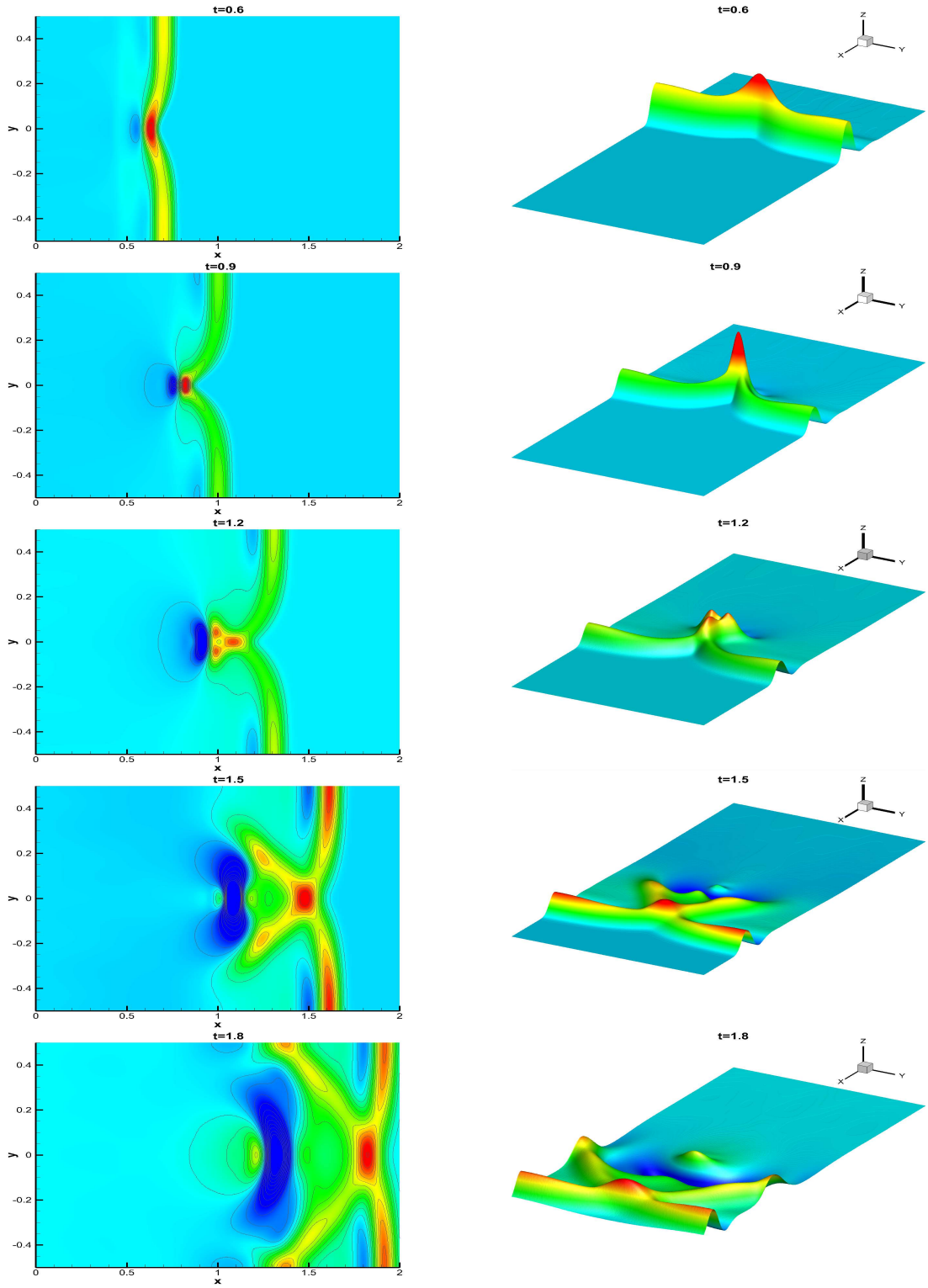

Figure 4.3 Example 1: Top (left) and three-dimensional (3-D) (right) views of the solution (w) computed by the well-balanced cell-vertex central-upwind scheme for $\varepsilon = 10^{-4}$.

The non well-balanced solution computed for $\varepsilon = 10^{-4}$ at time $t = 0.6$ on the same grid as before (with average cell areas $|M_j| = 7.78 \cdot 10^{-6}$) is shown in Figure 4.4. As one can see, the use of the non well-balanced scheme leads to spurious modes appearing at the plateau area which deform the solution. When we take a coarser mesh with average cell areas $|M_j| = 2.24 \cdot 10^{-5}$, the non well-balanced solution is severely deformed and is completely incorrect, see Figure 4.4 (middle). The well-balanced scheme applied on the same coarse mesh leads, on the contrary, to oscillation-free results as it is shown in Figure 4.4 (right). This clearly demonstrates a crucial role of a well-balanced source term quadrature.

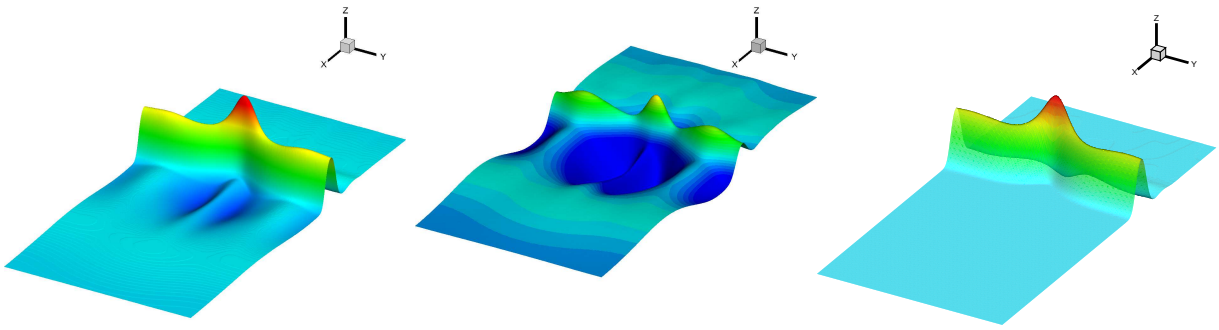

Figure 4.4 Example 1: Solutions (w) computed by the non well-balanced cell-vertex central-upwind scheme using the fine (left) and coarse (middle) grids and by the well-balanced cell-vertex central-upwind scheme using the coarse grid (right). Here, $\varepsilon = 10^{-4}$.

Example 2: Small Perturbation over Submerged Flat Plateau

In this example, we consider a submerged flat plateau as shown in Figure 4.5 (left). To further verify well-balanced and positivity preserving features of the proposed cell-vertex central-upwind scheme, we consider a slight modification of the test problem from Bryson et al. (2011), in which a small perturbation of the "lake at rest" steady state propagates over a submerged flat plateau located very close to the water surface.

The computational domain is $[0, 1] \times [-0.5, 0.5]$ and the bottom topography is given by

$$B(x, y) = \begin{cases} 1 - 2\varepsilon, & r \leq 0.1, \\ 10(1 - 2\varepsilon)(0.2 - r), & 0.1 \leq r \leq 0.2, \\ 0, & \text{otherwise}, \end{cases} \qquad r := \sqrt{(x - 0.5)^2 + y^2}. \tag{4.24}$$

The outflow boundary conditions are used in the x-direction, while the wall boundary conditions are imposed in the y-direction. As in Example 1, the initial conditions

correspond to a small perturbation of the "lake at rest" steady state:

$$w(x,y,0)=\begin{cases}1+\varepsilon, & 0.1<x<0.2,\\ 1, & \text{otherwise},\end{cases}\qquad u(x,y,0)\equiv v(x,y,0)\equiv 0.$$

We set $\varepsilon = 0.01$.

The solution is computed using the proposed cell-vertex central-upwind scheme with average cell areas $|M_j| = 2.24 \times 10^{-5}$. Figure 4.6 shows w computed at times $t = 0.2$, 0.35 and 0.65. As one can see, no oscillations are observed and the positivity of the water depth is preserved. We then compute the solution on the same grid at the same times, but using the non well-balanced central-upwind scheme described in Example 1. The obtained results are presented in Figure 4.7, where spurious deformations are clearly observed. These deformations lead to numerical oscillations and widely increase if a coarser mesh is used.

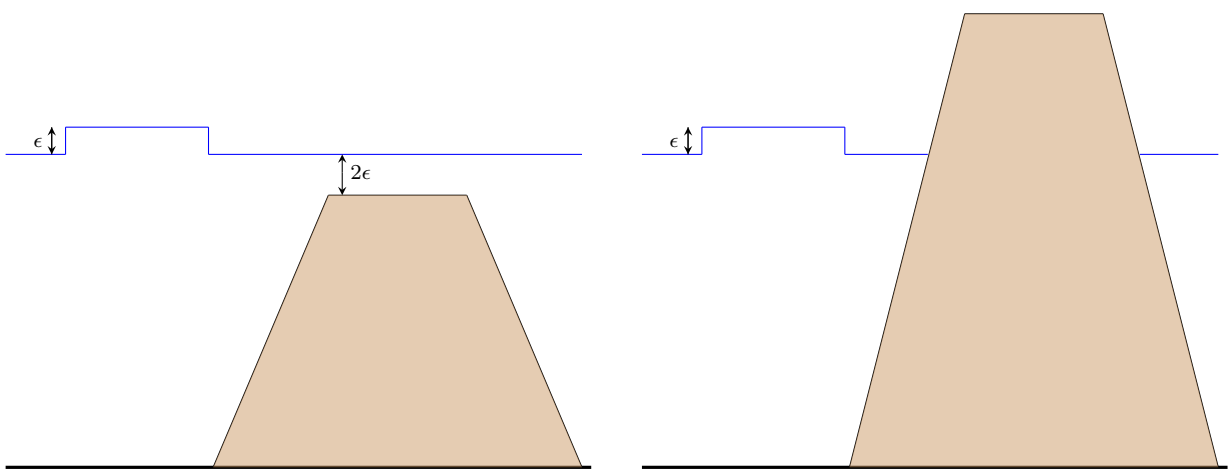

Figure 4.5 Examples 2 and 3: One-dimensional slices of the bottom topographies (4.24), left, and (4.25), right. These plots are not to scale.

Example 3: Small Perturbation Bending around a Round-Shape Island

This example, which is also a slight modification of the problem from Bryson et al. (2011), is designed to examine both well-balanced and positivity preserving properties of the studied scheme by testing its ability to handle a situation with a small perturbation of a "lake at rest" state propagating around an island. The round-shape island

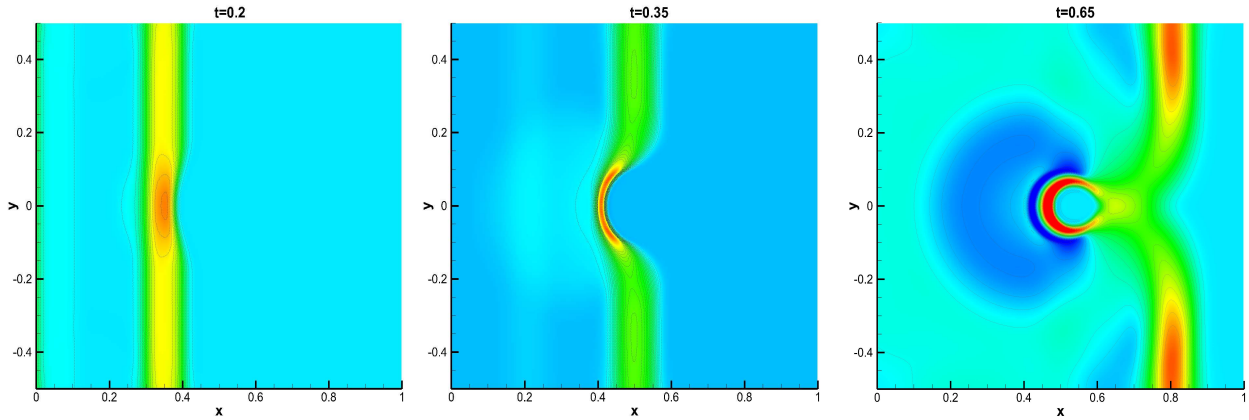

Figure 4.6 Example 2: Solution (w) computed by the well-balanced cell-vertex central-upwind scheme.

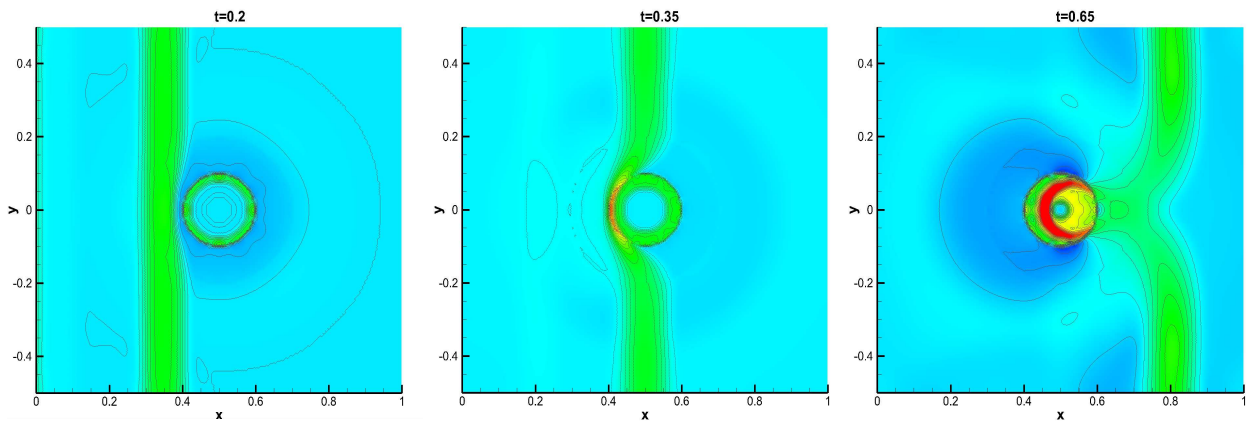

Figure 4.7 Example 2: Solution (w) computed by the non well-balanced cell-vertex central-upwind scheme.

(see Figure 4.5, right) is represented by the following bottom topography function:

$$B(x,y)=\begin{cases}1.1, & r\le 0.1,\\ 11(0.2-r), & 0.1<r<0.2,\\ 0, & \text{otherwise},\end{cases}\qquad r:=\sqrt{x^2+y^2}, \tag{4.25}$$

which is located in the center of the computational domain $[-0.5,0.5]\times[-0.5,0.5]$. The initial condition given by

$$w(x,y,0)=\begin{cases}1+\varepsilon, & -0.4<x<-0.3,\\ \max\Big(1,B(x,y)\Big), & \text{otherwise},\end{cases}\qquad u(x,y,0)\equiv v(x,y,0)\equiv 0.$$

As in Example 2, the outflow boundary conditions are used in the x-direction, while the wall boundary conditions are imposed in the y-direction.

The solution computed at times $t = 0.35$, 0.50 and 0.65 for $\varepsilon = 0.01$ using the proposed cell-vertex central-upwind scheme with average cell areas $|M_j| = 1.25 \times 10^{-5}$ is shown in Figure 4.8. The flow around the dry parts of the island is of a special interest. As one can see, the wave bends around the island without any oscillations. On the contrary, the results obtained using a non well-balanced discretization of the source terms contain large artificial waves, which develop completely different solution structure, see Figure 4.9.

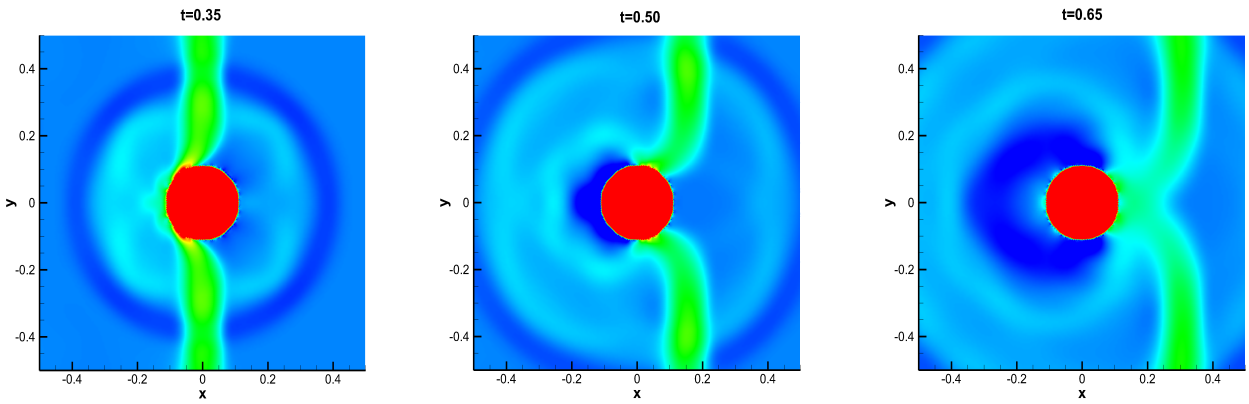

Figure 4.8 Example 3: Solution (w) computed by the well-balanced cell-vertex central-upwind scheme. The circle in the center of the computational domain represents the part of the bottom that is above the water surface.

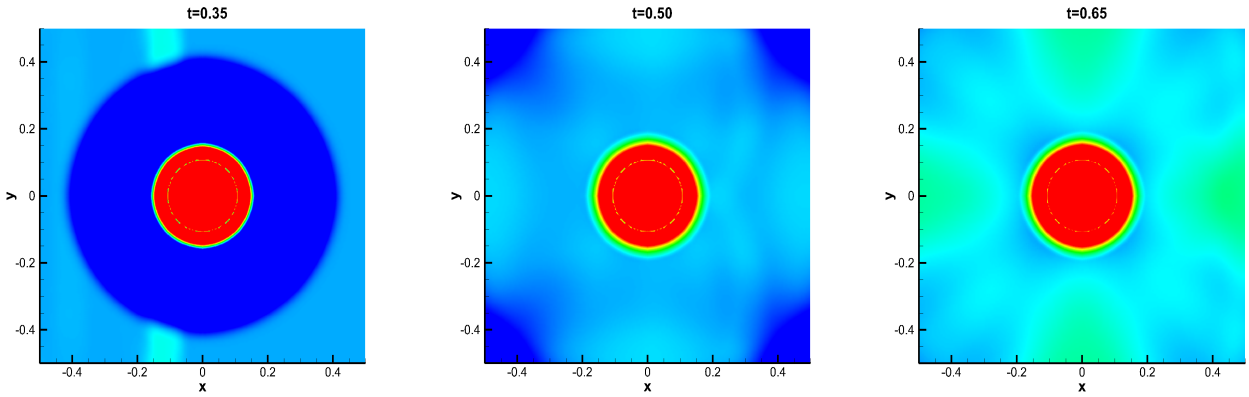

Figure 4.9 The same as Figure 4.8, but the solution (w) is computed by the non well-balanced cell-vertex central-upwind scheme.

Example 4: Dam Break over Discontinuous Topography

In the last example, we test the ability of the proposed cell-vertex central-upwind scheme to accurately resolve rapidly varying flows. We consider a one-dimensional dam break problem from Vukovic and Sopta (2002), see also Chen et al. (2013), which we solve using the 2-D code with outflow boundary conditions. In this problem, the bottom topography is given by

$$B(x,y) = \begin{cases} 8, & |x-750| \leq 187.5, \\ 0, & \text{otherwise}, \end{cases}$$

and the initial conditions are

$$w(x,y,0) = \begin{cases} 20, & x < 750, \\ 15, & \text{otherwise}, \end{cases} \qquad u(x,y,0) \equiv v(x,y,0) \equiv 0.$$

In Figure 4.10, we show the solutions computed using the cell-vertex central-upwind scheme at time $t = 15$ using the grid with average cell areas $M_j = 2.30$ and $M_j = 0.50$. The shock and rarefaction waves reach the discontinuities in the bottom topography at about $t \approx 17$. We then compute the solutions on the same two grids at time $t = 55$, at which the developed wave structure is much more complicated than at time $t = 15$. The obtained results are shown in Figure 4.11. The solutions computed by the proposed central-upwind scheme are in a good agreement with the solutions reported in Chen et al. (2013) and Vukovic and Sopta (2002), they are oscillation-free, and the achieved resolution is very high.

4.7 Conclusions

In this paper, we have introduced a new well-balanced, positivity preserving central-upwind scheme on unstructured cell-vertex grids for the Saint-Venant system of shallow water equations with variable bottom topography. We have proposed a novel non-oscillatory reconstruction in which the gradient of each variable is computed using a modified minmod-type method to ensure stability. The water surface reconstruction has been corrected to guarantee the positivity of the water depth over the entire computational cell. The well-balanced property of the scheme has been ensured by a special discretization of the source term cell averages.

The performance of the proposed cell-vertex central-upwind scheme was tested on a number of numerical examples. We have used the scheme to compute small perturbations of "lake at rest" steady-state solutions over several different bottom topographies, including the one that correspond to the island modeling. The proposed scheme has been also validated in the case of a rapidly varying flow over discontinuous bottom to-

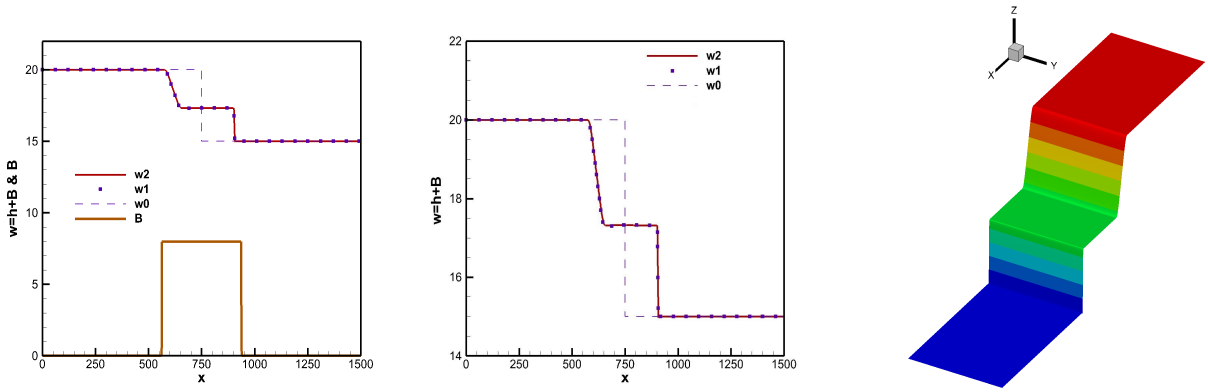

Figure 4.10 Example 4: Solution at time $t = 15$ computed by the proposed cell-vertex central-upwind scheme. In the left and middle graphs, a 1-D slice of the solution along the line $y = 0$ is shown. There, w_2 and w_1 are the water surface elevations computed using average cell areas $|M_j| = 0.50$ and $|M_j| = 2.30$, respectively, w_0 is the initial condition. The bottom topography B is plotted at the left. The 3-D view of the computed water surface is on the right.

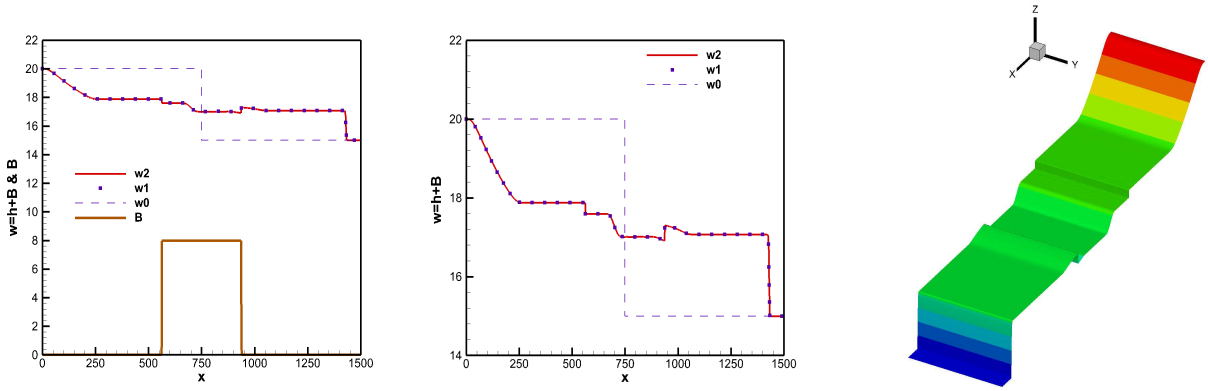

Figure 4.11 The same as Figure 4.10, but at a later time $t = 55$.

pography. The reported numerical experiments demonstrate the ability of the proposed method to avoid oscillations. The obtained numerical results confirm the well-balanced and the positivity preserving properties of the developed cell-vertex central-upwind scheme.

Even though an unstructured triangulation was used as a primary grid, an extension to more general polygon-type primary grids can be made in a straightforward manner.

CHAPITRE 5 Analyse des méthodes de volumes finis non structurées pour les écoulements en eaux peu profondes en utilisant le pseudo spectre

Analysis of Unstructured Finite Volume Methods for Shallow Water Flows using Pseudospectra

Résumé

La discrétisation des équations en eaux peu profondes dans le cas des maillages non structurés peut conduire à des modes numériques qui causent des problèmes de stabilité. Il convient de signaler que les conditions de stabilité asymptotique des méthodes numériques sont liées au comportement asymptotique des solutions. Ces conditions ne fournissent pas assez d'information sur le comportement des solutions pour les temps finis. Dans ce chapitre, on propose une nouvelle approche pour l'analyse de stabilité des méthodes numériques pour le cas des équations en eaux peu profondes en utilisant la notion du pseudo spectre. L'opérateur linéaire de la forme discrète du schéma numérique constitue le paramètre principal utilisé dans l'analyse.

Un maillage non structuré conduit en général à un opérateur discret non normal du schéma numérique. Dans certains cas, les vecteurs propres de l'opérateur peuvent être loin de l'orthogonalité, ce qui peut être source d'amplifications des solutions ou des problèmes de stabilité. Dans ces situations, même si le spectre correspondant à l'opérateur discret du schéma numérique a toutes ses valeurs propres à l'intérieur du cercle unité, le schéma peut conduire à des amplifications numériques de la solution. De grandes amplifications des solutions peuvent être observées même si la méthode numérique est stable au sens de Lax–Richtmyer. La nouvelle approche basée sur le pseudo spectre des matrices est efficace pour la vérification de la stabilité des méthodes de volumes finis pour les équations en eaux peu profondes. Les résultats de l'analyse peuvent être utiles pour le choix du type de maillage, des emplacements appropriés des variables primitives au niveau des cellules du maillage et de la méthode de discrétisation qui est stable pour une large gamme de modes. Pour les méthodes des volumes finis qui utilisent la méthode de Crank-Nicolson comme schéma temporel, on montre qu'il est important de considérer l'emplacement de l'ensemble des variables primitives au centre de chaque cellule de calcul.

5.1 Introduction

Pseudospectra have been used in several works (e.g. Demmel, 1990; Trefethen, 1997, 1999; Trefethen et al., 2001; Reddy, 1994). We recommend the 2005 book Spectra and Pseudospectra: The Behavior of Non-Normal Matrices and Operators by Trefethen and Embree, which provides more references and details about pseudospectra and the

behavior of non-normal matrices. The discrete operator of the numerical scheme which has a set of orthogonal eigenvectors is considered as normal (Trefethen and Emberee, 2005) . The condition number of this operator is 1. However, most discrete operators of schemes in shallow water are non-normal since they have non-orthogonal eigenvectors. Unfortunately for systems which are governed by a non-normal operator, the solutions may have large amplifications for finite times, and the eigenvalues of the operator are sensitive to perturbations. Pseudospectra are less employed in stability studies of numerical schemes. High-order finite-difference methods were analyzed by Zingg (1997) and Zingg and Lederle (2005) using spectra, pseudospectra, and singular value decomposition. The authors stated the importance of the use of these methods for detecting the instability of finite-difference schemes.

In this study, we focus on the stability analysis of selected unstructured finite volume methods for shallow water equations (SWEs) using pseudospectra. This is the first study in which these stability analysis techniques are used in shallow water. It should be mentioned that the conditions of the asymptotic stability are related to the asymptotic behavior of the solutions. These conditions do not provide any information on the behavior of the solution for finite times. In general, the discrete operators of most available schemes for SWEs are non-normal. In such cases, even if the spectrum corresponding to the fully-discrete form of the scheme has all eigenvalues within the unit circle, the scheme can have an unstable behavior or may lead to numerical amplification of the solution for finite times. Lax-Richtmyer stability is the commonly used definition of stability. However, as will be shown in this paper for finite volume methods on unstructured grids, there are some cases in which a scheme is Lax-Richtmyer stable but one still faces some unstable modes. This may cause numerical oscillations if the scheme is applied to simulate physical phenomena, which includes different types of waves such as long and short waves in shallow water equations. Since there are many solutions of SWEs that may have a very rich wave structure, stability analysis for a wide range of waves is required. In this paper, we will demonstrate the advantage of using pseudospectra to detect instabilities of finite volume schemes over unstructured grids. Several aspects are considered, such as the geometry of control volumes, the spatial methods used for the integration of the continuity and momentum equations and the boundary conditions. Although the spatial schemes considered in this study are not comprehensive, they illustrate various strategies to analyze the stability and to choose the suitable unstructured finite volume methods. In our analysis, we use five spatial methods combined with the Crank-Nicolson scheme.

The outline of this paper is as follows. In Section 5.2, SWEs and the numerical schemes are presented. In Section 5.3, we present the discrete operators of the finite volume methods and the stability analysis techniques. The numerical stability tests are performed in Section 5.4 for SWEs using pseudospectra. In Section 5.5, numerical tests are performed using β-plane waves in order to confirm the results of the analysis. Some

concluding remarks complete the study.

5.2 Discretization of Shallow Water Equations

5.2.1 Shallow water equations

The inviscid linear SWEs are written in Cartesian coordinates as (Vreugdenhil, 1994):

$$\begin{aligned} &\mathbf{u}_t + f\mathbf{k} \times \mathbf{u} + g\nabla\eta = 0 \\ &\eta_t + H\nabla \cdot \mathbf{u} = 0, \end{aligned} \tag{5.1}$$

where H is the mean depth, η represents the water surface elevation with respect to the reference plane, $\mathbf{u} = (u, v)^T$ is the vector composed of the depth-averaged velocity components in the x- and y-directions respectively, f is the Coriolis parameter, g is the gravitational acceleration, $\mathbf{k}$ is the unit vector in the vertical direction, and $(H+\eta)$ is the total water depth. We consider the β-plane approximation to the Coriolis parameter ($f = \beta y$), where β is the linear coefficient of variation of f with respect to y. The variable y is the meridional distance from the equator (positive northward), where $\beta = 2\tilde{\omega}/\mathbf{R} = 2.29 \times 10^{-11}\, m^{-1}s^{-1}$ and $\tilde{\omega}$ and $\mathbf{R}$ are the angular speed of the Earth's rotation and the mean radius of the Earth respectively ($\tilde{\omega} = 7.29 \times 10^{-5}\, rad\, s^{-1}, \mathbf{R} = 6371\, km$).

5.2.2 Finite volume schemes

a- Unstructured grid implementation

The analysis will be applied to some finite volume methods based on five grid configurations with respect to the location of the primitive variables and the control volume for each variable. The control volumes considered in this paper are shown in Figure 5.1. The numerical tests are performed for the five grids using the Crank-Nicolson scheme. For each grid of index **i**, the finite volume method is denoted by **i-CN**. The primitive variables are located at the geometric centers of triangles, at the vertices, or at the midpoints of the edges of triangles, as shown in Figure 5.1. The continuity equation and the momentum equations are integrated using different control volumes for the finite volume methods based on the grids 1, 2, and 5. We denote by Ω_η and Ω_u the control volumes used for the continuity equation and the momentum equations respectively, and we consider the same control volume ($\Omega_\eta = \Omega_u$) for the finite volume methods using Grids 3 and 4.

b- Spatial discretization

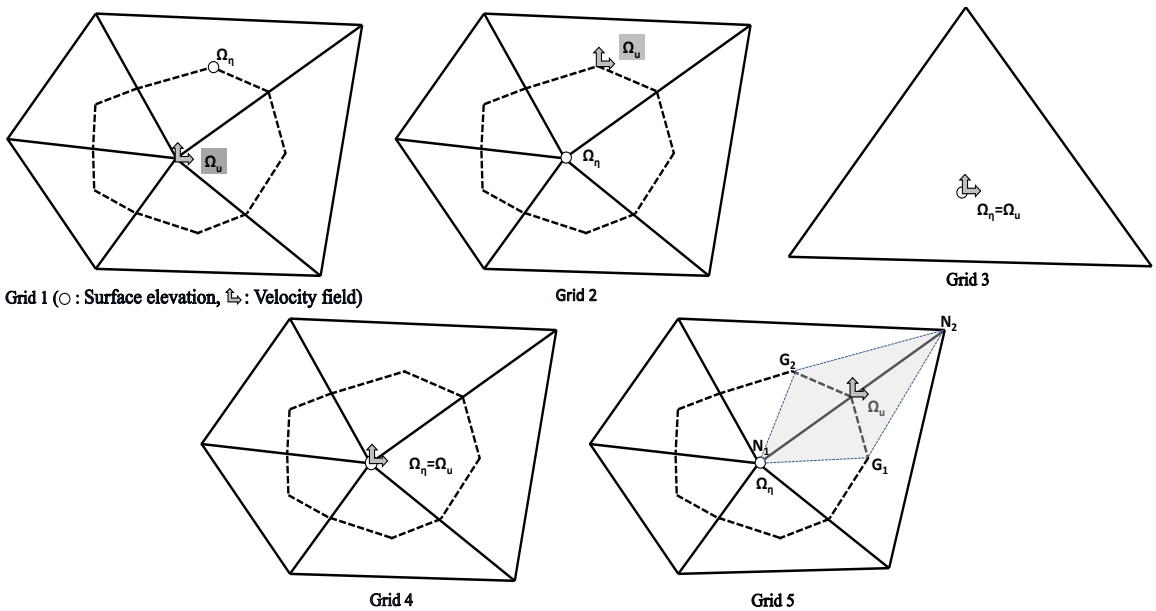

Figure 5.1 Sample of the unstructured grids used in the analysis. Grid i for method i-CN, where $i = 1, 2, 3, 4, 5$.

The SWEs are integrated over all control volumes Ω_η and Ω_u as follows:

$$\begin{aligned} &\int_{\Omega_\eta} (\eta_t + H\nabla \cdot \mathbf{u}) d\Omega_\eta = 0 \\ &\int_{\Omega_\mathbf{u}} (\mathbf{u}_t + \nabla \cdot \mathbf{F} - \mathbf{S})\, d\Omega_\mathbf{u} = 0, \end{aligned} \tag{5.2}$$

where $\mathbf{F} = (\mathbf{E}, \mathbf{G})^T$ is the flux vector with $\mathbf{E} = (g\eta, 0)^T$ and $\mathbf{G} = (0, g\eta)^T$, and $\mathbf{S} = (fv, -fu)^T$ is the source term. The Gauss divergence theorem is used to convert the surface integrals to the boundary integrals:

$$\begin{aligned} &\int_{\Omega_\eta} H\nabla \cdot \mathbf{u} d\Omega_\eta = \int_{\Gamma_\eta} H\mathbf{u} \cdot \mathbf{n}\, d\Gamma_\eta \\ &\int_{\Omega_\mathbf{u}} \nabla \cdot \mathbf{F} d\Omega_\mathbf{u} = \int_{\Gamma_\mathbf{u}} \mathbf{F} \cdot \mathbf{n}\, d\Gamma_\mathbf{u}, \end{aligned} \tag{5.3}$$

where $\Gamma_\mathbf{u}$ and Γ_η are the boundaries of the control volumes, and $\mathbf{n}$ is the unit outward normal vector to the boundary of each control volume considered. Then, Equations (5.2) and (5.3) lead to:

$$\begin{aligned} &\frac{d}{dt}\int_{\Omega_\mathbf{u}} \mathbf{u} d\Omega_\mathbf{u} = -\int_{\Gamma_\mathbf{u}} \mathbf{F} \cdot \mathbf{n}\, d\Gamma_\mathbf{u} + \int_{\Omega_\mathbf{u}} \mathbf{S} d\Omega_\mathbf{u} \\ &\frac{d}{dt}\int_{\Omega_\eta} \eta d\Omega_\eta = -\int_{\Gamma_\eta} H\mathbf{u} \cdot \mathbf{n}\, d\Gamma_\eta. \end{aligned} \tag{5.4}$$

For all cases of the control volumes considered, the boundary integral $\int_{\Gamma_{\mathbf{u}}} \mathbf{F} \cdot \mathbf{n}\, d\Gamma_{\mathbf{u}}$ in (5.4) may be approximated by a summation over the cell edges using

$$\int_{\Gamma_{\mathbf{u}}} \mathbf{F} \cdot \mathbf{n}\, d\Gamma_{\mathbf{u}} = \sum_{k=1}^{N} \int_{\Gamma_k} \mathbf{F} \cdot \mathbf{n}\, d\Gamma_k = \sum_{k=1}^{N} \left(\mathbf{F}_k \cdot \mathbf{n}_k\right)\ l_k, \tag{5.5}$$

where Γ_k, $\mathbf{F}_k$, $\mathbf{n}_k$, and l_k, $k = 1, 2, ..N$, are respectively the control volume edges, the outward fluxes, the unit outward normal vectors, and the lengths corresponding to the edges of the computational cells. The same form of Equation (5.5) is obtained for the case of the surface elevation by replacing the flux $\mathbf{F}$ by the parameter $H\mathbf{u}$ and considering the geometrical characteristics of the computational cell used for the continuity equation.

In the finite volume method based on Grid 1, the momentum equations are integrated over the Voronoi cells and the continuity equation is integrated over the triangles. The required interface values on the edges of the control volumes are calculated as averages of the values of the primitive variables at the two ends of the edge. A similar method is used for Grid 2 by considering the Voronoi cells as control volumes for the surface elevation and the triangles as control volumes for the velocity. The finite volume methods based on Grids 3 and 4 use the centered scheme to obtain the values at the interfaces of the triangles and Voronoi cells respectively. For Grid 5, the continuity equation is integrated over the boundary of the Voronoi cell, where the values of the velocity components at the middle of the triangle edges are known. For the momentum equations, the control volume is the quadrilateral $N_1G_1N_2G_2$ in which the surface elevations are known at the vertices N_1 and N_2. To obtain the values of the surface elevation at the vertices G_1 and G_2, a two-dimensional linear reconstruction on each side of the edge N_1N_2 is applied. The divergence theorem is applied over the two triangles with the centers G_1 and G_2 to obtain the gradients of the surface elevation used in the reconstructions.

Equations (5.4) and (5.5) are used to obtain the following discrete form for each computational cell:

$$\begin{aligned}
\frac{\partial \eta_j}{\partial t} &= \sum_{i \in I_j} \delta_i u_i + \sum_{i \in I_j} \gamma_i v_i \\
\frac{\partial u_j}{\partial t} &= \sum_{i \in J_j} \mu_i \eta_i + \beta y_j v_j \\
\frac{\partial v_j}{\partial t} &= \sum_{i \in K_j} \nu_i \eta_i - \beta y_j u_j,
\end{aligned} \tag{5.6}$$

where $(\eta_j, u_j, v_j)^T$ is the approximation of the cell averages of the solution, and I_j, J_j and K_j are the set of indices of the points used in the explicit formulation of the right-hand sides of Equations (5.4). Note that the continuity equation and the momentum

equations have different indices j for the schemes based on Grids 1, 2, and 5. The parameters δ_i, γ_i, μ_i, and ν_i depend on the finite volume method and the geometry of the computational cells. In our study we use the primary triangular grids with a skewness parameter, defined in Chapter 2, of value less than 0.50.

5.3 Discrete Operators of the Schemes and Analysis of Stability

5.3.1 Temporal discretization

We use the notations Δt and $t^n := n\Delta t$ respectively for the time step and the time at step n. The approximation at time t^n of the cell averages of the water surface elevation and the components of the velocity are denoted respectively by η_j^n, u_j^n and v_j^n. The Crank-Nicolson method is applied to discretize Equations (5.6)

$$\begin{aligned}
\frac{\eta_j^{n+1}-\eta_j^n}{\Delta t} &= \alpha \sum_{i\in I_j} \delta_i u_i^n + (1-\alpha)\sum_{i\in I_j} \delta_i u_i^{n+1} + \alpha \sum_{i\in I_j} \gamma_i v_i^n + (1-\alpha)\sum_{i\in I_j} \gamma_i v_i^{n+1} \\
\frac{u_j^{n+1}-u_j^n}{\Delta t} &= \alpha \sum_{i\in J_j} \mu_i \eta_i^n + (1-\alpha)\sum_{i\in J_j} \mu_i \eta_i^{n+1} + \alpha\beta y_j v_j^n + (1-\alpha)\beta y_j v_j^{n+1} \\
\frac{v_j^{n+1}-v_j^n}{\Delta t} &= \alpha \sum_{i\in K_j} \nu_i \eta_i^n + (1-\alpha)\sum_{i\in K_j} \nu_i \eta_i^{n+1} - \alpha\beta y_j u_j^n - (1-\alpha)\beta y_j u_j^{n+1}.
\end{aligned} \tag{5.7}$$

The global system can be written in the following form of dimension $M = p + 2q$, using the variable $\mathbf{U}^n := (\bar{\eta}^n, \bar{u}^n, \bar{v}^n)^T$ with $\bar{\eta}^n := (\eta_1^n, \eta_2^n, ...\eta_p^n)$, $\bar{u}^n := (u_1^n, u_2^n, ...u_q^n)$ and $\bar{v}^n := (v_1^n, v_2^n, ...v_q^n)$

$$\mathbf{A}\mathbf{U}^{n+1} = \mathbf{B}\mathbf{U}^n, \tag{5.8}$$

where $\mathbf{A}$ and $\mathbf{B}$ are the matrices of dimension $M \times M$ deduced from Equations (5.6) and the boundary conditions, and p and q represent respectively the numbers of control volumes used to compute the water surface elevation and the velocity. We obtain a generalized eigenvalue problem with the parameter-dependent matrix $\mathbf{B} - \lambda\mathbf{A}$, where λ denotes the variable of the eigenvalues. We obtain the equation $\mathbf{U}^{n+1} = \mathbf{C}\mathbf{U}^n$, by taking $\mathbf{C} = \mathbf{A}^{-1}\mathbf{B}$ which has the same pseudospectra as that of $\mathbf{B} - \lambda\mathbf{A}$ (Trefethen and Emberee, 2005).

5.3.2 Stability analysis techniques

a- Stability

The spectrum and the spectral radius for any arbitrary matrix $\mathbf{A}$ are denoted by $\sigma(\mathbf{A})$ and $\rho(\mathbf{A})$, respectively. In general we will use the L^2-norm, which is the most important norm for the pseudospectra, denoted by $\|\mathbf{A}\| = max_{\|x\|=1} \|\mathbf{A}x\|$. If other

norms are used, they will be specified in the notation. Following the previous section, if we consider a uniform time step Δt, the operator $\mathbf{C}$ will be independent of time and will depend only on the structure and size of the different grids and on the time step Δt. A theory was developed by Lax and Richtmyer (1956) for stability, which led to the equivalence theorem. They proved that for a finite-difference approximation with an operator $\mathbf{C}$ satisfying consistency, stability is a necessary and sufficient condition to ensure the convergence of the method. Additionally, they gave a definition of stability: the approximation based on the operator $\mathbf{C}$ is Lax-Richtmyer stable if for any fixed value of time T, there is a constant $\tilde{\mathbf{C}} \geq 1$ such that

$$\|\mathbf{V}^n\| \leq \tilde{\mathbf{C}}\|\mathbf{V}^0\|, \tag{5.9}$$

for all $n \geq 0$, where $n\Delta t \in [0, T]$, the constant $\tilde{\mathbf{C}}$ is independent of step n and $\mathbf{V}^0$ is the initial condition. This definition is extended by considering that the bound of the set $\{\tilde{\mathbf{C}}^n\}$ which is a function of Δt in some interval, $\Delta t \in [0, \tau]$, is generally continuous. Then, the new stability condition can be stated as: the approximation based on the operator $\mathbf{C}$ is Lax-Richtmyer stable if the set $\{\tilde{\mathbf{C}}^n\}$ is uniformly bounded for $0 < \Delta t \leq \tau$ and $0 \leq n\Delta t \leq T$. Other definitions of stability are introduced in the literature, such as in Carpenter et al. (1994), Beam and Warming (1993) and Gustafsson et al. (1972). Another concept is asymptotic stability, where we consider the behavior of the solution for large times, which requires that the power of the operator $\mathbf{C}$ (i.e. $\mathbf{C}^n$) is bounded for infinite time. Unfortunately, the solution for finite times may have excessive amplifications, especially in the case of non-normal matrices, and therefore the spectral radius is not necessarily a good indicator of the behavior of the scheme for finite times.

b- Pseudospectra

There are several definitions for the pseudospectra of a matrix. In this paper we employ the definition based on the singular values and the L^2-norm. The concept of singular value decomposition (SVD) is defined in general for complex matrices. In our case we recall the necessary parameters for our study, where we consider real matrices. Any matrix $\mathbf{Q} \in \mathbb{R}^M$ can be written in the form:

$$\mathbf{Q} = \mathbf{W}_1\mathbf{S}\mathbf{W}_2^T, \tag{5.10}$$

where the matrix $\mathbf{S}$ is given as:

$$\mathbf{S} = \begin{pmatrix} \mathbf{D} & \mathbf{Z_1} \\ \mathbf{Z_2} & \mathbf{Z_3} \end{pmatrix}, \tag{5.11}$$

with $\mathbf{W}_1^T\mathbf{W}_1 = I$, $\mathbf{W}_2^T\mathbf{W}_2 = I$, I is the identity matrix, and $\mathbf{Z}_i$, $i = 1, 2, 3$ are zero matrices. The matrix $\mathbf{Q}\mathbf{Q}^T$ is real and symmetric, so it is diagonalizable and all its eigenvalues are positives. The strictly positive eigenvalues are denoted by s_i^2. The

diagonal matrix used in the SVD method is $\mathbf{D} = \text{Diag}(s_1, s_2, ...s_p)$, and the parameters s_i are the singular values arranged in descending order $s_1 \geq s_2 \geq ... \geq s_p > 0$. The smallest singular value of matrix $\mathbf{Q}$ is denoted by $s_{min}(\mathbf{Q})$. The ε-pseudospectrum of the operator $\mathbf{C}$ is the set $\sigma_\varepsilon(\mathbf{C})$, defined as:

$$\sigma_\varepsilon(\mathbf{C}) = \{z \in \mathbb{C} : s_{min}(zI - \mathbf{C}) < \varepsilon\}. \tag{5.12}$$

If the matrix $\mathbf{C}$ is normal, the ε-pseudospectrum $\sigma_\varepsilon(\mathbf{C})$ is the union of the open disks of radius ε which have the points of the spectrum as their centers. The ε-pseudospectrum is given in this case by

$$\sigma_\varepsilon(\mathbf{C}) = \sigma(\mathbf{C}) + \Delta_\varepsilon, \tag{5.13}$$

where Δ_ε is the open disk in $\mathbb{C}$ of radius ε and center $z = 0$, and each complex number in $\sigma(\mathbf{C}) + \Delta_\varepsilon$ is the sum of two complex numbers from $\sigma(\mathbf{C})$ and Δ_ε. In this case all eigenvalues have a condition number equal to 1, and for any uniform perturbation one will observe uniform evolution without any bulge in the ε-pseudospectrum. For a non-normal matrix, the ε-pseudospectrum can be large.

5.4 Numerical Tests of Stability

5.4.1 Dimensionless form of SWEs

The discrete operators $\mathbf{C}$ of the schemes are obtained by using the dimensionless form of SWEs. Equations (5.1) are converted into a dimensionless form on the equatorial β-plane using the variables $\tilde{x} = x/\bar{L}$, $\tilde{y} = y/\bar{L}$, $\tilde{\eta} = \eta/\bar{H}$, $\tilde{u} = u/\bar{U}$ and $\tilde{v} = v/\bar{U}$. The reference values of the depth, time, length and velocity scales are $\bar{H} = H$, $\bar{T} = \beta^{-1/2}(gH)^{-1/4}$, $\bar{L} = (\beta\bar{T})^{-1}$, and $\bar{U} = \bar{L}/\bar{T}$, respectively. Using the above reference scales, the dimensionless system is obtained by setting $H = g = 1$ in the original system. A non-dimensional domain $[0, L] \times [0, L]$ is used with $L = 3$. In our analysis we use the ratios $\Delta t/d_m$, where d_m is the smallest distance between the locations of the primitive variables. This ratio corresponds to the Courant-Friedrichs-Lewy number for the dimensionless system. In our numerical tests, we consider an operator $\mathbf{C}$ of dimension 1200×1200 by using 400 triangles for the domain.

5.4.2 Test cases using the Crank−Nicolson method

First (test a) for the finite volume method 1-CN, we consider periodic boundary conditions in both the x- and y-directions and the time step $\Delta t = 0.5d_m$. The spectrum of this scheme is inside the unit circle, which confirms that the scheme is asymptotically stable. Following the numerical tests, we conclude that the power of the discrete operator of the scheme is bounded. This method is an example of schemes in which

the spectrum is inside the unit circle and the power of the operator is bounded but its behavior, as shown in Figure 5.2 (left), influences the stability of the solution. The pseudospectra of the scheme are shown in Figure 5.2 (right) in which a bulge is observed near the largest eigenvalues. Instability clearly appears in the case of the wall boundary conditions in y-direction while keeping the periodicity along the x- axis (test b). Figure 5.3 (left) shows the pseudospectra for this case using a time step $\Delta t = 0.3d_m$, where a large bulge is observed.

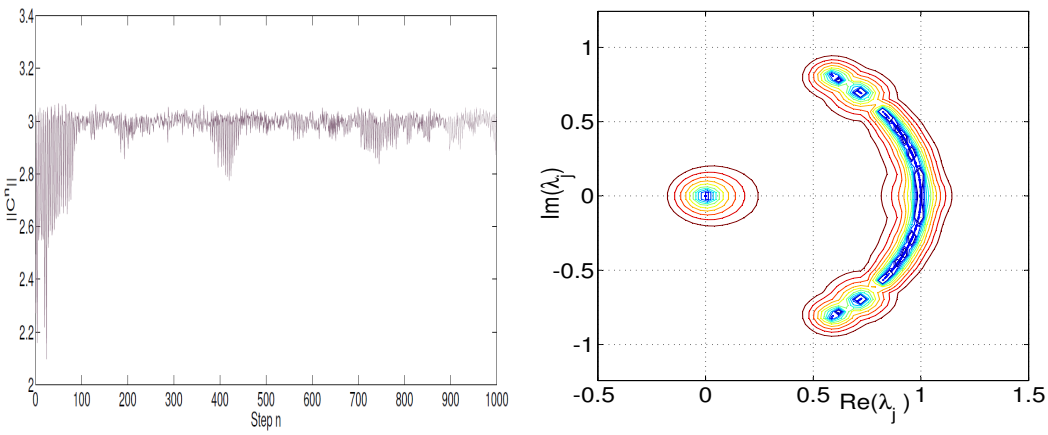

Figure 5.2 Left: Evolution of the parameter $\|\mathbf{C}^n\|$ using method 1-CN (test a) with periodic boundary conditions in the x- and y-directions. Right: Eigenvalue spectrum and pseudospectra of the same method with the same boundary conditions

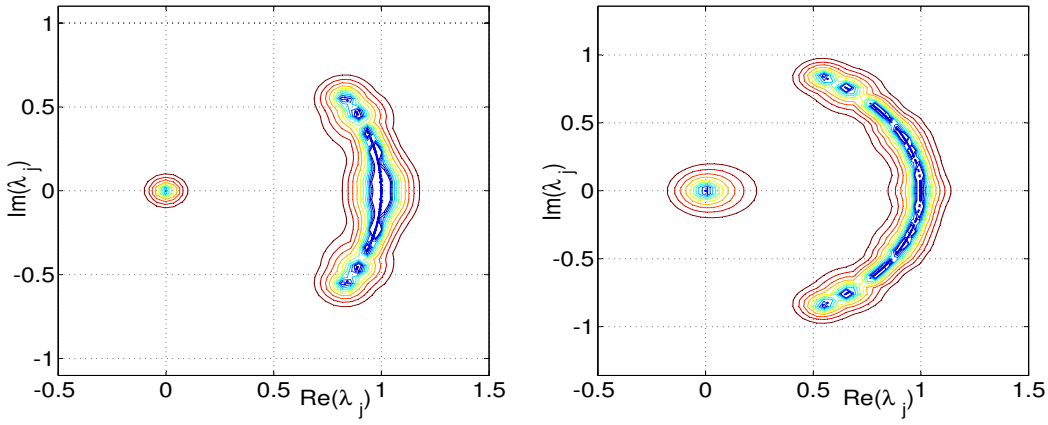

Figure 5.3 Left : Eigenvalue spectrum and pseudospectra of method 1-CN (test b) using periodic boundary condition in the x-direction and wall boundary condition in the y-direction. Right : Eigenvalue spectrum and pseudospectra of method 2-CN using periodic boundary conditions in the x- and y-directions.

For the method 2-CN, we consider the periodic boundary conditions in both x- and y-directions, and we use the time step $\Delta t = 0.5d_m$. The spectrum is inside the unit circle, and this method is asymptotically stable. The pseudospectra of the scheme are shown in Figure 5.3 (right), where we observe a small bulge near the largest eigenvalues compared to the pseudospectra of the scheme 1-CN. The use of a small time step can reduce these signs of instability for scheme 2-CN more than for scheme 1-CN.

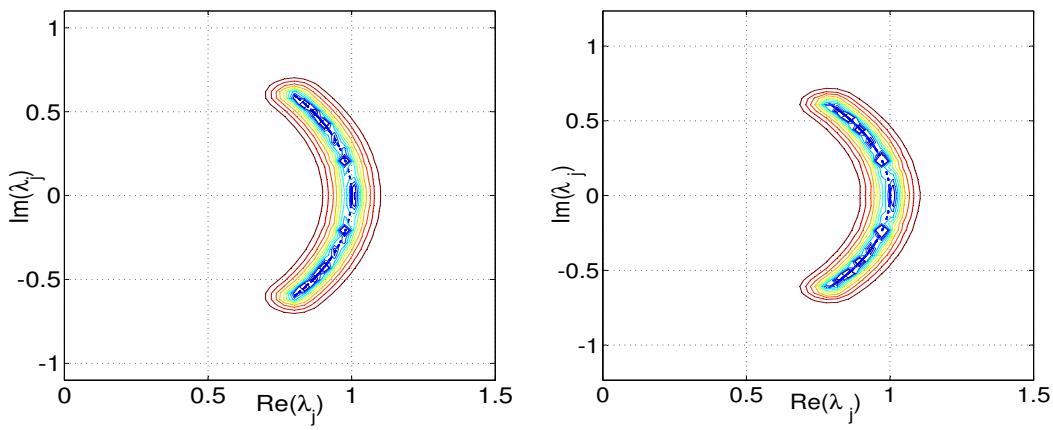

Figure 5.4 Left : Eigenvalue spectrum and pseudospectra of method 3-CN (test a) using uniform unstructured grid and periodic boundary conditions. Right : Eigenvalue spectrum and pseudospectra of method 3-CN (test b) using general unstructured grid and periodic boundary conditions.

The finite volume method 3-CN and the method 4-CN, presented below, are the best schemes in terms of stability among those considered in this paper. When an unstructured grid is considered with a uniform geometry of triangles and under periodic boundary conditions we obtain an operator $\mathbf{C}$ with negligible distance from normality. For this case, the discrete operator of the scheme is Lax-Richtmyer stable ($\|\mathbf{C}^n\| \leq 1$). As shown in Figure 5.4 (left), the entire spectrum of this method lies inside the unit circle, and following its pseudospectra shown in the same figure, no bulge is observed near the largest eigenvalues. For the second test, we use a general geometry of triangles and numerically study the behavior of $\|\mathbf{C}^n\|$ for several sizes of $\mathbf{C}$, and according to the numerical tests, this parameter is uniformly bounded. Therefore, the approximation based on this operator of the finite volume scheme is Lax-Richtmyer stable. The pseudospectra of the scheme show good results in Figure 5.4 (right) since we observe uniform evolution near the largest eigenvalue outside the unit circle.

For the finite volume method 4-CN, the behavior of the parameter $\|\mathbf{C}^n\|$ is studied for various cases by using discrete operators of dimensions 768×768, 1200×1200, and 2352×2352. This parameter is uniformly bounded and satisfies the condition $\|\mathbf{C}^n\| \leq \|\mathbf{C}\|$, ($\|\mathbf{C}\| = 2.0009$). Therefore, the finite volume method 4-CN is Lax-

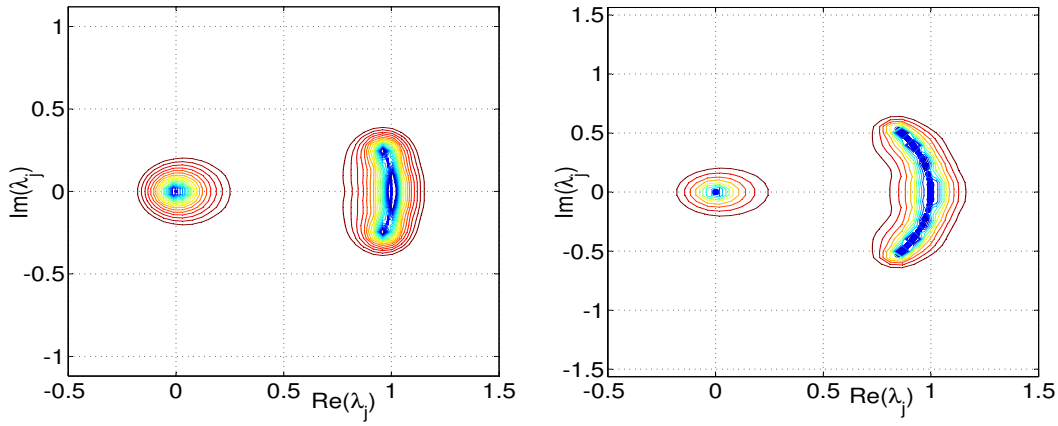

Figure 5.5 Left: Eigenvalue spectrum and pseudospectra of method 4-CN using unstructured grid and periodic boundary conditions. Right : Eigenvalue spectrum and pseudospectra of method 5-CN (test a) using uniform unstructured grid and periodic boundary conditions.

Richtmyer stable. The entire spectrum of this method lies inside the unit circle, as shown in Figure 5.5 (left), and following its pseudospectra shown in the same figure, no bulge is observed near the largest eigenvalues.

The pseudospectra of the scheme 5-CN are shown in Figure 5.5 (right), where a slight bulge is observed. The discrete operator of this scheme is very far from normality, and some amplifications of the solutions can be observed in finite times.

5.5 Numerical tests for Kelvin waves

In this section, numerical tests are performed for linear β-plane SWEs using Kelvin waves which are in perfect geostrophic balance in the meridional direction since the meridional flow, v, is zero, and they are propagated in the eastward direction. These waves are exact solutions of Equations (5.1).

$$\begin{aligned} u(x,y,t) &= -\cos\left(2\pi\frac{x-t}{X}\right)e^{-(y-L/2)^2/2}, \\ v(x,y,t) &= 0, \\ \eta(x,y,t) &= u(x,y,t). \end{aligned} \tag{5.14}$$

We consider a domain $[0,L]\times[0,L]$ with $L=6$ (nondimensional) and $X=L/4$. These solutions are in a steady state in a moving frame. When they are considered as an initial condition, they must be preserved, and in particular their total energy over one spatial period must remain constant. The evolution of the total energy over time will be analyzed, since it gives an idea of the instability of the numerical schemes. Figure

5.6 shows the evolution of the Kelvin total energy of the finite volume methods 1-CN, 2-CN, 3-CN, 4-CN, and 5-CN for two time periods. The methods 3-CN and 4-CN are stable, and they perform very well in the conservation of energy. The methods 1-CN and 5-CN have a bounded total energy but it increases in finite time, which is in agreement with our analysis using the pseudospectra, where it was demonstrated that there are some particular modes which present amplification for finite times. The numerical method 2-CN has some signs of instability when the pseudospectra are used, which is not visible in the behavior of the Kelvin total energy. For this scheme, some signs of oscillation will be observed in the contours of the solutions, as explained below.

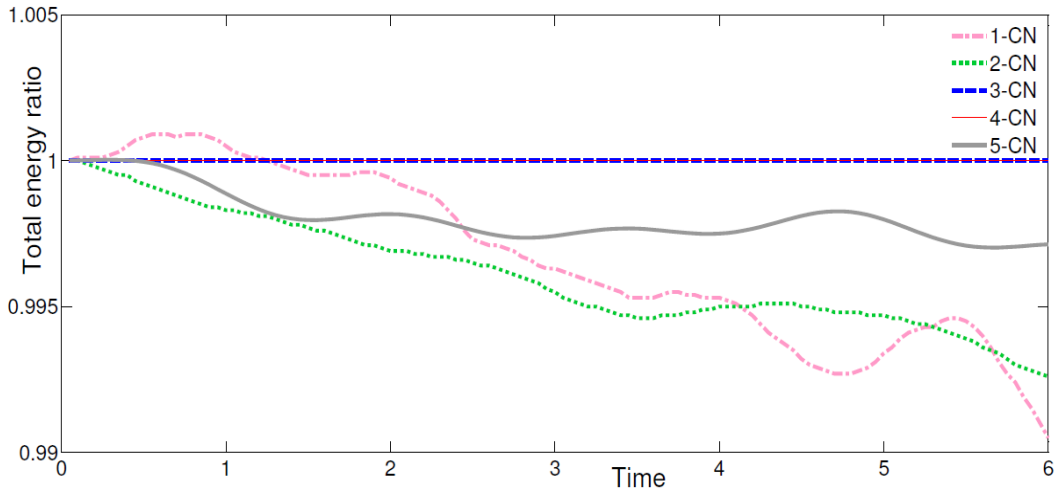

Figure 5.6 Change in total energy for Kelvin waves using the schemes i-CN, i=1, 2, 3, 4, and 5 for two time periods with $CFL = 0.5$

Figures 5.7 and 5.8 show the isolines of the surface elevation of Kelvin waves for one time period. The numerical oscillations are significant for the scheme 1-CN. For the schemes 2-CN and 5-CN, we observe very small oscillations. For these schemes, the oscillations are barely visible on the graphics. The pseudospectra were able to detect the instability of the two schemes, which confirms the efficiency of those methods for stability analysis. Finally, Figure 5.7 (left) shows the isolines of the surface elevation of Kelvin waves for the scheme 3-CN. The three-dimensional view of the solutions obtained by using the schemes 3-CN and 4-CN are shown in Figure 5.9. For the two schemes, the obtained solutions are oscillation-free, which confirms the results of our analysis using pseudospectra.

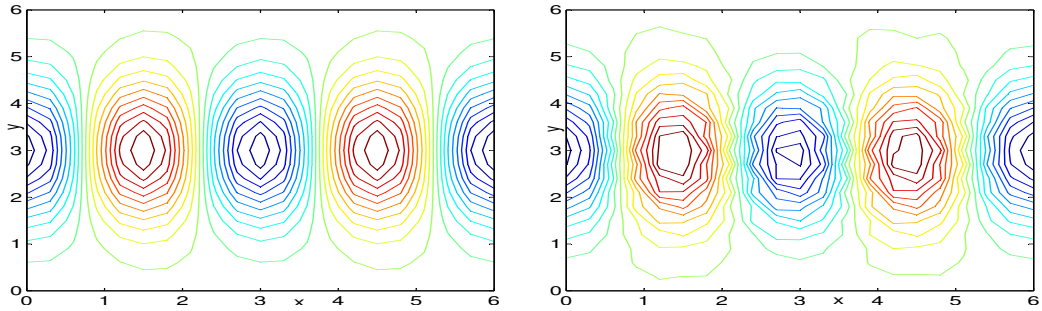

Figure 5.7 The isolines of the schemes 3-CN (left) and 1-CN (right) at one time period

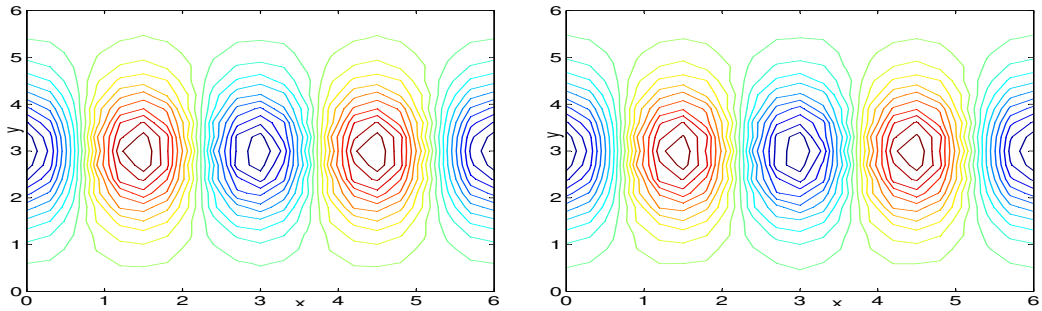

Figure 5.8 The isolines of the schemes 2-CN (left) and 5-CN (right) at one time period

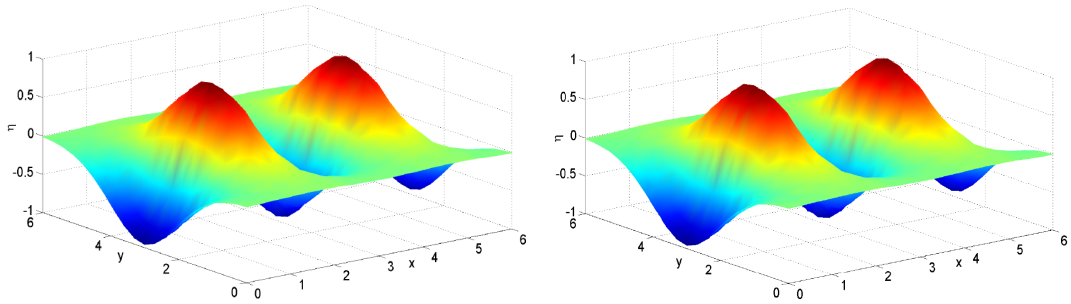

Figure 5.9 Three dimensional view of the water surface elevation using the schemes 3-CN (left) and 4-CN (right) at one time period

5.6 Conclusions

The discrete form of finite volume methods on unstructured grids for shallow water equations can lead to spurious modes which cause stability problems. The appearance of these oscillations is mainly due to the structure of the mesh, the placement of the primitive variables on the grid, and/or the employed finite volume method. Unstructured grids have large impacts on the structure of the discrete operators of finite volume

methods, which leads to non-normal matrices. For these matrices, the asymptotic stability and the Lax-Richtmyer stability are not sufficient to obtain a perfect stability for all modes. In this paper, pseudospectra were employed for detecting the instability of finite volume methods for shallow water equations over unstructured grids. The analysis has shown that the pseudospectra are very effective when studying the stability of finite volume methods. For the Crank-Nicolson method, it is shown that it is important to consider the placement of the primitive variables together at the center of the computational cell. Grids 3 and 4, which correspond to the cell-centered and node-centered methods, lead to the best results for stability. The finite volume methods using Grids 3 and 4 and the Crank-Nicolson method as a temporal scheme are the most interesting methods. The discrete operators of these schemes are almost normal under periodic boundary conditions when uniform grids are used. These schemes are also effective for the case of wall boundary conditions. The performance of schemes 3-CN and 4-CN is confirmed when the Kelvin waves are taken as the initial condition, and the solutions are stable and more accurate compared to the other schemes.

CHAPITRE 6 Une méthode des volumes finis respectant la condition de compatibilité géométrique pour les lois de conservation sur des surfaces courbes

A geometry-preserving finite volume method for conservation laws on curved geometries[1]

Résumé

Dans ce chapitre, une méthode des volumes finis qui satisfait la condition de compatibilité géométrique est développée pour les lois de conservation sur la sphère. Les travaux de Ben Artzi et LeFloch (2007) et Ben-Artzi, Falcovitz, et LeFloch (2009) sont pris comme base pour développer cette méthode en considérant la résolution du problème de Riemann généralisé. L'opérateur de divergence de la loi de conservation est discrétisé sous une forme qui satisfait la condition de compatibilité géométrique. L'approche du «splitting» directionnel en latitude et en longitude de la sphère ainsi que la méthode Runge-Kutta d'ordre trois (TVDRK3) sont utilisées pour l'intégration temporelle. L'approche du «splitting» directionnel utilisée pour simplifier la résolution du problème de Riemann n'a pas d'influence sur la propriété de compatibilité géométrique de l'opérateur de divergence sous sa forme discrète.

Dans le nouveau schéma numérique, une reconstruction linéaire est proposée en se basant sur les valeurs de la solution aux centres des cellules de calcul et sur les valeurs des solutions du problème de Riemann aux interfaces des cellules. Ces dernières sont obtenues en utilisant les approximations de second ordre basées sur la résolution du problème de Riemann généralisé. Dans la formulation du schéma numérique proposé, les dimensions géométriques sont considérées de manière analytique et la forme discrète du schéma respecte exactement la propriété de compatibilité géométrique. On obtient un lien très fort entre la forme semi-discrète du schéma et l'équation du système.

La méthode des volumes finis proposée est stable et elle est précise pour le cas des solutions discontinues de grands chocs et amplitudes en comparaison avec des schémas numériques très connus. Cette méthode est d'ordre deux dans l'espace et la méthode TVDRK3 combinée avec l'approche du «splitting» directionnel constitue une méthode efficace pour l'intégration temporelle. L'ordre trois de la méthode TVDRK3 est moins influencé par la méthode du «splitting» directionnel utilisée.

Le comportement asymptotique des solutions, que les méthodes analytiques disponibles ne permettent pas en général de connaître, est étudié numériquement dans le présent

[1]Cet article est réalisé en collaboration avec P.G. LeFloch et A. Mohammadian, et soumis pour publication sous la forme: A. Beljadid, P.G. LeFloch, A. Mohammadian, 2014, A geometry-preserving finite volume method for conservation laws on curved geometries. Advances in Computational Mathematics.

chapitre. Quelques propriétés mathématiques relatives à l'évolution des solutions du système hyperbolique sont analysées. Le comportement asymptotique et les propriétés des solutions discontinues sont analysés en fonction de la nature du flux. Une nouvelle classification des flux est proposée en introduisant les notions de flux feuilletés et de flux génériques. Les résultats des analyses ont montré que la nouvelle classification introduite et le caractère de linéarité du flux sont suffisants pour prédire le comportement asymptotique des solutions.

Pour le cas du flux feuilleté de type linéaire, les solutions sont simplement transportées le long des lignes de niveau. Le flux générique génère de fortes variations dans les solutions qui convergent vers des valeurs constantes dans des domaines indépendants sur la sphère. Pour le flux feuilleté qui a un comportement non linéaire, la solution converge vers des constantes, le long des lignes de niveau, qui représentent sa moyenne le long de chaque ligne. Les flux feuilletés non linéaires sont utilisés pour la construction des solutions stationnaires non triviales du système hyperbolique étudié. Pour ces flux, les solutions qui sont constantes le long des lignes de niveau sont des solutions stationnaires non triviales du système. Ces solutions sont utilisées dans les tests numériques pour valider les performances du schéma numérique proposé. Les résultats ont montré la capacité et le potentiel de la méthode proposée dans le cas des solutions discontinues avec de grands chocs et amplitudes pour les lois de conservation sur la sphère. Ce schéma numérique pourrait être étendu au cas des équations de Saint-Venant sur la sphère.

6.1 Introduction

This paper is devoted to nonlinear hyperbolic problems involving conservation laws or, more generally, balance laws and which are posed on curved geometries such as a surface. Our objective is to design robust and efficient numerical approximation methods which allow to compute discontinuous solutions and preserve the fundamental structure of the partial differential equations, especially geometry-related properties. Hence, our goal is to design and numerically investigate geometry-preserving, high-order accurate, finite volume methods. We advocate the use of a (geometric) formulation of the finite volume method based on the intrinsic (or covariant) form of the equations, rather then the coordinate expression which is more commonly used. In this manner, by properly taking into account the effects induced by the geometry, we can design methods that are, both, accurate and robust.

We do not a priori restrict ourself to a specific discretization technique, but, rather, we aim at comparing various strategies such as generalized Riemann solvers, second-order centered schemes, etc. Compressible fluid dynamics provide a large variety of problems which involve geometrical features. The prototype example is the system of shallow water on the sphere with topography, which describe fluid flows on the surface of the

Earth for instance, in connections with weather predictions Haltiner (1971).

Motivated by numerous applications in fluid dynamics, the study of hyperbolic conservation laws posed on curved manifolds were recently initiated in the mathematical and numerical literature. We build here on the work by Ben-Artzi and LeFloch (2007) who proposed to rely on an analogue of the inviscid Burgers equation for curved geometry and, more generally, various classes of hyperbolic conservation laws on manifolds. Since Burgers equations has played such an important role in the development of shock-capturing schemes for compressible fluid problems, it is also expected that the class of "geometric Burgers models" should provide an ideal simplified setup in order to design and test geometric-preserving shock-capturing scheme. The mathematical properties of entropy solutions to conservation laws on manifolds (including on spacetimes, that is, with time-dependent (Lorentzian) metrics) were then extensively investigated by LeFloch and co-authors (e.g. Amorim et al., 2005, 2008; Ben-Artzi and LeFloch, 2007; Ben-Artzi et al., 2009; LeFloch, 2011; LeFloch and Okutmustur, 2008).

Subsequently, hyperbolic conservation laws on evolving surfaces were studied by Dziuk, Kroöner, and Müller and by Giesselmann (2009). More generally, computational methods for evolving surfaces were developped in Dziuk and Elliott (2007) and the references therein.

Scalar conservation laws will thus be our starting point in the present work and, next, will extend our methodology and conclusions to other hyperbolic equations, such as the shallow water system. In the present paper, we thus focus on **geometric Burgers models** of the form

$$\partial_t u + \mathrm{div} F(\cdot, u) = 0, \qquad u = u(t) : S^2 \to \mathbb{R}, \tag{6.1}$$

with unknown u defined on a curved space which we take to be the two-dimensional sphere S^2 and "div" is the divergence operator. The flux $F(\cdot, u)$ is a prescribed vector field defined on S^2, which depends on the unknown variable u as a parameter (See Section 6.2, below). We adopt the methodology proposed in Ben-Artzi et al. (2009) which relies on second–order approximations based on generalized Riemann problems. We propose a scheme which uses a new piecewise linear reconstruction based on the values of the solution at the center of the computational cells and the values of the Riemann solutions at the cell interfaces, using the second-order approximations based on a generalized Riemann solver. In the proposed scheme, we use a total variation diminishing Runge-Kutta method (TVDRK3) with operator splitting for the temporal integration.

This **geometric finite volume method** therein is further developed and numerically investigated. We observe that certain global quantities are conserved by entropy solutions to scalar conservation laws posed on curved geometries. Our aim is therefore to exhibit these global invariants and investigate to which extend they are preserved

by (or remain monotone decreasing for) the approximation solutions generated by the geometric method. We are also interested in investigating the large-time asymptotics of solutions, which is not understood by analytical method and will be here studied numerically. As we show it in this paper, by distinguishing between several classes of flux vector fields and initial conditions, we can exhibit a variety of nonlinear wave phenomena.

Our analysis below shows that the proposed method is consistent with the **maximum principle**

$$\|u(t')\|_{L^\infty(S^2)} \leq \|u(t)\|_{L^\infty(S^2)}, \qquad t' \geq t, \tag{6.2}$$

the **entropy stability property**

$$\|u(t')\|_{L^p(S^2)} \leq \|u(t)\|_{L^p(S^2)}, \qquad t' \geq t, \tag{6.3}$$

for all exponents $p \in [1, +\infty)$, as well as with the **time-variation diminishing property**

$$\|\partial_t u\|_{\mathcal{M}(S^2)}(t') \leq \|\partial_t u\|_{\mathcal{M}(S^2)}(t), \qquad t' \geq t, \tag{6.4}$$

where $\mathcal{M}(S^2)$ denotes the space of bounded measures defined on S^2. On the other hand, the **contraction property** (for any two entropy solutions u, v)

$$\|v(t) - u(t)\|_{L^1(S^2)} \leq \|v(0) - u(0)\|_{L^1(S^2)} \tag{6.5}$$

is violated by the scheme and is satisfied only by its first-order version.

Our analysis will distinguish between **foliated flux fields** and "generic" (or fully coupled) flux fields. Geometric conservation laws with foliated flux are a combination of linear transport and nonlinear hyperbolic equations, in the sense that the solutions are simply transported within the level sets and (exact) solutions can be defined in each level set, independently of the level parameter. Under some assumption concerning the transport speed along the level sets (as shown in Test 1-a, below), the solutions can be globally preserved within the entire sphere S^2 in a suitably defined "moving frame". For foliated flux fields satisfying a suitable nonlinearity condition, the solutions converge to their constant average in each level set (as shown in Test 5, below). This latter case includes, in particular, solutions which converge to constant values in independent domains on the sphere (as illustrated in Tests 2 and 3, below).

For fully coupled flux fields, the solutions converge to constant values in independent domains on the sphere. The number of constants depends on the existence of the curves that split the sphere in independent parts, as we define below. In the case of the generic flux, the scalar function of the gradient can be decomposed in several homogeneous terms. The term is said to be homogeneous if it corresponds to a foliated flux field. The behavior of the solution is greatly influenced by the homogeneous terms of high order. The asymptotic convergence to constant values is influenced on the nature of

the terms of the generic flux and the initial condition.

Finite volume methods based on the geometric structure of the problem and especially the method under study in this paper have definite advantages:

- A strong link between the numerical scheme and the governing equation. The geometric dimensions are considered in an analytical way which leads to a discrete form of the scheme that respects exactly the geometric compatibility property,
- The quality of the numerical solutions is largely improved by using the proposed piecewise linear reconstruction based on the values of the function u at the centers of cells and the values of the solutions of Riemann problem at the cell interfaces obtained using the resolution of the generalized Riemann problem with second-order accuracy,
- The use of a Total Variation Diminution (TVD) temporal integration technique (see below) for the temporal scheme improves the quality of the numerical solutions. This is due to two reasons: on one hand, we obviously guarantee higher accuracy in time and, on the other hand, it is very important that the temporal accuracy allow us to improve the overall quality of the numerical solution (even in space).
- The scheme is second-order accurate in space and the splitting approach combined with the TVDRK3 method was found to be accurate and efficient for the time integration. The third order of TVDRK3 is less influenced by the splitting approach.

The authors are currently working on an extension of this method to the shallow water equations posed on the sphere. An outline of this paper is as follows. In Section 6.2, we describe the geometric conservation laws and the properties of their entropy solutions. In Section 6.3, we propose a new classification of flux fields. Section 6.4 is devoted to classes of solutions of particular interest. In Section 6.5, we give the description of the geometric finite volume method. Section 6.6 presents the analysis of the spatial and temporal orders of accuracy of the proposed scheme. In Section 6.7, numerical tests are performed for nonlinear foliated fluxes, fully coupled flux vector fields, as well as further tests in order to study the asymptotic convergence of solutions. Finally, Section 6.8 contains concluding remarks.

6.2 Geometric Burgers models on the sphere

6.2.1 Geometric hyperbolic conservation laws on manifolds

We are primarily interested in nonlinear hyperbolic equations posed on the sphere, but since the mathematical theory supporting the study in the present paper has been

developed for general manifolds endowed with a volume form, as we now explain, so we introduce it at this level of generality first. Fix any compact n-manifold M endowed with a volume form ω with L^∞ regularity. Given a **flux vector field** $F = F(x,u) \in T_xM$ depending on the real parameter u, where x is an arbitrary point on M and T_xM is the tangent space to M at the point x, we consider the **geometric hyperbolic balance law**

$$\partial_t u + \mathrm{div}_\omega F(\cdot, u) = 0 \quad \text{ in } \mathbb{R}_+ \times M, \tag{6.6}$$

with unknown $u : \mathbb{R}_+ \times M \to \mathbb{R}$, where (with some abus of notation) $\mathrm{div}_\omega X = \frac{1}{\omega}\partial_j(\omega X^j)$ with $\omega = \omega dx^1 dx^2 \dots dx^n$ in local coordinates $x = (x^j)_{1\le j\le n}$ and $X = (X^j)$ is an arbitrary vector field, where we use the short-hand notation $\partial_j := \partial/\partial x^j$. We impose that the flux is **geometry-compatible,** in the sense that

$$\mathrm{div}_\omega F(\cdot, \overline{u}) = 0, \tag{6.7}$$

where $\overline{u} \in \mathbb{R}$ is an arbitrary constant which is equivalent to saying that constants are (trivial) solutions of the conservation law. Then, weak solutions are understood in the following sense: for every test-function $\theta = \theta(t,x)$,

$$\iint_{\mathbb{R}_+\times M} \Big(\partial_t\theta(t,x)\, u(t,x) + \partial_j\theta(t,x) F^j(x, u(t,x)\Big)\, \omega(x) dt dx = 0, \tag{6.8}$$

where F^j denote the components of the vector field F in an arbitrary coordinate chart $x = (x^j)_{1\le j\le n}$. Here, we have identified the volume form ω with its expression ωdx in local coordinates (and, for simplicity in order to state (6.8), we have assumed that the manifold is covered by a single chart).

To any equation (6.6) with flux field satisfying the condition (6.7), we can associate a unique **semi–group of entropy solutions** characterized as follows: given any $u_0 \in L^\infty(M)$, there exists a unique entropy solution $u \in L^\infty(\mathbb{R}_+ \times M)$ to the initial value problem

$$\begin{aligned} &\partial_t U(u) + \mathrm{div}_\omega G(\cdot, u) \le 0, \qquad U'' \ge 0, \\ &u(0) = u_0, \end{aligned} \tag{6.9}$$

in which for every convex function $U : \mathbb{R} \to \mathbb{R}$ we have introduced the corresponding entropy flux $G = G(x,u) \in T_xM$ such that $\partial_u G := U' \partial_u F$. The inequalities in (6.9) are refered to as the **entropy inequalities**.

Moreover, this semi-group of entropy solutions satisfies several fundamental properties:

- **The entropy stability property:** for all $p \in [1,\infty)$ and $t \ge 0$

$$\|u(t)\|_{L^p_\omega(M)} \le \|u(0)\|_{L^p_\omega(M)}, \tag{6.10}$$

which also implies the **maximum principle** (by letting $p \to +\infty$):

$$\|u(t)\|_{L^\infty(M)} \leq \|u(0)\|_{L^\infty(M)}. \tag{6.11}$$

- **The L^1 contraction property**: given any two entropy solutions u, v and for all times $t \geq 0$

$$\|v(t) - u(t)\|_{L^1_\omega(M)} \leq \|v(0) - u(0)\|_{L^1_\omega(M)}. \tag{6.12}$$

- **The time-variation diminishing property:** given any entropy solution u

$$\|\partial_t u\|_{\mathcal{M}}(t) \leq \|\partial_t u\|_{\mathcal{M}}(0), \qquad t \geq 0. \tag{6.13}$$

We thus have a natural generalization of Kruzkov's theory (1970) to a manifold (Amorim et al., 2005, 2008; LeFloch, 2011; LeFloch and Okutmustur, 2008). Geometry-independent bounds hold, which are very useful in designing and testing discrete approximation schemes.

The low regularity of the volume form allows us to also include *shock wave in the geometry* (which is relevant to model earthquakes in the context of the shallow water system, for instance).

6.2.2 The models of interest in this paper

In the applications, the manifold M is often defined via an embedding in the higher-dimensional Euclidian space $\mathbb{R}^N$. For simplicity, in the rest of this paper we concentrate on surfaces and, specifically, the two-dimensional sphere endowed with a volume form ω and embedded in $\mathbb{R}^3$. We denote by S^2 the unit sphere embedded in $\mathbb{R}^3$ and endowed with the canonical volume form induced by the Euclidian metric.

By denoting by ∇_ω the covariant derivative operator on the sphere $S^2 \subset \mathbb{R}^3$, we now express the conservation law in the form

$$\partial_t u + \nabla_\omega \cdot (F(\cdot, u)) = 0, \tag{6.14}$$

or equivalently, in local coordinates, we can pose the problem on the unit sphere with a weight function $\omega = \omega(x)$

$$\partial_t u(t,x) + \frac{1}{\omega(x)} \nabla \cdot \Big(\omega(x)\, F(x, u(t,x))\Big) = 0. \tag{6.15}$$

Flux vector tangent to the sphere can always be expressed in the form

$$F(x,u) = n(x) \wedge \Phi(x,u), \qquad x \in S^2,\ u \in \mathbb{R}, \tag{6.16}$$

where $\Phi = \Phi(x, u)$ is a u-dependent vector field defined in the ambient space $\mathbb{R}^3$, $n = n(x)$ denotes the unit normal vector to the sphere and the symbol $\wedge$ denotes the cross product. As explained earlier, we are primarily interested in geometry-compatible flux vectors satisfying, by definition,

$$\nabla \cdot (F(\cdot, \bar{u})) = 0, \tag{6.17}$$

where $\overline{u}$ is an arbitrary real constant.

Especially, the broad class of **gradient-type flux vector fields** is defined by

$$\Phi(x, \bar{u}) = \nabla h(x, \bar{u}), \qquad x \in S^2, \quad \bar{u} \in \mathbb{R}, \tag{6.18}$$

in which $h = h(x, \bar{u})$ is an arbitrary scalar function and ∇ denotes the gradient operator in $\mathbb{R}^3$. Under these conditions, the flux vector field reads

$$F(x, \bar{u}) = n(x) \wedge \nabla h(x, \bar{u}), \qquad x \in S^2, \quad \bar{u} \in \mathbb{R} \tag{6.19}$$

and we then refer to (6.14) as the **geometric Burgers equations on the sphere** and are determined by a scalar function $h : \mathbb{R}^3 \times \mathbb{R} \to \mathbb{R}$. We will refer to the function h as the **scalar potential** of the equation. For instance, if h is chosen to be a linear function in the space variable, then Φ is independent of x but its projection on the tangent space of the sphere is still "non-trivial".

6.3 Classes of flux vector fields

6.3.1 Foliated flux vector fields

A flux field $F(x, u)$ depends on both the state variable u and the space variable x. For convenience, we adopt the notation $x_j = x^j$ from now on. Roughly speaking, the dependency in x drives the propagation of the waves, while the dependency in u may induce the formation of shocks in the solutions. Some aspects of the influence of the parameters x and u on the evolution of the solution are observed and analyzed in various cases studied in the numerical tests.

Let us illustrate with two examples, which we will later investigate numerically. In Test 1-a discussed below, the directions of propagation depend on the space variable x only, which is achieved by choosing the potential function $h(x, u) = -x_3 u$. The solutions are simply transported so that the directions of this transportation (or level sets) are defined by the curves $x_3 = c$, where $c \in [-1, 1]$ is any real constant. Having here a potential which is linear in u, no shock wave can form during the evolution (from regular data, say).

On the other hand, when the potential function is chosen to be $h(x, u) = -x_3 u^2/2$ (as will be investigated in Test 5, below), again the directions of evolution are given by

the curves $x_3 = c$, where $c \in [-1, 1]$ is any real constant. For this case, shock waves do form in finite time from "general" initial data: this feature is due to the nonlinear dependency of h in the variable u. These shocks generate rather large variations in solutions along the level sets. We still have formation of shocks within each level set and the solution can be computed independently in each of these lines. We will observe later that each solution converges asymptotically to a *constant value* on each line.

To conduct a rigorous numerical analysis and with the examples above in mind, it is useful to introduce some new definitions, which allows us to have the classification of the flux vectors and the type of evolution of solutions. Consider first the dependency in the variable $x \in S^2$. Our analysis has found that the following parameterized level sets $\Gamma_{C,u} = \left\{ x \in \mathbb{R}^3 \Big/ h(x, u) = C \right\}$ play a central role and that the following definition is most relevant.

Definition 6.3.1 *A gradient flux vector field $F(x, u) = n(x) \wedge \nabla h(x, u)$ defined on the sphere S^2 and associated with a potential function h is called a* **foliated flux field** *if the associated family of level sets $\left\{ \Gamma_{C,u} \right\}_{C \in \mathbb{R}}$ in $\mathbb{R}^3$ is independent of the parameter u, in the sense that for any two u_1, u_2 one can find C_1, C_2 such that $\Gamma_{C_1,u_1} = \Gamma_{C_2,u_2}$.*

As will be confirmed later by our numerical tests, when the foliated condition above holds, the directions of propagation associated with Equation (6.15) depend on the spatial variable x only, and are independent on the variable u; hence, the level sets are determined by the spatial variable only and remain unchanged over time, even under the evolution of the solution.

A typical subclass of interest is obtained when h has the following splitting form.

Definition 6.3.2 *All gradient flux vector field $F(x, u) = n(x) \wedge \nabla h(x, u)$ defined on the sphere S^2 and associated with a potential function h of the form*

$$h(x, u) = \overline{h}(x) f(u) \tag{6.20}$$

(for an arbitrary $\overline{h}$) are foliated and are referred to as **foliated flux field based on splitting.**

Flux vectors of the form above will be investigated later in numerical tests. In particular, we use the **x-linear potential functions**, defined by $h(x, u) = (x \cdot a) f(u)$, where $x \cdot a$ denotes the scalar product of the vector x and some constant vector $a \in \mathbb{R}^3$. In the latter situation, we have a very natural slicing of the sphere S^2 by planes in $\mathbb{R}^3$. In all the above cases, we obtain decoupled "dynamics" on each level set. If the family of level sets is locally a family of curves, then the conservation laws reduces to a family of one-dimensional equations on each curve.

6.3.2 Notion of independent domains

When the flux is *not* foliated, we will consider that we are in a "generic" situation and will use the terminology **"generic flux field"** and, in this case, the potential function $h = h(x, u)$ does not have the specific structure exhibited above. Yet, this function can be decomposed into some homogeneous terms and the evolution of the solution is influenced by all those terms, especially, the direction of propagation changes during the evolution, until the solution finally converges asymptotically to some limiting state.

The following notion of "independent domain" on the sphere, presented now, will be of importance in our forthcoming study of the asymptotic convergence of solutions.

Definition 6.3.3 *Given a gradient flux field, a subset of the sphere S^2 is called an* **independent domain** *if within the family of level sets $\left\{\Gamma_{C,u}\right\}_{C\in\mathbb{R}}$, one can find one level set that is independent of the parameter u and coincides with the boundary of this domain.*

Such independent domains may exist for foliated flux field as well as generic flux fields. For example, the circle on the sphere defined by $x_1 = 0$ splits the sphere in two independent domains for the foliated flux based on the potential function $h_1(x, u) = x_1 u^2$. The same is true for the generic flux field based on the potential function $h_2(x, u) = x_1 u^2 + x_1 x_2 u^3$.

6.3.3 Genuine nonlinearity and late-time asymptotics

Consider now the dependency of the scalar potential h in u. A special situation is obtained when the function h is linear in u, and in which case we use the terminology **"linear flux".** The classification that we introduced to distinguish between foliated flux and generic flux, and the character of linearity of the flux are expected to be sufficient to predict the late-time asymptotic behavior of the solutions. The following will be validated numerically concerning the asymptotic behavior of solutions. Under the notation and assumptions (6.16) and (6.18) and for any initial condition, three late-time asymptotic behaviors are expected for entropy solutions of Equation (6.15):

- For a linear foliated flux, the solutions are simply transported within the level sets.
- If the flux is foliated with nonlinear behavior, the solution converges to its constant average in each level set.
- The generic flux generates large variations in solutions, which finally converge to constants within independent domains on the sphere.

The late-time asymptotic behavior of the solutions for linear foliated flux is numerically studied in Section 6.7 using Test 1-a and Test 1-b. According to our analysis, we concluded that for this type of flux, the solution is transported along each level set. The propagation speeds of the solution along the sets depend on the variation of the scalar potential function according to the spatial variable.

We examine the case of nonlinear foliated flux using Test 2-b, Test 3-a and Test 5 presented in Section 6.7. In general for this type of flux the solution converges to its constant average in each level set as examined in Test 5. Some particular behaviors can be numerically observed according to the structure of the computational grid for nonlinear foliated fluxes in which the solution converges to constant values in independent domains on the sphere. The lines which split these domains are part of the level sets. As examples, the solution of Test 2-b converges asymptotically to two constant values in two independent domains on the sphere and the solution of Test 3-a converges to one constant value on the entire sphere.

For a generic flux, as will be shown by the tests performed in Section 6.7, the behavior of the solution is largely influenced by the dependency of the scalar potential $h(x, u)$, both on the spatial variable x and the value of the function u. In those tests we consider generic fluxes in which the scalar potential is composed of different homogeneous fluxes. Following the tests performed in this paper, for this type of flux, the solution converges to constant values in independent domains on the sphere. The system of equations is conservative. Thus the constant values, which are the asymptotic limits of convergence, represent the averages of the function taken as initial condition in those domains.

6.4 Special classes of solutions

6.4.1 Wave structure

There are many solutions of particular interest which may have a very rich wave structure, including spatially periodic solutions and steady state solutions. Since for the foliated flux, the system of equations of interest can be reduced to a family of one-dimensional equations on level sets, this type of flux is considered to construct some particular and interesting solutions. The foliated flux with linear behavior is used to obtain the spatially periodic solutions. The foliated flux with nonlinear behavior is employed to construct large families of stationary solutions which are communally used in the numerical tests to check the well-balanced property. We will see in our numerical tests that for a nonlinear foliated flux, the level sets introduced in this paper can be used to improve the numerical schemes by considering a suitable choice of the mesh in order to preserve the stationary solutions. More precisely, in order to preserve numerically the steady state solutions, the lines of the computational grids should be a part of the level sets and their equipotential curves which are orthogonal to them.

6.4.2 Spherical coordinates

The two-dimensional spherical coordinate system is considered here. The position of each point on the sphere is specified by its longitude $\lambda \in [0, 2\pi]$ and its latitude $\phi \in [-\pi/2, \pi/2]$. The coordinates are singular at the south and north poles, corresponding to $\phi = -\pi/2$ and $\phi = \pi/2$, respectively. The Cartesian coordinates are denoted by $x = (x_1, x_2, x_3)^T \in \mathbb{R}^3$ with the corresponding standard basis vectors i_1, i_2 and i_3. The spherical coordinates under consideration lead to the following unit normal vector to the sphere.

$$n(x) = \cos\phi\cos\lambda i_1 + \cos\phi\sin\lambda i_2 + \sin\phi i_3,$$

and for each point on the sphere with coordinates (λ, ϕ), we obtain the following unit tangent vectors in the directions of longitude and latitude

$$\begin{aligned} i_\lambda &= -\sin\lambda i_1 + \cos\lambda i_2, \\ i_\phi &= -\sin\phi\cos\lambda i_1 - \sin\phi\sin\lambda i_2 + \cos\phi i_3. \end{aligned}$$

The equation of conservation law (6.15), can be rewritten using the spherical coordinates in the following form:

$$\partial_t u + \frac{1}{\cos\phi}\left(\frac{\partial}{\partial\phi}(F_\phi\cos\phi) + \frac{\partial F_\lambda}{\partial\lambda}\right) = 0, \tag{6.21}$$

where F_λ and F_ϕ are the flux components in spherical coordinates. They are given for each three-dimensional flux of the form $\Phi(x,u) = \tilde{\boldsymbol{f}}_{\mathbf{1}}(x,u)i_1 + \tilde{\boldsymbol{f}}_{\mathbf{2}}(x,u)i_2 + \tilde{\boldsymbol{f}}_{\mathbf{3}}(x,u)i_3$ as follows:

$$\begin{aligned} F(x,u) &= F_\lambda(\lambda,\phi,u)i_\lambda + F_\phi(\lambda,\phi,u)i_\phi, \\ F_\lambda(\lambda,\phi,u) &= \tilde{\boldsymbol{f}}_{\mathbf{1}}(x,u)\sin\phi\cos\lambda + \tilde{\boldsymbol{f}}_{\mathbf{2}}(x,u)\sin\phi\sin\lambda - \tilde{\boldsymbol{f}}_{\mathbf{3}}(x,u)\cos\phi, \\ F_\phi(\lambda,\phi,u) &= -\tilde{\boldsymbol{f}}_{\mathbf{1}}(x,u)\sin\lambda + \tilde{\boldsymbol{f}}_{\mathbf{2}}(x,u)\cos\lambda. \end{aligned} \tag{6.22}$$

6.4.3 Solutions for linear foliated flux

We consider the family of linear fluxes defined on the basis of the scalar potential $h(x,u) = \overline{h}(x)u$ with $\overline{h}(x) = -cx_3^d$ for an integer $d \geq 1$ and a real number c chosen arbitrarily. Under these considerations, the three-dimensional flux reads

$$\Phi(x,u) = \nabla h(x,u) = -cdx_3^{d-1}ui_3. \tag{6.23}$$

The components of the flux in spherical coordinates can be deduced by using the explicit formulas (6.22) as follows

$$\begin{aligned} F_\lambda(\lambda,\phi,u) &= cdx_3^{d-1}\cos\phi u, \\ F_\phi(\lambda,\phi,u) &= 0. \end{aligned} \tag{6.24}$$

Finally, it is easy to derive the analytical solution for any initial condition $u_0(\lambda, \phi)$

$$\begin{aligned} u(x,t) &= u_0(\lambda - c_\phi t, \phi), \\ c_\phi &= c d x_3^{d-1}. \end{aligned} \tag{6.25}$$

The level sets of this type of fluxes are the circles on the sphere defined by constant latitudes. For $d = 1$ the solution is simply transported within those level sets with the same angular speed and it is globally preserved in a rotating frame. For $d > 1$ the solution is transported within the level sets with different angular speeds and it is preserved in a moving frame along each level set but the solution is not globally preserved. We note that more general forms of the solutions for linear fluxes can be obtained by considering other functions $\overline{h}(x)$.

6.4.4 Non-trivial steady state solutions

In this section, we present some general classes of non-trivial steady state solutions which will be used in the numerical tests. As mentioned before, foliated fluxes are used to construct non-trivial stationary solutions. More precisely, based on the expected asymptotic behavior for non-linear foliated flux, the solution of Equation (6.15) evolves along each level set and for a long period of time, this solution converges asymptotically to a constant value on each level set. Thus, it is straightforward (but fundamental) to deduce that for this type of flux, any stationary solution of Equation (6.15) must be constant along each level set. This result is important and further simplifies the problem to find the stationary solutions. However, for the solutions which are used in our numerical tests, it will be proved that they are stationary. We will be particularly interested, in this section, to a linear splitting flux vector defined on the basis of the scalar potential $h(x,u) = (x \cdot a) f(u)$, where as already mentioned $x.a$ denotes the scalar product of the vector x and some constant vector $a = (a_1, a_2, a_3)^T \in \mathbb{R}^3$. For this case the corresponding flux is obtained as

$$F(x,u) = f(u) n(x) \wedge a.$$

The level sets of this flux are the circles defined as the intersections of the sphere with the planes defined as $x.a = c$, where $|c| \leq \|a\|_2$. These level sets will be parametrized by the real constant c and denoted by Γ_c. The following Corollary describes for the above-mentioned type of flux, a family of non-trivial steady state solutions

Corollary 6.4.1 ***(A family of steady state solutions).*** *Consider the foliated flux vector $F(x,u) = f(u)n(x) \wedge a$, where a is some constant vector in $\mathbb{R}^3$. For any function $\tilde{u}$ which depends on one variable, the function defined as $u_0(x) = \tilde{u}(x \cdot a) = \tilde{u}(a_1 x_1 + a_2 x_2 + a_3 x_3)$ is a stationary solution to the conservation law* (6.15) *associated to the flux $F(x,u)$.*

Proof. In order to prove that the function $u_0(x)$ is a stationary solution we use the claim 3.2 in Ben-Artzi et al. (2009). To conduct this, we consider the function $H(x) = H_0(a_1x_1 + a_2x_2 + a_3x_3)$, where $H_0(\mu) = \int_{\mu_0}^{\mu} f(\tilde{u}(\mu))d\mu$ for some reference value μ_0. It is clear that $h(x, u) = (x \cdot a)f(u)$ is a smooth function in $\mathbb{R}^3$, particularly in a neighborhood of S^2. The following results are obtained

$$\nabla_y h(y, u_0(x))_{|y=x} = (a_1i_1 + a_2i_2 + a_3i_3)f(\tilde{u}(a_1x_1 + a_2x_2 + a_3x_3))$$

and $\nabla_y h(y, u_0(x))_{|y=x} = \nabla H(x)$, where ∇ is the standard gradient operator defined using the variable x and ∇_y is the gradient operator defined using the variable y. Therefore, all hypothesis of claim 3.2 in Ben-Artzi et al. (2009) are satisfied. Finally, the function $u_0(x)$ is a stationary solution of the conservation law (6.15). □

Since $u_0(x) = \tilde{u}(x \cdot a)$, then the function u_0 is constant on each level set Γ_c. We are interested in discontinuous solutions. The results of Corollary 6.4.1 will be used to construct discontinuous stationary solutions for some selected flux vectors. In particular, if the same assumptions of the Corollary 6.4.1 are considered with $f(u) = u^2/2$, then for any function $\tilde{u}$ which depends on one variable, the function defined as $u_0(x) = \chi(x \cdot a)\tilde{u}(x \cdot a)$ is a stationary solution to the conservation law (6.15), where $\chi(x \cdot a)$ is a discrete function which depends on the variable $x \cdot a$ and takes the values ± 1 .

Particular values of the vector a will be used in order to construct several forms of foliated flux which will be used in the numerical tests. In the second test, we consider the flux of the form $F(x, u) = f(u)n(x) \wedge i_1$ (i.e $a = i_1$). For this flux any function which depends on the first coordinate x_1 only, is a steady state solution of Equation (6.15). In the third test, the vector $a = i_1 + i_2 + i_3$ is considered and for this case we obtain a steady state solution in a spherical cap of the form $u_0(x) = \tilde{u}(x_1 + x_2 + x_3)$, where $\tilde{u}$ is an arbitrary real function depending on one variable.

6.5 Geometric finite volume method on the sphere

6.5.1 Discrete form of the divergence operator

Following Ben-Artzi et al. (2009), we design a Godunov-type, finite volume scheme that is based on an intrinsic approach and provides an accurate treatment of the geometry. Second-order accuracy is obtained with the technique developed by Ben-Artzi and Falcovitz (2003), LeFloch and Raviart (1988), and Bourgeade et al. (1989). Earlier work was done by Berger et al. (2009) and Rossmanith et al. (2004); Rossmanith (2006) based on high resolution schemes and approximate Riemann solvers, but by embedding the sphere in a "cubic mesh" in $\mathbb{R}^3$.

In the following, we present the discrete form of the geometry-compatible finite volume scheme which was formulated in Ben-Artzi et al. (2009). In order to ensure a suitable

discrete form, an important condition obtained from the theory established by Ben-Artzi and LeFloch (2007) called the "zero-divergence" was used in the construction of the scheme.

The general structures of the cells used in the numerical scheme are shown in Figure 6.1. When we go from the equator to the north or south poles, for some special latitude circles, the cell is changed by a ratio of 2 in order to reduce the number of cells, to respect the condition of stability and to have a homogeneous precision in the entire domain of the sphere. The domain of each cell is defined as $\Omega := \{(\lambda, \phi), \lambda_1 \leqslant \lambda \leqslant \lambda_2, \phi_1 \leqslant \phi \leqslant \phi_2\}$. A cell near the north or south poles has three sides which is a special case of the standard cell shown in Figure 6.1 with zero length for one side.

The divergence operator is discretized using the geometry compatibility condition and the flux is approximated using the following formula:

$$(\nabla \cdot (F(x, u(t, x))))^{approx} = \frac{I_i}{\omega_i}, \tag{6.26}$$

where $I_i = (\oint_{\partial\Omega} F(x, u) \cdot \nu(x) ds)^{approx}$ which is obtained using the divergence theorem, $\nu(x)$ is the unit normal vector to the boundary $\partial\Omega$ of the cell, ds is the arc length along $\partial\Omega$, and ω_i is the area of the cell. The parameter I_i is calculated for each side e of the cell in terms of the scalar potential h using the following expression:

$$\begin{aligned}
&\left(\oint_{e_1}^{e_2} F(x, u) \cdot \nu(x) ds\right)^{approx} = \oint_{e_1}^{e_2} (n(x) \wedge \Phi(x, u)) \cdot \nu(x) ds \\
&= -\oint_{e_1}^{e_2} \Phi(x, u) \cdot (n(x) \wedge \nu(x)) ds \\
&= -\oint_{e_1}^{e_2} \nabla h(x, u) \cdot \tau(x) ds = -\oint_{e_1}^{e_2} \nabla_{\partial\Omega} h(x, u) ds \\
&= -(h(e_2, u_m) - h(e_1, u_m)),
\end{aligned} \tag{6.27}$$

where e_1 and e_2 are, respectively, the initial and final endpoints of the edge e, $\tau(x)$ is the unit tangent vector to the boundary $\partial\Omega$, u_m is the solution of the Riemann problem in the orthogonal direction to the interface e, and the operator $\nabla_{\partial\Omega}$ is the derivative along the boundary $\partial\Omega$.

Observe that the grid structure shown in Figure 6.1 is favorable in order to apply the standard *splitting approach,* which will be described in the section below. We can then solve generalized Riemann problems at each interface of discontinuity by using the variables λ and ϕ separately. This grid structure also provides a discrete form of the scheme which *exactly satisfies* the "null-divergence" condition.

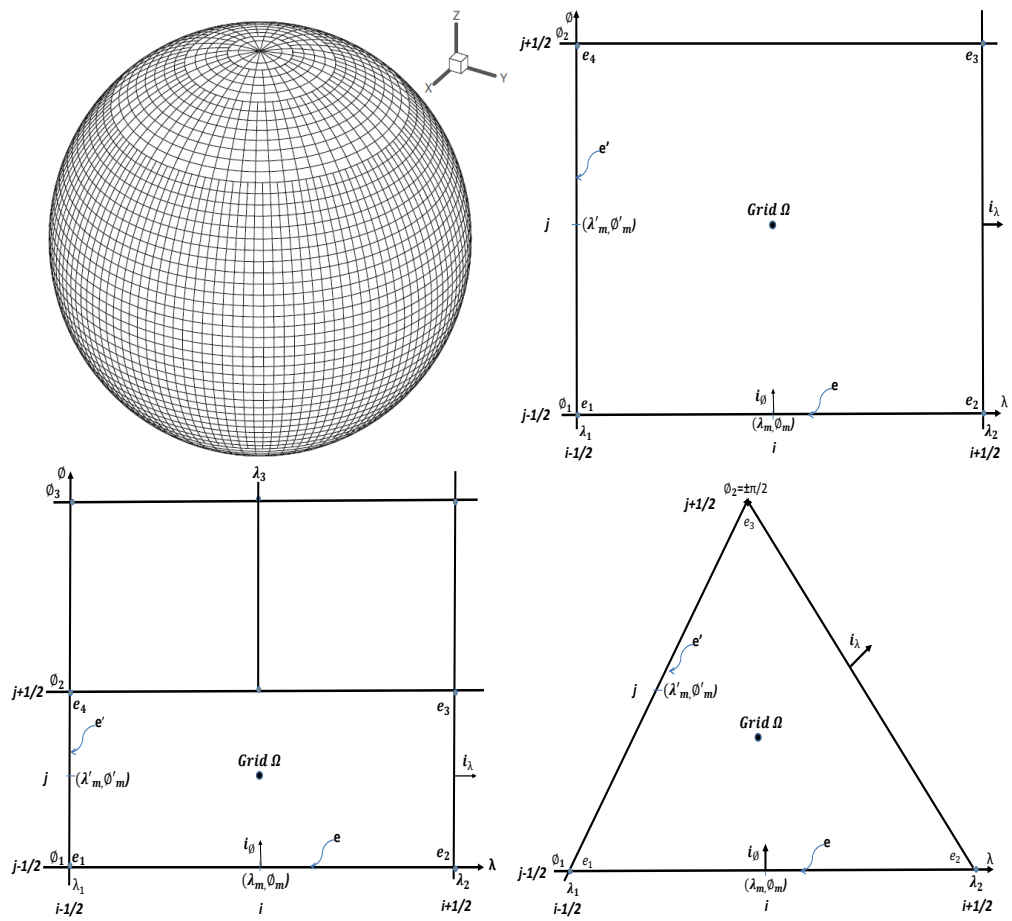

Figure 6.1 Types of grids used on the sphere

6.5.2 Equations for the splitting approach

The different approximations used in the numerical scheme are based on the splitting of the equations. Without loss of generality, the following scalar potentials are used to explain the procedure and the different equations of the splitting approach

$$h(x,u) = h_1(x)f_1(u) + h_2(x)f_2(u) + h_3(x)f_3(u), \tag{6.28}$$

which leads to the corresponding gradient flux vector:

$$\begin{aligned}\Phi(x,u) &= \nabla h(x,u)\\ &= \sum_{j=1}^{3}\frac{\partial h_j(x)}{\partial x_1}f_j(u)i_1 + \sum_{j=1}^{3}\frac{\partial h_j(x)}{\partial x_2}f_j(u)i_2 + \sum_{j=1}^{3}\frac{\partial h_j(x)}{\partial x_3}f_j(u)i_3.\end{aligned} \tag{6.29}$$

Using Claim 2.2 in Ben-Artzi et al. (2009), the above expression of $\Phi(x,u)$ as a gradient ensures the validity of the geometry compatibility condition. Equations (6.22) are used

to obtain the following flux components in spherical coordinates:

$$\begin{aligned} F_\lambda(\lambda,\phi,u) &= \Phi_1(x,u)\sin\phi\cos\lambda + \Phi_2(x,u)\sin\phi\sin\lambda - \Phi_3(x,u)\cos\phi, \\ F_\phi(\lambda,\phi,u) &= -\Phi_1(x,u)\sin\lambda + \Phi_2(x,u)\cos\lambda, \end{aligned} \tag{6.30}$$

where $\Phi_i(x,u) = \sum_{j=1}^{3} \frac{\partial h_j(x)}{\partial x_i} f_j(u), i = 1,2,3.$

The geometry compatibility condition is equivalent to the following relation in spherical coordinates which is valid for any constant value $\bar{u} \in \mathbb{R}$:

$$\frac{\partial(F_\phi(\lambda,\phi,\bar{u})\cos\phi)}{\partial\phi} + \frac{\partial F_\lambda(\lambda,\phi,\bar{u})}{\partial\lambda} = 0. \tag{6.31}$$

From (6.30) and (6.31) we derive

$$\begin{aligned} &-\sin\lambda \sum_{j=1}^{3} \frac{\partial h'_{j1}(x)\cos\phi}{\partial\phi} f_j(\bar{u}) + \cos\lambda \sum_{j=1}^{3} \frac{\partial h'_{j2}(x)\ cos\phi}{\partial\phi} f_j(\bar{u}) \\ &+ \sin\phi \sum_{j=1}^{3} \frac{\partial h'_{j1}(x)\cos\lambda}{\partial\lambda} f_j(\bar{u}) \\ &+ \sin\phi \sum_{j=1}^{3} \frac{\partial h'_{j2}(x)\sin\lambda}{\partial\lambda} f_j(\bar{u}) - \cos\phi \sum_{j=1}^{3} \frac{\partial h'_{j3}(x)}{\partial\lambda} f_j(\bar{u}) = 0, \end{aligned} \tag{6.32}$$

where $h'_{ji}(x) = \frac{\partial h_j(x)}{\partial x_i}, \quad i = 1,2,3 \quad j = 1,2,3.$ Using the conservation law in spherical coordinates (6.21), the flux components given by Equations (6.30), and the geometry-compatibility property formulated by Equation (6.32), we establish the following equivalent formulation of the conservation law, which is easier to work with:

$$\begin{aligned} &\frac{\partial u}{\partial t} - \sum_{j=1}^{3} h'_{j1}(x) \frac{\partial f_j(u)}{\partial\phi} \sin\lambda + \sum_{j=1}^{3} h'_{j2}(x) \frac{\partial f_j(u)}{\partial\phi} \cos\lambda \\ &+ \tan\phi \Big(\sum_{j=1}^{3} h'_{j1}(x) \frac{\partial f_j(u)}{\partial\lambda} \cos\lambda + \sum_{j=1}^{3} h'_{j2}(x) \frac{\partial f_j(u)}{\partial\lambda} \sin\lambda\Big) \\ &- \sum_{j=1}^{3} h'_{j3}(x) \frac{\partial f_j(u)}{\partial\lambda} = 0. \end{aligned} \tag{6.33}$$

The longitude and the latitude of the midpoint of the cell interface are denoted by λ_m

and ϕ_m, respectively. From Equation (6.33) we obtain the following "λ split" equations:

$$
\begin{aligned}
&\frac{\partial u}{\partial t} + \frac{\partial g(x,u)}{\partial \lambda} = S_\lambda, \\
&g(x,u) = \tan\phi_m \left[\sum_{j=1}^{3} h'_{j1}(x) f_j(u) \cos\lambda + \sum_{j=1}^{3} h'_{j2}(x) f_j(u) \sin\lambda \right] \\
&\qquad - \sum_{j=1}^{3} h'_{j3}(x) f_j(u), \\
&S_\lambda = \tan\phi_m \left[\sum_{j=1}^{3} f_j(u) \frac{\partial h'_{j1}(x) \cos\lambda}{\partial \lambda} + \sum_{j=1}^{3} f_j(u) \frac{\partial h'_{j2}(x) \sin\lambda}{\partial \lambda} \right] \\
&\quad - \sum_{j=1}^{3} f_j(u) \frac{\partial h'_{j3}(x)}{\partial \lambda},
\end{aligned}
\tag{6.34}
$$

while the "ϕ split" equations are

$$
\begin{aligned}
&\frac{\partial u}{\partial t} + \frac{\partial \kappa(x,u)}{\partial \phi} = S_\phi, \\
&\kappa(x,u) = -\sum_{j=1}^{3} h'_{j1}(x) f_j(u) \sin\lambda + \sum_{j=1}^{3} h'_{j2}(x) f_j(u) \cos\lambda, \\
&S_\phi = -\sum_{j=1}^{3} f_j(u) \sin\lambda \frac{\partial h'_{j1}(x)}{\partial \phi} + \sum_{j=1}^{3} f_j(u) \cos\lambda \frac{\partial h'_{j2}(x)}{\partial \phi}.
\end{aligned}
\tag{6.35}
$$

The right-hand side terms S_λ and S_ϕ of the previous equations are the result of the explicit differentiation of the flux functions $g(x,u)$ and $\kappa(x,u)$. Equations (6.34) and (6.35) are integrated using TVDRK3.

6.5.3 Second-order approximations based on generalized Riemann problems

We now present the algorithm used for the second-order finite volume method based on generalized Riemann solver. Since the same method is applied for the equations (6.34) and (6.35), we present the procedure for the case of the "λ split" equations. In the second-order method based on the general Riemann problem, it is assumed that for any time step t_n, the solution is approximated by a piecewise linear function.

Subject to the initial data for u, the proposed linear reconstruction is used as explained in detail in the next section to obtain the boundary values of u denoted by u_L and u_R and its λ-slopes denoted by $u_{L,\lambda}$ and $u_{R,\lambda}$ at each cell interface of midpoint of coordinates (λ_m, ϕ_m). These values are used to obtain the solution $\tilde{u}^m$ of the Riemann problem. The general Riemann problem method uses a linear temporal approximation to obtain the value of the solution of Riemann problem u^m at time $t_n + \Delta t/2$. The

approximation of this solution is obtained as:

$$u^m = \tilde{u}^m + \frac{\partial u}{\partial t}(\lambda_m, \phi_m, t_n)\frac{\Delta t}{2}. \tag{6.36}$$

In Equation (6.36), the derivative term is obtained by using the value of the slope of u in the longitude direction at the cell interface which is denoted by $u_{m,\lambda}$

$$\frac{\partial u}{\partial t}(\lambda_m, \phi_m, t_n) = -u_{m,\lambda}\frac{\partial g(x,u)}{\partial u}\mid_{(\lambda_m,\phi_m,\tilde{u}^m)}. \tag{6.37}$$

The parameter $u_{m,\lambda}$ is obtained by the associated Riemann problem. In the following we recall briefly the procedure used to solve the Riemann problem to obtain the values of $\tilde{u}^m$ and the slope $u_{m,\lambda}$.

For $u_L \leq u_R$, we consider the "convex envelope" of g and the solution $\tilde{u}^m$ is obtained as follows:

$$\tilde{u}^m = arg_{v\in[u_L,u_R]}\min(g(x,v) \mid (\lambda_m, \phi_m)). \tag{6.38}$$

There are three cases for this solution and for the slope $u_{m,\lambda}$:

(i) A wave moving to the right:

$$\tilde{u}^m = u_L, \qquad u_{m,\lambda} = u_{L,\lambda}. \tag{6.39}$$

(ii) A wave moving to the left:

$$\tilde{u}^m = u_R, \qquad u_{m,\lambda} = u_{R,\lambda}. \tag{6.40}$$

(iii) A sonic point:

$$u_L < \tilde{u}^m < u_R, \qquad \partial_u g(x,u)\mid_{(\lambda_m,\phi_m,\tilde{u}^m)} = 0. \tag{6.41}$$

Note that for the sonic case, it is easy to conclude, using the equations (6.37) and (6.41), that the time-derivative of u reduces to $\frac{\partial u}{\partial t}(\lambda_m, \phi_m, t_n) = 0$. The geometry-compatibility condition remains valid also for the second-order scheme. Indeed, if we consider the condition $u \equiv const$ in the computational cell and its neighbors, the slopes and the time-derivatives of the generalized Riemann solution vanish and the solution remains constant. Under the condition $u_L > u_R$, the same procedure is used by considering the "concave envelope" of g and the value $\tilde{u}^m$ that maximizes the function $g(x,v)$ with $v \in [u_R, u_L]$. Finally, the same procedure is used for the case of the "ϕ split" equations using the boundary values of u and its ϕ-slopes $u_{L,\phi}$ and $u_{R,\phi}$ at each cell interface.

6.5.4 The proposed piecewise linear reconstruction

In this section, we describe the proposed piecewise linear reconstruction. At each time step t_n, data cell average values $u^n_{i,j}$ in each cell of center (λ_i, ϕ_j) are locally replaced by a piecewise linear function. The indices i and j are used along the longitude and latitude respectively to locate the position of the centers of cells. The proposed reconstruction leads to the following local function:

$$u^n_{i,j}(\lambda, \phi) = u^n_{i,j} + (\lambda - \lambda_i)\alpha^n_{i,j} + (\phi - \phi_j)\beta^n_{i,j}, \tag{6.42}$$

where $\alpha^n_{i,j}$ and $\beta^n_{i,j}$ are the slopes in the directions of longitude and latitude, respectively. We consider the notation $u^{n\pm}_{i+\frac{1}{2},j}$ for the corresponding right and left values of u at the interface $(i+\frac{1}{2}, j)$ in the longitude direction which are obtained by using the proposed piecewise linear reconstruction at time t_n. In the same way the notation $u^{n\pm}_{i,j+\frac{1}{2}}$ is used for the right and left values at the interface $(i, j+\frac{1}{2})$ in the latitude direction for the function u at time t_n. These values are used to obtain the associated Riemann values which are denoted by $u^n_{i+\frac{1}{2},j}$ and $u^n_{i,j+\frac{1}{2}}$ at the interfaces $(i+\frac{1}{2}, j)$ and $(i, j+\frac{1}{2})$, respectively.

The slope $\alpha^n_{i,j}$ is obtained by using the following steps:

$$\begin{aligned}
\tilde{u}^n_{i+\frac{1}{2},j} &= u^{n-1}_{i+\frac{1}{2},j} + (\frac{\partial u}{\partial t})^{n-1}_{i+\frac{1}{2},j}(t_n - t_{n-1}),\\
\tilde{\alpha}^n_{i,j} &= \frac{1}{\Delta\lambda}(\tilde{u}^n_{i+\frac{1}{2},j} - \tilde{u}^n_{i-\frac{1}{2},j}),\\
\alpha^n_{i,j} &= \frac{1}{\Delta\lambda}\ minmod((u^n_{i+1,j} - u^n_{i,j})), \Delta\lambda\tilde{\alpha}^n_{i,j}, (u^n_{i,j} - u^n_{i-1,j})),
\end{aligned} \tag{6.43}$$

where $u^{n-1}_{i+\frac{1}{2},j}$ is the solution of the Riemann problem at the interface $(i+\frac{1}{2}, j)$ which is obtained in the previous step t_{n-1} using the procedure explained in Section 6.5.3. The value of $\tilde{u}^n_{i+\frac{1}{2},j}$ computed using the first equation in (6.43) based on the data from step t_{n-1} is an approximation of the solution of the Riemann problem at the interface $(i+\frac{1}{2}, j)$ at time t_n. The instantaneous time-derivative $(\frac{\partial u}{\partial t})^{n-1}_{i+\frac{1}{2},j}$ in (6.43) is computed in the previous step t_{n-1} using Equation (6.37). In the third equation of (6.43) we use the multivariable minmod function defined as

$$\begin{aligned}
&minmod(\sigma_1, \sigma_2, \sigma_3)\\
&= \begin{cases} \sigma \min(|\sigma_1|, |\sigma_2|, |\sigma_3|), & if\ sign(\sigma_1) = sign(\sigma_2) = sign(\sigma_3) = \sigma,\\ 0. & \text{otherwise.} \end{cases}
\end{aligned} \tag{6.44}$$

In the same way the slope $\beta^n_{i,j}$ is obtained by using the following steps:

$$\begin{aligned}
\tilde{u}^n_{i,j+\frac{1}{2}} &= u^{n-1}_{i,j+\frac{1}{2}} + (\frac{\partial u}{\partial t})^{n-1}_{i,j+\frac{1}{2}}(t_n - t_{n-1}),\\
\tilde{\beta}^n_{i,j} &= \frac{1}{\Delta\phi}(\tilde{u}^n_{i,j+\frac{1}{2}} - \tilde{u}^n_{i,j-\frac{1}{2}}),\\
\beta^n_{i,j} &= \frac{1}{\Delta\phi}\ minmod((u^n_{i,j+1} - u^n_{i,j})), \Delta\phi\tilde{\beta}^n_{i,j}, (u^n_{i,j} - u^n_{i,j-1})),
\end{aligned} \tag{6.45}$$

where $u^{n-1}_{i,j+\frac{1}{2}}$ is the solution in the previous step t_{n-1} of the Riemann problem at the interface $(i, j+\frac{1}{2})$ which is obtained by using the same procedure explained in Section 6.5.3 for the case of the "ϕ split" equations. The value of $\tilde{u}^n_{i,j+\frac{1}{2}}$ is obtained by using the first equation in (6.45). This value is an approximation of the solution of the Riemann problem at the interface $(i, j+\frac{1}{2})$ at time t_n. The value of the instantaneous time-derivative $(\frac{\partial u}{\partial t})^{n-1}_{i,j+\frac{1}{2}}$ is obtained in the previous step t_{n-1} using an equation similar to (6.37), and by using the slope $u_{m,\phi}$ of u in the latitude direction at the cell interface.

It should be mentioned that, in the first step, we do not have data from any previous step. For this case, in the proposed piecewise linear reconstruction, the third equations in (6.43) and (6.45) are used to obtain the slopes $\alpha^n_{i,j}$ and $\beta^n_{i,j}$, respectively, without using the parameters $\tilde{\alpha}^n_{i,j}$ and $\tilde{\beta}^n_{i,j}$. In the new third equations in (6.43) and (6.45), we use a two-variable minmod function.

6.6 Analysis of the spatial and temporal orders of accuracy of the scheme

6.6.1 Spatial order of the scheme

Two numerical tests are performed in order to determine experimentally the spatial order of accuracy of the proposed scheme. We used the nonlinear foliated flux defined from the scalar potential $h(x,u) = (x.a)f(u)$ with $f(u) = u^2/2$. In the first test, we consider the flux with $a = i_1$ and, as initial data, the three (steady state) functions: $u_1(0,x) = x_1$, $u_2(0,x) = x_1\cosh(x_1)$, and $u_3(0,x) = x_1^3\sin x_1$. For the second test, we use the flux with $a = i_1 + i_2 + i_3$ and, as initial data, the three steady state solutions: $v_1(0,x) = \sinh\theta/(1+\theta^2)$, $v_2(0,x) = (1-\theta)e^\theta$, and $v_3(0,x) = \theta^3$ with $\theta = x_1 + x_2 + x_3$.

To determine numerically the spatial convergence rate, we use a very small time step $\Delta t = 0.0001$ in order to render the temporal errors negligible. Different sizes of the computational cells are used to obtain the evolution of the error. We use the longitude step $\Delta\lambda$ and the latitude step $\Delta\phi$ with the same order, and to evaluate the errors, we use the mesh-size weighted L^1-norm. Figure 6.2 shows the evolution of the L^1-error in a log-log scale up to the time $T = 5$ for the first and second tests. For both choices of flux and for all the initial conditions under consideration, we observe that for small sizes of the computational cells the spatial convergence rate is approximately equal to

2. This result confirms the second-order of spatial accuracy of the scheme studied in this paper.

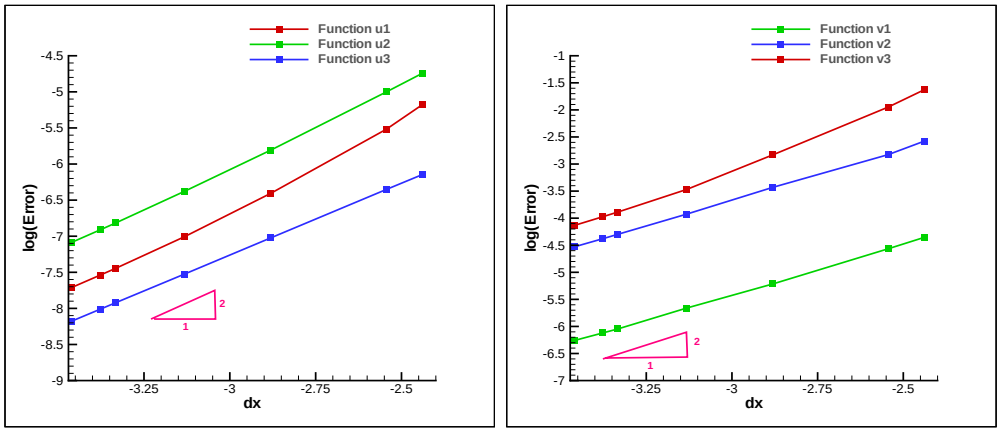

Figure 6.2 Evolution of L^1-error in log-log scale until time $T = 5$ ($dx = \log(\Delta\lambda)$) for the nonlinear foliated flux defined on the basis of the scalar potential $h(x, u) = (x.a)f(u)$. Left: For the nonlinear foliated flux defined by using $a = i_1$. Right: For the nonlinear foliated flux defined by using $a = i_1 + i_2 + i_3$.

6.6.2 Impact of the splitting approach on the temporal order of the scheme

In this section we study the impact of the splitting approach on the third-order of the TVD Runge-Kutta method used for time integration in the proposed scheme. The experimental determination of the temporal accuracy is more challenging since the combined effects of spatial and temporal errors are, in general, difficult to distinguish. The use of a very small size of the computational cells to reduce the spatial errors is not possible since the stability condition limiting the ratio value $\vartheta\Delta t/L_{cell}$ should be satisfied for each cell, where L_{cell} is the minimum distance inside the cell from its center, and ϑ is the wave speed. We will use the algorithm used by Bona et al. (1995) to numerically obtain the order of temporal accuracy of the proposed method.

For a fixed size of mesh $\Delta\lambda = \Delta\phi$, we consider a reference solution at time T denoted by u_{ref} which is obtained by using a very small time step Δt_{ref}. This reference solution will differ from the exact solution by an error that is almost purely from the spatial discretization. This solution is used in the numerical tests in order to cancel the spatial errors. For a fixed spatial size $\Delta\lambda$ of the computational cell, a modified error at time T denoted by $\tilde{E}$ is defined using the L^1-norm. This modified error is computed for values of time steps Δt that are larger than Δt_{ref}, using the following formulas:

$$\tilde{E}(T, \Delta t) = \left\| u^{(n)}(\Delta\lambda, \Delta t) - u^{(n)}_{ref}(\Delta\lambda, \Delta t_{ref}) \right\|_{L^1} / \left\| u^{(a)} \right\|_{L^1}, \tag{6.46}$$

where $u^{(a)}$ is the analytical solution and $u^{(n)}$ is the numerical solution obtained by

using a time step Δt and the same spatial step $\Delta\lambda$ used for the reference solution $u_{ref}^{(n)}$.

For small values of Δt which are larger than Δt_{ref}, the temporal rate of convergence of the proposed scheme can be visible because when we subtract the reference solution $u_{ref}^{(n)}(\Delta\lambda, \Delta t_{ref})$ from the approximate solution $u^{(n)}(\Delta\lambda, \Delta t)$, the spatial errors are almost canceled. In order to experimentally check the order of accuracy of the proposed temporal method which combines the method TVDRK3 with the splitting approach a numerical test is performed using the nonlinear foliated flux defined based on the scalar potential $h(x, u) = (x.a)f(u)$ with $f(u) = u^2/2$ and $a = i_1$. The proposed scheme is applied to the system (6.21) subject to the initial condition (steady state solution) $u(0, x) = x_1^3$.

The temporal rate of convergence at time $T = 5$ is shown in Table 6.1 for $\Delta t_{ref} = 0.0001$. The results confirm that the splitting method has less impact on the third order accuracy of the TVD Runge-Kutta method used for temporal integration.

Table 6.1 Temporal rate of convergence of the proposed scheme at $T = 5$

Δt	1E-03	2E-03	4 E-03	6 E-03	1 E-02
Rate	3.00	3.00	3.00	2.99	2.98

6.7 Numerical experiments

6.7.1 First test case with linear foliated flux

Referring to Section 6.4.3, here we perform two tests cases using linear fluxes based on different scalar potentials $h(x, u) = \overline{h}(x)u$. We consider a grid with an equatorial longitude step $\Delta\lambda = \pi/96$ and a latitude step $\Delta\phi = \pi/96$, and a time step $\Delta t = 0.01$. In the first numerical test (Test 1-a), the function $\overline{h}(x) = -x_3$ is considered, which leads to the following flux vector:

$$F_\lambda(\lambda, \phi, \lambda) = \cos\phi u, \qquad F_\phi(\lambda, \phi, \lambda) = 0. \tag{6.47}$$

We consider the initial condition with a discontinuity along the curve $x_1 = 0$, defined as:

$$u(0, x) = \begin{cases} \cos\phi, & x_1 \geq 0, \\ -\cos\phi, & \text{otherwise.} \end{cases} \tag{6.48}$$

For this case, the solution is transported with the same angular speed along the level sets which are the circles defined by $\phi = \phi_c$, where $\phi_c \in [-\pi/2, \pi/2]$. Figure 6.3, on the left, shows the solution at time $t = 50$ and confirms that it is globally preserved in rotating frame on the sphere.

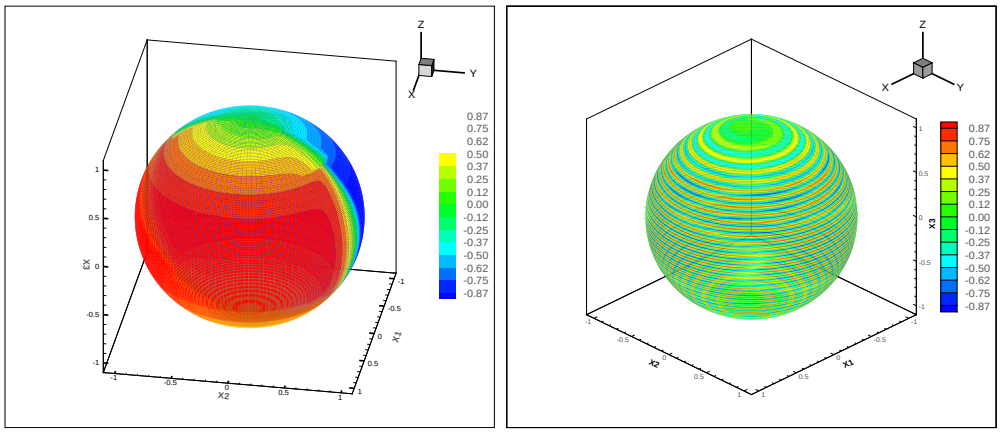

Figure 6.3 Solutions of Test 1-a (left) and Test 1-b (right) at time $t = 50$ with $\Delta t = 0.01$, $\Delta\lambda = \pi/96$, and $\Delta\phi = \pi/96$.

Now we consider the second test (test 1-b) in which the flux vector is defined on the basis of the scalar potential $h(x, u) = -x_3^2 u$, and the same initial condition is used as in the first test. For this test, again the solution is transported along the same level sets but with different angular speeds. As shown in Figure 6.3, on the right, the solution is preserved at time $t = 50$ on each level set in a moving frame but not globally preserved on the sphere.

6.7.2 Second test case with nonlinear foliated fluxes

In this section several aspects will be analyzed for a nonlinear foliated flux of the form $F(x, u) = f(u)n(x) \wedge i_1$ with $f(u) = u^2/2$. The evolution of L^1- error of the proposed scheme is analyzed using discontinuous steady state solutions. The entropy stability property (6.3), the time-variation diminishing property (6.4), and the contraction property (6.5), are analyzed for the first and second order of the scheme. The late-time asymptotic behaviors of the solutions are analyzed using this flux and different initial conditions.

First, we consider the following discontinuous steady state solution of Equation (6.15) which is taken as an initial condition (test 2-a).

$$u_1(0, x) = \begin{cases} 1, & x_1 \leq 0, \\ -1, & \text{otherwise.} \end{cases} \tag{6.49}$$

In this test we compute the numerical solution by using the computational cell with equatorial longitude step $\Delta\lambda = \pi/96$ and latitude step $\Delta\phi = \pi/96$, and a time step $\Delta t = 0.03$.

A two-dimensional view of the solution at time $t = 100$ is presented in Figure 6.4 (left) which confirms that the solution remains unchanged over the entire sphere. The

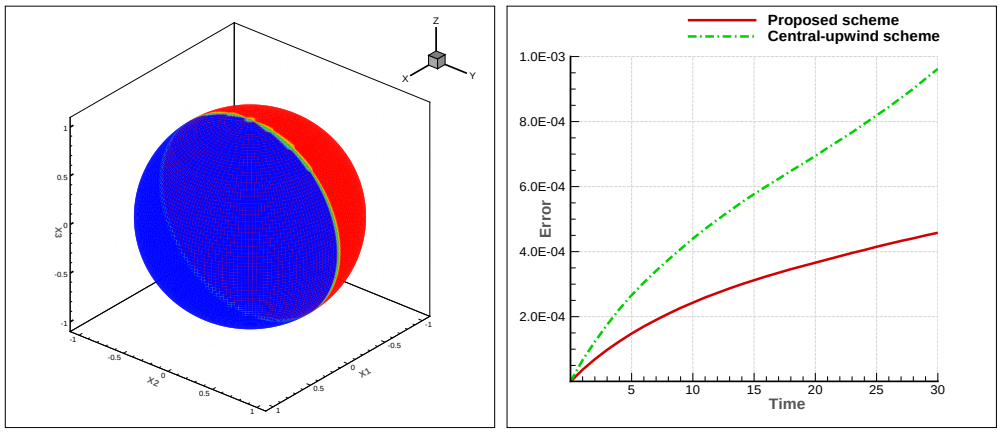

Figure 6.4 Solution of Test 2-a on the entire sphere (left) at time $t = 100$. Evolution of L^1-error of the solution until time $t = 30$ for Test 2-b (right) using the proposed scheme and Central-upwind scheme with $\gamma = 0.1$. The grid with $\Delta\lambda = \pi/96$ and $\Delta\phi = \pi/96$ is used for the two tests cases.

numerical solution is in good agreement with the initial condition and it remains steady state. For this example, since the discontinuity coincides with the level set ($x_1 = 0$), there is no numerical impact of the mesh structure on the solution which is preserved for a large simulation time. More generally, if we use the assumption that the sides of the mesh coincide with the level sets of the flux and their orthogonal equipotential curves, the second-order scheme based on the geometric-compatible property can correctly compute the discontinuous steady sate solutions based on constant values on partially closed domains on the sphere. One can say that the scheme satisfies the discontinuous geometric-compatible property since it is capable to capture all discontinuous steady sate solution based on constant values on closed domains which form a partition on the sphere. But this property is numerically valid if a subset of the level sets is taken as a lines of computational cells which is very important for the performance of the numerical schemes in the case of nonlinear foliated flux.

Now we consider a new test (Test 2-b) using the following steady state solution, with more discontinuities, which is defined in three domains separated by two closed curves on the sphere defining these discontinuities.

$$u_2(x) = \begin{cases} \gamma x_1^3, & -1 \leq x_1 \leq -0.5, \\ 0.5\gamma x_1^2, & -0.5 < x_1 < 0.5, \\ -0.25\gamma x_1, & 0.5 \leq x_1 \leq 1, \end{cases} \tag{6.50}$$

where the parameter γ is introduced in order to control the amplitude and shocks of the solution.

The numerical solution is computed by using the same computational grid as that

considered in the previous test and a time step $\Delta t = 0.03$. Figure 6.4, on the right, shows the evolution of L^1-error of the solution (6.50) with $\gamma = 0.1$ until time $t = 30$ using the proposed scheme and the extension of the central-upwind scheme developed by Kurganov and Petrova (2005) to the spherical case. The proposed scheme performs much better than the central-upwind scheme especially for the discontinuous solutions with large amplitudes and shocks. For $\gamma = 0.3$ at time $t = 5$, we obtain the L^1-error of 9.3×10^{-4} for the proposed method and the error of 1.7×10^{-3} for the central-upwind scheme.

Figure 6.5 shows the numerical solution on the equator of the sphere at time $t = 5$ using the initial condition (6.50) for the two cases $\gamma = 0.1$ (left) and $\gamma = 0.3$ (right). The numerical solution is in good agreement with the analytical steady-state solution (6.50).

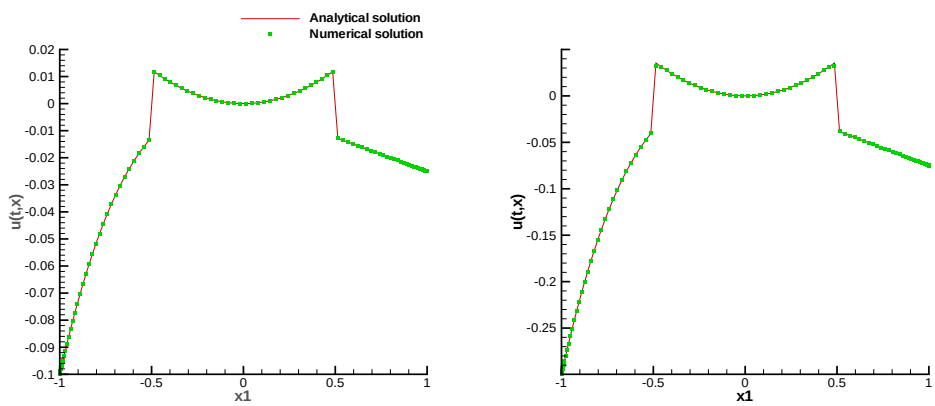

Figure 6.5 Solutions of Test 2.b at time $t = 5$ for the cases $\gamma = 0.1$ (left) and $\gamma = 0.3$ (right) using the proposed method with $\Delta t = 0.03$, $\Delta\lambda = \pi/96$, and $\Delta\phi = \pi/96$.

According to the scalar potential $h(x, u) = x_1 u^2/2$, the circle defined on the sphere by $x_1 = 0$ splits the sphere into two independent domains, the first domain includes the points with $x_1 \geq 0$ and the second includes the points with $x_1 < 0$. The average values of the initial condition in the first and second domains are denoted by $\overline{u}_I$ and $\overline{u}_{II}$, respectively. Figure 6.6, at the left, shows that the parameters $\|u(t) - \overline{u}_I\|_{L^1}$ and $\|u(t) - \overline{u}_{II}\|_{L^1}$ in the first and second domains of the sphere, respectively, are decreasing over time and tend to zero for a large simulation time. The solution converges asymptotically to different constant values in those domains and the convergence is faster in the second domain than the first domain. Figure 6.6, on the right, presents the two-dimensional view of the solution for a large simulation time and shows that the solution has almost reached an asymptotic convergence.

Following our numerical experiments, the parameter $\|u(t)\|_{L^p_\omega(M)}$ using L^p norm for $p = 1, 2, 3, 4, 5, 10$ and ∞, is decreasing with time which confirms that the entropy stability property (6.3) is verified for all those norms.

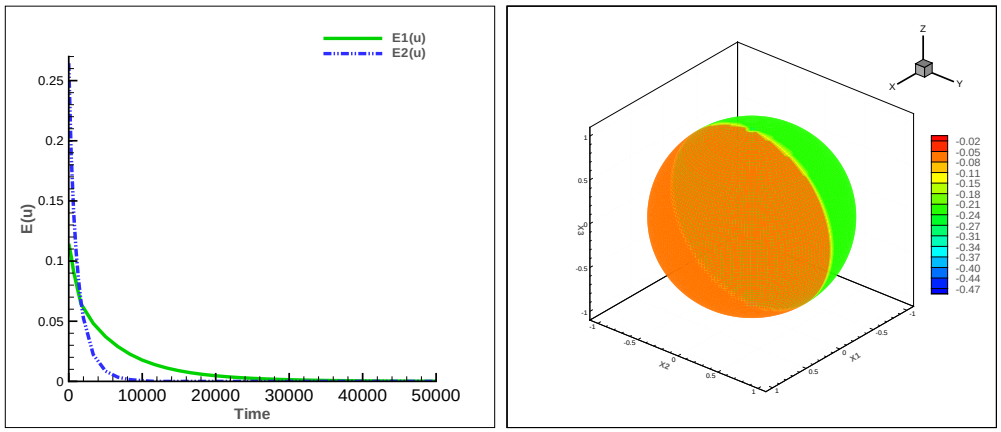

Figure 6.6 Convergence for Test 2-b using $\gamma = 1$. Left: Evolution of the parameters $E_1(u) = \|u(t) - \overline{u}_I\|_{L^1}$ in domain 1 and $E_2(u) = \|u(t) - \overline{u}_{II}\|_{L^1}$ in domain 2 with $\Delta t = 0.05$, $\Delta\lambda = \pi/96$ and $\Delta\phi = \pi/96$. Right: Two-dimensional view of the solution at time $t = 50000$.

Several tests were performed using the following functions in order to verify the time-variation diminishing property (6.4). As shown in Figure 6.7, this property holds for both cases of the first and second order of the scheme.

$$u_1(0,x) = \begin{cases} \sin\lambda, & x_1 \geq 0, \\ -\sin\lambda, & \text{otherwise}, \end{cases} \tag{6.51}$$

$$u_2(0,x) = \begin{cases} x_3, & \lambda \leq \pi, \\ x_3\cos\lambda, & \text{otherwise}, \end{cases} \tag{6.52}$$

$$u_3(0,x) = \begin{cases} x_2, & x_1 \leq 0, \\ -x_2 e^{x_1}, & \text{otherwise}, \end{cases} \tag{6.53}$$

$$u_4(0,x) = \begin{cases} \frac{1}{\theta-1}, & \theta < 0, \\ \frac{1}{1+\theta^2}, & 0 \leq \theta \leq 2/\sqrt{3}, \\ -3/7, & \text{otherwise}, \end{cases} \tag{6.54}$$

where $\theta = x_1 + 2x_2 + x_3$.

We now proceed to the analysis of the contraction property (6.5) for the numerical scheme using the L^1-norm. We start by giving an example of two functions which verify the contraction property for the first-order scheme but they do not verify this property for the second-order method. We consider the functions v_1 and w_1 defined

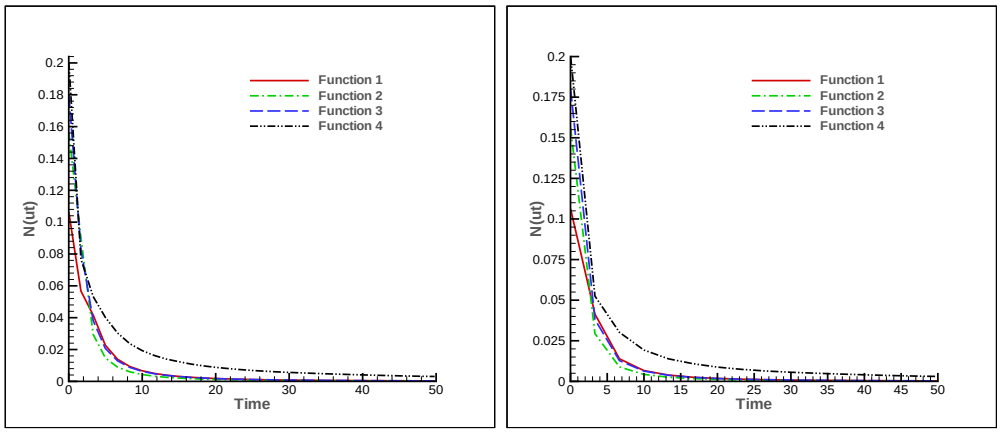

Figure 6.7 Time-variation diminishing property (6.4) for the first order of the scheme (left) and for the second order of the scheme (right) until time $t = 50$ with $\Delta t = 0.01$, $\Delta\lambda = \pi/96$ and $\Delta\phi = \pi/96$.

by:

$$v_1(0,x) = \begin{cases} x_1 + x_3^2, & x_1 > 0, \\ -x_1 - x_3^2, & \text{otherwise}, \end{cases} \tag{6.55}$$
$$w_1(0,x) = |x_1|.$$

Figure 6.8, on the left, shows the evolution of the ratio $E(v,w)$ defined by the following formula and confirms that this parameter is decreasing only in the case of the first-order scheme. Therefore, the contraction property (6.5) is valid only for the case of the first-order scheme.

$$E(v,w) = \|v(t) - w(t)\|_{L^1_\omega(M)} / \|v(0) - w(0)\|_{L^1_\omega(M)}. \tag{6.56}$$

Several tests are performed to verify the contraction property (6.5) for the first-order scheme using the five pairs of functions given in Appendix I-1.

Figure 6.8, on the right, shows the evolution of the ratio $E(v,w)$ for the five pairs of functions. This parameter is decreasing for all cases, which confirms that the contraction property (6.5) is valid for all pairs of functions considered.

6.7.3 Third test case with nonlinear foliated fluxes –an alternative form

We consider the nonlinear foliated flux $F(x,u) = f(u)n(x) \wedge (i_1 + i_2 + i_3)$ which corresponds to the scalar potential $h(x,u) = (x_1 + x_2 + x_3)f(u)$ with $f(u) = u^2/2$. Following the Corollary 6.4.1 for this flux any function of the form $u_0(x) = \tilde{u}(x_1 + x_2 + x_3)$ is a steady state solution of Equation (6.15), where $\tilde{u}$ is an arbitrary real function depending on one variable. In this section the tests are performed using the following

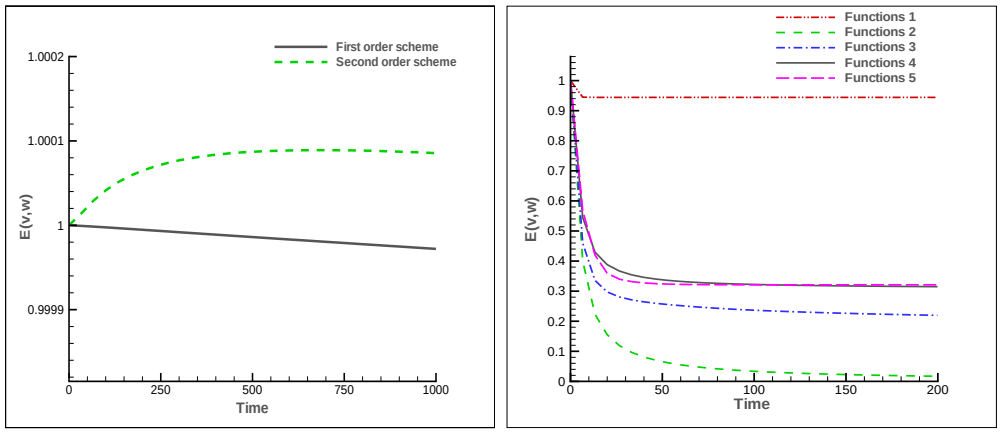

Figure 6.8 Contraction property (6.5) with $\Delta t = 0.01$, $\Delta\lambda = \pi/96$ and $\Delta\phi = \pi/96$. Left : for the case of the first and second order schemes. Right: for the case of the first-order scheme using several functions.

discontinuous steady state solution:

$$u_1(0,x) = \begin{cases} k\frac{\theta^3+1}{5-\theta}, & -\sqrt{3} \leq \theta \leq 0, \\ k(\theta^2 - \theta - 1/5), & \text{otherwise}, \end{cases} \tag{6.57}$$

where $\theta = x_1 + x_2 + x_3$. Using the inequality $(x_1 + x_2 + x_3)^2 \leq 3(x_1^2 + x_2^2 + x_3^2)$ we obtain $\theta \in [-\sqrt{3}, \sqrt{3}]$. The parameter k is used to control the amplitude and shocks of the solution (6.57).

The present test (3-a) is performed with the above function as an initial condition, using a grid with an equatorial longitude step $\Delta\lambda = \pi/96$, a latitude step $\Delta\phi = \pi/96$ and a time step $\Delta t = 0.02$. Figure 6.9 (left) shows the evolution of L^1-error of the solution until time $t = 30$ using three cases of values of the parameter k and confirms that the proposed scheme performs well for discontinuous solutions. Figure 6.9, on the right, shows the evolution of the parameter $\|u_1(t) - \overline{u}_1\|_{L^1}$ using the solution (6.57) with $k = 1$, where $\overline{u}_1$ is the average value of this solution on the sphere. For a large simulation time, the numerical solution converges to a constant value in the entire sphere.

The entropy stability property (6.3) is analyzed using the entropy solution (6.57) with the L^p-norm for $p = 1, 2, 3, 4, 5, 10$ and ∞. As shown in Figure 6.10, this property is checked for all those norms. The time-variation diminishing property (6.4) is now checked for the first and second order schemes using the L^1-norm and the following

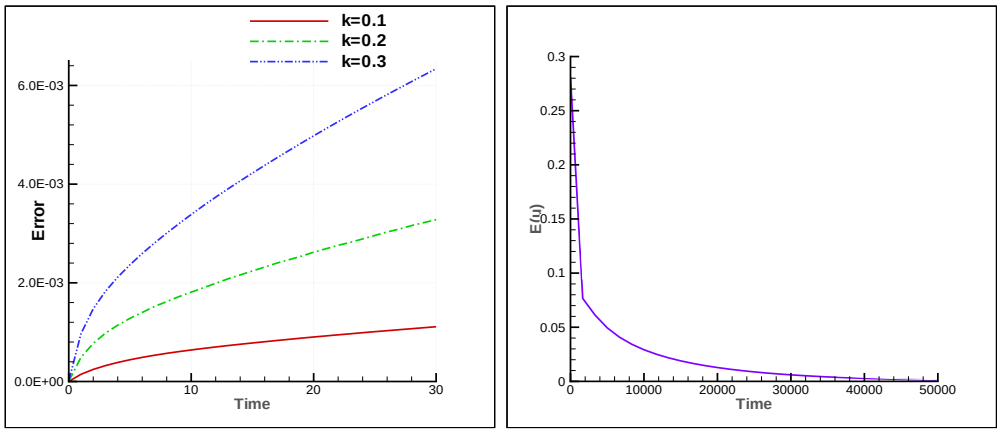

Figure 6.9 Left: Evolution of L^1-error for Test 3-a until time $t = 30$ with $\Delta t = 0.02$, $\Delta\lambda = \pi/96$ and $\Delta\phi = \pi/96$ for the three cases $k = 0.1$, $k = 0.2$ and $k = 0.3$. Right: Evolution of $\|u_1(t) - \overline{u}_1\|_{L^1}$ for large simulation time with $\Delta t = 0.05$, $\Delta\lambda = \pi/96$ and $\Delta\phi = \pi/96$

functions:

$$u_1(0,x) = \begin{cases} x_2 + \theta x_1, & -\sqrt{3} \le \theta \le 0, \\ -x_2 + \theta x_3, & \text{otherwise}, \end{cases}$$

$$u_2(0,x) = \begin{cases} x_1 + x_2 \cos\lambda, & 0 \le \lambda \le \pi/2, \\ -x_1, & \pi/2 < \lambda < \pi, \\ x_1 + x_3 \sin\lambda, & \text{otherwise}, \end{cases} \tag{6.58}$$

$$u_3(0,x) = \begin{cases} \theta e^{x_3} + e^{\theta}, & -\sqrt{3} \le \theta \le 0, \\ -1 + \theta \log\theta, & \text{otherwise}, \end{cases}$$

$$u_4(0,x) = \begin{cases} x_2 \sinh(x_1) + \frac{1}{x_1^2+4}, & -1 \le x_1 \le 0, \\ -\cosh(x_1)/4, & \text{otherwise}. \end{cases} \tag{6.59}$$

Figure 6.10, on the right, presents the evolution of the parameter $\|\partial_t u\|_{\mathcal{M}}(t)$ for the first-order scheme. Figure 6.11, on the left, shows the evolution of this parameter for the second-order scheme. This parameter decreases over time, which shows that the time-variation diminishing property (6.4) is valid for all the functions arbitrarily chosen for the first and second order schemes.

The contraction property (6.5) is now validated for the first-order scheme using the five pairs of functions given in Appendix I-2. As shown in Figure 6.11, on the right, the ratios $E(v,w)$ defined by Equation (6.56) are decreasing, which confirms that the contraction property (6.5) is valid for all those pairs of functions.

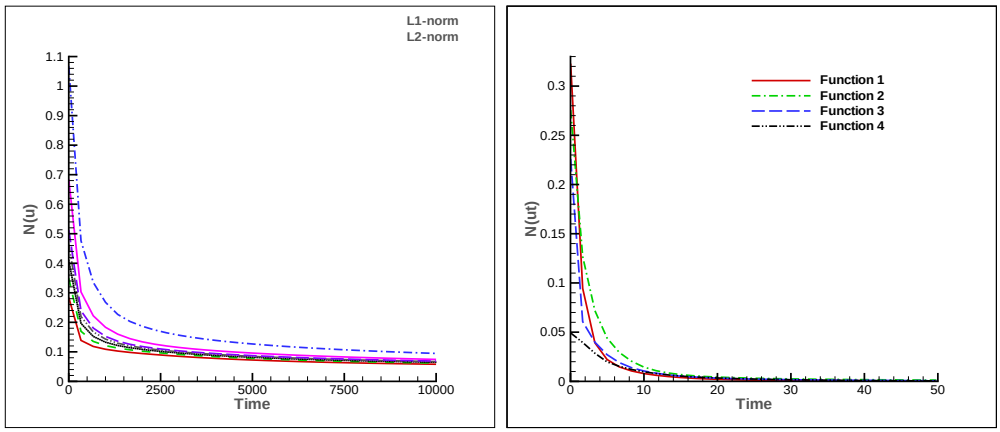

Figure 6.10 Stability property for Test 3-a (left) and property (6.4) for the first-order scheme until time $t = 50$ (right) with $k = 1$, $\Delta t = 0.01$, $\Delta\lambda = \pi/96$ and $\Delta\phi = \pi/96$.

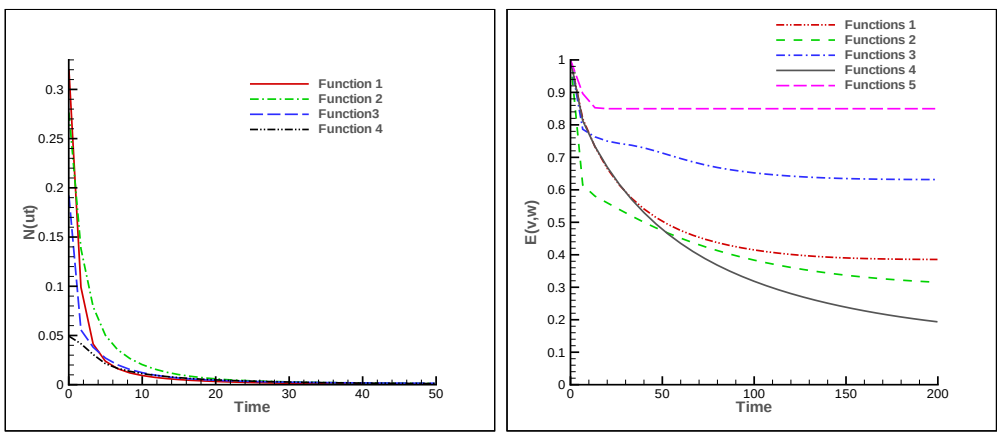

Figure 6.11 Property (6.4) for the second-order scheme until time $t = 50$ (left) and contraction property (6.5) for the first-order scheme (right) with $\Delta t = 0.01$, $\Delta\lambda = \pi/96$ and $\Delta\phi = \pi/96$.

6.7.4 Fourth test case with fully coupled flux vector fields

In this section, we consider a generic flux defined on the basis of the parameterized scalar potential and composed of two different terms in order to ensure the generic behavior.

$$\begin{aligned} h(x,u) &= h_1(x)f_1(u) + h_2(x)f_2(u), \\ h_1(x) &= x_1, \qquad h_2(x) = x_2, \\ f_1(u) &= su^2/2, \;\; f_2(u) = \mu u^3/3, \quad s + \mu = 1. \end{aligned} \tag{6.60}$$

Setting the value of the parameter s allows us to observe several characters of solutions and to study the impact of each part of the scalar potential on the evolution of these solutions and their late-time asymptotic behaviors. We present a synthesis of the tests

performed. First we consider the initial condition u_2 given in Test 2-b defined by Equation (6.52). Test 2-b corresponds to the particular case $s = 1$, where the solution evolves to two constant values in two independent domains. If the parameter $s = 0.95$ is used with the same initial condition u_2, both terms of the potential flux have an impact on the solution which converges to one constant value in the entire sphere.

Figure 6.12 shows the convergence curves related to the evolution of the parameter $\|u_2(t) - \overline{u}_2\|_{L^1}$ for different values of the parameter s. For all values of the parameter s presented in Figure 6.12, the solution converges to one constant value in the entire sphere. The evolution of the solutions and their asymptotic convergence are highly influenced by the magnitude of the different components of the potential function and the initial condition.

The entropy stability property (6.3) is now verified for the generic flux corresponding to $s = 0.5$. Figure 6.12, on the right, shows the evolution of the parameter $\|u_2(t)\|_{L^p_\omega(M)}$ for $p = 1, 2, 3, 4, 5, 10$ and $p = \infty$. According to this figure, we conclude that the solution u_2 satisfies the entropy stability property for all L^p norms considered.

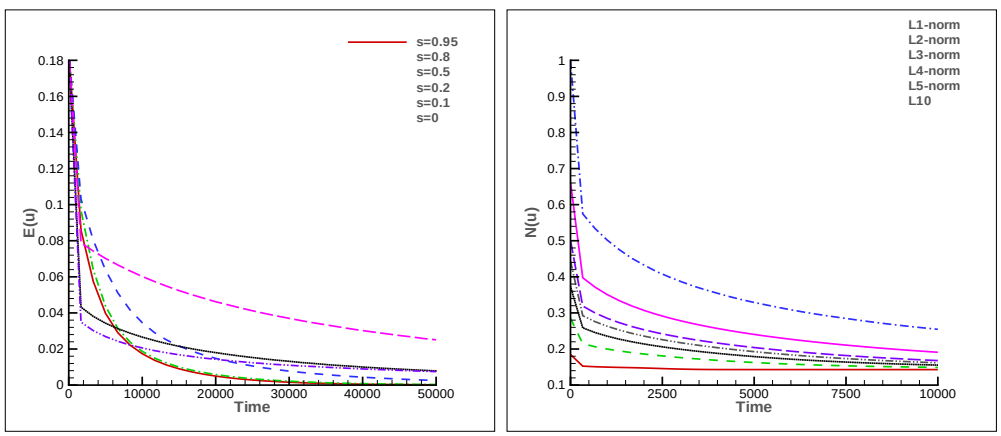

Figure 6.12 Convergence and stability for Test 4 with $\Delta\lambda = \pi/96$ and $\Delta\phi = \pi/96$. Left: Evolution of the parameter $E(u) = \|u_2(t) - \overline{u}_2\|_{L^1}$ for generic flux with different values of the parameter s with $\Delta t = 0.05$. Right: Entropy stability property (6.3) for the generic flux with $\Delta t = 0.01$.

For the generic flux with $s = 0.5$, the time-variation diminishing property (6.4) is verified for the second-order scheme using the five initial conditions u_0, u_1, u_2, u_3 and u_4, defined previously in the second test. Figure 6.13, on the left, presents the evolution of the parameter $\|\partial_t u\|_{\mathcal{M}}(t)$ and confirms that this parameter decreases with time, which shows that the time-variation diminishing property holds for those functions. The five pairs of functions defined in Appendix I-1 are used to check the contraction property (6.5) for the first-order scheme using the generic flux with $s = 0.5$. Figure 6.13 shows that the ratio parameter $E(v, w)$ is decreasing for the five pairs of functions which confirms that the contraction property holds for these pairs of functions.

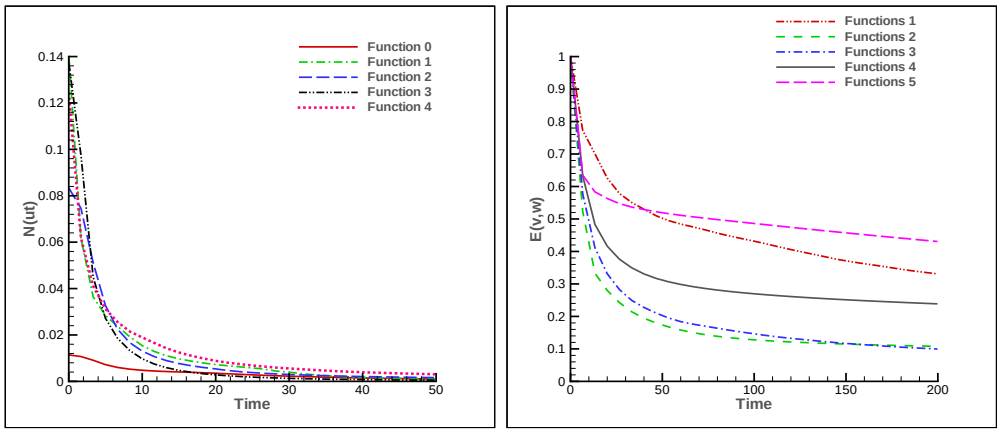

Figure 6.13 Time-variation diminishing property (6.4) (left) and contraction property (6.5) for the first-order scheme (right) for the generic flux with $\Delta t = 0.01$, $\Delta\lambda = \pi/96$ and $\Delta\phi = \pi/96$.

6.7.5 Fifth test case: revisiting the asymptotic convergence property

The aim of this section is to complete the analysis of the behavior of the solutions of Equation (6.15) for a nonlinear foliated flux. We present some examples in order to cover all cases of asymptotic convergence of the solutions for nonlinear foliated fluxes.

In the cases already seen in Test 2 and Test 3 for the asymptotic convergence of the solutions for nonlinear foliated flux, we observe numerically that the solution converges to constant values on independent domains on the sphere. This is a particular case of what is generally observed if the numerical scheme is more accurate and capable of simulating the solutions for a large simulation time. Then, in order to cover the other behaviors, we give more attention to the accuracy of the used scheme. To achieve this goal we consider the same second-order geometry-compatible finite volume scheme by considering a computational cell in which the sides are a part of the level sets and their equipotential curves.

We consider the nonlinear foliated flux defined on the basis of the scalar potential function $h(x,u) = -x_3u^2/2$. For this flux the level sets are the curves defined by $\phi = \phi_c$, where ϕ_c is a real constant in $[-\pi/2, \pi/2]$. The obtained level sets coincide with the grid used which ensures a very good accuracy of the scheme. In the following, we present the results for two cases with different forms of the functions considered as initial conditions. The first initial condition $u_1(0,x)$ is defined as follows:

$$u_1(0,x) = \begin{cases} x_1 + x_2 \sin\lambda, & 0 \le \lambda \le \pi, \\ -x_1 + x_2 \sin\lambda, & \text{otherwise}, \end{cases} \tag{6.61}$$

and the second initial condition $u_2(0,x)$ is defined by:

$$u_2(0,x) = \begin{cases} x_2 + \sinh(x_1)\cos\lambda, & 0 \le \lambda \le \pi/2, \\ -x_2, & \pi/2 < \lambda < \pi, \\ x_2 + \cosh(x_3)\sin\lambda, & \text{otherwise.} \end{cases} \tag{6.62}$$

Figure 6.14 shows the two-dimensional view of the first initial condition u_1 and the corresponding solution for a long simulation time. The two-dimensional view of the second initial condition u_2 and the corresponding solution for a long simulation time are shown in Figure 6.15. According to the numerical results shown in those figures for the two cases studied, the solution remains unchanged after a certain time. The solution converges to a nontrivial stationary solution which is constant on each level set. From the numerical tests one can conclude that, given any nonlinear foliated flux and any arbitrary function taken as an initial condition, the solution of Equation (6.15) converges to a steady state solution which is in general nontrivial and constant on the level sets. Each constant convergence value is the average value on the corresponding level set of the function taken as an initial condition.

We conclude this section by noting our important result which has a very significant positive effect for the performance of the numerical schemes in the case of nonlinear foliated fluxes. For this type of flux we recommend the use of a computational cell which is compatible with the level sets of the flux. More precisely, we can use a subset of the level sets and their orthogonal equipotentials of the nonlinear foliated flux as the constructing lines of computational cells. Using a choice of this kind, a very good conservation of nontrivial stationary solutions is guaranteed which enormously improves the accuracy of the schemes when more general solutions are considered.

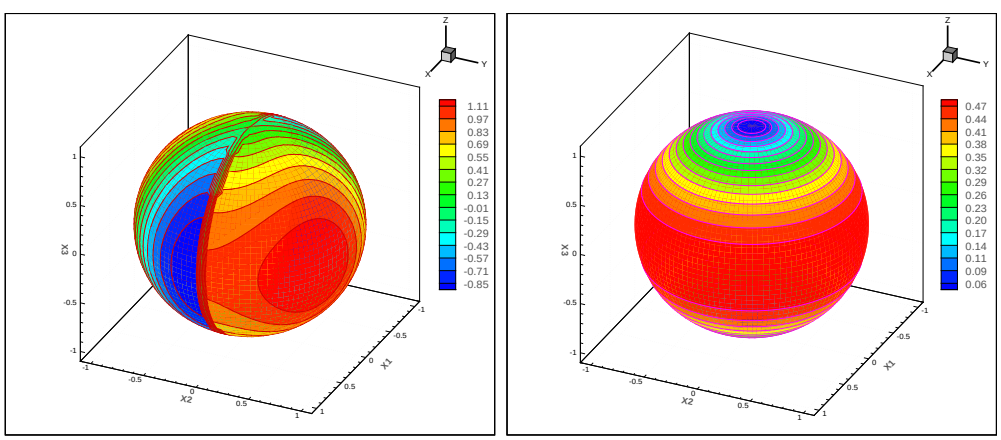

Figure 6.14 Initial condition u_1 (left) and solution u_1 at time $t = 50000$ (right) with $\Delta t = 0.05$, $\Delta\lambda = \pi/96$ and $\Delta\phi = \pi/96$.

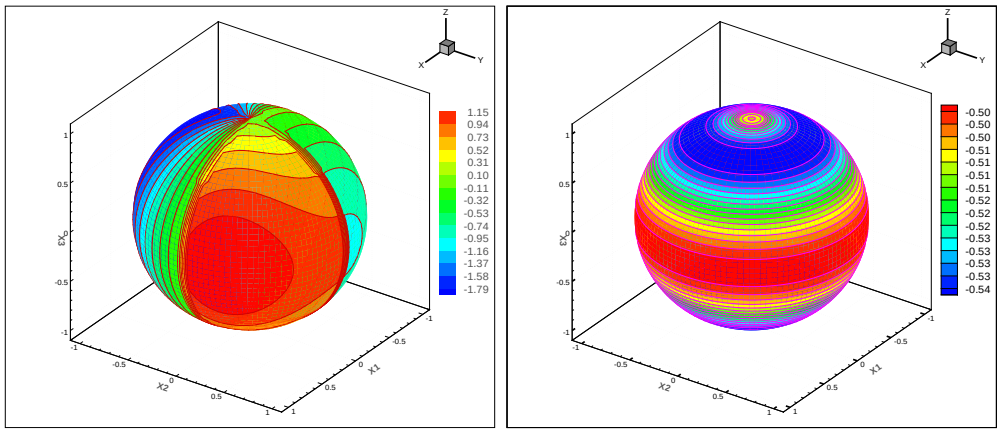

Figure 6.15 Initial condition u_2 (left) and solution u_2 at time $t = 50000$ (right) with $\Delta t = 0.05$, $\Delta\lambda = \pi/96$ and $\Delta\phi = \pi/96$.

6.8 Concluding remarks

In this paper, we have analyzed a class of nonlinear hyperbolic conservation laws posed on the sphere. We propose a second-order scheme using a geometry-compatible finite volume based on a generalized Riemann solver and a total variation diminishing Runge-Kutta method with an operator splitting approach for time integration. In the proposed piecewise linear reconstruction, we used the values at the centers of cells and the values of the solutions of the Riemann problem at the interfaces of cells which are obtained using the second-order approximations based on a generalized Riemann solver. The numerical solutions are largely improved using the proposed reconstruction and the method performs well in terms of accuracy and stability for discontinuous solutions with large amplitude and shocks compared to some well-known schemes. The numerical analysis shows the second order accuracy of the scheme and that the employed splitting approach has a less impact on the third order of the accuracy of the TVD Runge-Kutta method used for temporal integration.

The first- and second-order versions of the proposed geometry-compatible finite volume scheme were investigated, and we numerically established several important properties enjoyed by discontinuous solutions defined on a curved geometry, including the contraction, time-variation monotonicity, and the entropy monotonicity properties. Furthermore, we carefully investigated the late-time asymptotic behavior of solutions, by distinguishing various types of flux potential. The following main conclusions were established for the class of nonlinear hyperbolic conservation laws and the finite volume schemes under consideration:

- The entropy stability property is valid in all L^p norms with $p \in [1, +\infty)$, and the time-variation diminishing property is satisfied by the first- and second-order

schemes.

- The contraction property is satisfied by the first-order scheme but, as might be have been expected, this property is not valid for the second-order method.
- Two classes of flux were distinguished according to the structure of the flux potential. We introduced the notions of foliated flux and generic flux. The late-time asymptotic behavior of solutions was found to strongly depend on the flux (foliated or generic) as well as its (linearity or) nonlinearity. Specifically, when the flux is foliated and linear, the solutions are transported in time within the level sets of the potential. When the flux is foliated and is genuinely nonlinear, the solutions converge to their (constant) average within each level set.
- For generic flux, the solutions evolve with large variations which depend on the geometry and converge to constant values within certain "independent" domains defined on the sphere. The number of constant values depends on curves that "split" the sphere into possibly several independent domains.

We end with a main conclusion concerning the performance of the numerical schemes for the foliated flux, for which the level sets play an essential role in understanding the evolution of solutions; this is especially true for genuinely nonlinear foliated flux. For such flux, we strongly recommend the use of a suitable computational mesh, not only in terms of space constraints and the desired accuracy which are commonly used for the choice of the mesh, but also according to the physical phenomenon studied. The latter is reflected by the flux. More precisely, we recommend the use of a subset of the level sets of the nonlinear foliated flux as the construction lines of the computational grids. When this adjustment is respected, the steady state solutions for the nonlinear foliated flux can be captured with more accuracy and better results can be obtained for general solutions.

Appendix I

In this appendix we present the pairs of functions used in Section 6.7 in order to analyze the contraction property (6.5) for the first-order scheme

I-1: Pairs of functions used in the second test case to check the contraction property (6.5)

$$\begin{aligned} v_1(0,x) &= \begin{cases} x_2\sinh(x_1) - x_3\cosh(x_1), & x_1 \le 0, \\ x_3 + x_1, & \text{otherwise}, \end{cases} \\ w_1(0,x) &= \begin{cases} x_2x_1^3 + e^{x_1}, & x_1 \le 0, \\ -\cos x_1 + x_3^2x_1, & \text{otherwise}, \end{cases} \end{aligned} \tag{6.63}$$

$$v_2(0,x) = \begin{cases} x_2 \cos x_1, & x_1 \le 0, \\ -x_2, & \text{otherwise}, \end{cases}$$
$$w_2(0,x) = \begin{cases} x_3 \cos x_1, & x_1 \le 0, \\ -x_3 + x_3 x_1 \log(x_1), & \text{otherwise}, \end{cases} \tag{6.64}$$

$$v_3(0,x) = \begin{cases} x_3^2 \cos(\pi x_1) + x_2, & x_1 \le 1/2, \\ -x_2 + x_3(2x_1^2 - x_1), & \text{otherwise}, \end{cases}$$
$$w_3(0,x) = \begin{cases} x_2 x_3 \cos(\pi x_1) + x_3(x_1^4 - x_1), & x_1 \le 1, \\ x_2 x_3, & \text{otherwise}, \end{cases} \tag{6.65}$$

$$v_4(0,x) = \begin{cases} x_2 e^{x_1 + x_3}, & x_1 \le 0, \\ -x_2 e^{x_3}, & \text{otherwise}, \end{cases}$$
$$w_4(0,x) = \begin{cases} \frac{x_2 x_3}{1 - x_1}, & x_1 \le 0, \\ x_1^2 - x_2 x_3, & \text{otherwise}, \end{cases} \tag{6.66}$$

$$v_5(0,x) = \begin{cases} \cosh(x_1 + x_2), & x_1 \le 0, \\ -\cosh(x_2), & \text{otherwise}, \end{cases}$$
$$w_5(0,x) = \begin{cases} \cosh(x_3), & x_1 \le 0, \\ -\cosh(x_3), & \text{otherwise}. \end{cases} \tag{6.67}$$

I-2: Pairs of functions used in the third test case to check the contraction property (6.5)

$$v_1(0,x) = \begin{cases} x_2\theta + \cosh(\theta), & -\sqrt{3} \le \theta \le 0, \\ x_2\theta - \cosh(\theta), & \text{otherwise}, \end{cases}$$
$$w_1(0,x) = \begin{cases} x_2, & -\sqrt{3} \le \theta \le 0, \\ -x_2, & \text{otherwise}, \end{cases} \tag{6.68}$$

$$v_2(0,x) = \begin{cases} 1, & 0 \le \lambda \le \pi, \\ \cos\lambda, & \text{otherwise}, \end{cases}$$
$$w_2(0,x) = \begin{cases} \arcsin(x_2), & 0 \le \lambda \le \pi, \\ \cos\lambda \arcsin(x_2), & \text{otherwise}, \end{cases} \tag{6.69}$$

$$v_3(0,x) = \begin{cases} e^{\theta} x_1 x_2 x_3, & -\sqrt{3} \le \theta \le 0, \\ -x_1 x_2 x_3, & \text{otherwise}, \end{cases}$$
$$w_3(0,x) = \begin{cases} x_2 x_3, & -\sqrt{3} \le \theta \le 0, \\ -x_2 x_3 + e^{-1/\theta}, & \text{otherwise}, \end{cases} \tag{6.70}$$

$$v_4(0,x) = \begin{cases} \frac{\theta^2+1}{2-\theta}, & -\sqrt{3} \leq \theta \leq 0, \\ x_2\theta - 1/2, & \text{otherwise}, \end{cases}$$
$$w_4(0,x) = \begin{cases} \frac{e^\theta}{\theta-1}, & -\sqrt{3} \leq \theta \leq 0, \\ \frac{\cosh(\theta)}{1+\theta} + x_3\theta, & \text{otherwise}, \end{cases} \tag{6.71}$$

$$v_5(0,x) = \begin{cases} \theta, & \theta \leq 1, \\ \frac{x_1 \log(\theta)}{\theta} - \theta, & \text{otherwise}, \end{cases}$$
$$w_5(0,x) = \begin{cases} \theta - 2\theta^3, & \theta \leq 1, \\ \frac{1}{\theta} + x_2 \log(\theta), & \text{otherwise}. \end{cases} \tag{6.72}$$

CHAPITRE 7 Un schéma numérique efficace respectant la condition de compatibilité géométrique pour les lois de conservation sur la sphère

An efficient geometry-preserving scheme for conservation laws on the sphere[1]

Résumé

Dans ce chapitre, une nouvelle méthode des volumes finis partiellement centrée est développée pour les lois de conservation scalaires sur la sphère. Les schémas partiellement centrés sont simples vu qu'ils ne font pas appel à la résolution du problème de Riemann. Ces schémas sont appliqués avec succès aux systèmes hyperboliques sur une géométrie plane (Bollermann, 2013; Kurganov et Levy, 2002; Kurganov et Petrova, 2005, 2007). Aucune méthode partiellement centrée n'est développée pour ces systèmes dans le cas des surfaces courbes.

Dans notre démarche, on considère les équations de Burgers pour le développement et la validation de la nouvelle méthode proposée. La méthode développée dans le chapitre 6, pour le même système objet du présent chapitre, utilise un «splitting» directionnel afin de simplifier la résolution du problème de Riemann. L'approche du «splitting» directionnel et la résolution du problème de Riemann présentent un coût important en temps de calcul. L'objectif est de développer un schéma de volumes finis sans solveur de Riemann, non diffusif et qui ne fait pas recours à l'approche du «splitting» directionnel. L'opérateur de divergence du système est discrétisé sous une forme qui satisfait la condition de compatibilité géométrique. Pour construire la forme semi-discrète du nouveau schéma numérique, les étapes de reconstruction, d'évolution et de projection ont été suivies. Dans la formulation du nouveau schéma, une hypothèse a été adoptée dans laquelle on suppose que les dérivées spatiales de la fonction du système sont bornées indépendamment du pas de temps considéré. Cette hypothèse est utilisée pour construire une forme simplifiée du schéma numérique sans faire appel à la résolution du problème de Riemann au niveau des interfaces des cellules de calcul. Les tests numériques montrent que cette hypothèse est plus adaptée pour le cas des solutions discontinues avec des amplitudes et chocs moyens. Les dimensions géométriques de la sphère sont considérées de manière analytique et la forme semi-discrète du schéma numérique respecte la propriété de compatibilité géométrique. Les valeurs extrêmes des vitesses locales de propagation des ondes au niveau des interfaces de chaque cellule de calcul sont utilisées dans la formulation de la méthode proposée.

Les principaux avantages du nouveau schéma numérique sont sa simplicité puisque

[1]Cet article est réalisé en collaboration avec P.G. LeFloch, et soumis pour publication sous la forme: A. Beljadid, P.G. LeFloch, 2015, A central-upwind geometry-preserving method for hyperbolic conservation laws on the sphere. Journal of Computational Physics (Elsevier)

la résolution du problème de Riemann est évitée, et le lien fort entre sa forme semi-discrète et l'équation du système. Une reconstruction non-oscillatoire est proposée dans laquelle les composantes du gradient sont calculées en utilisant la fonction Minmod. La méthode TVD Runge-Kutta d'ordre 3 est utilisée pour l'intégration temporelle. Les flux feuilletés non linéaires qui sont introduits au chapitre 6 sont utilisés pour construire des solutions stationnaires non triviales du système. Ces solutions sont utilisées dans les tests numériques pour valider les performances de cette méthode en termes d'efficacité et de précision. Les résultats confirment la stabilité de la méthode proposée et montrent sa capacité à préserver les solutions stationnaires non triviales avec une bonne précision pour les lois de conservation hyperboliques non linéaires sur la sphère.

On note que la précision de la méthode proposée peut être améliorée davantage en utilisant la quadrature de Gauss pour l'intégration spatiale. Cette extension n'a aucun impact sur la condition de compatibilité géométrique du schéma numérique.

La méthode des volumes finis proposée est moins chère (en temps de calcul) en comparaison avec des méthodes décentrées qui utilisent le «splitting» directionnel et la résolution du problème de Riemann aux interfaces des cellules de calcul. Le schéma partiellement centré développé pour les lois de conservation sur la sphère peut être étendu au cas du système hyperbolique multidimensionnel et du système de Saint-Venant sur la sphère.

7.1 Introduction

The solutions of hyperbolic partial differential equations may develop discontinuities in finite times even for smooth initial conditions. The methods used for this case are called shock capturing schemes. Upwind and central schemes have been used to numerically solve these equations. Generally, it can be stated that the difference between these schemes is that upwind methods use characteristic information while central methods don't. The use of characteristic information in upwind schemes can improve the results but renders these schemes, in some cases, computationally expensive. The central schemes are widely used (e.g. Russo, 2002) after the pioneering work of Nessyahu and Tadmor (1990), where a second order finite volume central method on a staggered grid in space-time was proposed. This method offers a high resolution with the simplicity of the Riemann-solver free approach. Following Kurganov and Tadmor (2000a), this scheme suffers from excessive numerical viscosity when a small time step is considered.

In order to improve the performance of central schemes, some characteristic information can be used. Kurganov et al. (2001) proposed the central-upwind schemes which are based on information obtained from the local speeds of the wave propagation. The central-upwind schemes can be considered as the generalization of the central schemes developed by Kurganov and Tadmor (2000a,b), Kurganov and Levy (2000),

and Kurganov and Petrova (2001). The central-upwind schemes are simple since there are no Riemann solvers, and they have proven their effectiveness in multiple studies as shown in Kurganov and Petrova (2006, 2007, 2008, 2009); Kurganov et al. (2007); Kurganov and Levy (2002). Kurganov and Petrova (2005) extended the central-upwind scheme to triangular grids for solving two-dimensional Cartesian systems of conservation laws.

Several studies have been recently developed for hyperbolic conservation laws posed on curved manifolds. The solutions of conservation laws including the systems on manifolds and on spacetimes were studied in Rossmanith et al. (2004); Morton and Sonar (2007) and by LeFloch and co-authors (e.g. Amorim et al., 2005, 2008; Ben-Artzi and LeFloch, 2007; Ben-Artzi et al., 2009; LeFloch, 2011; LeFloch and Okutmustur, 2008). More general studies for hyperbolic conservation laws for evolving surface are developed by Dziuk, Kroöner, and Müller, Giesselmann (2009), and Dziuk and Elliott (2007). Ben-Artzi and LeFloch (2007) established the well-posedness theory for conservation laws on manifolds.

The Burgers equations are considered as an hyperbolic system which is ideal and simple to develop and validate numerical schemes. These equations have been widely used to develop shock-capturing schemes. Beljadid, LeFloch and Mohammadian (2014b) used the Burgers equations and adopted the methodology proposed in Ben-Artzi et al. (2009) which uses the second–order approximations based on generalized Riemann problems. A scheme was proposed using a piecewise linear reconstruction based on the values of the solution at the center of the computational cells and the values of the Riemann solutions at the cell interfaces. A second-order approximations based on a generalized Riemann solver is employed and a total variation diminishing Runge-Kutta method (TVDRK3) with an operator splitting was used for the temporal integration.

The finite volume methods developed in Beljadid et al. (2014b) and Ben-Artzi et al. (2009) are strongly connected to the governing equation. The geometric dimensions are considered in an analytical way which leads to a discrete forms of the schemes that respect the geometric compatibility property. The splitting approach which is used in these schemes simplifies the resolution of the Riemann problem but it increases the computational cost. In this study, we will propose a scheme which is less expensive. This scheme is Riemann-problem-solver-free and its resolution does not use any splitting approach which is widely used in upwind schemes to simplify the resolution of the Riemann problem.

The Burgers equations will be used in the present study to develop and validate the new geometric-preserving method. We will propose a geometry-compatible central-upwind scheme for scalar nonlinear hyperbolic conservation laws posed on the sphere. This system has a simple appearance but it generates solutions that have different wave structures in the presence of the geometry effects and its solutions are effective

for testing the proposed method. Our goal is to develop and validate a Riemann-problem-solver-free finite volume method which respects the geometric compatibility (divergence free) condition in the discrete level. The proposed scheme should be efficient and accurate for discontinuous solutions and with negligible geometric effects on the solutions.

An outline of the paper is as follows: In Section 7.2, the governing equations related to this study are presented. Section 7.3 is devoted to the derivation of the semi-discrete form of the new geometry-compatible central-upwind scheme. In Section 7.4, the coordinate system and the non-oscillatory reconstruction used in the proposed method are described. In Section 7.5, we present the geometry-compatible flux vectors and some particular nontrivial steady state solutions of our system which will be used to validate the performance of the proposed scheme. In Section 7.6, we demonstrate the high resolution and robustness of the new scheme in a series of numerical experiments. Finally, some remarks are provided to conclude the study

7.2 Governing equations

We are interested in nonlinear hyperbolic equations posed on the sphere S^2 based on the flux vector $F := F(x, u)$, depending on the real parameter $u(t, x)$ and the spatial variable x. This flux is assumed to respect the following geometric compatibility condition: For any arbitrary constant value $\bar{u} \in \mathbb{R}$

$$\nabla \cdot (F(\cdot, \bar{u})) = 0, \tag{7.1}$$

and that the flux can be written in the form

$$F(x, u) = n(x) \wedge \Phi(x, u). \tag{7.2}$$

The function $\Phi(x, u)$ is a vector field in $\mathbb{R}^3$ which is restricted to S^2 and is defined by

$$\Phi(x, u) = \nabla h(x, u), \tag{7.3}$$

where $h \equiv h(x, u)$ is a smooth function depending on the space variable x and the state variable $u(t, x)$.

Using Claim 2.2 in Ben-Artzi et al. (2009), the conditions (7.2) and (7.3) for the flux vector are sufficient to ensure the validity of the geometric compatibility condition (7.1).

We will develop and validate a new geometry-preserving central-upwind scheme which approximate the solutions of the following conservation law

$$\partial_t u + \nabla \cdot F(x, u) = 0, \quad (x, t) \in S^2 \times \mathbb{R}_+ \tag{7.4}$$

For some data u_0 on the sphere, we consider the following initial condition for the unknown function $u := u(t, x)$

$$u(x, 0) = u_0(x), \qquad x \in S^2 \tag{7.5}$$

Equation (7.4) can be rewritten in the form

$$\partial_t u + \frac{1}{\sqrt{|g|}} \partial_j (\sqrt{|g|} F^j(x, u)) = 0, \tag{7.6}$$

or

$$\partial_t (\sqrt{|g|} u) + \partial_j (\sqrt{|g|} F^j(x, u)) = 0, \tag{7.7}$$

where in local coordinates $x = (x^j)$, the derivatives are denoted by $\partial_j := \frac{\partial}{\partial x^j}$, F^j are the components of the flux vector and g is the metric. The conservation law (7.4) becomes as follows

$$\partial_t (v) + \partial_j (\sqrt{|g|} F^j(x, v/\sqrt{|g|})) = 0, \tag{7.8}$$

where $v = u\sqrt{|g|}$. This form will be used in the derivation of the semi-discrete form of the proposed scheme.

7.3 Derivation of the proposed scheme

7.3.1 Notations

The derivation of the new central-upwind scheme will be described in detail for the three steps: reconstruction, evolution, and projection. We will develop and give a semi-discrete form of the proposed method for general computational grid used to discretize the sphere. We assume the discretization of the sphere $S^2 := \bigcup_{j=1}^{j=N} C_j$, where C_j are the computational cells with area $|C_j|$. We denote by m_j the number of cell sides of C_j and by $C_{j1}, C_{j2}, ...C_{jm_j}$ the neighboring computational cells that share with C_j the common sides $(\partial C_j)_1, (\partial C_j)_2, ...(\partial C_j)_{m_j}$, respectively. The length of each cell-interface $(\partial C_j)_k$ is denoted by l_{jk}. The discrete value of the state variable $u(t, x)$ inside the cell C_j is denoted by $\mathbf{u}_j^n$ at a point $G_j \in C_j$. The longitude and latitude coordinates of G_j are presented in Section 7.4.2 since they should be chosen according to the reconstruction used in the proposed scheme. Finally, we use the notations Δt and $t_n := n\Delta t$ for the time step and the time at step n, respectively. Note that the development of the first order in time is sufficient to have the exact semi-discrete form of the proposed scheme. The resulting ODE can be numerically solved using a higher-order SSP ODE solver as Runge-Kutta of multistep methods. In the numerical experiments, the third TVD Runge-Kutta method proposed by Shu and Osher (1988) is used.

7.3.2 Discretization of the divergence operator

In this section, we will present a general form of the discretization of the divergence operator for general computational grid on the sphere. The approximation of the flux divergence can be written using the divergence theorem as

$$\begin{aligned} [\nabla \cdot F(x,u)]^{approx} &= \frac{I_j}{|C_j|}, \\ I_j &= [\oint_{\partial C_j} F(x,u) \cdot \nu(x) ds]^{approx}, \end{aligned} \tag{7.9}$$

where $\nu(x)$ is the unit normal vector to the boundary ∂C_j of the computational cell C_j and ds is the arc length of ∂C_j.

The scalar potential function h is used to obtain the following approximation along each side of the computational cell C_j

Claim 7.3.1 *For a three-dimensional flux $\Phi(x,u)$ given by (7.3), where $h \equiv h(x,u)$ is a smooth function in the neighboring of the sphere S^2, the total approximate flux through the cell interface e is given by*

$$\oint_{e^1}^{e^2} F(x,u) \cdot \nu(x) ds = -(h(e^2, u_j) - h(e^1, u_j)), \tag{7.10}$$

where e^1 and e^2 are the initial and final endpoints of the side e using the sense of integration and u_j is the estimate value of the variable u along the side e.

Proof. The flux vector is written in the form

$$F(x,u) = n(x) \wedge \Phi(x,u)$$

Then we derive the approximation of the integral along each cell side of C_j

$$\begin{aligned} &\oint_{e^1}^{e^2} F(x,u) \cdot \nu(x) ds = \oint_{e^1}^{e^2} (n(x) \wedge \Phi(x,u)) \cdot \nu(x) ds \\ &= -\oint_{e^1}^{e^2} \Phi(x,u) \cdot (n(x) \wedge \nu(x)) ds = -\oint_{e^1}^{e^2} \nabla h(x,u) \cdot \tau(x) ds \\ &= -\oint_{e^1}^{e^2} \nabla_{\partial C_j} h(x,u) ds = -(h(e^2, u_j) - h(e^1, u_j)), \end{aligned} \tag{7.11}$$

where $\tau(x)$ is the unit vector tangent to the boundary ∂C_j. □

Remark 1. Using the discrete approximations based on Claim 7.3.1, if a constant

value of the state variable $u(t,x) = u_j = \bar{u}$ is considered we obtain

$$\begin{aligned}[\nabla \cdot F(x,u)]^{approx} &= \frac{1}{|C_j|}[\oint_{\partial C_j} F(x,u)\cdot \nu(x) ds]^{approx} \\ &= -\sum_{e\in\partial C_j} (h(e^2,\bar{u}) - h(e^1,\bar{u})) = 0,\end{aligned} \tag{7.12}$$

which confirms that the discrete approximation of the divergence operator respects the divergence free condition.

7.3.3 Reconstruction

In the following, we will present the reconstruction of the proposed scheme. The semi-discrete form for Equation (7.4) will be derived by using the approximation of the cell averages of the solution. At each time $t = t_n$ the computed solution is

$$\mathbf{u}_j^n \approx \frac{1}{|C_j|}\int_{C_j} u(x,t_n) dV_g, \tag{7.13}$$

where $dV_g = \sqrt{g} dx^1 dx^2$.

The discrete values $\mathbf{u}_j^n$ of the solution at time $t = t_n$ are used to construct a conservative piecewise polynomial function with possible discontinuities at the interfaces of the computational cells C_j

$$\tilde{u}^n(x) = \sum_j w_j^n(x)\chi_j(x), \tag{7.14}$$

where $w_j^n(x)$ is a polynomial in two variables, and χ_j is the characteristic function which is defined using the Kronecker symbol and for any point of spatial coordinate x inside the computational cell C_k we consider $\chi_j(x) := \delta_{jk}$.

The maximum of the directional local speeds of propagation of the waves inward and outward the *kth* interface of the computational cell C_j are denoted by a_{jk}^{in} and a_{jk}^{out}, respectively. When the solution evolves over a time step Δt, the discontinuities move inward and outward the *kth* interface of the computational cell C_j with maximum distances $a_{jk}^{in}\Delta t$ and $a_{jk}^{out}\Delta t$, respectively. These distances of propagation at each cell interface are used to delimit different areas in which the solution still smooth and the areas in which the solution may not be smooth when it evolves from the time level t_n to t_{n+1}.

We define the domain D_j as the part inside the cell C_j in which the solution still smooth, see Figure 7.1. Two other types of domains are defined, the first type includes the rectangular domains D_{jk}, $k = 1, 2...m_j$, along each side of C_j of width $(a_{jk}^{out}+a_{jk}^{in})\Delta t$ and length $l_{jk} + O(\Delta t)$ and the second type includes the domains denoted by E_{jk}, $k = 1, 2...m_j$, around the cell vertices of computational cells. These domains are

decomposed into two sub-domains $D_{jk} = D_{jk}^{+} \cup D_{jk}^{-}$ and $E_{jk} = E_{jk}^{+} \cup E_{jk}^{-}$, where the sub-domains with superscript "+" and "−" are the domains inside and outside of the cell C_j, respectively. For purely geometrical reasons, the areas of the three types of sub-domains are of orders $|D_j| = O(1)$, $|D_{jk}| = O(\Delta t)$ and $|E_{jk}| = O(\Delta t^2)$.

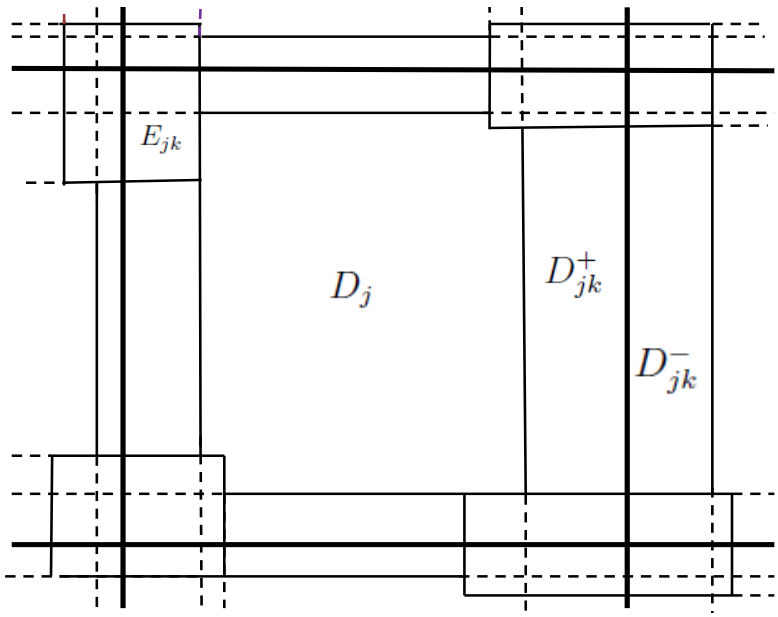

Figure 7.1 Schematic view of the decomposition of the control volume

We consider the projection of the flux vector F according to the orthogonal to the *kth* cell interface $(\partial C_j)_k$.

$$f_{jk} = N_{jk} \cdot F, \tag{7.15}$$

where N_{jk} is the unit normal vector to the cell interface $(\partial C_j)_k$.

The one-sided local speeds of propagation of the waves at the *kth* cell interface $(\partial C_j)_k$, inward and outward the computational cell C_j, are estimated by

$$\begin{aligned} a_{jk}^{out} &= \max\{\frac{\partial f_{jk}}{\partial v}(M_{jk}, u_j(M_{jk})), \frac{\partial f_{jk}}{\partial v}(M_{jk}, u_{jk}(M_{jk})), 0\}, \\ a_{jk}^{in} &= -\min\{\frac{\partial f_{jk}}{\partial v}(M_{jk}, u_j(M_{jk})), \frac{\partial f_{jk}}{\partial v}(M_{jk}, u_{jk}(M_{jk})), 0\}, \end{aligned} \tag{7.16}$$

where $u_j(M_{jk})$ and $u_{jk}(M_{jk})$ are the values of the state variable u at the midpoint M_{jk} of $(\partial C_j)_k$, which are obtained from the non-oscillatory reconstructions respectively for the computational cell C_j and its neighboring cell C_{jk}.

7.3.4 Evolution and projection

The computed cell averages $\bar{\mathbf{u}}_j^{n+1}$ of the numerical solution at time step t_{n+1} over the computational cells C_j are used to obtain the piecewise linear reconstruction $\tilde{w}^{n+1}$

$$\bar{\mathbf{u}}_j^{n+1} = \frac{1}{|C_j|}\int_{C_j} \tilde{w}^{n+1}(x)dV_g. \tag{7.17}$$

The function $\tilde{w}^{n+1}$ is smooth inside the domain D_j and its average over this domain will be denoted by $\bar{w}^{n+1}(D_j)$

$$\bar{w}^{n+1}(D_j) = \frac{1}{|D_j|}\int_{D_j} \tilde{w}^{n+1}(x)dV_g. \tag{7.18}$$

Note that it is possible to derive the fully discrete form of the proposed scheme but it is impractical to use and for simplicity, we will develop the semi-discrete form of the scheme. The ODE for approximating the cell averages of the solutions is derived by tending the time step Δt to zero. This eliminates some terms because of their orders and we keep the more consistent terms

$$\begin{aligned}
&\frac{d\bar{\mathbf{u}}_j}{dt}(t_n) = \lim_{\Delta t\to 0}\frac{\bar{\mathbf{u}}_j^{n+1} - \bar{\mathbf{u}}_j^{n}}{\Delta t}\\
&= \lim_{\Delta t\to 0}\frac{1}{\Delta t}[\frac{1}{|C_j|}\int_{D_j} \tilde{w}^{n+1}(x)dV_g + \frac{1}{|C_j|}\sum_{k=1}^{m_j}\int_{D_{jk}^+} \tilde{w}^{n+1}(x)dV_g\\
&+ \frac{1}{|C_j|}\sum_{k=1}^{m_j}\int_{E_{jk}^+} \tilde{w}^{n+1}(x)dV_g - \bar{\mathbf{u}}_j^{n}].
\end{aligned} \tag{7.19}$$

Since the areas of domains E_{jk} with $k = 1, 2, ..., m_j$ are of order Δt^2 we obtain

$$\int_{E_{jk}^+} \tilde{w}^{n+1}(x)dV_g = O(\Delta t^2). \tag{7.20}$$

This approximation allows us to deduce that the third term in the right-hand side of Equation (7.19) is of order Δt^2 and the result for the limit of this term vanishes for the ODE.

The second term in Equation (7.19), in which we use the rectangular domains D_{jk}^+, will be estimated by using the assumption that the spatial derivatives of $\tilde{w}^{n+1}$ are bounded independently of Δt. Under this assumption the following Claim gives an estimation of this term with an error of order Δt^2 for each $k \in [1, m_j]$.

Claim 7.3.2 *Consider the reconstruction given by* (7.14)*, its evolution* $\tilde{w}^{n+1}$ *over the global domain, and the definitions given in Section 7.3.3 for the domains* D_{jk} *and* D_{jk}^+*.*

If we assume that the spatial derivatives of $\tilde{w}^{n+1}$ are bounded independently of Δt, then

$$\int_{D^+_{jk}} \tilde{w}^{n+1}(x)dV_g = |D^+_{jk}|\bar{w}^{n+1}(D_{jk}) + O(\Delta t^2). \tag{7.21}$$

The proof of this Claim is provided in Appendix I.

Using Equation (7.21) in Claim 7.3.2 we obtain

$$\begin{aligned}\frac{1}{|C_j|}\sum_{k=1}^{m_j}\int_{D^+_{jk}} \tilde{w}^{n+1}(x)dV_g &= \frac{1}{|C_j|}\sum_{k=1}^{m_j}|D^+_{jk}|\bar{w}^{n+1}(D_{jk}) + O(\Delta t^2)\\ &= \frac{\Delta t}{|C_j|}\sum_{k=1}^{m_j} a^{in}_{jk}(l_{jk} + O(\Delta t))\bar{w}^{n+1}(D_{jk}) + O(\Delta t^2).\end{aligned} \tag{7.22}$$

Therefore Equation (7.19) can be written as

$$\frac{d\bar{\mathbf{u}}_j}{dt}(t_n) = \lim_{\Delta t\to 0}\frac{1}{\Delta t}[\frac{|D_j|}{|C_j|}\bar{w}^{n+1}(D_j) - \bar{\mathbf{u}}^n_j] + \sum_{k=1}^{m_j}\lim_{\Delta t\to 0}\frac{|D^+_{jk}|}{\Delta t|C_j|}\bar{w}^{n+1}(D_{jk}), \tag{7.23}$$

where

$$\bar{w}^{n+1}(D_{jk}) = \frac{1}{|D_{jk}|}\int_{D_{jk}} \tilde{w}^{n+1}(x)dV_g. \tag{7.24}$$

In order to derive the semi-discrete form of the proposed finite volume method from Equation (7.23), one needs to compute the average values $\bar{w}^{n+1}(D_{jk})$ and $\bar{w}^{n+1}(D_j)$. To compute $\bar{w}^{n+1}(D_{jk})$, Equation (7.4) is integrated over the space-time control volume $D_{jk} \times [t_n, t_{n+1}]$. After integration by parts and applying the divergence theorem to transform the surface integral of the divergence operator to the boundary integral, the following equations are obtained

$$\begin{aligned}\bar{w}^{n+1}(D_{jk}) &= \frac{1}{|D_{jk}|}[\int_{D^+_{jk}} w^n_j(x)dV_g + \int_{D^-_{jk}} w^n_{jk}(x)dV_g]\\ &- \frac{1}{|D_{jk}|}\int_{t_n}^{t_{n+1}}\int_{D_{jk}} \nabla\cdot F(x,u)dV_g,\end{aligned} \tag{7.25}$$

and

$$\begin{aligned}\int_{D_{jk}} \nabla\cdot F(x,u)dV_g = [\int_{\partial D_{jk}} F(x,u)\cdot\nu(x)ds]^{approx} = \sum_{i=1}^{i=4}\int_{(\partial D_{jk})_i} F(x,u)\cdot\nu(x)ds =\\ -[-h(e^2_{jk}, u_j(M_{jk})) + h(e^1_{jk}, u_j(M_{jk})) + h(e^2_{jk}, u_{jk}(M_{jk})) - h(e^1_{jk}, u_{jk}(M_{jk}))] + O(\Delta t),\end{aligned} \tag{7.26}$$

where $(\partial D_{jk})_i$, $i = 1, 2, 3, 4$, are the four edges of the rectangular domain D_{jk}, e^2_{jk} and e^1_{jk} are the initial and final endpoints of the cell interface $(\partial C_j)_k$, and as mentioned before w^n_j and w^n_{jk} are the piecewise polynomial reconstructions in the computational

cells C_j and C_{jk} respectively.

The term in the right-hand side of Equation (7.26) of order $O(\Delta t)$ corresponds to the global result of the integration along the two edges of the domain D_{jk} having the length $(a_{jk}^{int} + a_{jk}^{out})\Delta t$, and the rest of the integration due to the difference between the length of the domain D_{jk} and the length of the cell interface $(C_j)_k$.

In order to compute the spatial integrals in Equation (7.25), the Gaussian quadrature can be applied. In our case, the midpoint rule is used for simplicity

$$\int_{D_{jk}^+} w_{jk}^n dV_g + \int_{D_{jk}^-} w_{jk}^n dV_g \approx l_{jk}\Delta t[a_{jk}^{in} u_j(M_{jk}) + a_{jk}^{out} u_{jk}(M_{jk})]. \tag{7.27}$$

Equations (7.26) and (7.25) lead to

$$\begin{aligned}\lim_{\Delta t\to 0} \bar{w}^{n+1}(D_{jk}) &= \frac{l_{jk}}{a_{jk}^{in} + a_{jk}^{out}}[a_{jk}^{in} u_j(M_{jk}) + a_{jk}^{out} u_{jk}(M_{jk})] \\ &+ \frac{1}{a_{jk}^{in} + a_{jk}^{out}}[-h(e_{jk}^2, u_j(M_{jk})) + h(e_{jk}^1, u_j(M_{jk})) + h(e_{jk}^2, u_{jk}(M_{jk})) - h(e_{jk}^1, u_{jk}(M_{jk}))].\end{aligned} \tag{7.28}$$

Therefore

$$\begin{aligned}\lim_{\Delta t\to 0} \sum_{k=1}^{m_j} \frac{\left|D_{jk}^+\right|}{\Delta t\,|C_j|} \bar{w}^{n+1}(D_{jk}) &= \sum_{k=1}^{m_j} \frac{a_{jk}^{in} l_{jk}}{|C_j|\,(a_{jk}^{in} + a_{jk}^{out})}[a_{jk}^{in} u_j(M_{jk}) + a_{jk}^{out} u_{jk}(M_{jk})] \\ &+ \sum_{k=1}^{m_j} \frac{a_{jk}^{in}}{|C_j|\,(a_{jk}^{in} + a_{jk}^{out})}[-h(e_{jk}^2, u_j(M_{jk})) + h(e_{jk}^1, u_j(M_{jk})) + h(e_{jk}^2, u_{jk}(M_{jk})) - h(e_{jk}^1, u_{jk}(M_{jk}))].\end{aligned} \tag{7.29}$$

Now, the average value $\bar{w}^{n+1}(D_j)$ will be computed. Equation (7.4) is integrated over the space-time control volume $D_j \times [t_n, t_{n+1}]$ and after integration by parts and using the divergence theorem to transform the surface integral to boundary integral we obtain

$$\begin{aligned}\bar{w}^{n+1}(D_j) &= \frac{1}{|D_j|}\int_{D_j} w_j^n dV_g - \frac{1}{|D_j|}\int_{t_n}^{t_{n+1}}\int_{D_j} \nabla\cdot F(x,u) dV_g \\ &= \frac{1}{|D_j|}\int_{D_j} w_j^n dV_g - \frac{\Delta t}{|D|}\sum_{k=1}^{m_j}[-h(e_{jk}^2, u_j(M_{jk})) + h(e_{jk}^1, u_j(M_{jk}))].\end{aligned} \tag{7.30}$$

Using the previous equality we obtain

$$\begin{aligned}&\frac{1}{\Delta t}[\frac{|D_j|}{|C_j|}\bar{w}^{n+1}(D_j) - \bar{u}_j^n] = \\ &\frac{1}{\Delta t}\{\frac{1}{|C_j|}\int_{D_j} w_j^n dV_g - \frac{\Delta t}{|C_j|}\sum_{k=1}^{m_j}[-h(e_{jk}^2, u_j(M_{jk})) + h(e_{jk}^1, u_j(M_{jk}))] - \bar{u}_j^n\},\end{aligned} \tag{7.31}$$

which leads to

$$\lim_{\Delta t\to 0}\frac{1}{\Delta t}[\frac{|D_j|}{|C_j|}\bar{w}^{n+1}(D_j)-\bar{u}_j^n]= \\ -\frac{1}{|C_j|}\sum_{k=1}^{m_j}a_{jk}^{in}l_{jk}u_j(M_{jk})-\frac{1}{|C_j|}\sum_{k=1}^{m_j}[-h(e_{jk}^2,u_j(M_{jk}))+h(e_{jk}^1,u_j(M_{jk}))] \tag{7.32}$$

Equations (7.29) and (7.32) are used together to obtain the following semi-discrete form

$$\frac{d\bar{\mathbf{u}}_j}{dt}=-\frac{1}{|C_j|}\sum_{k=1}^{m_j}a_{jk}^{in}l_{jk}u_j(M_{jk})-\frac{1}{|C_j|}\sum_{k=1}^{m_j}[-h(e_{jk}^2,u_j(M_{jk}))+h(e_{jk}^1,u_j(M_{jk}))]+ \\ \sum_{k=1}^{m_j}\frac{a_{jk}^{in}l_{jk}}{|C_j|\,(a_{jk}^{in}+a_{jk}^{out})}[a_{jk}^{in}u_j(M_{jk})+a_{jk}^{out}u_{jk}(M_{jk})]+ \\ \sum_{k=1}^{m_j}\frac{a_{jk}^{in}}{|C_j|\,(a_{jk}^{in}+a_{jk}^{out})}[-h(e_{jk}^2,u_j(M_{jk}))+h(e_{jk}^1,u_j(M_{jk}))+h(e_{jk}^2,u_{jk}(M_{jk}))-h(e_{jk}^1,u_{jk}(M_{jk})). \tag{7.33}$$

This equation can be rewritten in the following form

$$\frac{d\bar{\mathbf{u}}_j}{dt}=\frac{1}{|C_j|}\sum_{k=1}^{m_j}\frac{a_{jk}^{in}a_{jk}^{out}l_{jk}}{a_{jk}^{in}+a_{jk}^{out}}(u_{jk}(M_{jk})-u_j(M_{jk}))+\frac{a_{jk}^{in}a_{jk}^{out}}{|C_j|\,(a_{jk}^{in}+a_{jk}^{out})}\{a_{jk}^{in}[h(e_{jk}^2,u_j(M_{jk})) \\ -h(e_{jk}^1,u_j(M_{jk}))]+a_{jk}^{out}[h(e_{jk}^2,u_{jk}(M_{jk}))-h(e_{jk}^1,u_{jk}(M_{jk}))]\}, \tag{7.34}$$

which can be rewritten in the form

$$\frac{d\bar{\mathbf{u}}_j}{dt}=-\frac{1}{|C_j|}\sum_{k=1}^{m_j}\frac{a_{jk}^{in}\mathbf{H}(u_{jk}(M_{jk}))+a_{jk}^{out}\mathbf{H}(u_j(M_{jk}))}{a_{jk}^{in}+a_{jk}^{out}}+\frac{1}{|C_j|}\sum_{k=1}^{m_j}\frac{a_{jk}^{in}a_{jk}^{out}l_{jk}}{a_{jk}^{in}+a_{jk}^{out}}[u_{jk}(M_{jk})-u_j(M_{jk})], \tag{7.35}$$

where $\mathbf{H}(u_j(M_{jk}))$ and $\mathbf{H}(u_{jk}(M_{jk}))$ are given by

$$\mathbf{H}(u_j(M_{jk}))=-[h(e_{jk}^2,u_j(M_{jk}))-h(e_{jk}^1,u_j(M_{jk}))] \\ \mathbf{H}(u_{jk}(M_{jk}))=-[h(e_{jk}^2,u_{jk}(M_{jk}))-h(e_{jk}^1,u_{jk}(M_{jk}))]. \tag{7.36}$$

The function $\mathbf{H}$ is defined in the form 7.36 in order to be consistent with the total approximate flux through the cell interface as presented by Equation 7.10 in Claim 7.3.1.

Remark 2. If the value of $a_{jk}^{\text{in}}+a_{jk}^{\text{out}}$ in Equation (7.35) is zero or very close to zero

(smaller than 10^{-8} in our numerical experiments), we avoid division by zero or by a very small number using the following approximations

$$\frac{a_{jk}^{in}\mathbf{H}(u_{jk}(M_{jk})) + a_{jk}^{out}\mathbf{H}(u_j(M_{jk}))}{a_{jk}^{in} + a_{jk}^{out}} \approx \frac{1}{2}[\sum_{k=1}^{m_j}\mathbf{H}(u_j(M_{jk})) + \sum_{k=1}^{m_j}\mathbf{H}(u_{jk}(M_{jk}))],$$
$$\frac{a_{jk}^{in}a_{jk}^{out}}{|C_j|\,(a_{jk}^{in} + a_{jk}^{out})}\sum_{k=1}^{m_j} l_{jk}[u_{jk}(M_{jk}) - u_j(M_{jk})] \approx 0.$$

This approximation is obtained using similar extreme distances of the propagation of the discontinuity at the cell interface inward and outward the computational cell to define the domains D_j, D_{jk} and E_{jk}

The semi-discretization (7.35) and (7.36) is a system of ODEs, which has to be integrated in time using an accurate and stable temporal scheme. In our numerical examples reported in Section 7.6, we used the third-order total variation diminishing Runge-Kutta method (see, Appendix II).

7.3.5 Geometry-compatible condition

In the semi-discrete form of the proposed scheme (7.35) and (7.36), if we consider a constant value of the function $u \equiv \bar{u}$, the second term in the right hand side of Equation (7.35) vanishes. For this constant function we obtain for each interface cell k

$$u_j(M_{jk}) = u_{jk}(M_{jk}) = \bar{u},$$

and

$$\mathbf{H}(u_j(M_{jk})) = \mathbf{H}(u_{jk}(M_{jk}))$$

The first term in the right hand side of Equation (7.35) becomes

$$-\frac{1}{|C_j|}\sum_{k=1}^{m_j}\frac{a_{jk}^{in}\mathbf{H}(u_{jk}(M_{jk})) + a_{jk}^{out}\mathbf{H}(u_j(M_{jk}))}{a_{jk}^{in} + a_{jk}^{out}} = -\frac{1}{|C_j|}\sum_{k=1}^{m_j}\mathbf{H}(u_j(M_{jk}))$$

Since we have

$$\sum_{k=1}^{m_j}\mathbf{H}(u_j(M_{jk})) = \sum_{k=1}^{m_j}\mathbf{H}(u_{jk}(M_{jk})) = -\sum_{k=1}^{m_j}[h(e_{jk}^2, \bar{u}) - h(e_{jk}^1, \bar{u})] = 0,$$

we conclude that the first term in the right-hand side of Equation (7.35) will be canceled which confirms that the proposed scheme respects the geometry-compatibility condition.

Remark 3. In the proposed scheme, the midpoint rule was used to compute the spatial integrals. In order to improve the accuracy of the proposed scheme, the Gaussian quadrature can be used instead of the midpoint rule. The Gaussian quadrature will not have any impact on the geometry-compatibility condition of the proposed scheme.

7.4 The proposed scheme using the latitude-longitude grid on the sphere

The geometry-compatible scheme was developed in the previous section for scalar non-linear hyperbolic conservation laws using general grids on the sphere. However, in order to prevent oscillations an appropriate piecewise linear reconstruction should be proposed according to the computational grids used in the proposed method. In the following, we will present the computational grid and the non-oscillatory piecewise linear reconstruction used in our numerical experiments.

7.4.1 Computational grid on the sphere

The position of each point on the sphere can be represented by its longitude $\lambda \in [0, 2\pi]$ and its latitude $\phi \in [-\pi/2, \pi/2]$. The grid considered in our numerical experiments is shown in Figure 7.2. The coordinates are singular at the south and north poles, corresponding to $\phi = -\pi/2$ and $\phi = \pi/2$, respectively. The Cartesian coordinates are denoted by $x = (x_1, x_2, x_3)^T \in \mathbb{R}^3$ for a standard orthonormal basis vectors $\mathbf{i_1}, \mathbf{i_2}$, and $\mathbf{i_3}$.

The unit tangent vectors in the directions of longitude and latitude at each point of coordinates (λ, ϕ) are given as follows

$$\begin{aligned} i_\lambda &= -\sin\lambda \mathbf{i_1} + \cos\lambda \mathbf{i_2}, \\ i_\phi &= -\sin\phi\cos\lambda \mathbf{i_1} - \sin\phi\sin\lambda \mathbf{i_2} + \cos\phi \mathbf{i_3}. \end{aligned} \tag{7.37}$$

The unit normal vector to the sphere at the same point is given by

$$n(x) = \cos\phi\cos\lambda \mathbf{i_1} + \cos\phi\sin\lambda \mathbf{i_2} + \sin\phi \mathbf{i_3}. \tag{7.38}$$

In spherical coordinates, for any vector field F represented by $F := F_\lambda \mathbf{i}_\lambda + F_\phi \mathbf{i}_\phi$, the equation of conservation law (7.4), can be rewritten as

$$\partial_t u + \frac{1}{\cos\phi}\left(\frac{\partial}{\partial\phi}(F_\phi \cos\phi) + \frac{\partial F_\lambda}{\partial\lambda}\right) = 0. \tag{7.39}$$

The three general structures of the cells used as part of the discretization grid on the sphere are shown in Figure 7.3. When we go from the equator to the north or south poles, the cells are changed by a ratio of 2 at some special latitude circles to reduce

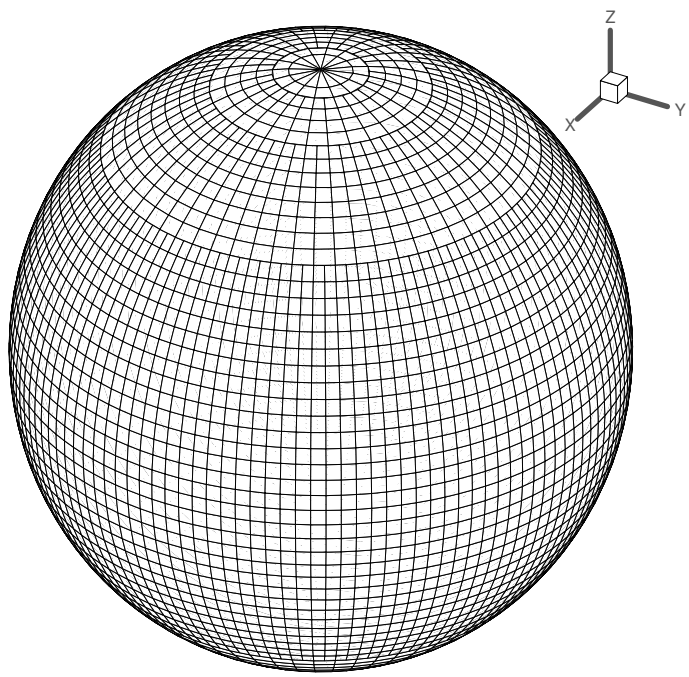

Figure 7.2 Type of grid used on the sphere

the number of cells in order to satisfy the stability condition and to ensure consistency of precision in the entire domain of the sphere.

The domain of each cell Ω is defined as

$$\Omega := \{(\lambda, \phi),\ \lambda_1 \leqslant \lambda \leqslant \lambda_2, \phi_1 \leqslant \phi \leqslant \phi_2\}\,. \tag{7.40}$$

Near the north or south poles, a triangular cell is considered which is a special case of the standard rectangular cell shown in Figure 7.3 with zero length for the side located on the pole.

7.4.2 Non-oscillatory piecewise linear reconstruction

In this section, we describe the piecewise linear reconstruction used in the proposed scheme. For simplicity, in the notations we will use the indices i and j for the cell centers along the longitude and latitude, respectively (see, Figure 7.3). At each time step t_n, data cell average values $u^n_{i,j}$ in each cell of center (λ_i, ϕ_j) are locally replaced by a piecewise linear function. The obtained reconstruction is as follows

$$u^n_{i,j}(\lambda, \phi) = u^n_{i,j} + (\lambda - \lambda_i)\mu^n_{i,j} + (\phi - \phi_j)\sigma^n_{i,j}, \tag{7.41}$$

where $\mu^n_{i,j}$ and $\sigma^n_{i,j}$ are the slopes in the directions of longitude and latitude, respectively. To prevent oscillations, we propose the following minmod-type reconstruction to obtain the slopes in the longitude and latitude directions

$$\begin{aligned}\mu^n_{i,j} &= minmod[\frac{u^n_{i+1,j}-u^n_{i,j}}{\lambda_{i+1}-\lambda_i}, \frac{u^n_{i+1,j}-u^n_{i-1,j}}{\lambda_{i+1}-\lambda_{i-1}}, \frac{u^n_{i,j}-u^n_{i-1,j}}{\lambda_i-\lambda_{i-1}})],\\ \sigma^n_{i,j} &= minmod[\frac{u^n_{i,j+1}-u^n_{i,j}}{\phi_{j+1}-\phi_j}, \frac{u^n_{i,j+1}-u^n_{i,j-1}}{\phi_{j+1}-\phi_{j-1}}, \frac{u^n_{i,j}-u^n_{i,j-1}}{\phi_j-\phi_{j-1}})],\end{aligned} \tag{7.42}$$

where the *minmod* function is

$$\begin{aligned}&minmod(\kappa_1,\kappa_2,\kappa_3)\\ &= \begin{cases}\kappa\min(|\kappa_1|,|\tau_2|,|\kappa_3|), & if\ \kappa = sign(\kappa_1)=sign(\kappa_2)=sign(\kappa_3),\\ 0, & \text{otherwise.}\end{cases}\end{aligned} \tag{7.43}$$

At each step we compute the average values of the state variable u in the computational cells. The same values correspond to the values of u at the cell centers of coordinates (λ_i, ϕ_j). The suitable points, inside the cells which respect these conditions for the linear reconstruction used in this study, should have the following spherical coordinates

$$\begin{aligned}\lambda_i &= \frac{\lambda_1+\lambda_2}{2},\\ \phi_j &= \frac{\phi_2\sin(\phi_2)-\phi_1\sin(\phi_1)+\cos\phi_2-\cos\phi_1}{\sin\phi_2-\sin\phi_1},\end{aligned} \tag{7.44}$$

where λ_1, λ_2, ϕ_1, and ϕ_2 correspond to the longitude and latitude coordinates of the cell nodes as shown in Figure 7.3.

7.5 Geometry-compatible flux vectors and particular solutions of interest

We have introduced in Beljadid et al. (2014b), two classes of flux vector fields for Equation (7.39). In this classification, the structure of the potential function $h(x,u)$ was used to distinguish between foliated and generic fluxes. The dependency of the potential function on the spatial variable x generates the propagation of the waves, while the dependency on the state variable u leads to the formation of shocks in the solutions. The foliated flux with linear behavior generates the spatially periodic solutions while the foliated flux with nonlinear behavior can generate nontrivial stationary solutions. From our analysis in Beljadid et al. (2014b), we have concluded that the new classification introduced and the character of linearity of the flux are sufficient to predict the late-time asymptotic behavior of the solutions. For a linear foliated flux, the solutions are simply transported along the level sets. The generic flux generates large variations in solutions, which converge to constant values within independent

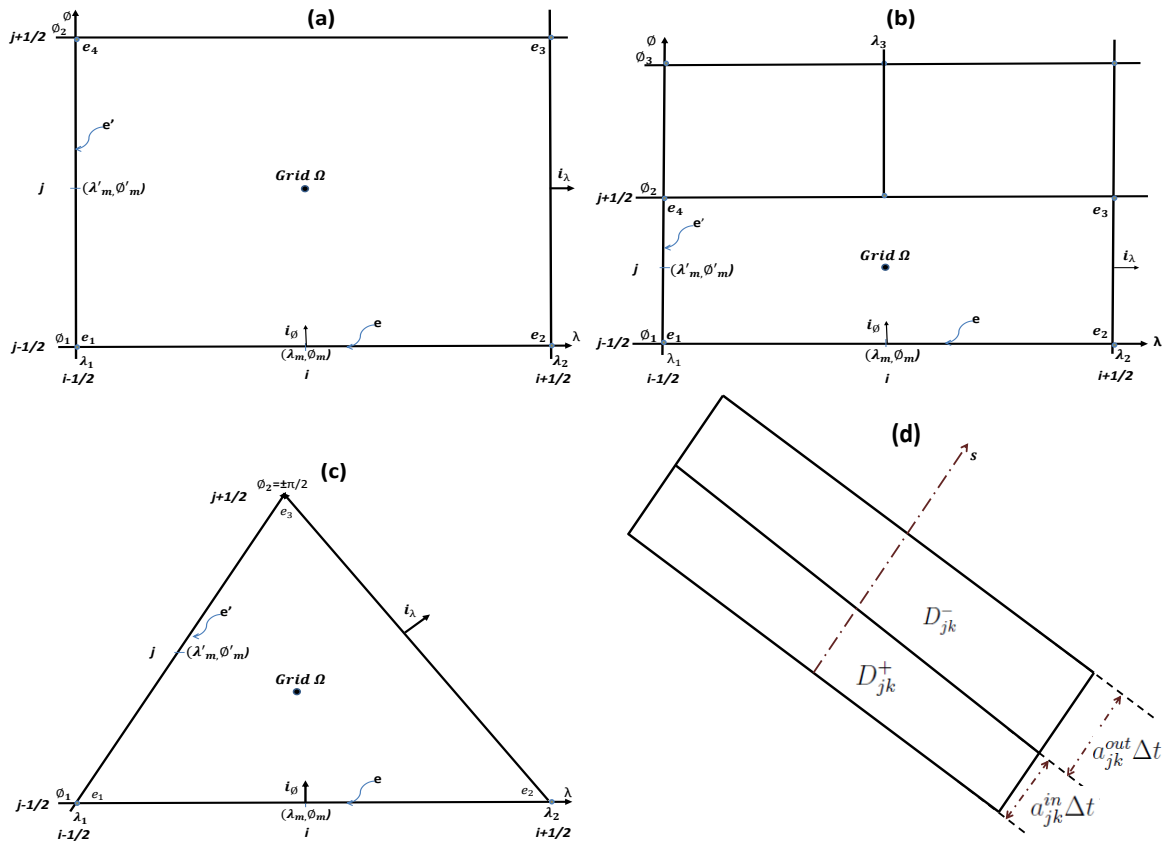

Figure 7.3 (a)-(b)-(c): Types of grids used on the sphere. (d): The domain $D_{jk} = D_{jk}^{+} \cup D_{jk}^{-}$.

domains on the sphere. For the nonlinear foliated flux, the solution converges to its constant average in each level set. For this flux, any steady state solution should be constant along each level set. This type of nontrivial stationary solutions are used in our numerical experiments to demonstrate the performance of the proposed method. We recommend Beljadid et al. (2014b) and Ben-Artzi et al. (2009) which provide more information and details about the construction of nontrivial steady state solutions of Equation (7.39).

The non-trivial steady state solutions which will be used in our numerical experiments are obtained using some particular nonlinear foliated fluxes. We will be particularly interested to use a linear splitting flux vector fields. We consider the nonlinear foliated flux (see Beljadid et al., 2014b; Ben-Artzi et al., 2009) based on the scalar potential function $h(x,u) = (x \cdot a)f(u)$, where $x \cdot a$ is the scalar product of the vector x and some constant vector $a = (a_1, a_2, a_3)^T \in \mathbb{R}^3$, and f is a function of the state variable u. The flux is obtained as

$$F(x,u) = f(u)n(x) \wedge a.$$

For this foliated flux vector, for any function $\tilde{u}$ which depends on one variable, the function defined as $u_0(x) = \tilde{u}(x \cdot a) = \tilde{u}(a_1x_1 + a_2x_2 + a_3x_3)$ is a steady state solution to the conservation law (7.39) associated to the flux vector $F(x,u)$. Arbitrary values of the vector a are used to construct nonlinear foliated fluxes and the corresponding nontrivial stationary solutions. First, we will consider the flux of the form $F(x,u) = f(u)n(x) \wedge i_1$. For this flux any function which depends on x_1 only, is a nontrivial steady state solution of Equation (7.39). Another form of foliated flux is obtained by using $a = i_1 + i_2 + i_3$ and the steady state solutions are of the form $u_0(x) = \tilde{u}(x_1 + x_2 + x_3)$.

Since we are more interested to study the discontinuous solutions, we will consider the flux vector using the function $f(u) = u^2/2$. For this foliated flux, the function defined as $u_0(x) = \chi(x \cdot a)\tilde{u}(x \cdot a)$ is a stationary solution of Equation (7.39), where $\chi(x \cdot a) = \pm 1$.

7.6 Numerical experiments

In this section, we demonstrate the performance of the proposed scheme on a variety of numerical examples. First, we consider the nonlinear foliated flux of the form $F(x,u) = f(u)n(x) \wedge i_1$ with $f(u) = u^2/2$. We take the following discontinuous steady state solution of Equation (7.39) as initial condition (Test 1).

$$u_2(x) = \begin{cases} \gamma x_1^3, & -1 \leq x_1 \leq 0.5, \\ -\gamma x_1^2/(2x_1+1), & 0.5 \leq x_1 \leq 1, \end{cases} \tag{7.45}$$

where γ is an arbitrary constant which controls the amplitude and shocks of the solution. This solution has a single closed curve of discontinuity on the sphere.

The numerical solution is computed using a grid with an equatorial longitude step $\Delta\lambda = \pi/96$ and a latitude step $\Delta\phi = \pi/96$, and a time step $\Delta t = 0.04$. Figure 7.4, on the left, shows the numerical solution with $\gamma = 0.1$ which is computed using the proposed scheme at a global time $t = 5$. The numerical solution remains nearly unchanged in time using the proposed scheme. The numerical solution error defined by using the L^2-norm is computed by summation over all grid cells on the sphere. For Test 1, the error is $u_{error} = 1.5 \times 10^{-4}$ at time $t = 5$, which is small compared to the full range of the numerical solution $u_{max} - u_{min} = 0.11237$.

Another test is performed using the steady state solution (7.45) as initial condition with $\gamma = 0.5$ (Test 2) and the same computational grid used in Test 1 and a time step $\Delta t = 0.04$. As is shown in Figure 7.4, on the right for Test 2, the solution remains nearly unchanged up to a global time $t = 5$. The error using the L^2-norm is $u_{error} = 2.7 \times 10^{-3}$, which is small compared to the full range of the solution $u_{max} - u_{min} = 0.56185$.

Now we consider a new test (Test 3) using the following steady state solution, with more discontinuities, which is defined in three domains separated by two closed curves on the sphere

$$u_2(x) = \begin{cases} \gamma x_1^4, & -1 \leq x_1 \leq -0.5, \\ 0.5\gamma x_1^3, & -0.5 < x_1 < 0.5, \\ -0.25\gamma x_1^2, & 0.5 \leq x_1 \leq 1. \end{cases} \tag{7.46}$$

The numerical solution is computed using a time step $\Delta t = 0.04$ and the same grid on the sphere used in the previous tests. As shown in Figure 7.5, on the left, the numerical solution which is obtained at time $t = 5$ using the proposed method based on the initial condition (7.46) with $\gamma = 0.1$ remains nearly unchanged. The error is $u_{error} = 9.6 \times 10^{-5}$, which is small compared to the full range $u_{max} - u_{min} = 0.12488$.

For $\gamma = 0.5$ (Test 4) we used the same computational grid and a time step $\Delta t = 0.04$. As is shown in Figure 7.5, on the right, again for this test the numerical solution at time $t = 5$ remains nearly unchanged. The error using the L^2-norm is $u_{error} = 1.9 \times 10^{-3}$, which is small compared to the full range of the solution $u_{max} - u_{min} = 0.62441$.

In the following, the performance of the proposed finite volume method will be analyzed using some particular steady state solutions in a spherical cap. The scalar potential function $h(x, u) = (x_1 + x_2 + x_3)f(u)$ is considered with $f(u) = u^2/2$. This leads to the nonlinear foliated flux $F(x, u) = f(u)n(x) \wedge (i_1 + i_2 + i_3)$. The function of the form $u(x) = \chi(\theta)\tilde{u}(\theta)$ is a steady state solution of Equation (7.39), where $\tilde{u}$ is an arbitrary real function depending on one variable and $\theta = x_1 + x_2 + x_3$.

In this numerical example (Test 5), the following discontinuous steady state solution

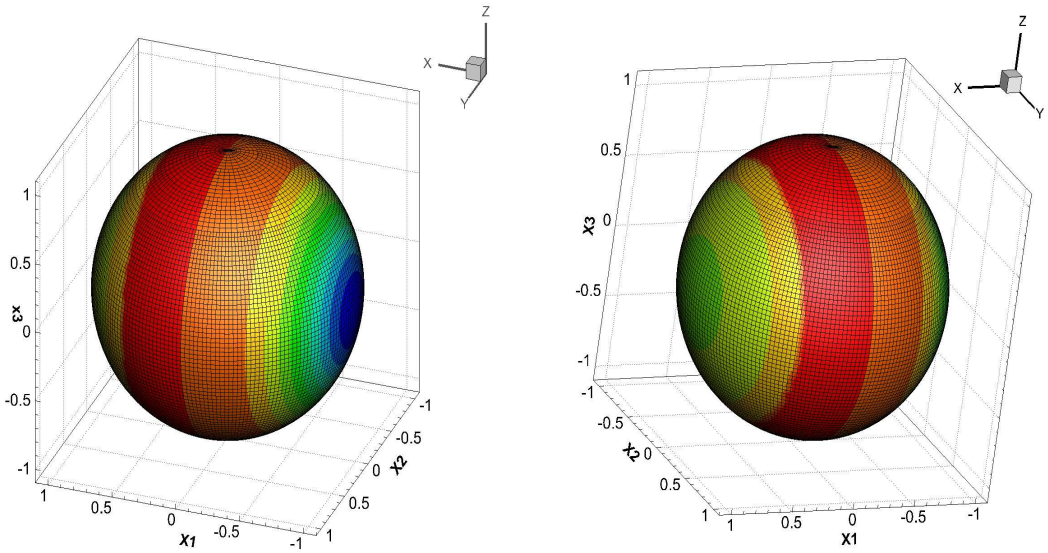

Figure 7.4 Solutions on the entire sphere at time $t = 5$ for Test 1 (left) and Test 2 (right)

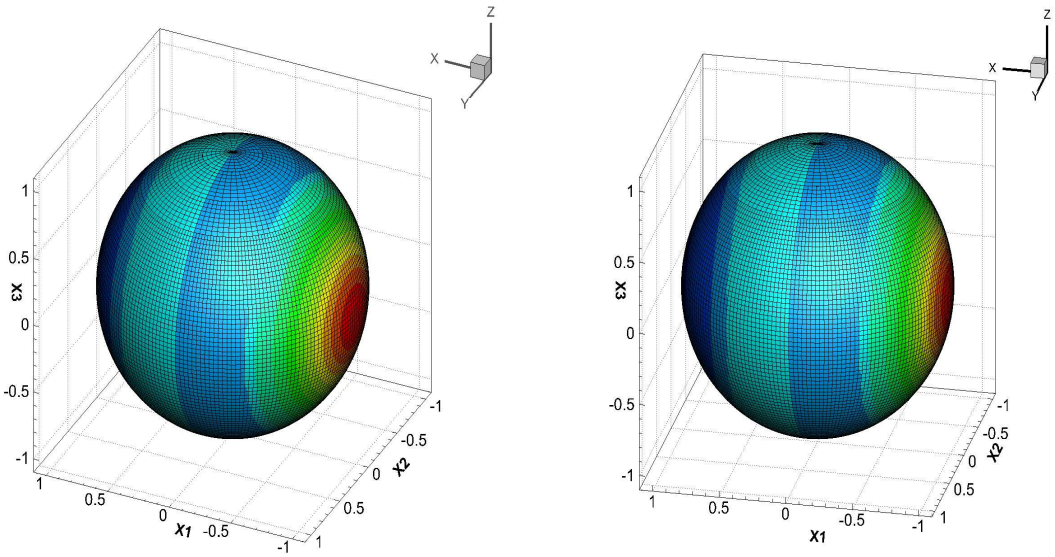

Figure 7.5 Solutions on the entire sphere at time $t = 5$ for Test 3 (left) and Test 4 (right)

is considered as initial condition

$$u(0,x) = \begin{cases} 0.1/(\theta+2), & 0 \leq \theta, \\ -0.1/(\theta+2), & \text{otherwise.} \end{cases} \tag{7.47}$$

The numerical solution is computed by using a grid with an equatorial longitude step $\Delta\lambda = \pi/96$ and a latitude step $\Delta\phi = \pi/96$, and a time step $\Delta t = 0.02$. Figure 7.6, on the left, shows the numerical solution which remains nearly unchanged in time after being subjected to integration up to a global time $t = 5$ by the proposed scheme. The numerical solution error defined by using the L^2-norm is $u_{error} = 1.3 \times 10^{-3}$, which is small compared to the full range $u_{max} - u_{min} = 0.1$.

The following numerical example (Test 6) is performed using the same nonlinear foliated flux considered in Test 5 and the steady state solution with more discontinuities defined by

$$u(0,x) = \begin{cases} 0.2\theta^3, & 0.5 \leq \theta, \\ 0.1\theta^2, & \theta \leq -0.5, \\ -0.025, & \text{otherwise.} \end{cases} \tag{7.48}$$

The numerical solution is computed using the same grid used in Test 5 and a time step $\Delta t = 0.02$. Figure 7.6, on the right, shows the numerical solution at time $t = 5$ which remains stationary with the error $u_{error} = 1.8 \times 10^{-3}$ which is negligible compared to the full range of the solution $u_{max} - u_{min} = 1.0638$.

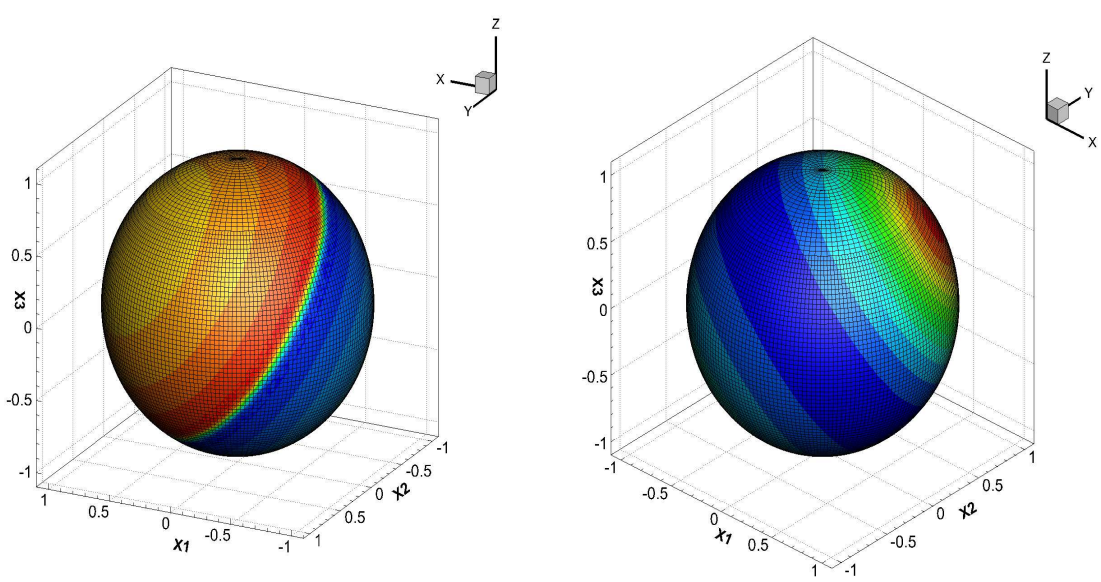

Figure 7.6 Solutions on the entire sphere at time $t = 5$ for Test 5 (left) and Test 6 (right)

7.7 Concluding remarks

In this paper, we have introduced a new efficient geometry-preserving scheme for conservation laws on the sphere. The main advantage of the proposed scheme is its simplicity since there is no Riemann solver. The proposed scheme is strongly connected to the analytical properties of the equation and the geometry of the sphere. In the proposed scheme, in order to improve the accuracy, the Gaussian quadrature can be used instead of the midpoint rule to compute the spatial integrals. The Gaussian quadrature will not have any impact on the geometry-compatibility condition of the proposed scheme. The semi-discrete form of the proposed method using the Gaussian quadrature will remain strongly connected to the analytic properties of the equation and the geometry of the sphere.

In the proposed method, a non-oscillatory reconstruction is used in which the gradient of each variable is computed using a minmod-function to ensure stability. Our numerical experiments demonstrate the ability of the proposed scheme to avoid oscillations. The performance of the second-order version of the designed scheme is tested using numerical examples. The results clearly demonstrated the proposed scheme's potential and ability to resolve the discontinuous solutions of conservation laws on the sphere.

Note that the formulation of the semi-discrete form of the proposed method is based on some approximations and assumptions. The proposed scheme is more suitable for discontinuous solutions with shocks of average amplitude. However, the proposed scheme has the advantage of simplicity compared to upwind schemes. As previously mentioned, the first advantage is that the proposed scheme is Riemann-problem-solver-free. The second advantage is related to the resolution, where the proposed scheme does not use any splitting approach which is widely used in upwind schemes to simplify the resolution of the Riemann problem. This again renders the proposed numerical scheme less expensive compared to upwind methods. The scheme developed for scalar nonlinear hyperbolic conservation laws could be extended to multidimensional hyperbolic conservation laws and shallow water models posed on the sphere.

Appendix I : Proof of Claim 7.3.2

Proof. It is obvious that for the case $\left|D_{jk}^{+}\right| = 0$ or $\left|D_{jk}^{-}\right| = 0$ equation (7.21) is valid. We assume that $\left|D_{jk}^{+}\right|\left|D_{jk}^{-}\right| \neq 0$ and we consider

$$R = \int_{D_{jk}^{+}} \tilde{w}^{n+1}(x) dV_g - |D_{jk}^{+}| \bar{w}^{n+1}(D_{jk}).$$

We have

$$
\begin{aligned}
R &= \int_{D_{jk}^+} \tilde{w}^{n+1}(x)dV_g - \frac{\left|D_{jk}^+\right|}{|D_{jk}|}(\int_{D_{jk}^+} \tilde{w}^{n+1}(x)dV_g + \int_{D_{jk}^-} \tilde{w}^{n+1}(x)dV_g), \\
&= \frac{|D_{jk}^+|}{|D_{jk}|}[\frac{|D_{jk}^-|}{|D_{jk}^+|}\int_{D_{jk}^+} \tilde{w}^{n+1}(x)dV_g - \int_{D_{jk}^-} \tilde{w}^{n+1}(x)dV_g], \\
&= \frac{|D_{jk}^+|}{|D_{jk}|}[\frac{a_{jk}^{out}}{a_{jk}^{in}}\int_{-a_{jk}^{in}\Delta t}^{0} \tilde{w}^{n+1}(s)\tilde{l}_{jk}ds - \int_0^{a_{jk}^{out}\Delta t} \tilde{w}^{n+1}(s)\tilde{l}_{jk}ds],
\end{aligned}
\tag{7.49}
$$

where $\tilde{l}_{jk}$ is the length of the domain D_{jk} and s is a variable according to the orthogonal outward axis to the kth cell interface.

After the change of variable in the first integral of the last equality in (7.49), one obtains

$$
R = \frac{|D_{jk}^+|}{|D_{jk}|}\tilde{l}_{jk}\int_0^{a_{jk}^{out}\Delta t}(\tilde{w}^{n+1}(-\frac{a_{jk}^{in}}{a_{jk}^{out}}s) - \tilde{w}^{n+1}(s))ds.
$$

Using the mean value theorem to the function $\tilde{w}^{n+1}$ we obtain

$$
R = -\frac{|D_{jk}^+|}{|D_{jk}|}\tilde{l}_{jk}\int_0^{a_{jk}^{out}\Delta t}\frac{a_{jk}^{in}+a_{jk}^{out}}{a_{jk}^{out}}s\frac{\partial\tilde{w}^{n+1}}{\partial s}(c_s)ds,
$$

where $c_s \in [min(s, -sa_{jk}^{in}/a_{jk}^{out}), max(s, -sa_{jk}^{in}/a_{jk}^{out})]$. We denote by M the upper bound value of the spatial derivative of the function $\tilde{w}^{n+1}$ over the domain D_{jk}. Therefore we obtain

$$
|R| \leq Ml\frac{|D_{jk}^+|}{|D_{jk}^-|}\int_0^{a_{jk}^{out}\Delta t} sds = \frac{Ml}{2}|D_{jk}^+||D_{jk}^-|.
$$

Since $\tilde{l}_{jk} = l_{jk} + O(\Delta t)$, and both the areas $|D_{jk}^+|$ and $|D_{jk}^-|$ are of order Δt we obtain $R = O(\Delta t^2)$. □

Appendix II : Time integration methods

Consider an ODE defined by

$$
\frac{d\mathbf{u}}{dt} = \mathbf{L}(\mathbf{u}),\ \mathbf{u}(t_0) = \mathbf{u}^0, \tag{7.50}
$$

where $\mathbf{L}$ is a spatial operator. The TVDRK3 method (Shu and Osher, 1988) is performed via three stages to solve Equation (7.50).

$$
\begin{aligned}
\mathbf{u}^{(1)} &= \mathbf{u}^n + \Delta t\mathbf{L}(\mathbf{u}^n) \\
\mathbf{u}^{(2)} &= \frac{3}{4}\mathbf{u}^n + \frac{1}{4}\mathbf{u}^{(1)} + \frac{1}{4}\Delta t\mathbf{L}(\mathbf{u}^{(1)}) \\
\mathbf{u}^{n+1} &= \frac{1}{3}\mathbf{u}^n + \frac{2}{3}\mathbf{u}^{(2)} + \frac{2}{3}\Delta t\mathbf{L}(\mathbf{u}^{(2)})
\end{aligned}
\tag{7.51}
$$

CHAPITRE 8 Conclusions et recommandations

Le projet de recherche a pour objectif le développement et l'évaluation de nouvelles méthodes numériques pour les écoulements peu profonds. De nouvelles techniques sont développées pour les discrétisations spatiales et temporelles des équations de Saint-Venant et des lois de conservation sur les surfaces courbes. L'objectif est de construire des schémas numériques stables, non diffusifs, et qui présentent des performances intéressantes en termes de précision et temps de calcul. Les résultats de la thèse peuvent être récapitulés comme suit :

- Deux nouvelles méthodes numériques sont proposées pour la modélisation des processus lents et rapides dans les écoulements océaniques et atmosphériques,
- Un nouveau schéma numérique efficace sans solveur de Riemann est proposé pour les écoulements peu profonds sur une topographie variable,
- Une nouvelle approche est proposée pour l'analyse de stabilité des méthodes numériques pour les écoulements en eaux peu profondes,
- Deux nouveaux schémas numériques sont développés pour les lois de conservation sur des surfaces courbes qui ont de larges potentiels d'être appliqués pour le cas des écoulements peu profonds sur la sphère.

Dans la première partie de la thèse, une méthode des volumes finis explicite sur un maillage non structuré est proposée pour les écoulements de grande échelle avec un terme source qui comprend l'effet de Coriolis. Le schéma proposé est efficace pour les écoulements à grande échelle en présence des ondes lentes (ondes de Rossby) et des ondes rapides (ondes de gravité). L'approche du «splitting» directionnel et la méthode d'Adams d'ordre quatre sans aucune itération sur le correcteur sont utilisées pour l'intégration temporelle. Ces méthodes sont suffisantes pour supprimer le bruit numérique des ondes courtes sans amortissement des ondes longues dans les écoulements à grande échelle. De bons résultats sont obtenus à la fois pour les ondes de gravité et les ondes de Rossby.

En présence des termes sources dans le système qui génèrent des ondes de grandes vitesses de propagation, l'utilisation des méthodes implicites est nécessaire. Ces méthodes sont aussi nécessaires dans le cas où le système génère de large gamme de fréquences des ondes. C'est le cas des écoulements atmosphériques où il est important d'utiliser les schémas semi-implicites afin d'utiliser un pas de temps optimal. Dans le cas des modèles de prévision numérique atmosphérique, les schémas implicites sont utilisés pour traiter les termes linéaires responsables de la propagation des ondes rapides et le traitement explicite est utilisé pour les termes non linéaires qui sont responsables

des ondes lentes. Dans l'objectif d'améliorer la stabilité de la méthode actuellement utilisée par Environnement Canada dans son modèle numérique pour les prévisions météorologiques, nous avons proposé une nouvelle classe de schémas semi-implicites semi-lagrangiens. Cette classe est introduite pour remédier aux problèmes d'instabilité liés au traitement du terme non linéaire pour le cas des solutions qui ont un caractère oscillatoire. Cette classe de schémas présente de bonnes qualités de stabilité, de précision et de convergence en comparaison avec les méthodes actuellement utilisées en prévision numérique atmosphérique.

Dans le projet de recherche, une autre méthode des volumes finis est développée pour le cas du système de Saint-Venant avec une topographie variable. La méthode proposée est efficace et elle ne fait pas appel au solveur de Riemann. Elle assure l'équilibre entre le terme de flux et le terme source dû à la topographie et elle préserve la positivité de la hauteur d'eau. La méthode est stable et présente de bonne qualité de convergence, en particulier pour le cas des faibles perturbations des états d'équilibre. La méthode proposée peut être appliquée pour le système de Saint-Venant avec une topographie discontinue ou hautement variable et sur des domaines complexes où il est nécessaire d'utiliser des maillages non structurés.

Dans la thèse, une nouvelle approche est introduite pour analyser la stabilité des méthodes des volumes finis appliquées aux écoulements peu profonds. Dans cette approche, on utilise la notion du pseudo spectre des matrices. Cette méthode est efficace pour analyser la stabilité en comparaison avec la stabilité asymptotique et la stabilité de Lax-Richtmyer. Cette approche est utile pour le choix du type de maillage, des emplacements adéquats des variables primitives du système, et de la technique de discrétisation la plus stable.

Les équations de Burgers sont considérées pour développer des méthodes numériques robustes pour les lois de conservation sur des surfaces courbes. Ces équations constituent un modèle simple mais très significatif pour valider les schémas numériques proposés. Deux méthodes numériques sont développées dans le cas des systèmes hyperboliques sur des surfaces courbes. Le premier schéma est proposé en utilisant la résolution du problème de Riemann généralisé au niveau des interfaces des cellules de calcul. Les principaux avantages de cette méthode des volumes finis sont:

- Les dimensions géométriques de la sphère sont considérées de manière analytique dans la forme semi-discrète du schéma numérique et on obtient une formulation qui respecte exactement la condition de compatibilité géométrique,

- La reconstruction linéaire proposée a permis d'améliorer nettement la qualité des résultats dans le cas des solutions discontinues,

- Le schéma de volumes finis proposé est d'ordre deux dans l'espace. La méthode du «splitting» directionnel et la méthode TVDRK3 constituent un schéma efficace

pour l'intégration temporelle. L'ordre trois de la méthode TVDRK3 est moins influencé par la méthode du «splitting» directionnel utilisée,

- Le schéma proposé est efficace dans le cas des solutions discontinues de grands chocs et amplitudes.

La méthode proposée est utilisée pour étudier le comportement asymptotique des solutions. Une classification des flux est proposée où nous avons introduit la notion de flux feuilletés et celle de flux génériques. Cette classification et la linéarité des flux constituent un concept très important et suffisant pour prédire le comportement asymptotique des solutions du système. Les résultats obtenus pour le cas des flux feuilletés non linéaires présentent un intérêt particulier pour construire des solutions stationnaires non triviales. Pour ces flux, les solutions qui sont constantes le long des lignes de niveau sont des solutions stationnaires non triviales du système. Ces solutions sont utilisées pour évaluer les performances de la méthode proposée.

Un nouveau schéma partiellement centré est proposé pour les lois de conservation scalaires hyperboliques sur la sphère. Ce schéma présente les avantages suivants:

- La caractéristique principale du schéma numérique est qu'il permet d'éviter la résolution du problème de Riemann aux interfaces des cellules de calcul,
- Aucun «splitting» directionnel n'est utilisé,
- La méthode est moins coûteuse en termes de temps de calcul en comparaison avec les méthodes qui utilisent le «splitting» directionnel pour simplifier la résolution du problème de Riemann,
- Le schéma proposé est efficace dans le cas des solutions discontinues d'amplitudes et chocs moyens,
- La précision de la méthode peut être nettement améliorée en utilisant la quadrature de Gauss pour l'intégration spatiale sans aucun impact sur la condition de compatibilité géométrique du schéma numérique.

Comme extensions des travaux du présent projet de recherche, on présente les recommandations suivantes :

- La classe de schémas semi-implicites semi-lagrangiens potentiellement applicables aux modèles atmosphériques dépend du paramètre de décentrement au niveau du correcteur en deuxième étape. On recommande une étude détaillée sur la base des cas réels relatifs aux simulations atmosphériques pour le choix du décentrement le plus adéquat de point de vue stabilité et précision,

- La classe de schémas semi-implicites semi-lagrangiens potentiellement applicables aux modèles atmosphériques est proposée en se basant sur la méthode BDF2 qui a été analysée par Dharmaraja (2007). On recommande d'effectuer des analyses similaires pour quelques familles de méthodes à deux étapes et d'étudier la possibilité de construire des méthodes semi-implicites semi-lagrangiennes à deux étapes applicables aux modèles atmosphériques,

- Le schéma numérique partiellement centré développé pour le système de Saint-Venant en deux dimensions peut être étendu au cas des écoulements peu profonds à trois dimensions,

- Le schéma de volumes finis basé sur la résolution du problème de Riemann généralisé pour les lois de conservation scalaires hyperboliques non linéaires sur la sphère peut être étendu au système de Saint-Venant sur la sphère,

- Le schéma partiellement centré développé pour les lois de conservation scalaires sur la sphère peut être étendu au système de Saint-Venant sur la sphère. Ce schéma numérique peut aussi être étendu au système hyperbolique multidimensionnel sur la sphère.

RÉFÉRENCES

P. Amorim, M. Ben-Artzi, et P. G. LeFloch, “Hyperbolic conservation laws on manifolds: Total variation estimates and finite volume method”, *Meth. Appl. Analysis*, vol. 12, pp. 291–324, 2005.

P. Amorim, P. G. LeFloch, et B. Okutmustur, “Finite volume schemes on Lorentzian manifolds”, *Comm. Math. Sc.*, vol. 6, pp. 1059–1086, 2008.

R. M. Beam et R. F. Warming, “The asymptotic spectra of banded toeplitz and quasi-toeplitz matrices”, *SIAM J. Scientific Computing*, vol. 14, pp. 971–1006, 1993.

A. Beljadid et P. G. LeFloch, “An efficient geometry-preserving scheme for conservation laws on the sphere”, Will be submitted in January 2015.

A. Beljadid, A. Mohammadian, et H. Qiblawey, “Numerical simulation of rotation dominated linear shallow water flows using finite volume methods and fourth order Adams scheme”, *Computers & Fluids*, vol. 62, pp. 64–70, 2012.

——, “An unstructured finite volume method for large-scale shallow flows using the fourth-order Adams scheme”, *Computers & Fluids*, vol. 88, pp. 579–589, 2013a.

——, “An accurate finite volume method using fourth order Adams scheme on triangular grids for the Saint-Venant System”, *4th Specialty Conference on Coastal, Estuary and Offshore Engineering, Montreal, Quebec*, no. May 29 to June 1, 2013b.

A. Beljadid, A. Mohammadian, M. Charron, et C. Girard, “Theoretical and numerical analysis of a class of semi-implicit semi-lagrangian schemes potentially applicable to atmospheric models”, *Monthly Weather Review - American Meteorological Society*, vol. 142, no. 12, pp. 4458–4476, 2014a.

A. Beljadid, P. G. LeFloch, et A. Mohammadian, “A geometry-preserving finite volume method for conservation laws on curved geometries”, *Advances in Computational Mathematics. Submitted*, 2014b.

A. Beljadid, A. Mohammadian, et A. Kurganov, “Well-balanced positivity preserving cell-vertex central-upwind scheme for shallow water flows”, *Applied Mathematical Modelling*, Submitted, 2014c.

M. Ben-Artzi et J. Falcovitz, *Generalized Riemann problems in computational fluid dynamics*. Cambridge University Press, London, 2003.

M. Ben-Artzi et P. G. LeFloch, “The well posedness theory for geometry compatible hyperbolic conservation laws on manifolds”, *Ann. Inst. H. Poincaré Anal. Non Linéaire*, vol. 24, pp. 989–1008, 2007.

M. Ben-Artzi, J. Falcovitz, et P. G. LeFloch, “Hyperbolic conservation laws on the sphere. A geometry-compatible finite volume scheme”, *J. Comput. Phys.*, vol. 228, pp. 5650–5668, 2009.

M. J. Berger, D. A. Calhoun, C. Helzel, et R. J. LeVeque, “Logically rectangular finite volume methods with adaptive refinement on the sphere”, *Philos. Trans. R. Soc. Lond. Ser. A*, vol. 367, pp. 4483–4496, 2009.

J. Blazek, *Computational Fluid Dynamics: Principles and Applications.* Elsevier, Amsterdam, 2006.

A. Bollermann, G. Chen, A. Kurganov, et S. Noelle, “A well-balanced reconstruction of wet/dry fronts for the shallow water equations”, *J. Sci. Comput.*, vol. 56, no. 2, pp. 267–290, 2013.

J. L. Bona, V. A. Dougalis, O. A. Karakashian, et W. R. McKinney, “Conservative high-order numerical schemes for the generalized Korteweg-de Vries equation”, *Phil. Trans. Roy. Soc. London Ser. A*, vol. 351, pp. 107–164, 1995.

L. Bonaventura, “A semi-implicit, semi-Lagrangian scheme using the height coordinate for a nonhydrostatic and fully elastic model of atmospheric flows”, *Journal of Computational Physics.*, vol. 158, pp. 186–213, 2000.

A. Bourgeade, P. G. LeFloch, et P. A. Raviart, “An asymptotic expansion for the solution of the generalized Riemann problem. Part II: Application to the gas dynamics equations”, *Ann. Inst. H. Poincaré, Nonlin. Anal.*, vol. 6, pp. 437–480, 1989.

J. P. Boyd, “Equatorial solitary waves. Part 1: Rossby solitons”, *Journal of Physical Oceanography*, vol. 10, pp. 1699–1717, 1980.

——, “Equatorial solitary waves. Part 3: Westward-travelling modons”, *Journal of Physical Oceanography*, vol. 15, pp. 46–54, 1985.

S. Bryson, Y. Epshteyn, A. Kurganov, et G. Petrova, “Well-balanced positivity preserving central-upwind scheme on triangular grids for the Saint-Venant system”, *M2AN Math. Model. Numer. Anal.*, vol. 45, no. 3, pp. 423–446, 2011.

M. H. Carpenter, D. Gottlieb, et S. Abarbanel, “Time-stable boundary conditions for finite-difference schemes solving hyperbolic systems: Methodology and application to high-order compact schemes”, *J.Comp. Phys*, vol. 111, pp. 220–236, 1994.

M. J. Castro, J. A. López, et C. Parés, “Finite volume simulation of the geostrophic adjustment in a rotating shallow-water system”, *SIAM Journal on Scientific Computing*, vol. 31(1), pp. 444–477, 2008.

Y. Chen, A. Kurganov, M. Lei, et Y. Liu, “An adaptive artificial viscosity method for the Saint-Venant system. In lectures presented at a Workshop at the Mathematical Research Institute Oberwolfach, Germany, Jan 15-21, 2012; R. Ansorge et al. (Eds.): Recent Developments in the Numerics of Nonlinear Conservation Laws, Series : Notes on Numerical Fluid Mechanics and Multidisciplinary Design”, *Springer-Verlag Berlin Heidelberg*, vol. 120, pp. 125–141, 2013.

A. Chertock, A. Kurganov, et Y. Liu, “Central-upwind schemes for the system of shallow water equations with horizontal temperature gradients”, *Numerische Mathematik.*, vol. 127, pp. 595–639, 2014.

C. Clancy et J. A. Pudykiewicz, “A class of semi-implicit predictor-corrector schemes for the time integration of atmospheric models”, *Journal of Computational Physics*, vol. 250, pp. 665–684, 2013.

P. Colella et P. R. Woodward, “The piecewise parabolic method (PPM) for gas dynamical simulations”, *Journal of Computational Physics*, vol. 54, pp. 174–210, 1984.

M. J. P. Cullen, “Alternative implementations of the semi-Lagrangian semi-implicit schemes in the ECMWF model”, *Q. J. R. Meteorol. Soc*, vol. 127, pp. 2787–2802, 2001.

A. I. Delis, I. K. Nikolos, et M. Kazolea, “Performance and comparison of cell-centered and node-centered unstructured finite volume discretizations for shallow water free surface flows”, *Arch. Comput. Methods Eng.*, vol. 18, no. 1, pp. 57–118, 2011.

J. Demmel, “Nearest defective matrices and the geometry of ill-conditioning”, *In M.G. Cox and S. Hammarling, editors, Reliable Numerical Computation*, vol. 44, pp. 35–55, 1990.

S. Dharmaraja, *An analysis of the TR-BDF2 integration scheme*, M.S. in Computation for Design and Optimization éd. Massachusetts Institute of Technology, 2007.

D. R. Durran, *Numerical Methods for Fluid Dynamics: with Applications to Geophysics*, springer, 2nd éd. Texts in Applied Mathematics 32, 2010.

D. R. Durran et P. A. Reinecke, “Instability in a class of explicit two-time-level semi-Lagrangian schemes”, *Q. J. R. Meteorol. Soc*, vol. 130, pp. 365–369, 2004.

D. Dziuk, D. Kroöner, et T. Müller, “Scalar conservation laws on moving hypersurfaces”, *Interf. Free Bound., to appear.*

G. Dziuk et C. M. Elliott, “Finite elements on evolving surfaces”, *IMA J. Numer. Anal.*, vol. 27, pp. 262–292, 2007.

M. G. G. Foreman, “A two-dimensional dispersion analysis of selected methods for solving the linearized shallow water equations”, *Journal of Computational Physics*, vol. 56(2), pp. 287–323, 1984.

T. Gallouet, J. M. Hérard, et N. Seguin, “Some approximate Godunov schemes to compute shallow-water equations with topography”, *Computers & Fluids*, vol. 32 (4), pp. 479–513, 2003.

J. Giesselmann, “A convergence result for finite volume schemes on Riemannian manifolds”, *M2AN Math. Model. Numer. Anal.*, vol. 43, pp. 929–955, 2009.

S. K. Godunov, “Finite difference method for numerical computation of discontinuous solutions of the equations of fluid dynamics”, *Matematicheskii Sbornik*, vol. 47(3), pp. 271–306, 1959.

S. K. Godunov, A. W. Zabrodyn, et G. P. Prokopov, “A difference scheme for two-dimensional unsteady problems of gas dynamics and computation of flow with a detached shock wave”, *Z. Vycisl. Mat. i Mat. Fiz*, vol. 1(6), pp. 1020–1050, 1961.

I. G. Gospodinov, V. G. Spiridonov, et J. F. Geleyn, “Second-order accuracy of two-time-level semi-Lagrangian schemes”, *Q. J. R. Meteorol. Soc*, vol. 127, pp. 1017–1033, 2001.

S. Gottlieb, C. W. Shu, et E. Tadmor, “Strong stability-preserving high-order time discretization methods”, *SIAM Rev.*, vol. 43, pp. 89–112, 2001.

S. Gottlieb, D. Ketcheson, et C. W. Shu, *Strong stability preserving Runge-Kutta and multistep time discretizations.* World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2011.

B. Gustafsson, H. O. Kreiss, et A. Sundstrom, “Stability theory of difference approximation for mixed initial boundary value problems II”, *Math. Comp.*, vol. 26, pp. 649–686, 1972.

G. J. Haltiner, *Numerical weather prediction.* John Wiley Press, 1971.

E. Hanert, D. Y. Le-Roux, V. Legat, et E. Deleersnijder, "Advection schemes for unstructured grid ocean modelling", *Ocean Modelling*, vol. 7, pp. 39–58, 2004.

——, "An efficient Eulerian finite element method for the shallow water equations", *Ocean Modelling*, vol. 10, pp. 115–136, 2005.

E. Hanert, R. A. Walters, D. Y. Le-Roux, et J. D. Pietrzak, "A tale of two elements: P1NC-P1 and RT0", *Ocean Modelling*, vol. 28, pp. 24–33, 2009.

M. Hortal, "The development and testing of a new two-time-level semi-Lagrangian scheme (SETTLS) in the ECMWF forecast model", *Q. J. R. Meteorol. Soc*, vol. 128, pp. 1671–1687, 2002.

S. Karni, A. Kurganov, et G. Petrova, "A smoothness indicator for adaptive algorithms for hyperbolic systems", *J. Comput. Phys.*, vol. 178, pp. 323–341, 2002.

B. Khouider et A. Majda, "A non-oscillatory balanced scheme for an idealized tropical climate model. Part I: Algorithm and validation", *Theor. Comput. Fluid Dyn*, vol. 19, pp. 331–354, 2005.

A. Kurganov et D. Levy, "A third-order semi-discrete central scheme for conservation laws and convection-diffusion equations", *SIAM J. Sci. Comput.*, vol. 22, pp. 1461–1488, 2000.

——, "Central-upwind schemes for the Saint-Venant system", *Mathematical Modelling and Numerical Analysis.*, vol. 36, pp. 397–425, 2002.

A. Kurganov et J. Miller, "Central-upwind scheme for savage-hutter type model of submarine landslides and generated tsunami waves", *Computational Methods in Applied Mathematics*, vol. 14, pp. 177–201, 2014.

A. Kurganov et G. Petrova, "A third-order semi-discrete genuinely multidimensional central scheme for hyperbolic conservation laws and related problems", *Numer Math*, vol. 88, pp. 683–729, 2001.

——, "Central-upwind schemes on triangular grids for hyperbolic systems of conservation laws", *Numerical Methods for Partial Differential Equations*, vol. 21, pp. 536–552, 2005.

——, "Adaptive central-upwind schemes for Hamilton-Jacobi equations with nonconvex Hamiltonians", *Journal of Scientific Computing*, vol. 27, pp. 323–333, 2006.

——, “A second-order well-balanced positivity preserving central-upwind scheme for the Saint-Venant system”, *Commun. Math. Sci.*, vol. 5, pp. 133–160, 2007.

——, “A central-upwind scheme for nonlinear water waves generated by submarine landslides”, *Hyperbolic problems: Theory, Numerics, Applications (Lyon 2006). Editor: S. Benzoni-Gavage and D. Serre, Springer*, pp. 635–642, 2008.

——, “Central-upwind schemes for two-layer shallow equations”, *SIAM J. Sci. Comput.*, vol. 31, pp. 1742–1773, 2009.

A. Kurganov et E. Tadmor, “Solution of two-dimensional Riemann problems for gas dynamics without Riemann problem solvers”, *Numer. Methods Partial Differential Equations*, vol. 18, pp. 584–608, 2002.

——, “New high resolution central schemes for nonlinear conservation laws and convection-diffusion equations”, *J. Comput. Phys.*, vol. 160, pp. 241–282, 2000a.

——, “New high-resolution semi-discrete central schemes for Hamilton- Jacobi equations”, *J. Comput. Phys.*, vol. 160, pp. 720–742, 2000b.

A. Kurganov, S. Noelle, et G. Petrova, “Semi-discrete central-upwind scheme for hyperbolic conservation laws and Hamilton-Jacobi equations”, *SIAM J. Sci. Comput.*, vol. 23, pp. 707–740, 2001.

A. Kurganov, G. Petrova, et B. Popov, “Adaptive semi-discrete central-upwind schemes for nonconvex hyperbolic conservation laws”, *SIAM J. Sci. Comput.*, vol. 29, pp. 2381–2401, 2007.

P. D. Lax et R. D. Richtmyer, “Survey of the stability of linear finite difference equations”, *Comm. Pure Appl. Math*, vol. 9, pp. 267–293, 1956.

D. Y. Le-Roux et B. Pouliot, “Analysis of numerically induced oscillations in two-dimensional finite-element shallow-water models. Part II: Free planetary waves”, *SIAM Journal on Scientific Computing*, vol. 30, pp. 1971–1991, 2008.

D. Y. Le-Roux, M. Dieme, et A. Sene, “Time discretization schemes for Poincaré waves in finite-element shallow-water models”, *SIAM Journal on Scientific Computing*, vol. 33(5), pp. 2217–2246, 2011.

P. G. LeFloch, “Hyperbolic conservation laws on spacetimes, in “Nonlinear conservation laws and applications”, *IMA Vol. Math. App, Springer, New York*, vol. 153, pp. 379–391, 2011.

P. G. LeFloch et B. Okutmustur, “Hyperbolic conservation laws on spacetimes. A finite volume scheme based on differential forms”, *Far East J. Math. Sci.*, vol. 31, pp. 49–83, 2008.

P. G. LeFloch et P. A. Raviart, “An asymptotic expansion for the solution of the generalized Riemann problem. Part I: General theory”, *Ann. Inst. H. Poincaré, Nonlinear Analysis*, vol. 5, pp. 179–207, 1988.

R. J. LeVeque, “Balancing source terms and flux gradients in high-resolution Godunov methods: The quasi-steady wave-propagation algorithm”, *J. Comput. Phys*, vol. 146, pp. 346–365, 1998.

G. F. Lin, J. S. Lai, et W. D. Guo, “Finite-volume component-wise TVD schemes for 2D shallow water equations”, *Advances in Water Resources*, vol. 26, pp. 861–873, 2003.

D. J. Mavriplis, “Unstructured-mesh discretizations and solvers for computational aerodynamics”, *AIAA journal*, vol. 46, no. 6, pp. 1281–1298, 2008.

A. McDonald et J. R. Bates, “ Improving the estimate of the departure point in a two-time-level semi-Lagrangian and semi-implicit model”, *Monthly Weather Review - American Meteorological Society*, vol. 115, pp. 737–739, 1987.

A. McDonald et J. Haugen, “A two-time-level, three-dimensional semi-Lagrangian, semi-implicit, limited-area gridpoint model of the primitive equations”, *Monthly Weather Review - American Meteorological Society*, vol. 120, pp. 2603–2621, 1992.

A. Mohammadian et D. Y. Le-Roux, “Simulation of shallow flows over variable topography using unstructured grid”, *International Journal for Numerical Methods in Fluids*, vol. 48(10), pp. 1149–74, 2006.

——, “Fourier analysis of a class of upwind schemes in shallow water systems for Gravity and Rossby waves”, *International Journal for Numerical Methods in Fluids*, vol. 57(4), pp. 389–416, 2008.

A. Mohammadian, D. Y. Le-Roux, M. Tajrishi, et K. Mazaheri, “A mass conservative scheme for simulating shallow flows over variable topography using unstructured grid”, *Advances in Water Resources*, vol. 28, pp. 523–537, 2005.

K. W. Morton et T. Sonar, “Finite volume methods for hyperbolic conservation laws”, *Acta Numer.*, vol. 16, pp. 155–238, 2007.

H. Nessyahu et E. Tadmor, “Non-oscillatory central differencing for hyperbolic conservation laws”, *J. Comput. Phys.*, vol. 87, no. 2, pp. 408–463, 1990.

I. K. Nikolos et A. I. Delis, “An unstructured node-centered finite volume scheme for shallow water flows with wet-dry fronts over complex topography”, *Comput. Methods Appl. Mech. Engrg.*, vol. 198, no. 47-48, pp. 3723–3750, 2009.

S. C. Reddy, “Pseudospectra of the convection-diffusion operator”, *SIAM J. Appl. Math*, vol. 54, pp. 1634–1649, 1994.

H. Ritchie, “Application of the semi-Lagrangian method to a multilevel spectral primitive equations model”, *Q. J. R. Meteorol. Soc*, vol. 117, pp. 91–106, 1991.

H. Ritchie et C. Beaudoin, “Approximations and Sensitivity Experiments with a Baroclinic semi-Lagrangian Spectral Model”, *Monthly Weather Review - American Meteorological Society*, vol. 122, pp. 2391–2399, 1994.

H. Ritchie, C. Temperton, A. Simmons, M. Hortal, T. Davies, D. Dent, et M. Hamrud, “ Implementation of the semi-Lagrangian method in a high-resolution version of the ECMWF forecast model”, *Monthly Weather Review - American Meteorological Society*, vol. 123, pp. 489–514, 1995.

A. Robert, “A stable numerical integration scheme for the primitive meteorological equations”, *Atmos Ocean*, vol. 19, pp. 35–46, 1981.

P. L. Roe, “Approximate Riemann solvers, parameter vectors, and difference schemes”, *Journal of Computational Physics*, vol. 43, pp. 357–372, 1981.

M. I. Roldán, L. Valenzuela, et E. Zarza, “Thermal analysis of solar receiver pipes with superheated steam”, *Applied Energy*, vol. 103, pp. 73–84., 2013.

A. Rossmanith, “A wave propagation method for hyperbolic systems on the sphere”, *J. Comput. Phys.*, vol. 213, pp. 629–658, 2006.

A. Rossmanith, D. S. Bale, et R. J. LeVeque, “A wave propagation algorithm for hyperbolic systems on curved manifolds”, *J. Comput. Phys.*, vol. 199, pp. 631–662, 2004.

G. Russo, “Central schemes and systems of balance laws”, *Vieweg, Braunschweig, Hyperbolic partial differential equations (Hamburg, 2001)*, pp. 59–114, 2002.

C. W. Shu, “A survey of strong stability preserving high-order time discretization”, *In Estep, D. and Tavener, S. Collected Lectures on the Preservation of Stability under Discretization SIAM*, pp. 51–65, 2002.

C. W. Shu et S. Osher, “Efficient implementation of essentially non-oscillatory shock-capturing schemes”, *Journal of Computational Physics*, vol. 77, pp. 439–471, 1988.

P. Smolarkiewicz et J. Pudykiewicz, “A class of semi-Lagrangian approximations for fluids”, *J. Atmos. Sci.*, vol. 49, pp. 2082–2096, 1992.

R. J. Spiteri et S. J. Ruuth, “A new class of optimal high-order strong-stability-preserving time discretization methods”, *SIAM J. Numer. Anal*, vol. 40, pp. 469–491, 2002.

A. Staniforth et J. Côté, “ Semi-Lagrangian integration schemes for atmospheric models - A review”, *Monthly Weather Review - American Meteorological Society*, vol. 119, pp. 2206–2223, 1991.

A. L. Stewart, P. J. Dellar, et E. R. Johnson, “Numerical simulation of wave propagation along a discontinuity in depth in a rotating annulus”, *Computers & Fluids*, vol. 46(1), pp. 442–447, 2011.

M. Tanguay, A. Robert, et R. Laprise, “A semi-implicit semi-lagrangian fully compressible regional forecast model”, *Monthly Weather Review - American Meteorological Society*, vol. 118, pp. 1970–1980, 1990.

C. Temperton et A. Staniforth, “An efficient two-time-level semi-Lagrangian semi-implicit integration scheme”, *Q. J. R. Meteorol. Soc*, vol. 113, pp. 1025–1039, 1987.

C. Temperton, M. Hortal, et A. Simmons, “A two-time-level semi-Lagrangian global spectral model”, *Q. J. R. Meteorol. Soc*, vol. 127, pp. 111–126, 2001.

W. C. Thacker, “Some exact solutions to the nonlinear shallow-water wave equations”, *J. Fluid Mech*, vol. 107, pp. 499–508, 1981.

L. N. Trefethen, “Pseudospectra of linear operators”, *SIAM Review*, vol. 39, pp. 383–406, 1997.

——, “Computation of pseudospectra”, *Acta Numerica*, vol. 8, pp. 247–295, 1999.

L. N. Trefethen et M. Emberee, *Spectra and Pseudospectra: The Behavior of Non-Normal Matrices and Operators.* Princeton University Press, Princeton and Oxford, 2005.

L. N. Trefethen, M. Contedini, et M. Embree, “Spectra, pseudospectra, and localization for random bidiagonal matrices”, *Comm. Pure. Math*, vol. 54, pp. 595–623, 2001.

B. Van Leer, “Towards the ultimate conservative difference scheme. V. A second-order sequel to godunov’s method”, *Journal of Computational Physics*, vol. 32, pp. 101–136, 1979.

M. E. Vázquez-Cendón, “Improved treatment of source terms in upwind schemes for the shallow water equations in channels with irregular geometry”, *Journal of Computational Physics*, vol. 148, pp. 497–526, 1999.

C. B. Vreugdenhil, *Numerical methods for shallow-water flow.* Kluwer Academic Publishers, 1994.

S. Vukovic et L. Sopta, “ENO and WENO schemes with the exact conservation property for one-dimensional shallow water equations”, *J. Comput. Phys.*, vol. 179, no. 2, pp. 593–621, 2002.

R. A. Walters et G. F. Carey, “Analysis of spurious oscillation modes for the shallow water and Navier–Stokes equations”, *Computers & Fluids*, vol. 11, pp. 51–68, 1983.

R. A. Walters, E. M. Lane, et E. Hanert, “Useful time-stepping methods for the Coriolis term in a shallow water model”, *Ocean Modelling*, vol. 28, pp. 66–74, 2009.

J. B. White-III et J. J. Dongarra, “High-performance high-resolution semi-Lagrangian tracer transport on a sphere”, *Journal of Computational Physics*, vol. 230, pp. 6778–6799, 2011.

D. W. Zingg, “Aspects of linear stability analysis for higher-order finite-difference methods”, *AIAA 97-1939*, 1997.

D. W. Zingg et M. Lederle, “On linear stability analysis of high-order finite-difference methods”, *AIAA Paper 2005-5249*, 2005.