UNIVERSITÉ PARIS 13

Institut Galilée
UNIVERSITÉ PARIS 13

École Doctorale 146

THÈSE

pour obtenir le grade de docteur délivré par

Université Paris 13

Spécialité doctorale "Mathématiques-Analyse numérique"

présentée et soutenue publiquement par

Minh Hieu Do

le 19 Décembre 2017

Analyse mathématique de schémas volume finis pour la simulation des écoulements quasi-géostrophiques à bas nombre de Froude

JURY

M. Fayssal Benkhaldoun,	Professeur, Université Paris 13	President
M. François Bouchut,	Directeur de recherche CNRS, Université Paris EST	Rapporteur
M. Christophe Berthon,	Professeur, Université de Nantes	Rapporteur
M. Vladimir Zeitlin,	Professeur, École Normale Supérieure Paris	Examinateur
Mme. Carine Lucas,	Maître de conférences HDR, Université de Orléans	Examinateur
M. Pascal Omnes,	Professeur associé, CEA et Université Paris 13	Directeur de thèse
M. Emmanuel Audusse,	Maître de conférences, Université Paris 13	Co-encadrant de thèse
M. Yohan Penel,	Chargé de recherche, CEREMA	Co-encadrant de thèse

Laboratoire Analyse, Géométrie et Application
UMR CNRS 7539, Villetaneusse, France

LAGA

It is with my deepest gratitude and appreciation that I dedicate this thesis

to the memories of my father;
to my beloved mother, brothers and sisters

for their constant source of love,
support and encouragement.

Résumé

Le système de Saint-Venant joue un rôle important dans la simulation de modèles océaniques, d'écoulements côtiers et de ruptures de barrages. Plusieurs sortes de termes sources peuvent être pris en compte dans ce modèle, comme la topographie, les effets de friction de Manning et la force de Coriolis. Celle-ci joue un rôle central dans les phénomènes à grande échelle spatiale car les circulations atmosphériques ou océaniques sont souvent observées autour de l'équilibre géostrophique qui correspond à l'équilibre du gradient de pression et de cette force. La capacité des schémas numériques à bien reproduire le lac au repos a été largement étudiée; en revanche, la question de l'équilibre géostrophique (incluant la contrainte de vitesse à divergence nulle) est beaucoup plus complexe et peu de travaux lui ont été consacrés.

Dans cette thèse, nous concevons des schémas volumes finis qui préservent les équilibres géostrophiques discrets dans le but d'améliorer significativement la précision des simulations numériques de perturbations autour de ces équilibres. Nous développons tout d'abord des schémas colocalisés et décalés sur des maillages rectangulaires ou triangulaires pour une linéarisation du modèle d'origine. Le point commun décisif de ces méthodes est d'adapter et de combiner les stratégies dites "topographie apparente", "bas Mach" et "pénalisation de divergence" pour contrôler l'effet de la diffusion numérique contenue dans les schémas, de telle sorte qu'elle ne détruise pas les équilibres géostrophiques. Enfin, nous étendons ces stratégies au cas non-linéaire et montrons des résultats prometteurs.

Mots Clés— équilibre géostrophique, bas nombre de Froude, système hyperbolique, méthode de volumes finis, schéma de Godunov, diffusion numérique, schéma équilibre, force de Coriolis.

Abstract

The shallow water system plays an important role in the numerical simulation of oceanic models, coastal flows and dam-break floods. Several kinds of source terms can be taken into account in this model, such as the influence of bottom topography, Manning friction effects and Coriolis force. For large scale oceanic phenomena, the Coriolis force due to the Earth's rotation plays a central role since the atmospheric or oceanic circulations are frequently observed around the so-called geostrophic equilibrium which corresponds to the balance between the pressure gradient and the Coriolis source term. The ability of numerical schemes to well capture the lake at rest, has been widely studied. However, the geostrophic equilibrium issue, including the divergence free constraint on the velocity, is much more complex and only few works have been devoted to its preservation.

In this manuscript, we design finite volume schemes that preserve the discrete geostrophic equilibrium in order to improve significantly the accuracy of numerical simulations of perturbations around this equilibrium. We first develop collocated and staggered schemes on rectangular and triangular meshes for a linearized model of the original shallow water system. The crucial common point of the various methods is to adapt and combine several strategies known as the Apparent Topography, the Low Mach and the Divergence Penalisation methods, in order to handle correctly the numerical diffusions involved in the schemes on different cell geometries, so that they do not destroy geostrophic equilibria. Finally, we extend these strategies to the non-linear case and show convincing numerical results.

Keywords— Geostrophic equilibrium, low Froude number, hyperbolic system, finite volume method, Godunov scheme, numerical diffusion, well-balanced scheme, Coriolis force.

Acknowledgments

> *None of us got to where we are alone. Whether the assistance we received was obvious or subtle, acknowledging someone's help is a big part of understanding the importance of saying thank you.*
>
> Harvey Mackay.

Time goes by so fast and it has been already three years since the first time I came to France for a huge turning point in my life. Living in foreign country with totally different culture and language is such a challenge but a really good experience for me since I have a great opportunity to learn many new things and meet a lot of wonderful people. Doing a PHD thesis in France in general and at University Paris 13 in particular has been one of the most enjoyable time in my life. Without the guidance of the committee members, the help from my friends and the encouragement of my family, my thesis would not have been possible. Therefore, I would like to thank all people who have contributed in a variety of ways to this dissertation.

First and foremost, I would like to express my deepest gratitude to my doctoral advisor Professor Pascal Omnes for his guidance and support in several years. It has been a great pleasure for me to work on an interesting PHD project and from which I have a great chance to improve my mathematical skills and work with various researchers from different institutions. He is such a great advisor who read all my reports meticulously and explained to me clearly what I got stuck in my thesis. I am thankful to him for listening patiently all my questions, sharing his knowledge to me and giving me a lot of outstanding advice to overcome numerous obstacles.

Next, I am truly grateful to another advisor Dr. Emmanuel Audusse for his patience, caring motivation and immense knowledge. He has spent a lot of time discussing with me during the last there years and gave me various useful suggestions. Not only that, he also gave me full of opportunities to figure out many interesting things in my PHD thesis. Moreover, he also provided an excellent atmosphere for doing research as well as many chance to attend a lot of useful scientific conferences.

A special thank to my another advisor Dr. Yohan Penel for his valuable guidance, patience and enthusiasm. He has been willing to help me whenever I need some helps and gave me many helpful discussions as well. I learned a lot from him not only math but also how to use Latex effectively and the way to organize a scientific paper. I am gratefully indebted his hard working in our research papers. He spent endless hours reading, checking and improving my documents.

I would like to send to another unofficial advisor Dr. Stephane Dellacherie my appreciation for his encouragement. His kindness gives me a sense of joyfulness and happiness. It was lucky for me to met and work with him during the internship of my master thesis and at the beginning of my PHD thesis.

I would also like to thank all members of the project team ANGE of INRIA Paris for being

nice to me. I am happy to receive a lot of benefits from this dynamic research center and I am additionally thankful to Dr. Martin Parisot for a lot of excellent comments and useful discussions in the context of my thesis and also another interesting approach for my work.

Beside my supervisors, I am deeply thankful to Professors Christophe Berthon and François Bouchut, the two reviewers of my dissertation. It is a great honor for me to have my PHD thesis evaluated by these experts. I also want to express my gratitude to the rest of my thesis committee: Professors Fayssal Benkhaldoun, Vladimir Zeitlin and Carine Lucas, for their insightful comments and encouragements. I appreciate all their hard questions which incented me to widen my research from various aspects.

Moreover, I really appreciate Carolin Japhet for giving me a great chance to participate the Cemracs 2016 which devoted to numerical challenges in parallel scientific computing. Being a member of the research group with a project namely *Schwarz for TrioCFD* provided me some basic knowledge of domain decomposition method which is an interesting topic I truly want to go deeper in my research career. I really want to thank Katia Ait Ameur and Thomas Rubiano for their discussion during the research project.

I owe my warmest affection to all members of the math department in University Paris 13. I especially thank Isabelle and Yolande for their kindness, helpfulness and patience to speak french with me.

Moreover, without hesitation, I would like to thank all my Vietnammese friends in University Paris 13 for their encouragement and support. They make my stay and study in this place more enjoyable. I am happy to share a lot of memorable moments with them.

Finally, I would also express my appreciation to my parents, sisters and brothers who provide unending inspiration. They are always supporting and encouraging me with their best effort.

Thank you very much, everyone !

Contents

List of Tables

List of Figures

Introduction

> “*Creativity requires the courage to let go of certainties .*”
> Erich Fromm

The shallow water equations (also called Saint-Venant equations), a hyperbolic system of partial differential equations, can be used to model many interesting phenomena in geophysical fluid mechanics. This system is an appropriate approximation for the atmospheric and oceanic flows since in these situations, the scale of horizontal motions is much larger than that of the vertical motion. At large scale motion, it is worth including in this model the effects of the geometry of the Earth and the Coriolis force induced by its rotation. We can incorporate the influence of these factors by including some additional source terms in the shallow water equations. Then, these rotating shallow water equations (RSWE) can be written as

$$\begin{cases} \partial_t h + \nabla \cdot (h\mathbf{u}) = 0, & \text{(1a)} \\ \partial_t (h\mathbf{u}) + \nabla \cdot (h\mathbf{u} \otimes \mathbf{u}) + \nabla \left(g \dfrac{h^2}{2} \right) = -gh\nabla b - h\Omega \mathbf{u}^{\perp}, & \text{(1b)} \end{cases}$$

where h and $\boldsymbol{u} = (u, v)$ are functions of time $t > 0$ and space $(x, y) \in \mathbb{R}^2$. These variables denote respectively the vertical height of the water and the horizontal velocity. In this model, the Coriolis parameter Ω stands for the angular velocity and $\mathbf{u}^{\perp} = (-v, u)$ is the orthogonal velocity. Let us mention that the complete detail for the derivation of the RSWE from the there dimensional rotating incompressible Euler or Navier-Stokes equations can be found in several textbooks [1, 2]. It is necessary to take into account the magnitudes of the parameters. Let us denote by U the typical velocity scale of the fluid flows, L the typical length scale, T the time scale and H the height scale. Then, we introduce the nondimensional variables by

$$\bar{\mathbf{x}} = \frac{\mathbf{x}}{L}, \quad \bar{t} = \frac{t}{T}, \quad \bar{h} = \frac{h}{H}, \quad \text{and} \quad \bar{\mathbf{u}} = \frac{\mathbf{u}}{U}.$$

We now begin with the nondimensionalization process by substituting all the dimensional variables in RSWE (1) by their nondimensional variables. By doing that, we obtain the following system with only nondimensional quantities

$$\begin{cases} \dfrac{H}{T} \partial_{\bar{t}} \bar{h} + \dfrac{HU}{L} \nabla_{\bar{\mathbf{x}}} \cdot (\bar{h}\bar{\mathbf{u}}) = 0, & \text{(2a)} \\ \dfrac{HU}{T} \partial_{\bar{t}} (\bar{h}\bar{\mathbf{u}}) + \dfrac{HU^2}{L} \nabla_{\bar{\mathbf{x}}} \cdot \left(\bar{h}\bar{\mathbf{u}} \otimes \bar{\mathbf{u}} \right) + \dfrac{gH^2}{L} \nabla_{\bar{\mathbf{x}}} \left(\dfrac{\bar{h}^2}{2} \right) = -\dfrac{gH^2}{L} \bar{h} \nabla_{\bar{\mathbf{x}}} b - H\Omega U \bar{h} \bar{\mathbf{u}}^{\perp}. & \text{(2b)} \end{cases}$$

For the sake of simplicity, we now drop the notation bar of the variables in (2) and multiply respectively the first and second equations with the quantities $\frac{L}{HU}$ and $\frac{L}{HU^2}$. As a result, we

derive the dimensionless shallow water equations in rotating frame given by

$$\begin{cases} \mathsf{St}\,\partial_t h + \nabla\cdot(h\mathbf{u}) = 0, & \text{(3a)} \\ \mathsf{St}\,\partial_t(h\mathbf{u}) + \nabla\cdot(h\mathbf{u}\otimes\mathbf{u}) + \dfrac{1}{\mathsf{Fr}^2}\nabla\left(\dfrac{h^2}{2}\right) = -\dfrac{1}{\mathsf{Fr}^2}h\nabla b - \dfrac{1}{\mathsf{Ro}}h\mathbf{u}^\perp, & \text{(3b)} \end{cases}$$

where dimensionless numbers St, Fr and Ro are known as the Strouhal, the Froude and the Rossby numbers respectively defined by

$$\mathsf{St} = \frac{L}{UT}, \quad \mathsf{Fr} = \frac{U}{\sqrt{gH}}, \quad \mathsf{Ro} = \frac{U}{\Omega L}.$$

The Froude number, given by the ratio between the typical fluid velocity U and the gravity wave speed $\sqrt{gH}$, is related to the compressibility effect of the shallow water equation. This is an analogue of the well known Mach number in the Euler equations.

The Rossby number is the ratio of the inertial force to the Coriolis force. A large Rossby number (because of small scale motion L, large speeds U or slow rotation Coriolis Ω) signifies that the system is dominated by inertial forces. On the contrary, a small Rossby number indicates that the system is strongly affected by the rotation.

Since we are interested in large scale oceanographic flows, we shall focus on cases where

$$\mathsf{Ro} = \mathcal{O}(M) \quad \text{and} \quad \mathsf{Fr} = \mathcal{O}(M)$$

with M a small parameter. The Strouhal number St has a strong relation to the time scale of the motions. In particular, we first consider the case where the reference time scale is set up to be equal to the convection time scale, *i.e.* $T = \frac{L}{U}$. In this long time scale, the Strouhal number is obviously equal to one.
We now consider solutions of system (3) with the help of an asymptotic expansion of the unknowns

$$f(t,x) = f_0(t,x) + Mf_1(t,x) + M^2 f_2(t,x) + \mathcal{O}(M^3) \tag{4}$$

where the order of magnitude is equal to the small Froude and Rossby numbers.
By inserting these expansions (4) into the non-dimensional shallow water and collecting the terms with power of M, the momentum equation gives us

$$\mathcal{O}(M^{-2}) \quad : \quad \nabla(h_0 + b) = 0, \tag{5}$$

$$\mathcal{O}(M^{-1}) \quad : \quad \nabla h_1 = -\boldsymbol{u}_0^\perp. \tag{6}$$

We now turn to short time scales when the Strouhal number is of order $\mathcal{O}(M^{-1})$ and we restrict our study to flat topography. The lake at rest (5) immediately leads to $\nabla h_0 = 0$ which means that h_0 is a constant in space. Moreover, by using periodic boundary conditions for the mass equation, we obtain that h_0 is also a constant in time. Therefore, we will denote that $h_0 = h_\star$, which allows us to write

$$h(t,x) = h_\star + Mh_1(x,t) + \cdots$$

Then, we obtain respectively from the mass and momentum equations of (3) the following relations

$$\mathcal{O}(1) \quad : \quad \partial_t h_1 + h_0\nabla\cdot\boldsymbol{u}_0 = 0, \tag{7}$$

$$\mathcal{O}(M^{-1}) \quad : \quad \partial_t \boldsymbol{u}_0 + \nabla h_1 = -\boldsymbol{u}_0^\perp. \tag{8}$$

0.1 Motivations and purposes of the thesis

When considering the homogeneous system (in the absence of source terms), there are several conservative numerical fluxes which can be used in numerical schemes, such that when the discretization steps tend to 0, the limits of these schemes are weak solutions of the system hyperbolic conservation law. We mention the textbooks [3, 4] for such kinds of Riemann solvers. However, with the presence of source terms, it is still a challenge to discretize them appropriately in order to ensure the preservation of some desirable properties of the continuous system, as well as to avoid stability issues.

In the absence of Coriolis force, an essential requirement for numerical schemes is that they should capture well the "lake at rest" particular solutions given by (5) because many observations are actually small perturbations around this steady state, and the accuracy of the numerical simulations is directly linked to the preservation of this steady state. Many studies are devoted to the preservation of the lake at rest by numerical schemes. For instance, the *hydrostatic reconstruction*, introduced in [5] based on a local reconstruction at the interfaces of cells, is one of the well-known methods for the shallow water equations with non-flat topography. This technique is then extended to schemes with an arbitrary order of accuracy in [6]. We also mention one its modification [7] and a new reconstruction [8] motivated by the wet-dry front. For the moving equilibrium case, the problem is more complicated and we mention the works [9, 10] that deal with the one dimensional case.

However, in the presence of Coriolis force, the question for the preservation of the geostrophic equilibrium (6) is a difficult problem and needs to be studied carefully. In particular, this non trivial equilibrium implies the divergence free constraint

$$\nabla \cdot \boldsymbol{u}_0 = 0, \tag{9}$$

and the steady state now becomes much more complex. There are very few works related to this topic. In the finite elements framework, we can point out the works of Le Roux et al. in [11, 12]. The authors study the dispersion relation and the possible spurious modes of several types of finite elements applied to the linearized shallow water equations. In the collocated finite volume framework, the authors in [13] adapt the *hydrostatic reconstruction* to the presence of Coriolis source term by introducing a new topography. This result leads to the so called *Apparent Topography method* and works well in the one dimensional case. This strategy is then extended to the two dimensional case on Cartesian meshes in [14].

However, as mentioned above, with the geostrophic equilibrium, we do not only have to deal with the balance between the pressure gradient and Coriolis force, but we also have to take into account the divergence free condition (9). This implies that the low Froude number situation (for the shallow water) will experience the same difficulties as the low Mach number situation (for Euler equations).

Fortunately, a substantial amount of research articles have been devoted to the low Mach number problem. Guillard and Viozat in [15] perform the analysis of the first order Roe scheme to show that in the low Mach number limit, the discrete equations imply pressure fluctuations of the order of the Mach number, while the solutions of the continuous equations have pressure fluctuations that scale with the square of the Mach number. They followed the preconditioning technique in [16, 17] to modify the viscosity matrix on the purpose of recovering the correct scaling of the pressure. In [18], Li and Gu introduced an All-Speed Roe type scheme by changing non-linear eigenvalues in the numerical expression of the Roe-type schemes. Their flux is quite simple and this scheme has the same behaviour in the low Mach number limit as the original continuous equations. More importantly, unlike the traditional preconditioned Roe scheme, the All-Speed-Roe scheme recover the divergence constraint of the zero order velocity at the discrete level.

In 2011, Felix Rieper introduced a low Mach number fix for Roe's approximate Riemann solver (LMRoe) [19] based on the modification of the characteristic variable. In particular, the correction is simply achieved by multiplying the jump in the normal velocity component with the local Mach number on the purpose of removing the accuracy problem. Moreover, Stéphane Dellacherie proposed in [20] a theoretical framework based on the Hodge decomposition to clearly explain the inaccuracy of the classical Godunov scheme applied to compressible Euler system on Cartesian meshes. He and his co-authors also explained in [21] that the behavior of the Godunov scheme not merely depends on the space dimension but also on the type of mesh. Then, the work in [22] proposed an *all Mach Godunov type schemes* applied to the compressible Euler system by using the strategy which is named *all Mach correction*.

The main goal of this thesis is to create stable numerical schemes with explicit time discretization, *i.e.*with no linear systems to solve, such that these schemes can preserve the geostrophic equilibrium including the divergence constraint, in order to improve the accuracy of the numerical schemes when the perturbations take place around this non trivial equilibrium. To do that, we follow the framework in [20] and pay attention to the following purposes:

- Explain the wrong behavior of the classical Godunov type schemes applied to the linear wave equation and point out the main reasons of the inaccuracy problem.
- Propose modified Godunov schemes which are able to capture the consistent discrete geostrophic equilibrium or at least are accurate around this steady state.
- Figure out the influence of the cell geometry on the Godunov scheme applied to the linear wave equation with Coriolis source term. In particular, we perform the analysis for the kernel of the Godunov schemes on rectangular and triangular grids. This is motivated by the fact that on triangular grids, all the jumps in the normal velocity components disappear at the cell faces
$$\Delta U = (\mathbf{u}_i - \mathbf{u}_j) \cdot \mathbf{n}_{ij} = 0$$
for discrete velocities that are divergence free in the sense to be specified; which is investigated in [21, 23, 24].
- Develop staggered type schemes with appropriate numerical diffusions and compare the obtained results with those obtained by collocated schemes in terms of dispersion relation, damping error, phase and group velocities.
- Extend the results in the linear case to the full non-linear shallow water equations with various convincing test cases.

0.2 Outline of the thesis

To illustrate all purposes mentioned above, this thesis is organized into three big parts and seven chapters.

Part I of this thesis is devoted to the study of numerical schemes applied to the linear wave equation with Coriolis source term in dimension one. In particularly, we first assume that the solution does not depend on the y direction and we consider the quasi-1D linear wave equation with Coriolis source term[1]

$$\begin{cases} \partial_t r + a_\star \partial_x u = 0, \\ \partial_t u + a_\star \partial_x r = \omega v, \\ \partial_t v = -\omega u \end{cases} \tag{10}$$

[1]For the sake of simplicity, we note $r = h_1$, $u = u_0$ and $v = v_0$ in (10).

where $a_\star$ and ω are constants of order one, respectively related to the wave velocity and to the rotating velocity. The stationary state corresponding to Equation (10) is the 1D version of the geostrophic equilibrium (6) and is called *1D geostrophic equilibrium*. It is such that

$$u = 0, \quad a_\star \partial_x r = \omega v. \tag{11}$$

The main goal of this part is to construct and analyze schemes that capture the *1D geostrophic equilibrium* (11) and this part consists of there chapters:

- In chapter 1, we use a Hodge decomposition to analyze the kernel of the modified equation associated to the Godunov scheme applied to (10). This work shows that unlike for the homogeneous system (no Coriolis force), the inaccuracy of the classical Godunov type schemes already appear in dimension one with the presence of the Coriolis source term. The numerical viscosity in the pressure equation is responsible for this problem. Because of it, the kernel of the modified equation becomes trivial and only includes the constant state. As a consequence, the kernel of the classical scheme is not rich enough to approximate (11). To recover the accuracy, this work proposes two new schemes which are named the collocated *Low Froude scheme* and *All Froude scheme* by respectively deleting the diffusion term on the pressure equation and keeping it small enough with size $\mathcal{O}(M)$. The first correction exactly capture the steady state (11) and although the second correction still does not have discrete kernel which discretizes well the continuous one, the *All Froude scheme* is proved to be accurate at Low Froude number locally in time, which means that when the initial condition is close to the discrete kernel, the numerical solution of this scheme is still close to this kernel within a simulation time $t = \mathcal{O}(1)$. Moreover, with an appropriate time discretization for the Coriolis force, both of these schemes are proved to be stable under some time step condition which turn out to be less restrictive than that of the classical one.

- Chapter 2 is motivated by [13, 25]; we adapt the Apparent Topography strategy presented in these works to the linear wave system (10) in order to construct a well-balanced scheme. Unlike the Low Froude and All Froude schemes, the discrete kernel of the Apparent Topography scheme is defined at the interfaces of the cells, instead of at the cell centers. It is another consistent discretization of the *1D geostrophic equilibrium* (11). In this chapter, we prove the discrete Hodge decomposition with the kernel and its orthogonal subspace defined at the interfaces of the cells, and from the numerical point of view, we prove that the Apparent Topography scheme is still accurate at low Froude number locally in time. This scheme has a larger damping rate than the well balanced schemes proposed in chapter 1. Due to the structure of the discrete kernel at the interface, we do not have various choices for the discretization in time of the Coriolis force. As a consequence, the time step of this scheme has a strong relation to the Coriolis parameter ω. However, in this chapter, by using a Von Neumann analysis, we also show that the optimal time step of the Apparent Topography scheme is just the combination of the classical one and the stability time step of the inertial oscillation.

- In Chapter 3, we are interested in the adaptation of Low Froude and Apparent Topography strategies on staggered meshes. A Fourier analysis is performed in this chapter to show that the staggered schemes have better dispersion relations than those of the collocated schemes. Particularly, only staggered schemes can ensure that the dispersion law is a monotone function like it is at the continuous level. This property has a strong impact on the accuracy of numerical schemes and helps us avoid numerical oscillations caused by the waves with shortest wavelength $2\Delta x$, see e.g [26]. This chapter also shows that the

Low Froude staggered scheme is really robust with respect to the time step and to the relation between the Rossby deformation radius and the space step. On the other hand, we also point out that, unlike the classical and Apparent Topography schemes, the Low Froude scheme is an *orthogonality preserving scheme*, which means that it can capture the orthogonal subspace of the kernel, in addition to the kernel itself. Therefore, the behaviors of the Apparent Topography and Low Froude schemes are totally different one from the other with respect to the transient states. Finally, the stability condition of the staggered type schemes is also obtained in this chapter by using a Von Neumann analysis.

Part II of this thesis is related to the preservation of the *2D geostrophic equilibrium*. Particularly, we focus on the following linear wave equation with Coriolis source term

$$\begin{cases} \partial_t r + a_\star \nabla \cdot \boldsymbol{u} = 0 \\ \partial_t \boldsymbol{u} + a_\star \nabla r = -\omega \boldsymbol{u}^\perp. \end{cases} \tag{12}$$

The steady state of this equations is clearly given by

$$\nabla . \cdot \boldsymbol{u} = 0, \quad a_\star \nabla r = -\omega \boldsymbol{u}^\perp. \tag{13}$$

This Part is composed of there chapters which emphasize on the accuracy of numerical schemes around the discrete version of the steady state (13).

- The objective of Chapter 4 in this manuscript is to derive modified collocated Godunov finite volume schemes applied to (12). The Hodge decomposition in 2D is introduced and the analysis of the modified equation shows us a new difficulty. Unlike the one dimensional case, troubles not only come from the pressure equation, but also from the velocity equations. As a consequence, naive extensions of the works in Part I are not enough to ensure the well-balanced property of numerical schemes in 2D. Of course, we are unable to delete all diffusion terms since the explicit schemes obtained that way are always unstable. To overcome this challenge, we study the extension of the Apparent Topography scheme by using the strategy which is named *Divergence Penalisation* based on the idea in [20]. Moreover, we also investigate the combination of these strategies on the purpose of combining their respective advantages. Due to the structure of the discrete kernel, we develop in this chapter two types of schemes: the cell-centered and vertex-based schemes. These schemes are then proved to be stable under some CFL conditions.

- Chapter 5 of the thesis presents how we can adapt the Apparent Topography and Divergence Penalisation techniques on staggered Cartesian meshes. In particular, we compute the velocity field on the primary cells and the pressure on dual cells. We also define some discrete operators which possess mimetic properties and we construct the discrete Hodge decomposition on staggered grids. Unlike the vertex based scheme on collocated grids, we can clearly define the discrete orthogonal subspace. On the other hand, we also perform the analysis for the discrete Fourier modes to investigate the behavior of the dispersion relation and damping error of the staggered type schemes. The CFL condition is also shown in this chapter for the staggered type schemes.

- In Chapter 6, we investigate the effect of the cell geometry on the Godunov type schemes applied to (12). Without Coriolis source term, the work [21] indicates that there is no problem related to the divergence constraint on triangular grids. However, for the case with Coriolis force, the study in this chapter points out some drawbacks of the collocated scheme on triangular grids. In particular, the analysis of the kernel of the collocated scheme leads to the fact that some gradient of a P_1 conforming function should be equal to the

gradient of some P_1 non-conforming function. Therefore, it is essential to use a staggered scheme on triangular grids to avoid this problem and also to keep the satisfaction of the free divergence condition. As a result, we only have to deal with the problem on the pressure equation to obtain numerical schemes which can capture the discrete kernel defined on triangular grid.

Part III of this manuscript is used for the study of the non-linear shallow water equation in a rotating framework.

- In Chapter 7, the final chapter in this thesis, we extend some satisfactory strategies developed in the linear case to the non-linear shallow water equation with Coriolis force. The results in [22] clearly shows that the Low Froude strategy in the non-linear case is not a good modification since the obtained scheme is unconditionally unstable. Therefore, in this chapter, for stability reasons, the Low Froude strategy will be replaced by the All Froude one. Moreover, we also point out that this modification is still good for the linear wave equation since the obtained schemes are accurate at low Froude number locally in time. Although we do not have theoretical evidence to show that the proposed schemes actually work well in the non-linear case, various good numerical results in this chapter indicate that the modified schemes are much better than the classical one.

The appendix A of this manuscript is devoted to the study of the inertial oscillation. It explains the behavior of the time discretization applied to the Coriolis force. This appendix clearly shows that the totally explicit scheme is always unstable and we have to use a time discretization of the Coriolis source term which is implicit enough. Therefore, it is highly recommended not to use an explicit treatment of the Coriolis source term in both the linear wave equation and shallow water equations.

Part I

Analysis of numerical schemes for the linear equation with Coriolis source term in 1D

CHAPTER 1

Analysis of Godunov type schemes for the linear wave equation with Coriolis source term

In theory there is no difference
between theory and practice.
In practice there is.

Lawrence "Yogui" Berra,1925
New York Yankees baseball player

This work has been done in collaboration with Emmanuel Audusse, Stephane Dellacherie, Pascal Omnes and Yohan Penel. It has been published in ESAIM: Proceedings and Surveys, Volume 58, 2017, pages 1-26.

Abstract

We propose a method to explain the behaviour of the Godunov finite volume scheme applied to the linear wave equation with Coriolis source term at low Froude number. In particular, we use the Hodge decomposition and we study the properties of the modified equation associated to the Godunov scheme. Based on the structure of the discrete kernel of the linear operator discretized by using the Godunov scheme, we clearly explain the inaccuracy of the classical Godunov scheme at low Froude number and we introduce a way to modify it to recover a correct accuracy.

Chapter content

1.1 Introduction

In this communication, we study some procedures to make finite volume Godunov type schemes accurate when solving perturbations around a steady-state. In what follows, we restrict the analysis to the quasi-1d linear wave equation with Coriolis term. Nevertheless, our future main objective is to derive accurate and stable finite volume collocated schemes for the dimensionless shallow water equations

$$\begin{cases} \mathsf{St}\,\partial_t h + \nabla\cdot(h\bar{\mathbf{u}}) = 0, & \text{(1.1a)} \\ \mathsf{St}\,\partial_t(h\bar{\mathbf{u}}) + \nabla\cdot(h\bar{\mathbf{u}}\otimes\bar{\mathbf{u}}) + \dfrac{1}{\mathsf{Fr}^2}\nabla\left(\dfrac{h^2}{2}\right) = -\dfrac{1}{\mathsf{Fr}^2}h\nabla b - \dfrac{1}{\mathsf{Ro}}h\bar{\mathbf{u}}^{\perp}, & \text{(1.1b)} \end{cases}$$

in a rotating frame when the flow is a perturbation around the so-called geostrophic equilibrium. In System (1.1) unknowns h and $\bar{\mathbf{u}}$ respectively denote the water depth and the velocity of the water column and function $b(x)$ denotes the topography of the considered oceanic basin and is a given function. Dimensionless numbers St, Fr and Ro respectively stand for the Strouhal, the Froude and the Rossby numbers defined by

$$\mathsf{St} = \frac{L}{UT}, \quad \mathsf{Fr} = \frac{U}{\sqrt{gH}}, \quad \mathsf{Ro} = \frac{U}{\Omega L}$$

where the parameter g and Ω denote the gravity coefficient and the angular velocity of the Earth. Constants U, H, L and T are some characteristic velocity, vertical and horizontal lengths and time. In the sequel, we shall focus on cases where

$$\mathsf{Ro} = \mathcal{O}(M) \quad \text{and} \quad \mathsf{Fr} = \mathcal{O}(M) \tag{1.2}$$

with M a small parameter. For large scale oceanographic flows, typical values lead to $M \sim 10^{-2}$ since

$$U \approx 1\,\mathrm{m}\cdot\mathrm{s}^{-1}, \quad L \approx 10^6\,\mathrm{m}, \quad H \approx 10^3\,\mathrm{m}, \quad \Omega \approx 10^{-4}\,\mathrm{rad}\cdot\mathrm{s}^{-1}.$$

In order to exhibit some asymptotic regimes for small Froude and Rossby numbers, we perform an expansion of the unknowns such that

$$f(t,x) = f_0(t,x) + Mf_1(t,x) + M^2 f_2(t,x) + \mathcal{O}(M^3) \tag{1.3}$$

given the orders of magnitude (1.2). We first focus on long time regimes, i.e. for Strouhal number of order $\mathcal{O}(1)$. At the leading order, solutions of equations (1.1) satisfy the so-called lake at rest equilibrium

$$\nabla(h_0 + b) = 0. \tag{1.4}$$

At the next order, the flow satisfies the so-called geostrophic equilibrium

$$\nabla h_1 = -\bar{\mathbf{u}}_0^{\perp}. \tag{1.5}$$

Note that this relation implies

$$\nabla\cdot\bar{\mathbf{u}}_0 = 0. \tag{1.6}$$

The ability of numerical schemes to well capture the particular solutions (1.4) and (1.5) is of great practical interest since it has a direct consequence on the accuracy of the numerical solution when perturbations around these equilibria are considered. A substantial amount of articles in the literature has been devoted to the preservation of the lake at rest equilibrium (1.4), see in particular [25] and references therein.

The question of the geostrophic equilibrium (1.5) including the divergence constraint (1.6) is more complex. It has been studied in a finite element framework by Le Roux [12]. The author

considers in his work the linearised version of System (1.1) and studies the behaviour of several types of finite elements. He shows that spurious modes are created, in particular when the number of degrees of freedom is not the same for height and velocity unknowns. In the finite volume framework, the nonlinear case has been studied in [25, 27, 28]. In particular, Bouchut and coauthors introduce in [13] the *apparent topography method* that allows to adapt to this problem the hydrostatic reconstruction method [5] that was developed to ensure the preservation of the lake at rest equilibrium (1.4).

Let us now focus on the behaviour of solutions of System (1.1) for short times, i.e. for Strouhal number of order $\mathcal{O}(M^{-1})$. Here the study is restricted to some flat topography and solutions independent of the y direction. The asymptotic expansion (1.3) is inserted in System (1.1). At the leading order, any solution of System (1.1) satisfies the quasi-1d linear wave equation with Coriolis source term[1]

$$\begin{cases} \partial_t r + a_\star \partial_x u = 0, \\ \partial_t u + a_\star \partial_x r = \omega v, \\ \partial_t v = -\omega u \end{cases} \tag{1.7}$$

where $a_\star$ and ω are constants of order one, respectively related to the wave velocity and to the rotating velocity. The stationary state corresponding to Equation (1.7) is the 1d version of the geostrophic equilibrium (1.5) and is called *1D geostrophic equilibrium*. It is such that

$$u = 0, \quad a_\star \partial_x r = \omega v. \tag{1.8}$$

Many works were devoted to the study of the homogeneous wave equations. In particular, in a serie of articles [20, 21], Dellacherie and coauthors studied the behavior of Godunov type schemes for the 2d linear wave equation. Their works are part of a more general study about the use of Godunov type schemes in the context of the incompressible limit for Euler equations, *i.e.* for low Mach number flows, see for example [15, 19, 23, 24, 29]. Similar works are related to low Froude flows [30, 31]. In the present work, we extend the aforementioned approach from Dellacherie and coauthors to take into account the Coriolis source term. First, in Section 1.2, we analyze the continuous case by using a Hodge type decomposition. Then, in Section 1.3 and 1.4, we study three Godunov type numerical schemes to compute approximate solutions of Equation (1.7):

- The Classical Godunov scheme;
- The *Low Froude* Godunov scheme;
- The *All Froude* Godunov scheme.

For each scheme, we study the kernel of the discrete operator and we compare it to the continuous kernel (1.8). Then, we study the accuracy of the scheme at low Froude number, *i.e.* when the initial solution is close to the kernel. This is done first for the modified equation associated to the scheme in Section 1.3 and then in the fully discrete case in Section 1.4. Moreover we study the stability of each discrete scheme by using Fourier analysis. Finally we present in Section 1.5 some numerical results to illustrate our purpose.

1.2 Properties of the linear wave equation with Coriolis source term

We first focus on the properties of the linear wave equation on the 1d torus $\mathbb{T}$. To begin with, we introduce the Hilbert space

$$\left(L^2(\mathbb{T})\right)^3 = \left\{ q = (r,u,v) \;\middle|\; \int_{\mathbb{T}} r^2 \; \mathrm{d}x + \int_{\mathbb{T}} \left(u^2+v^2\right) \; \mathrm{d}x < \infty \right\}$$

[1] For the sake of simplicity, we note $r = h_1$, $u = u_0$ and $v = v_0$ in (1.7).

equipped with the scalar product

$$\langle q_1, q_2 \rangle = \int_{\mathbb{T}} r_1 r_2 \ \mathrm{d}x + \int_{\mathbb{T}} (u_1 u_2 + v_1 v_2) \ \mathrm{d}x.$$

1.2.1 Structure of the kernel of the original model

Let us define the following space

$$\mathcal{E}_{\omega\neq 0} = \left\{ q = (r,u,v) \in \left(L^2(\mathbb{T})\right)^3 \,\middle|\, u = 0, \, \forall \phi \in C_c^\infty(\mathbb{T}), \, \int_{\mathbb{T}} a_\star r \partial_x \phi \ \mathrm{d}x = -\int_{\mathbb{T}} \omega v \phi \ \mathrm{d}x \right\}. \tag{1.9}$$

We then prove this preliminary result:

Lemma 1.1. *The orthogonal of $\mathcal{E}_{\omega\neq 0}$ is*

$$\mathcal{E}_{\omega\neq 0}^\perp = \left\{ q = (r,u,v) \in \left(L^2(\mathbb{T})\right)^3 \,\middle|\, \forall \varphi \in C_c^\infty(\mathbb{T}), \, \int_{\mathbb{T}} a_\star v \partial_x \varphi \ \mathrm{d}x = -\int_{\mathbb{T}} \omega r \varphi \ \mathrm{d}x \right\}.$$

Moreover, we have $\mathcal{E}_{\omega\neq 0} \oplus \mathcal{E}_{\omega\neq 0}^\perp = \left(L^2(\mathbb{T})\right)^3$. In other words, any $q \in \left(L^2(\mathbb{T})\right)^3$ can be uniquely decomposed into

$$q = \hat{q} + \tilde{q} \tag{1.10}$$

where $\hat{q} \in \mathcal{E}_{\omega\neq 0}$ and $\tilde{q} \in \mathcal{E}_{\omega\neq 0}^\perp$.

The Hodge decomposition (1.10) allows us to define the orthogonal projection

$$\mathbb{P} : \left\{ \begin{array}{ccc} \left(L^2(\mathbb{T})\right)^3 & \longrightarrow & \mathcal{E}_{\omega\neq 0} \\ q & \longmapsto & \hat{q} \end{array} \right. \tag{1.11}$$

Remark 1.1. *The kernel and its orthogonal set can be described in a simpler way due to the definition of Sobolev spaces, namely*

$$\mathcal{E}_{\omega\neq 0} = \left\{ q = (r,u,v) \in \left(L^2(\mathbb{T})\right)^3 \,\middle|\, r \in H^1(\mathbb{T}), \, u = 0, \, v = \frac{a_\star}{\omega} r' \right\},$$
$$\mathcal{E}_{\omega\neq 0}^\perp = \left\{ q = (r,u,v) \in \left(L^2(\mathbb{T})\right)^3 \,\middle|\, v \in H^1(\mathbb{T}), \, r = \frac{a_\star}{\omega} v' \right\}.$$

Moreover, the fact that we consider periodic functions implies that for $\hat{q} \in \mathcal{E}_{\omega\neq 0}$ and $\tilde{q} \in \mathcal{E}_{\omega\neq 0}^\perp$, we have

$$\int_{\mathbb{T}} \hat{v} \ \mathrm{d}x = 0 \quad \text{and} \quad \int_{\mathbb{T}} \tilde{r} \ \mathrm{d}x = 0$$

due to boundary conditions.

Proof. Our purpose is to prove that $\mathcal{E}_{\omega\neq 0}^\perp = A$ where

$$A = \left\{ q = \left(\frac{a_\star}{\omega} v', u, v \right) \,\middle|\, u \in L^2(\mathbb{T}), \, v \in H^1(\mathbb{T}) \right\}.$$

Firstly, let us prove that $A \subset \mathcal{E}_{\omega\neq 0}^\perp$. For $\tilde{q} \in A$, we have

$$\forall q \in \mathcal{E}_{\omega\neq 0}, \ \langle \tilde{q}, q \rangle = \int_{\mathbb{T}} r \frac{a_\star}{\omega} \tilde{v}' \ \mathrm{d}x + \int_{\mathbb{T}} \frac{a_\star}{\omega} r' \tilde{v} \ \mathrm{d}x = \frac{a_\star}{\omega} \left(\int_{\mathbb{T}} r \tilde{v}' \ \mathrm{d}x + \int_{\mathbb{T}} r' \tilde{v} \ \mathrm{d}x \right).$$

According to [32, Corollary 8.10] with $(r, \tilde{v}) \in \left(H^1(\mathbb{T})\right)^2$, we have $r\tilde{v} \in H^1(\mathbb{T})$ and $(r\tilde{v})' = r'\tilde{v} + r\tilde{v}'$. Therefore, we obtain, thanks to periodic boundary conditions on $\mathbb{T}$

$$\int_{\mathbb{T}} r \tilde{v}' \ \mathrm{d}x + \int_{\mathbb{T}} r' \tilde{v} \ \mathrm{d}x = \int_{\mathbb{T}} (r\tilde{v})' \ \mathrm{d}x = 0,$$

which leads to $\langle \tilde{q}, q\rangle = 0$, $\forall q \in \mathcal{E}_{\omega\neq 0}$. It means that $\tilde{q} \in \mathcal{E}_{\omega\neq 0}^{\perp}$.

Secondly, we prove that $\mathcal{E}_{\omega\neq 0}^{\perp} \subset A$. Let $\tilde{q} \in \mathcal{E}_{\omega\neq 0}^{\perp}$. Therefore

$$\forall r \in H^1(\mathbb{T}),\ \int_{\mathbb{T}} \tilde{r} r\ \mathrm{d}x + \int_{\mathbb{T}} \frac{a_\star}{\omega} \tilde{v} r'\ \mathrm{d}x = 0,$$

which implies

$$\forall r \in C_c^\infty(\mathbb{T}),\ \int_{\mathbb{T}} \tilde{v} r'\ \mathrm{d}x = -\frac{\omega}{a_\star} \int_{\mathbb{T}} \tilde{r} r\ \mathrm{d}x.$$

As a result, $\tilde{v} \in H^1(\mathbb{T})$ and $a_\star \tilde{v}' = \omega \tilde{r}$. We come to the conclusion that

$$\mathcal{E}_{\omega\neq 0}^{\perp} = A = \left\{ q = (r,u,v) \in \left(L^2(\mathbb{T})\right)^3 \,\middle|\, \forall \varphi \in C_c^\infty(\mathbb{T}), \int_{\mathbb{T}} a_\star v \partial_x \varphi\ \mathrm{d}x = -\int_{\mathbb{T}} \omega r \varphi\ \mathrm{d}x \right\}.$$

We eventually have to prove that

$$\mathcal{E}_{\omega\neq 0} \oplus \mathcal{E}_{\omega\neq 0}^{\perp} = \left(L^2(\mathbb{T})\right)^3.$$

We only have to check $\left(L^2(\mathbb{T})\right)^3 \subset \mathcal{E}_{\omega\neq 0} \oplus \mathcal{E}_{\omega\neq 0}^{\perp}$, because of the fact that $\mathcal{E}_{\omega\neq 0} \oplus \mathcal{E}_{\omega\neq 0}^{\perp} \subset \left(L^2(\mathbb{T})\right)^3$ is trivial.

For $q = (r,u,v) \in \left(L^2(\mathbb{T})\right)^3$, let us set

$$\begin{aligned}
&\hat{r} = \mu(r) - h, && \tilde{r} = r - \mu(r) + h,\\
&\hat{u} = 0, && \tilde{u} = u,\\
&\hat{v} = -\frac{a_\star}{\omega} \partial_x h, && \partial_x \tilde{v} = \frac{\omega}{a_\star}(r - \mu(r) + h) \ \text{ and } \int_{\mathbb{T}} \tilde{v}\ \mathrm{d}x = \int_{\mathbb{T}} v\ \mathrm{d}x,
\end{aligned}$$

where $\mu(r) = \frac{1}{|\mathbb{T}|}\int_{\mathbb{T}} r\ \mathrm{d}x$ and $h \in H^1(\mathbb{T})$ is the unique solution of the variational formulation

$$\forall\ \varphi \in H^1(\mathbb{T}),\ \int_T \partial_x h \partial_x \varphi\ \mathrm{d}x + \frac{\omega^2}{a_\star^2} \int_{\mathbb{T}} \varphi h\ \mathrm{d}x = -\frac{\omega}{a_\star} \int_{\mathbb{T}} v \partial_x \varphi\ \mathrm{d}x - \frac{\omega^2}{a_\star^2} \int_{\mathbb{T}} (r - \mu(r)) \varphi\ \mathrm{d}x.$$

The existence and uniqueness of $h \in H^1(\mathbb{T})$ results from the Lax-Milgram theorem for $\omega \neq 0$.

We easily check that $\hat{q} \in \mathcal{E}_{\omega\neq 0}$ and $\tilde{q} \in \mathcal{E}_{\omega\neq 0}^{\perp}$. To reach the conclusion, we have to check that $\hat{q} + \tilde{q} = q$. The equalities $\hat{r} + \tilde{r} = r$ and $\hat{u} + \tilde{u} = u$ are trivially verified. For v, we have:

$$\forall\ \Phi \in C^\infty(\mathbb{T}),\ \int_{\mathbb{T}} (v - \hat{v} - \tilde{v}) \partial_x \Phi\ \mathrm{d}x = \int_{\mathbb{T}} v \partial_x \Phi\ \mathrm{d}x + \frac{a_\star}{\omega} \int_{\mathbb{T}} \partial_x h \partial_x \Phi\ \mathrm{d}x + \frac{\omega}{a_\star} \int_{\mathbb{T}} (r - \mu(r) + h) \Phi\ \mathrm{d}x = 0$$

due to the choice of h. Using the density of $C^\infty(\mathbb{T})$ in $H^1(\mathbb{T})$, we obtain that $v - (\hat{v} + \tilde{v}) = c$. By using the fact that $\int_{\mathbb{T}} \hat{v} dx = 0$ and $\int_{\mathbb{T}} \tilde{v} dx = \int_{\mathbb{T}} v dx$, we get $c = 0$. Therefore, we have $\hat{v} + \tilde{v} = v$. □

1.2.2 Behaviour of the solution

By using Lemma 1.1, we obtain the following properties for the linear wave equation (1.7):

Proposition 1.1. *Let q be a solution of (1.7) on $\mathbb{T}$ with initial condition q^0. Then:*

i. $\forall q^0 \in \mathcal{E}_{\omega\neq 0}$, *we have* $q(t > 0, \cdot) = q^0 \in \mathcal{E}_{\omega\neq 0}$.

ii. $\forall q^0 \in \mathcal{E}_{\omega\neq 0}^{\perp}$, *we have* $q(t > 0, \cdot) \in \mathcal{E}_{\omega\neq 0}^{\perp}$.

Proof. We note that System (1.7) can be written as

$$\partial_t q + A\partial_x q + Bq = 0 \quad \text{where} \quad A = \begin{pmatrix} 0 & a_\star & 0 \\ a_\star & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -\omega \\ 0 & \omega & 0 \end{pmatrix}.$$

Due to the fact that matrix A has 3 real distinct eigenvalues ($\lambda = 0$, $\lambda = -a_\star$ and $\lambda = a_\star$), System (1.7) is strictly hyperbolic. Therefore this system has a unique solution [33, Th. 2.22]. And for any initial condition $q^0 = (r^0, u^0, v^0)$ in $\mathcal{E}_{\omega\neq 0}$, it is obvious that this unique solution is given by $q(t > 0, \cdot) = q^0$, which proves *(i)*.

Let $q^0 = (r^0, u^0, v^0) \in \mathcal{E}_{\omega\neq 0}^{\perp}$. We notice that

$$\begin{cases} r = r^0 - a_\star \int_0^t \partial_x u \ \mathrm{d}\tau, \\ u = u^0 - \int_0^t (a_\star \partial_x r - \omega v) \ \mathrm{d}\tau, \\ v = v^0 - \int_0^t \omega u \ \mathrm{d}\tau. \end{cases}$$

Therefore, for all $\hat{q} \in \mathcal{E}_{\omega\neq 0}$, we obtain

$$\begin{aligned} \langle q, \hat{q} \rangle &= \langle q^0, \hat{q} \rangle - a_\star \int_{\mathbb{T}} \int_0^t \partial_x u \hat{r} \ \mathrm{d}x \ \mathrm{d}\tau - \int_{\mathbb{T}} \int_0^t \omega u \hat{v} \ \mathrm{d}x \ \mathrm{d}\tau \\ &= \langle q^0, \hat{q} \rangle - \int_0^t \int_{\mathbb{T}} a_\star \partial_x u \hat{r} \ \mathrm{d}x \ \mathrm{d}\tau - \int_0^t \int_{\mathbb{T}} \omega u \hat{v} \ \mathrm{d}x \ \mathrm{d}\tau \\ &= \langle q^0, \hat{q} \rangle + \int_0^t \int_{\mathbb{T}} a_\star \partial_x \hat{r} u \ \mathrm{d}x \ \mathrm{d}\tau - \int_0^t \int_{\mathbb{T}} \omega \hat{v} u \ \mathrm{d}x \ \mathrm{d}\tau = \langle q^0, \hat{q} \rangle = 0. \end{aligned}$$

As a result, we conclude that $q(t > 0, \cdot) \in \mathcal{E}_{\omega\neq 0}^{\perp}$, which proves *(ii)*. □

Corollary 1.1. *Let q be the solution of (1.7) with initial condition q^0. Then, q can be decomposed into*

$$q = \mathbb{P}q^0 + (q - \mathbb{P}q^0) \in \mathcal{E}_{\omega\neq 0} \oplus \mathcal{E}_{\omega\neq 0}^{\perp}.$$

1.2.3 Evolution of the energy

Let us define the energy as $E = \langle q, q \rangle$.

Proposition 1.2. *Let q be the solution of (1.7) on $\mathbb{T}$. Then, the energy is conserved*

$$E(t > 0) = E(t = 0).$$

Proof. Because q is the solution of (1.7), we have

$$\begin{cases} \partial_t r = -a_\star \partial_x u, \\ \partial_t u = \omega v - a_\star \partial_x r, \\ \partial_t v = -\omega u \end{cases}$$

which allows to obtain

$$\begin{aligned} \frac{1}{2} \frac{\mathrm{d}}{\mathrm{d}t} \langle q, q \rangle &= a_\star \int_{\mathbb{T}} r(-\partial_x u) \ \mathrm{d}x + \int_{\mathbb{T}} u \left(\omega v - a_\star \partial_x r\right) \ \mathrm{d}x + \int_{\mathbb{T}} v \left(-\omega u\right) \ \mathrm{d}x \\ &= a_\star \int_{\mathbb{T}} r(-\partial_x u) \ \mathrm{d}x + a_\star \int_{\mathbb{T}} u(-\partial_x r) \ \mathrm{d}x = 0. \end{aligned}$$

Hence we have $E'(t) = 0$ which concludes the proof. □

Corollary 1.2. *For all times $t > 0$, we have $\|q(t,\cdot) - \mathbb{P}q^0\| = \|q^0 - \mathbb{P}q^0\|$.*

1.3 Properties of the first order modified equation associated to the Godunov finite volume scheme

It is well known that the *classical Godunov scheme* is not accurate at low Mach number (or low Froude number). With the homogeneous linear wave equation ($\omega = 0$), the problem appears only in the 2d case over rectangular meshes. The work in [20, 21] clearly points out the main reason of the inaccuracy. Shortly, this is because the *classical Godunov scheme* suffers from the loss of invariance of the well-prepared subspace $\mathcal{E}$ when the numerical diffusion related to the velocity equation is not equal to 0. However in our case with Coriolis source term, the problem appears already in 1d due to the numerical diffusion related to the pressure equation. We shall explain this point by studying the properties of the first-order modified equation associated to 1d Godunov like schemes which is given by

$$\begin{cases} \partial_t r + a_\star \partial_x u - \nu_r \partial_{xx}^2 r = 0, \\ \partial_t u + a_\star \partial_x r - \nu_u \partial_{xx}^2 u = \omega v, \\ \partial_t v = -\omega u, \end{cases} \tag{1.12}$$

where

$$\nu_r = \frac{\kappa_r |a_\star| \Delta x}{2}, \quad \nu_u = \frac{\kappa_u |a_\star| \Delta x}{2}, \tag{1.13}$$

for some mesh size $\Delta x > 0$ and viscosity parameters $\kappa_r > 0$ and $\kappa_u > 0$ (see [20] for more details). The classical Godunov scheme corresponds to $\kappa_r = \kappa_u = 1$. In the sequel, we rewrite (1.12) under a vector formulation

$$\begin{cases} \partial_t q + L_\nu q = 0, \\ q(t=0,x) = q^0(x) \end{cases} \tag{1.14}$$

where L_ν is the following spatial differential operator

$$L_\nu = L - B_\nu, \quad Lq = \begin{pmatrix} a_\star \partial_x u \\ a_\star \partial_x r - \omega v \\ \omega u \end{pmatrix} \quad \text{and} \quad B_\nu q = \begin{pmatrix} \nu_r & 0 & 0 \\ 0 & \nu_u & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \partial_{xx}^2 r \\ \partial_{xx}^2 u \\ 0 \end{pmatrix}.$$

1.3.1 Evolution of the energy

Lemma 1.2. *Let q_ν be the solution of System (1.14) on $\mathbb{T}$. Then:*

i. If we define the energy by $E_\nu = \langle q_\nu, q_\nu \rangle = \|r\|^2 + \|u\|^2 + \|v\|^2$, we obtain

$$E_\nu(t \geq 0) \leq E_\nu(t = 0)$$

which means that System (1.14) is dissipative.

ii. If we define the average of energy by $\bar{E}_\nu = \|\bar{r}\|^2 + \|\bar{u}\|^2 + \|\bar{v}\|^2$ with

$$\bar{r}(t) = \frac{1}{|\mathbb{T}|} \int_{\mathbb{T}} r(t,x) \ \mathrm{d}x, \qquad \bar{u}(t) = \frac{1}{|\mathbb{T}|} \int_{\mathbb{T}} u(t,x) \ \mathrm{d}x, \qquad \bar{v}(t) = \frac{1}{|\mathbb{T}|} \int_{\mathbb{T}} v(t,x) \ \mathrm{d}x,$$

we obtain

$$\bar{E}_\nu(t=0) = \bar{E}_\nu(t > 0).$$

iii. Moreover, we have

$$\forall t>0,\ \bar{E}_\nu(0)=\bar{E}_\nu(t)\leq E_\nu(t)\leq E_\nu(0).$$

Proof. We have

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|q_\nu\|^2(t)=-\langle Lq_\nu,q_\nu\rangle+\langle B_\nu q_\nu,q_\nu\rangle.$$

However,

$$\langle Lq_\nu,q_\nu\rangle=\langle a_\star\partial_x u,r\rangle+\langle a_\star\partial_x r-\omega v,u\rangle+\langle\omega u,v\rangle=0$$

and

$$\langle B_\nu q_\nu,q_\nu\rangle=\left\langle \nu_r\frac{\partial^2 r}{\partial x^2},r\right\rangle+\left\langle \nu_u\frac{\partial^2 u}{\partial x^2},u\right\rangle=-\nu_r\|\partial_x r\|^2-\nu_u\|\partial_x u\|^2.$$

For this reason, we obtain $E'_\nu(t)\leq 0$ which means that $E_\nu(t\geq 0)\leq E_\nu(t=0)$.

By integrating the first order modified equation over $\mathbb{T}$ and using periodic boundary conditions, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\bar{r}(t)=0,\qquad \frac{\mathrm{d}}{\mathrm{d}t}\bar{u}(t)=\omega\bar{v}(t)\qquad\text{and}\qquad \frac{\mathrm{d}}{\mathrm{d}t}\bar{v}(t)=-\omega\bar{u}(t)$$

which leads to

$$\frac{\mathrm{d}}{\mathrm{d}t}\bar{r}(t)^2=0,\qquad \frac{\mathrm{d}}{\mathrm{d}t}\bar{u}(t)^2=2\omega\bar{v}(t)\bar{u}(t)\qquad\text{and}\qquad \frac{\mathrm{d}}{\mathrm{d}t}\bar{v}(t)^2=-2\omega\bar{u}(t)\bar{v}(t).$$

As a result, we get

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[\bar{r}(t)^2+\bar{u}(t)^2+\bar{v}(t)^2\right]=0,$$

which means that $\bar{E}_\nu(t=0)=\bar{E}_\nu(t>0)$. It is interesting to note that

$$\begin{aligned}E_\star(t)&:=\int_{\mathbb{T}}(r-\bar{r})^2\ \mathrm{d}x+\int_{\mathbb{T}}(u-\bar{u})^2\ \mathrm{d}x+\int_{\mathbb{T}}(v-\bar{v})^2\ \mathrm{d}x\\&=\int_{\mathbb{T}}(r^2+u^2+v^2)\ \mathrm{d}x-2\bar{r}\int_{\mathbb{T}}r\ \mathrm{d}x-2\bar{u}\int_{\mathbb{T}}u\ \mathrm{d}x-2\bar{v}\int_{\mathbb{T}}v\ \mathrm{d}x+\int_{\mathbb{T}}(\bar{r}^2+\bar{u}^2+\bar{v}^2)\ \mathrm{d}x\\&=\int_{\mathbb{T}}(r^2+u^2+v^2)\ \mathrm{d}x-\int_{\mathbb{T}}(\bar{r}^2+\bar{u}^2+\bar{v}^2)\ \mathrm{d}x=E_\nu(t)-\bar{E}_\nu(t).\end{aligned}$$

Therefore, we obtain $E'_\star(t)=E'_\nu(t)\leq 0$ and $E_\nu(t)\geq\bar{E}_\nu(t)$ (since $E_\star(t)\geq 0$). □

1.3.2 Structure of the kernel of the modified equation

Interestingly, the structure of the kernel of the operator L_ν is deeply related to the value of ν_r. Indeed, we have:

Lemma 1.3.

i. When $\nu_r=0$, the subspace $\mathcal{E}_{\omega\neq 0}$ is also the kernel of the modified equation

$$\ker L_{\nu_r=0}=\mathcal{E}_{\omega\neq 0}.$$

Moreover, $\mathcal{E}^\perp_{\omega\neq 0}$ is invariant by the modified equation.

ii. When $\nu_r\neq 0$, the subspace $\mathcal{E}_{\omega\neq 0}$ is not invariant for the modified equation since

$$\ker L_{\nu_r\neq 0}=\{q:=(r,u,v)\,|\,r=const,u=0,v=0\}\subsetneq\mathcal{E}_{\omega\neq 0}.$$

Proof. With $\nu_r = 0$, it is easy to see that

$$\ker L_{\nu_r=0} = \mathcal{E}_{\omega\neq 0}.$$

As for the orthogonal space, the proof of Prop. 1.1 *(ii)* stands for $\nu_r = 0$.

We now focus on the case $\nu_r \neq 0$. Let us suppose that $q = (r,u,v) \in \ker L_\nu$. As $u = 0$, we have $a_\star \partial_x r - \omega v = 0$. Then, from $L_\nu q = 0$, we deduce

$$0 = \langle L_\nu q, q\rangle = \nu_r \|\partial_x r\|^2$$

which implies that $\partial_x r = 0$ or equivalently r is a constant. This leads to $v = 0$ and $q = (const, 0, 0)$. □

The result in Lemma 1.3 indicates that *the classical Godunov scheme* ($\kappa_r = 1$) does not capture all states $q \in \mathcal{E}_{\omega\neq 0}$ because of the fact that the corresponding kernel is a proper subset of $\mathcal{E}_{\omega\neq 0}$. This gives rise to the loss of invariance. However, when the numerical viscosity on the pressure vanishes ($\nu_r = 0$), we recover all states $q \in \mathcal{E}_{\omega\neq 0}$.

1.3.3 Behaviour of the solution of the modified equation

We recall that M is a small parameter. Let us introduce the following definitions:

Definition 1.1. *A state q^0 is said to be* well-prepared *if $\|q^0 - \mathbb{P}q^0\| = \mathcal{O}(M)$, where $\mathbb{P}$ is defined by (1.11).*

Definition 1.2. *The solution q_ν of System (1.14) is said to be* accurate at low Froude number at any time *if:*

$$\forall C_1 > 0, \exists C_2 > 0, \|q^0 - \mathbb{P}q^0\| \leq C_1 M \Longrightarrow \forall t \geq 0, \|q_\nu - \mathbb{P}q^0\|(t) \leq C_2 M,$$

where C_2 is a positive parameter that does not depend on M.

Definition 1.3. *The solution q_ν of System (1.14) is said to be* accurate at low Froude number locally in time *if:*

$$\forall C_1 > 0, \forall C_2 > 0 : C_2 = \mathcal{O}(1), \exists C_3 > 0, \|q^0 - \mathbb{P}q^0\| \leq C_1 M \Longrightarrow \forall t \leq C_2, \|q_\nu - \mathbb{P}q^0\|(t) \leq C_3 M,$$

where $C_3 = \mathcal{O}(1)$.

Remark 1.2. *We notice that if the solution is accurate at low Froude number, it is free of spurious acoustic waves (refer to [20] for more details).*

We have the following result. We recall that $\nu_\# = \frac{\kappa_\# |a_\star| \Delta x}{2}$.

Theorem 1.1. *Let q_ν is the solution of System (1.14). Then:*

i. When $\kappa_r = 0$, the solution is accurate at low Froude number at any time. Moreover, it satisfies $\|q_\nu - \mathbb{P}q^0\|(t) \leq \|q^0 - \mathbb{P}q^0\|$.

ii. When $\kappa_r = \mathcal{O}(M)$, the solution is accurate at low Froude number locally in time.

iii. When $\kappa_r = \mathcal{O}(1)$, the solution is accurate at low Froude number locally in time if

$$\Delta x = \mathcal{O}(M).$$

Remark 1.3. *From Point* (iii)*, we can state that for* $\kappa_r = \mathcal{O}(1)$*, it is enough to consider a very fine mesh to obtain accurate results. We shall see in the sequel that we actually need to consider fine meshes which is a strong restriction from the computational point of view.*

Proof. Let q_ν^a be the solution of

$$\begin{cases} \partial_t q + L_\nu q = 0, \\ q(t=0,x) = \mathbb{P}q^0(x) \end{cases}$$

and q_ν^b be the solution of

$$\begin{cases} \partial_t q + L_\nu q = 0, \\ q(t=0,x) = q^0(x) - \mathbb{P}q^0(x). \end{cases}$$

Then by linearity the solution of (1.14) is $q_\nu = q_\nu^a + q_\nu^b$. If we suppose that $\|q^0 - \mathbb{P}q^0\| = C_1 M$, then by applying Lemma 1.2, we obtain

$$\|q_\nu^b\|(t) \leq \|q_\nu^b\|(0) = \|q^0 - \mathbb{P}q^0\| = C_1 M. \tag{1.15}$$

We also notice that

$$\|q_\nu - \mathbb{P}q^0\|(t) = \|q_\nu^a + q_\nu^b - \mathbb{P}q^0\|(t) \leq \|q_\nu^a - \mathbb{P}q^0\|(t) + \|q_\nu^b\|(t). \tag{1.16}$$

If $\kappa_r = 0$, then $q_\nu^a = \mathbb{P}q^0$ according to Lemma 1.3 *(i)*. This proves Point *(i)*.

As for Points *(ii)* and *(iii)*, we set $\hat{q}^0 = (\hat{r}^0, \hat{u}^0, \hat{v}^0) := \mathbb{P}q^0$ and $q_\nu^a = (r_\nu^a, u_\nu^a, v_\nu^a)$. Then, we obtain

$$\begin{cases} \partial_t(r_\nu^a - \hat{r}^0) + a_\star \partial_x(u_\nu^a - \hat{u}^0) - \nu_r \partial_{xx}^2(r_\nu^a - \hat{r}^0) + a_\star \partial_x \hat{u}^0 - \nu_r \partial_{xx}^2 \hat{r}^0 = 0, \\ \partial_t(u_\nu^a - \hat{u}^0) + a_\star \partial_x(r_\nu^a - \hat{r}^0) - \nu_u \partial_{xx}^2(u_\nu^a - \hat{u}^0) + a_\star \partial_x \hat{r}^0 - \nu_u \partial_{xx}^2 \hat{u}^0 = \omega(v_\nu^a - \hat{v}^0) + \omega \hat{v}^0, \\ \partial_t(v_\nu^a - \hat{v}^0) + \omega(u_\nu^a - \hat{u}^0) + \omega \hat{u}^0 = 0. \end{cases} \tag{1.17}$$

On the other hand, since $\mathbb{P}q^0 \in \mathcal{E}_{\omega \neq 0}$, we have that $\hat{u}^0 = 0$ and $a_\star \partial_x \hat{r}^0 = \omega \hat{v}^0$. Therefore, (1.17) reduces to

$$\begin{cases} \partial_t(r_\nu^a - \hat{r}^0) + a_\star \partial_x(u_\nu^a - \hat{u}^0) - \nu_r \partial_{xx}^2(r_\nu^a - \hat{r}^0) - \nu_r \partial_{xx}^2 \hat{r}^0 = 0, \\ \partial_t(u_\nu^a - \hat{u}^0) + a_\star \partial_x(r_\nu^a - \hat{r}^0) - \nu_u \partial_{xx}^2(u_\nu^a - \hat{u}^0) = \omega(v_\nu^a - \hat{v}^0), \\ \partial_t(v_\nu^a - \hat{v}^0) = -\omega(u_\nu^a - \hat{u}^0). \end{cases} \tag{1.18}$$

Multiplying Equation (1.18) by $q_\nu^a - \hat{q}^0$, integrating over $\mathbb{T}$ and using periodic boundary conditions, we obtain

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|q_\nu^a - \mathbb{P}q^0\|^2 = -\nu_r \|\partial_x(r_\nu^a - \hat{r}^0)\|^2 - \nu_u \|\partial_x(u_\nu^a - \hat{u}^0)\|^2 + \nu_r \left\langle \partial_{xx}^2 \hat{r}^0, r_\nu^a - \hat{r}^0 \right\rangle$$

which yields

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|q_\nu^a - \mathbb{P}q^0\|^2 \leq \nu_r \|\partial_{xx}^2 \hat{r}^0\| \cdot \|r_\nu^a - \hat{r}^0\| \leq \nu_r \|\partial_{xx}^2 \hat{r}^0\| \cdot \|q_\nu^a - \mathbb{P}q^0\|.$$

This leads to

$$\frac{\mathrm{d}}{\mathrm{d}t}\|q_\nu^a - \mathbb{P}q^0\| \leq \nu_r \|\partial_{xx}^2 \hat{r}^0\|.$$

We deduce from the latter inequality that

$$\|q_\nu^a - \mathbb{P}q^0\|(t) \leq \nu_r t \|\partial_{xx}^2 \hat{r}^0\| \tag{1.19}$$

since $q_\nu^a(0) = \mathbb{P}q^0$. From (1.15), (1.16) and (1.19), we infer

$$\|q_\nu - \mathbb{P}q^0\|(t) \leq C_1 M + \nu_r t \|\partial_{xx}^2 \hat{r}^0\|.$$

Given (1.13), we deduce Points *(ii)* and *(iii)* respectively for $\kappa_r = \mathcal{O}(M)$ and $\Delta x = \mathcal{O}(M)$. □

1.3.4 Fourier analysis

To go further in the study of the accuracy of the numerical scheme, we perform a Fourier analysis to investigate diffusion and dispersion effects. Let us consider functions of the form

$$q(t,x) = e^{i(\tau t + kx)}\hat{q} \tag{1.20}$$

where k is the wave number and τ is the frequency of the wave. These functions can be solutions to the modified equation only under a *dispersion relation* between τ and k which is commonly written as $\tau = \tau(k)$. In general, this relation lies in the complex set: the real part $\Re(\tau)$ and the imaginary part $\Im(\tau)$ indicate respectively propagation and decay of Fourier modes.

Given a wave number k, we only consider mesh sizes satisfying

$$k < \frac{\pi}{\Delta x} \tag{1.21}$$

so that the associated wave is captured by the scheme.

Functions (1.20) are solutions to the modified equation (1.12) if

$$i\tau\hat{q} + A\hat{q} = 0, \qquad \text{where} \quad A(k,\nu_r,\nu_u,a_\star,\omega) = \begin{pmatrix} \nu_r k^2 & a_\star ik & 0 \\ a_\star ik & \nu_u k^2 & -\omega \\ 0 & \omega & 0 \end{pmatrix}. \tag{1.22}$$

This means that $-i\tau$ is an eigenvalue of A. We shall denote by λ the eigenvalues of A in the sequel. Hence the decay of Fourier mode k corresponds to $\Re(\lambda) \geq 0$.

Proposition 1.3. *Under Hypothesis (1.21), the damping of Fourier modes is parametrised by κ_r as follows.*

i. When $\kappa_r = 0$, the wave associated to the kernel of the wave operator is preserved ($\lambda = 0$).

ii. When $\kappa_r = \mathcal{O}(M)$, the wave resulting from $\lambda(\nu_r = 0) = 0$ is damped at an $\mathcal{O}(M)$ speed.

iii. When $\kappa_r = \mathcal{O}(1)$ and $\Delta x = \mathcal{O}(1)$, all Fourier modes are strongly damped at an $\mathcal{O}(1)$ speed.

Proof. The linear system (1.22) reads in terms of eigenvalues λ

$$\nu_r k^2 r + ika_\star u = \lambda r, \tag{1.23a}$$
$$ika_\star r + \nu_u k^2 u - \omega v = \lambda u, \tag{1.23b}$$
$$\omega u = \lambda v. \tag{1.23c}$$

The characteristic polynomial of Matrix A is

$$\chi(\lambda,\nu_r) := \lambda^3 - k^2(\nu_r + \nu_u)\lambda^2 + (\omega^2 + k^2 a_\star^2 + k^4\nu_r\nu_u)\lambda - k^2\omega^2\nu_r = 0. \tag{1.24}$$

It is a third order polynomial whose highest order coefficient is equal to one. It thus has either one real root and two complex conjugate roots (denoted respectively by λ_0, λ_c and $\bar{\lambda}_c$) or three real roots (denoted respectively by λ_0, λ_+ and λ_-).

It is possible to determine its three roots when $\nu_r = 0$:

$$\lambda_0(\nu_r = 0) = 0,$$
$$\lambda_c(\nu_r = 0) = \frac{1}{2}\left[k^2\nu_u + i\sqrt{4(\omega^2 + k^2a_\star^2) - k^4\nu_u^2}\right].$$

We mention that the term under the square root is actually positive under Hyp. (1.21) (see (1.13) for the definition of ν_u). Point *(i)* is proven.

We remark that $\partial_\lambda\chi$ does not vanish as soon as

$$k^2\Delta x^2\left((\kappa_r - \kappa_u)^2 + \kappa_r\kappa_u\right) < 12\left(1 + \frac{\omega^2}{a_\star^2k^2}\right). \tag{1.25}$$

Due to Hyp. (1.21), this inequality always holds for κ_r and κ_u in $[0,1]$. Hence by means of the implicit function theorem, we can define a function $\nu_r \mapsto \lambda_0(\nu_r)$ for ν_r small enough. Since coefficients multiplying λ^k in (1.24) are affine functions in ν_r, we infer that λ_0 is continuous and analytic with respect to ν_r [34]. This shows that

$$\lambda_0(\nu_r) \underset{\nu_r\to 0}{\sim} \lambda_0'(\nu_r = 0)\nu_r = -\frac{\partial_{\nu_r}\chi(0,0)}{\partial_\lambda\chi(0,0)}\nu_r = \frac{k^2\omega^2}{k^2a_\star^2 + \omega^2}\nu_r.$$

In particular, we deduce that if $\kappa_r = \mathcal{O}(M)$, then $\lambda_0(\nu_r) = \mathcal{O}(M)$. This proves Point *(ii)*.

Let us now provide other properties of the eigenvalues. We substitute (1.23c) into (1.23b) and then multiply (1.23a) by $\bar{r}$ and (1.23b) by $\bar{u}$ to obtain

$$\frac{1}{\lambda}\omega^2|u|^2 + \lambda(|r|^2 + |u|^2) = k^2(\nu_r|r|^2 + \nu_u|u|^2) + ika_\star(u\bar{r} + r\bar{u}). \tag{1.26}$$

On the one hand, the real part of (1.26)

$$\Re(\lambda)\left[\frac{\omega^2|u|^2}{|\lambda|^2} + |r|^2 + |u|^2\right] = k^2\left(\nu_r|r|^2 + \nu_u|u|^2\right),$$

shows that all eigenvalues have positive real parts (unless $k = 0$ for which eigenvalues are pure imaginary), which ensures the decay for all Fourier modes.

The three roots of (1.24) satisfy

$$\lambda_1 + \lambda_2 + \lambda_3 = k^2(\nu_r + \nu_u), \tag{1.27a}$$
$$\lambda_1\lambda_2 + (\lambda_1 + \lambda_2)\lambda_3 = \omega^2 + k^2a_\star^2 + k^4\nu_r\nu_u, \tag{1.27b}$$
$$\lambda_1\lambda_2\lambda_3 = k^2\omega^2\nu_r. \tag{1.27c}$$

Substituting $\lambda_1\lambda_2$ from (1.27c) into (1.27b), we get

$$(\lambda_1 + \lambda_2)\lambda_3 + \frac{k^2\omega^2\nu_r}{\lambda_3} = \omega^2 + k^2a_\star^2 + k^4\nu_r\nu_u. \tag{1.28}$$

Let us first focus on the case of a single real eigenvalue: we take $\lambda_1 = \lambda_c$, $\lambda_2 = \bar{\lambda}_c$ and $\lambda_3 = \lambda_0$. Eq. (1.28) yields

$$\lambda_0 \geq \frac{k^2\omega^2\nu_r}{\omega^2 + k^2a_\star^2 + k^4\nu_r\nu_u} \tag{1.29}$$

since it has been proven that $\Re(\lambda_c) \geq 0$.

We also notice that $\chi(0,\nu_r) = -k^2\omega^2\nu_r < 0$ and $\chi(k^2\nu_r,\nu_r) = k^4 a_\star^2 \nu_r > 0$. Hence, since there is a single real eigenvalue, this implies that $\lambda_0 \leq k^2\nu_r$ and we have

$$\frac{\omega^2}{\omega^2 + k^2 a_\star^2 + k^4\nu_r\nu_u} k^2\nu_r \leq \lambda_0 \leq k^2\nu_r. \tag{1.30}$$

As for the complex conjugate roots, we get from (1.27a) and (1.28) that $\mu := 2\Re(\lambda_c)$ verifies

$$f(\mu) := \mu^2 - k^2(\nu_r+\nu_u)\mu - \frac{k^2\omega^2\nu_r}{k^2(\nu_r+\nu_u)-\mu} + \omega^2 + k^2 a_\star^2 + k^4\nu_r\nu_u = 0. \tag{1.31}$$

We remark that $f(\mu) \geq g(\mu)$ where

$$g(\mu) := -k^2(\nu_r+\nu_u)\mu - \frac{k^2\omega^2\nu_r}{k^2(\nu_r+\nu_u)-\mu} + \omega^2 + k^2 a_\star^2 + k^4\nu_r\nu_u. \tag{1.32}$$

Since $f(0) = g(0) > 0$, this implies that any root μ of (1.31) is larger than the smallest positive root of (1.32).

Equation $g(\mu) = 0$ can be written as

$$k^2(\nu_r+\nu_u)\mu^2 - \Big[k^4(\nu_r+\nu_u)^2 + (\omega^2 + k^2 a_\star^2 + k^4\nu_r\nu_u)\Big]\mu + (\omega^2 + k^2 a_\star^2 + k^4\nu_r\nu_u)k^2(\nu_r+\nu_u) - k^2\omega^2\nu_r = 0. \tag{1.33}$$

Due to the fact that

$$\Delta = \Big[k^4(\nu_r+\nu_u)^2 - (\omega^2 + k^2 a_\star^2 + k^4\nu_r\nu_u)\Big]^2 + 4k^4(\nu_r+\nu_u)\nu_r\omega^2 > 0,$$

Equation (1.33) has two real positive solutions so that

$$2\Re(\lambda_c) \geq \frac{(\omega^2 + k^2 a_\star^2 + k^4\nu_r\nu_u)k^2(\nu_r+\nu_u) - k^2\omega^2\nu_r}{k^4(\nu_r+\nu_u)^2 + (\omega^2 + k^2 a_\star^2 + k^4\nu_r\nu_u)}. \tag{1.34}$$

In the case of three real roots, (1.29) holds for each of them by symmetry as they are all positive. Lower bounds (1.29) and (1.34) ensure that real parts of all eigenvalues are of order 1 when ν_r is of order 1. This proves Point *(iii)*. □

1.4 Analysis of fully discrete Godunov schemes

There are two main possible time strategies for Godunov type schemes applied to the linear wave equation with Coriolis source term. The first one is a classical splitting discretisation where one deals with the problem without source term in a first step and then the Coriolis source term is considered in a second step, which then consists in solving an ordinary differential equation. It is well known that this splitting strategy is not well adapted to preserve stationary states and then to compute small perturbations around them [25, 35], see also Appendix 1.A. Thus we focus on the analysis of the second strategy that consists in computing acoustic and Coriolis effects in a single step. As a matter of fact, there are many ways to take into account the Coriolis source term. For example, we can discretise this term using explicit, implicit and even Crank-Nicolson strategies. Hence, we introduce two new parameters θ_1 and θ_2 to parametrise the strategy.

1.4.1 Study of the discrete kernel of the one step Godunov scheme

We consider a homogeneous cartesian mesh $(x_i)_{1\leq i\leq N}$. The one step fully discrete Godunov scheme is given by

$$\begin{cases} \dfrac{r_i^{n+1}-r_i^n}{\Delta t}+a_\star\dfrac{u_{i+1}^n-u_{i-1}^n}{2\Delta x}-\nu_r\dfrac{r_{i+1}^n-2r_i^n+r_{i-1}^n}{\Delta x^2}=0,\\ \dfrac{u_i^{n+1}-u_i^n}{\Delta t}+a_\star\dfrac{r_{i+1}^n-r_{i-1}^n}{2\Delta x}-\nu_u\dfrac{u_{i+1}^n-2u_i^n+u_{i-1}^n}{\Delta x^2}=\omega\left[\theta_1 v_i^n+(1-\theta_1)v_i^{n+1}\right],\\ \dfrac{v_i^{n+1}-v_i^n}{\Delta t}=-\omega\left[\theta_2 u_i^n+(1-\theta_2)u_i^{n+1}\right] \end{cases} \tag{1.35}$$

for $i\in\{1,\dots,N\}$ and $0\leq\theta_1,\theta_2\leq 1$. Periodic boundary conditions read

$$q_0^{n+1}=q_N^{n+1},\ q_{N+1}^{n+1}=q_1^{n+1}. \tag{1.36}$$

We now investigate the kernel of the fully discrete one step scheme. It is strongly related to the value of the numerical viscosity κ_r. In particular we have the following result:

Lemma 1.4.

i. When $\nu_r=0$, the kernel of the one step scheme is

$$\mathcal{E}^h_{\omega\neq 0}:=\ker L_{\nu_r=0,h}=\left\{q=(r,u,v)\in\mathbb{R}^{3N}\;\middle|\;u_i=0,\ \frac{a_\star}{2\Delta x}(r_{i+1}-r_{i-1})=\omega v_i\right\}.$$

ii. When $\nu_r\neq 0$, the kernel of the one step scheme is

$$\ker L_{\nu_r\neq 0,h}=\left\{q=(r,u,v)\in\mathbb{R}^{3N}\;\middle|\;\exists C\in\mathbb{R}: r_i=C, u_i=0, v_i=0\right\}.$$

Proof. A stationary state verifies $r_i^{n+1}=r_i^n$, $u_i^{n+1}=u_i^n$ and $v_i^{n+1}=v_i^n$. Therefore, we easily obtain from (1.35) that

$$\begin{cases} a_\star\frac{u_{i+1}^n-u_{i-1}^n}{2\Delta x}-\nu_r\frac{r_{i+1}^n-2r_i^n+r_{i-1}^n}{\Delta x^2}=0, & \text{(1.37a)}\\ a_\star\frac{r_{i+1}^n-r_{i-1}^n}{2\Delta x}-\nu_u\frac{u_{i+1}^n-2u_i^n+u_{i-1}^n}{\Delta x^2}=\omega v_i^n, & \text{(1.37b)}\\ 0=-\omega u_i^n. & \text{(1.37c)} \end{cases}$$

Point *(i)* is straightforward: we get from (1.37c) that $u_i^n=0$, and then (1.37a) is trivially satisfied since $\nu_r=0$. Then, (1.37b) yields that

$$\frac{a_\star}{2\Delta x}(r_{i+1}^n-r_{i-1}^n)=\omega v_i^n.$$

Now we consider the case $\nu_r\neq 0$. According to (1.37c), $u_i^n=0$ for all i. Together with (1.37a) and $\nu_r\neq 0$, we get $r_{i+1}^n-r_i^n=r_i^n-r_{i-1}^n$. By induction we get $r_{N+1}^n-r_N^n=r_N^n-r_{N-1}^n=\ldots=r_2^n-r_1^n=r_1^n-r_0^n=c$ where c is a constant. This implies $r_N^n=r_0^n+Nc$. On the other hand, periodic conditions require to have $r_N^n=r_0^n$. Therefore, we get $c=0$ and $r_i^n=$ constant. This leads to $v_i^n=0$ by using (1.37b). Point *(ii)* is proven. □

1.4.2 Stability of the discrete one step Godunov scheme

For $0 \leq \theta_1, \theta_2 \leq 1$, let us denote

$$\Theta_1 = 1 - \theta_1 - \theta_2, \quad \Theta_2 = \theta_1\theta_2 + (1-\theta_1)(1-\theta_2) \in [0,1], \quad \Theta_3 = (1-2\theta_1)(1-2\theta_2) \in [-1,1].$$

Lemma 1.5. *For $\kappa_r = 0$ and $\kappa_u > 0$, we have:*

i. When $\theta_1 + \theta_2 > 1$, the one step scheme (1.35) is unstable.

ii. When $\theta_1 + \theta_2 \leq 1$, we consider two cases:

(a) If $\frac{\kappa_u^2 a_\star^2}{\omega^2 \Delta x^2} \leq \Theta_3$, the one step scheme (1.35) is stable provided that

$$\Delta t \leq \Delta t_a := \frac{\kappa_u \Delta x}{2|a_\star|} \frac{1}{\left(1 - \frac{\omega \Delta x}{|a_\star|}\sqrt{\Theta_1}\right)_+}; \tag{1.38a}$$

(b) If $\frac{\kappa_u^2 a_\star^2}{\omega^2 \Delta x^2} > \Theta_3$, the one step scheme (1.35) is stable provided that

$$\Delta t \leq \min\{\Delta t_a, \Delta t_b\} \quad \text{where} \quad \Delta t_b := \frac{\Delta x}{\kappa_u |a_\star|} \times \begin{cases} \frac{2\kappa_u^2 a_\star^2}{\omega^2 \Delta x^2 \Theta_3}\left[1 - \sqrt{1 - \frac{\omega^2 \Delta x^2}{\kappa_u^2 a_\star^2}\Theta_3}\right], & \text{if } \Theta_3 \neq 0, \\ 1, & \text{otherwise.} \end{cases} \tag{1.38b}$$

Remark 1.4. *The standard CFL condition for the homogeneous case ($\omega = 0$) reads [20]*

$$\Delta t \leq \Delta t_0 := \frac{\Delta x}{|a_\star|} \min\left\{\frac{\kappa_u}{2}, \frac{1}{\kappa_u}\right\}.$$

Inequality (1.38a) clearly shows that taking Coriolis forces into account requires a less restrictive CFL condition. It is also the case for (1.38b) when $\Theta_3 \geq 0$ thanks to the convexity of the function $x \mapsto 1 - \sqrt{1-x}$. We also notice that for the Crank-Nicolson scheme $\theta_1 = \theta_2 = \frac{1}{2}$, we recover the standard bound Δt_0.

Remark 1.5. *An asymptotic expansion for $\Delta x \ll 1$ in the bound Δt_a and Δt_b in (1.38a-1.38b) yields*

$$\Delta t_a = \frac{\kappa_u \Delta x}{2|a_\star|} + \mathcal{O}(\Delta x^2), \quad \Delta t_b = \frac{\Delta x}{\kappa_u |a_\star|} + \mathcal{O}(\Delta x^3)$$

and then one still recovers the classical bound Δt_0 for the homogeneous problem.

Remark 1.6. *For large values of the Coriolis parameter ω, the constraint (1.38a) is always satisfied ($\Delta t_a = +\infty$) while for the second constraint (1.38b), it depends on the sign of Θ_3:*

- *If $\Theta_3 \geq 0$, there is no constraint upon Δt for ω large enough;*
- *If $\Theta_3 < 0$, the asymptotic bound reads*

$$\Delta t_b \approx \frac{2}{\omega}.$$

We then recover the standard stability condition for the ODE system solved by means of a θ-scheme (1.A.1b).

Remark 1.7. *Figure 1.1 specifies the stability area. In the red zone, the scheme is unstable according to Point* (i). *In the green zone, the scheme is stable under a CFL-like constraint (characterised by* Δt_a *or* $\min(\Delta t_a, \Delta t_b)$*) that is less restrictive than the homogeneous bound* Δt_0 *while in the blue zone, the scheme is stable provided* Δt *is smaller than* $\min(\Delta t_a, \Delta t_b) \leq \Delta t_0$.

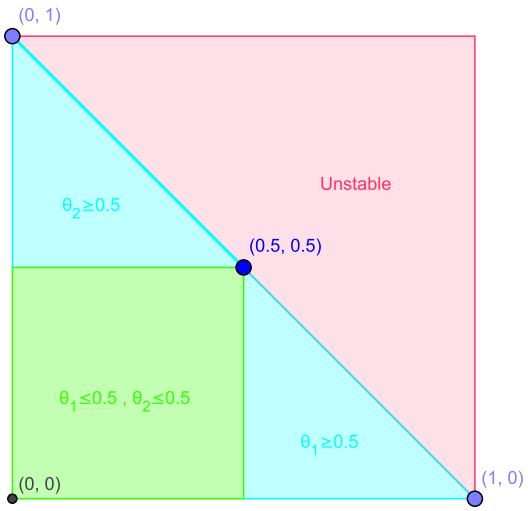


Figure 1.1: *Region of stability condition.*

Proof. We perform a Von Neumann analysis to investigate the stability condition for Scheme (1.35). Let us denote

$$\sigma = \frac{\Delta t}{\Delta x},\ \gamma = \omega \Delta t \quad \text{and} \quad s = \sin\left(\frac{k\Delta x}{2}\right).$$

We now substitute

$$q_j^n = \begin{pmatrix} r_j^n \\ u_j^n \\ v_j^n \end{pmatrix} = \begin{pmatrix} R_n \\ U_n \\ V_n \end{pmatrix} e^{ikj\Delta x}$$

into (1.35) in order to obtain

$$A q_j^{n+1} = B q_j^n \tag{1.39}$$

where the matrices A and B are given by

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -(1-\theta_1)\gamma \\ 0 & (1-\theta_2)\gamma & 1 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 1 - 2\kappa_r |a_\star| \sigma s^2 & -a_\star \sigma i \sin(k\Delta x) & 0 \\ -a_\star \sigma i \sin(k\Delta x) & 1 - 2\kappa_u |a_\star| \sigma s^2 & \theta_1 \gamma \\ 0 & -\theta_2 \gamma & 1 \end{pmatrix}.$$

In addition, we have

$$A^{-1} = \frac{1}{\Lambda(\theta_1, \theta_2)} \begin{pmatrix} \Lambda(\theta_1, \theta_2) & 0 & 0 \\ 0 & 1 & \gamma(1-\theta_1) \\ 0 & -\gamma(1-\theta_2) & 1 \end{pmatrix}$$

with

$$\Lambda(\theta_1, \theta_2) = 1 + \gamma^2 (1-\theta_1)(1-\theta_2). \tag{1.40}$$

Therefore, we can rewrite (1.39) as the following equation

$$q_j^{n+1} = C q_j^n$$

where the amplification matrix $C = A^{-1}B$ is given by

$$C = \frac{1}{\Lambda(\theta_1,\theta_2)} \begin{pmatrix} (1-2\kappa_r|a_\star|\sigma s^2)\Lambda(\theta_1,\theta_2) & -a_\star\sigma i\sin(k\Delta x)\Lambda(\theta_1,\theta_2) & 0 \\ -a_\star\sigma i\sin(k\Delta x) & 1-\gamma^2\theta_2(1-\theta_1)-2\kappa_u|a_\star|\sigma s^2 & \gamma \\ \gamma(1-\theta_2)a_\star\sigma i\sin(k\Delta x) & -\gamma[1-(1-\theta_2)2\kappa_u|a_\star|\sigma s^2] & 1-\gamma^2\theta_1(1-\theta_2) \end{pmatrix}, \tag{1.41}$$

whose characteristic polynomial will be denoted by $\mathcal{P}(\lambda)$. We now consider the modes which are constant in space ($k = 0$). In this case, the amplification matrix in (u, v) is given by

$$\frac{1}{1+\gamma^2(1-\theta_1)(1-\theta_2)} \begin{pmatrix} 1-\gamma^2\theta_2(1-\theta_1) & \gamma \\ -\gamma & 1-\gamma^2\theta_1(1-\theta_2) \end{pmatrix}.$$

Therefore the characteristic equation $\mathcal{P}(\lambda) = 0$ reduces to

$$\lambda^2 - \frac{2-\gamma^2(\theta_1+\theta_2-2\theta_1\theta_2)}{\Lambda(\theta_1,\theta_2)}\lambda + \frac{1+\gamma^2\theta_1\theta_2}{\Lambda(\theta_1,\theta_2)} = 0, \tag{1.42}$$

and the condition $|\lambda_1\lambda_2| \le 1$ is equivalent to

$$1+\gamma^2\theta_1\theta_2 \le 1+\gamma^2(1-\theta_1)(1-\theta_2),$$

that is fulfilled if and only if

$$\gamma^2[(\theta_1+\theta_2)-1] \le 0,$$

which leads to the condition $\theta_1 + \theta_2 \le 1$. This proves Point *(i)*.

Now we consider the case of interest $\kappa_r = 0$ (*c.f.* Lemma 1.4). The characteristic polynomial $\mathcal{P}(\lambda)$ reduces to

$$\mathcal{P}_0(\lambda) = (1-\lambda)\left[\lambda^2 - \frac{2-\gamma^2(\theta_1+\theta_2-2\theta_1\theta_2)-2\kappa_u|a_\star|\sigma s^2}{\Lambda(\theta_1,\theta_2)}\lambda + \frac{1+\gamma^2\theta_1\theta_2-2\kappa_u|a_\star|\sigma s^2+4a_\star^2\sigma^2 s^2(1-s^2)}{\Lambda(\theta_1,\theta_2)}\right]. \tag{1.43}$$

One root of this polynomial is $\lambda_0 = 1$ and the two others roots $\lambda_\pm$ are the solutions of the following second degree equation

$$\lambda^2 + \xi\lambda + \zeta = 0 \tag{1.44}$$

with

$$\xi = -\frac{2-\gamma^2(\theta_1+\theta_2-2\theta_1\theta_2)-2\kappa_u|a_\star|\sigma s^2}{\Lambda(\theta_1,\theta_2)} \quad \text{and} \quad \zeta = \frac{1+\gamma^2\theta_1\theta_2-2\kappa_u|a_\star|\sigma s^2+4a_\star^2\sigma^2 s^2(1-s^2)}{\Lambda(\theta_1,\theta_2)}.$$

In order to ensure that the roots of (1.44) are in the unit circle ($|\lambda_\pm| \le 1$), the coefficients ξ and ζ must satisfy

$$|\zeta| \le 1 \quad \text{and} \quad |\xi| \le 1+\zeta.$$

- Firstly, the condition $\zeta \le 1$ is equivalent to

$$\frac{1+\gamma^2\theta_1\theta_2-2\kappa_u|a_\star|\sigma s^2+4a_\star^2\sigma^2 s^2(1-s^2)}{\Lambda(\theta_1,\theta_2)} \le 1$$

which leads to

$$f_1(s^2) := -\gamma^2\Theta_1 - 2\kappa_u|a_\star|\sigma s^2 + 4a_\star^2\sigma^2 s^2(1-s^2) \le 0.$$

With s varying in $[-1,1]$, the previous condition holds provided $\max_{[0,1]} f_1 \le 0$. As Function f_1 is maximal over $\mathbb{R}$ at $X_1 := \frac{1}{2}\left(1 - \frac{\kappa_u}{2|a_\star|\sigma}\right)$, we deduce that

$$\max_{[0,1]} f_1 = \begin{cases} f_1(0), & \text{if } X_1 \le 0, \\ f_1(X_1), & \text{otherwise.} \end{cases}$$

If $X_1 \le 0$ which is equivalent to $\sigma \le \frac{\kappa_u}{2|a_\star|}$, the condition $f_1(0) \le 0$ is always satisfied. If $X_1 > 0$, $f_1(X_1) \le 0$ reads

$$\left(|a_\star|\sigma - \frac{\kappa_u}{2}\right)^2 \le \gamma^2\Theta_1 \iff \left(\frac{|a_\star|}{\Delta x} - \omega\sqrt{\Theta_1}\right)\Delta t \le \frac{\kappa_u}{2}.$$

Hence $\Delta t \le \Delta t_a$.

- Next, the condition $\zeta \ge -1$ can be written as

$$f_2(s^2) := \gamma^2\Theta_2 + 2(1 - \kappa_u|a_\star|\sigma s^2) + 4a_\star^2\sigma^2 s^2(1-s^2) \ge 0.$$

We shall see below that this constraint is weaker that another one ($f_3(s^2) \ge 0$) and needs not be taken into account.

- Let us now turn to the condition upon ξ. The first case $-\xi \le 1 + \zeta$ reads

$$2-\gamma^2(\theta_1+\theta_2-2\theta_1\theta_2)-2\kappa_u|a_\star|\sigma s^2 \le 2+\gamma^2[1-(\theta_1+\theta_2)+2\theta_1\theta_2]-2\kappa_u|a_\star|\sigma s^2+4a_\star^2\sigma^2 s^2(1-s^2)$$

which comes down to

$$-\gamma^2 - 4a_\star^2\sigma^2 s^2(1-s^2) \le 0.$$

The latter inequality always holds and does not imply an additional constraint upon Δt.

- Finally, we consider the case $\xi \le 1 + \zeta$. This leads to

$$-2+\gamma^2(\theta_1+\theta_2-2\theta_1\theta_2)+2\kappa_u|a_\star|\sigma s^2 \le 2+\gamma^2[1-(\theta_1+\theta_2)+2\theta_1\theta_2]-2\kappa_u|a_\star|\sigma s^2+4a_\star^2\sigma^2 s^2(1-s^2).$$

It follows that

$$f_3(s^2) := \gamma^2\Theta_3 + 4(1 - \kappa_u|a_\star|\sigma s^2) + 4a_\star^2\sigma^2 s^2(1-s^2) \ge 0.$$

From $\Theta_3 = 2\Theta_2 - 1$, we infer that $2f_2(s^2) - f_3(s^2) \ge 0$ over $[0,1]$. This implies that the condition $f_2(s^2) \ge 0$ is a consequence of $f_3(s^2) \ge 0$.

Function f_3 is maximal over $\mathbb{R}$ at $X_3 := \frac{1}{2}\left(1 - \frac{\kappa_u}{|a_\star|\sigma}\right) \le \frac{1}{2}$. The minimum over $[0,1]$ is reached for $s^2 = 1$ and the condition $f_3(s^2) \ge 0$ reduces to

$$0 \le f_3(1) = \omega^2\Theta_3\Delta t^2 - \frac{4\kappa_u|a_\star|}{\Delta x}\Delta t + 4 =: Q_3(\Delta t).$$

The resolution of the second order equation $Q_3(\Delta t) = 0$ leads to the stability condition (1.38b) depending on the sign of $\omega^2\Theta_3\Delta x^2 - \kappa_u^2 a_\star^2$.

□

Lemma 1.6 (Stability of the All Froude Godunov scheme)**.** *The CFL condition (1.38a-1.38b) obtained for $\kappa_r = 0$ still ensures the stability of the All Froude Godunov scheme, i.e. for the choice $\kappa_r = \mathcal{O}(M)$.*

Proof. The proof is obtained by using a classical continuity argument. The key point is to prove the modulus of the eigenvalue λ_0 is increasing when $\kappa_r \to 0^+$. Particularly, the characteristic polynomial $\mathcal{P}(\lambda)$ of the amplification matrix (1.41) is given by

$$\mathcal{P}(\lambda) = (1-\lambda-2\kappa_r|a_\star|\sigma s^2)\left(\lambda^2 - \frac{2-\gamma^2(\theta_1+\theta_2-2\theta_1\theta_2)-2\kappa_u|a_\star|\sigma s^2}{\Lambda(\theta_1,\theta_2)}\lambda + \frac{1+\gamma^2\theta_1\theta_2-2\kappa_u|a_\star|\sigma s^2}{\Lambda(\theta_1,\theta_2)}\right) + (1-\lambda)\frac{4a_\star^2\sigma^2 s^2(1-s^2)}{\Lambda(\theta_1,\theta_2)},$$

which can be decomposed as

$$\mathcal{P}(\lambda) = \mathcal{P}_0(\lambda) + \kappa_r \mathcal{P}_1(\lambda), \tag{1.45}$$

where $\mathcal{P}_0$ is given by (1.43) and

$$\mathcal{P}_1(\lambda) = -2|a_\star|\sigma s^2\left(\lambda^2 - \frac{2-\gamma^2(\theta_1+\theta_2-2\theta_1\theta_2)-2\kappa_u|a_\star|\sigma s^2}{\Lambda(\theta_1,\theta_2)}\lambda + \frac{1+\gamma^2\theta_1\theta_2-2\kappa_u|a_\star|\sigma s^2}{\Lambda(\theta_1,\theta_2)}\right).$$

Since the roots of polynomial $\mathcal{P}_0$ are simple – see the proof of Lemma 1.5 – a classical continuity argument [34] allows us to write the roots of the polynomial $\mathcal{P}$ by using an asymptotic expansion

$$\lambda = \lambda^{(0)} + \kappa_r \lambda^{(1)} + \mathcal{O}(\kappa_r^2) \tag{1.46}$$

where $\lambda^{(0)}$ is a root of $\mathcal{P}_0$. The stability of the scheme is obtained if the modulus of all roots of $\mathcal{P}$ is smaller than one. If $\lambda^{(0)} = \lambda_\pm$, the results is obvious since one can ensure $|\lambda_\pm| < 1$ by considering

$$\Delta t \leq K \min\{\Delta t_a, \Delta t_b\},$$

with $K < 1$ small enough and Δt_a, Δt_b given in (1.38a-1.38b). The case $\lambda^{(0)} = \lambda_0 = 1$ is a bit more tricky. By inserting the asymptotic expansion (1.46) into relation (1.45), we obtain

$$\mathcal{P}(\lambda) = \kappa_r\left[\lambda_1 \mathcal{P}_0'(\lambda_0) + \mathcal{P}_1(\lambda_0)\right] + \mathcal{O}(\kappa_r^2).$$

The condition $\mathcal{P}(\lambda) = 0$ thus implies

$$\lambda_1 = -\frac{\mathcal{P}_1(\lambda_0)}{\mathcal{P}_0'(\lambda_0)}.$$

Easy computations lead to

$$\begin{aligned}\mathcal{P}_1(\lambda_0) &= -2|a_\star|\sigma s^2\left(1 - \frac{2-\gamma^2(\theta_1+\theta_2-2\theta_1\theta_2)-2\kappa_u|a_\star|\sigma s^2}{\Lambda(\theta_1,\theta_2)} + \frac{1+\gamma^2\theta_1\theta_2-2\kappa_u|a_\star|\sigma s^2}{\Lambda(\theta_1,\theta_2)}\right)\\ &= -\frac{2|a_\star|\sigma s^2}{\Lambda(\theta_1,\theta_2)}\gamma^2 < 0.\end{aligned}$$

On the other hand, since $\mathcal{P}_0(\lambda) = (1-\lambda)\widetilde{\mathcal{P}_0}(\lambda)$, we have

$$\begin{aligned}\mathcal{P}_0'(1) = -\widetilde{\mathcal{P}_0}(1) &= -1 + \frac{2-\gamma^2(\theta_1+\theta_2-2\theta_1\theta_2)-2\kappa_u|a_\star|\sigma s^2}{\Lambda(\theta_1,\theta_2)} - \frac{1+\gamma^2\theta_1\theta_2-2\kappa_u|a_\star|\sigma s^2}{\Lambda(\theta_1,\theta_2)}\\ &- \frac{4a_\star^2\sigma^2 s^2(1-s^2)}{\Lambda(\theta_1,\theta_2)} = -\frac{\gamma^2+4a_\star^2\sigma^2 s^2(1-s^2)}{\Lambda(\theta_1,\theta_2)} < 0.\end{aligned}$$

It follows that

$$-2|a_\star|\sigma < \lambda_1 < 0,$$

and the scheme is stable. □

1.5 Numerical results

1.5.1 Test case with the initial condition close to the the kernel

Let us fix the parameters $a_\star = 1$, $\omega = 1$, $M = 10^{-3}$ and consider the initial condition

$$q_i^0 = \hat{q}_i^0 + M\frac{\tilde{q}_i^0}{\|\tilde{q}_i^0\|}$$

$$\text{with} \quad \hat{q}_i^0 = \begin{pmatrix} \sin(\omega x_i) \\ 0 \\ a_\star \cos(\omega x_i)\frac{\sin(\omega\Delta x)}{\omega\Delta x} \end{pmatrix} \in \mathcal{E}^h_{\omega\neq 0}, \quad \tilde{q}_i^0 = \begin{pmatrix} a_\star \cos(\omega x_i)\frac{\sin(\omega\Delta x)}{\omega\Delta x} \\ 1 \\ \sin(\omega x_i) \end{pmatrix} \in \mathcal{E}^{h,\perp}_{\omega\neq 0},$$

that is close to the kernel $\mathcal{E}^h_{\omega\neq 0}$ (see Lemma 1.4) up to a perturbation of order M.

We solve the 1D linear wave equation (1.7) by means of the schemes we analyzed in the previous sections, namely the *low Froude* scheme (1.35) for $\kappa_r = 0$, the *all Froude scheme* (1.35) for $\kappa_r = \mathcal{O}(M)$, and the *classical Godunov* scheme (1.35) for $\kappa_r = 1$. In a first step, we take $\theta_1 = 1$ and $\theta_2 = 0$.

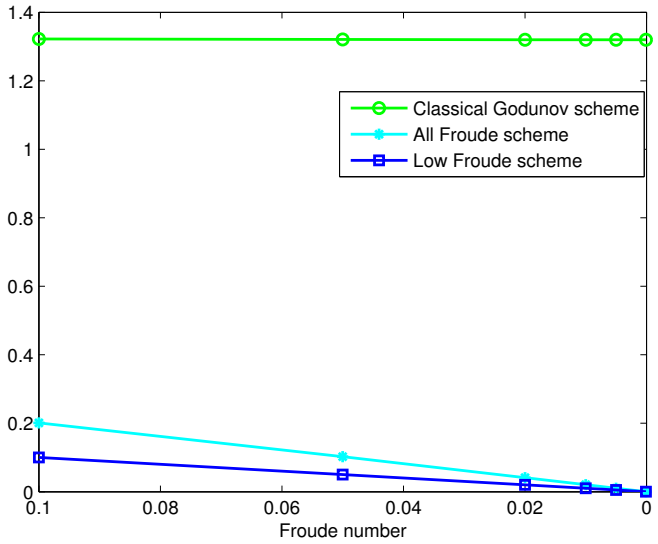
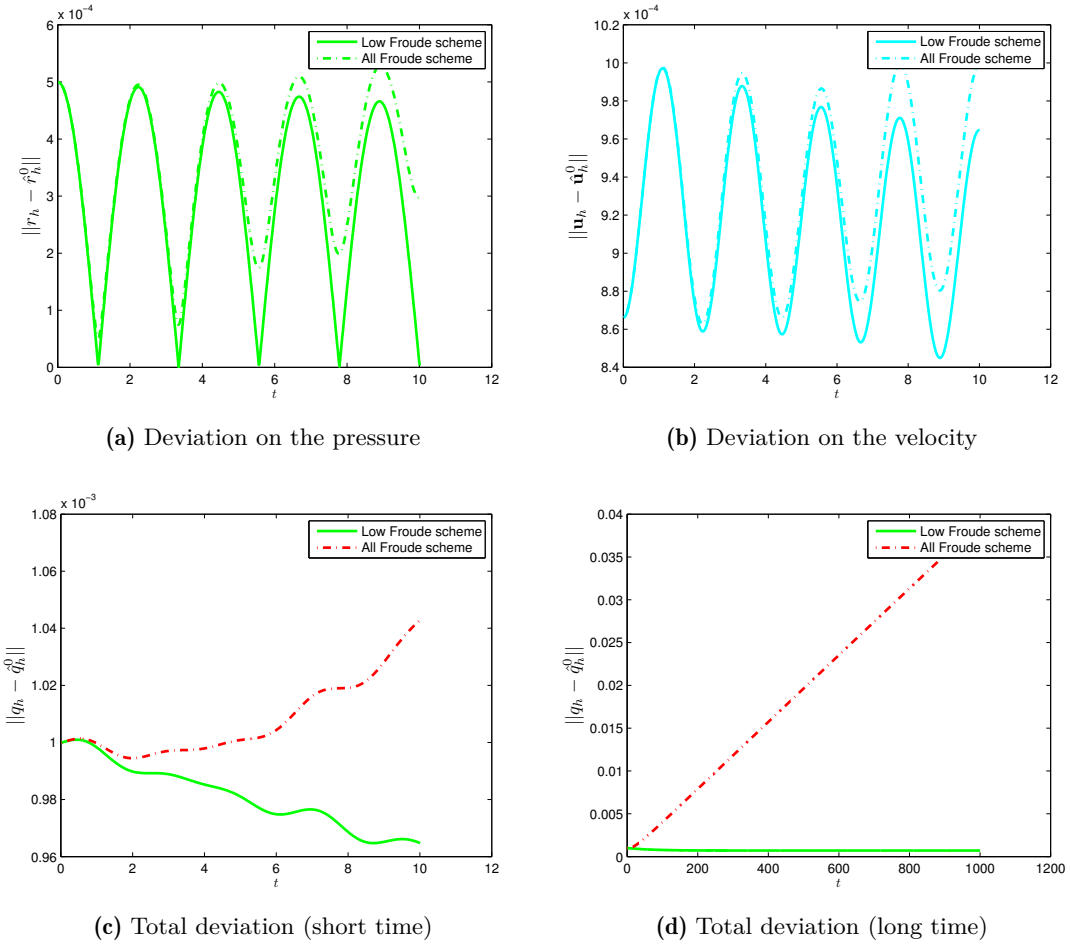


Figure 1.2: *Evolution of* $\max_t \|q_h - \mathbb{P}q_h^0\|(t)$ *for* $t = \mathcal{O}(1)$ *when the Froude number goes to* 0 *for the Low Froude Godunov, the All Froude Godunov and the Classical Godunov schemes.*

We observe on Figure 1.2 that the two schemes designed for the low Froude regime have the correct behaviour as the Froude number goes to 0, unlike the classical Godunov scheme which is not accurate as stated before.

We now investigate the accuracy with time at a fixed Froude number. As it was stated in Theorem 1.1*(i)*, we see on Figures 1.3(a)-(c) that the two aforementioned schemes are accurate for times $t = \mathcal{O}(1)$ since the numerical solutions remain close to the projection of the initial data onto the kernel (the norm of the difference is of order 10^{-3}). However for large times the all Froude scheme turns out be inaccurate as the corresponding solution is moving away from the kernel. It illustrates the result from Theorem 1.1*(ii)*.



(a) Deviation on the pressure

(b) Deviation on the velocity

(c) Total deviation (short time)

(d) Total deviation (long time)

Figure 1.3: *Comparisons of schemes: proximity to the discrete kernel as time increases.*

Next, we now focus on Theorem 1.1*(iii)* by means of Figures 1.4(a)-(b), where we see that the total deviation of the Classical Godunov scheme is of order M when the mesh is sufficiently refined ($\Delta x = \mathcal{O}(M)$). Note that even in this case, the behaviour of the low/all Froude Godunov schemes is better than that of the classical Godunov scheme.

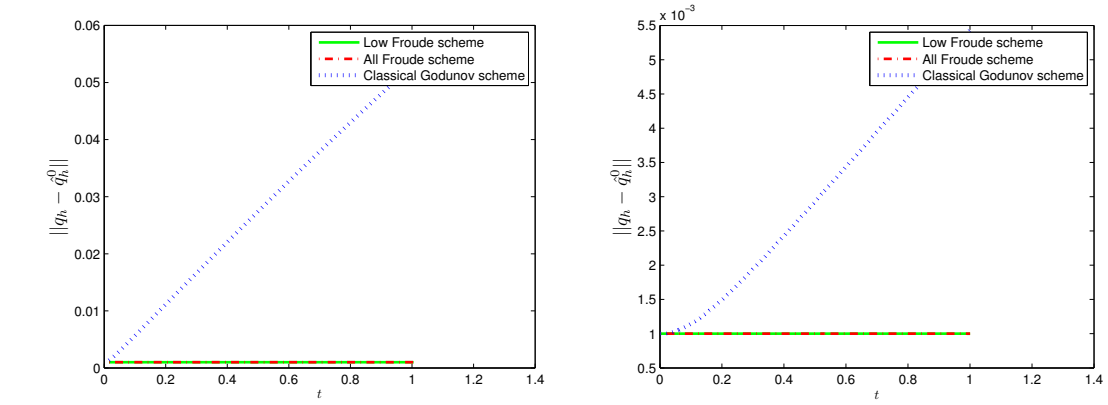
In Figures 1.5 and 1.6, we change the value θ_1 and θ_2 of the Low/All Froude schemes. These figures indicate that the total deviation depends on the value of $\theta_1 + \theta_2$.

1.5.2 Stability test case with discontinuous initial condition

In the second test case, we consider the initial condition given by

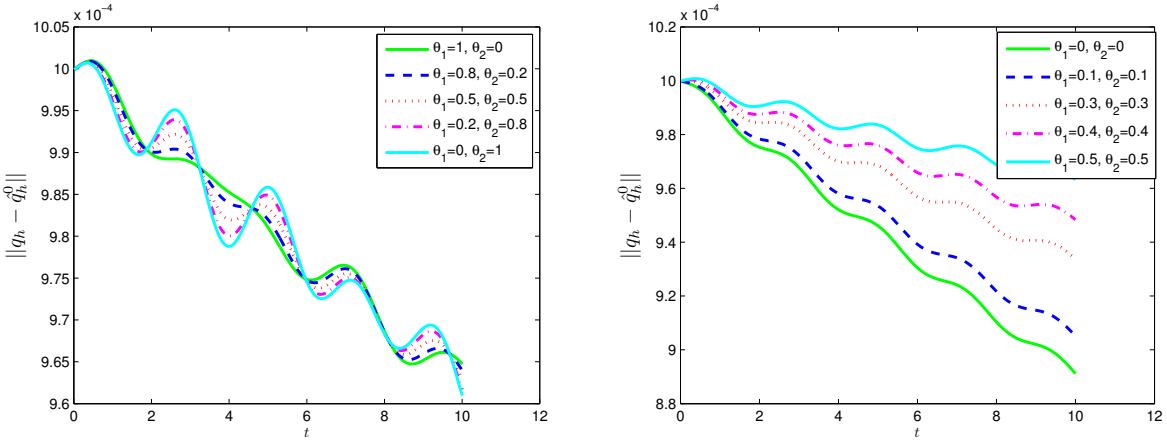
$$\begin{cases} r_i^0 = \chi_{[-\frac{1}{2},\frac{1}{2}]}(x_i), \\ u_i^0 = 1, \\ v_i^0 = 1. \end{cases} \tag{1.47}$$

In this test, we choose $\omega = 1$, $\Delta x = 0.01$ and $a_\star$ such that the Rossby deformation is equal to $R_d := \frac{a_\star}{\omega} = \Delta x$ and $\kappa_u = 1$. In Figure 1.7(a), we choose $\theta_1 = 0.5$ and $\theta_2 = 0$. Hence, in this case we have $\Theta_1 = 0.5$ and $\Theta_3 = 0$ which leads to $\Delta t_a = \frac{0.5}{1-\sqrt{0.5}}$, $\Delta t_b = 1$ and $\Delta t_0 = 0.5$. Therefore, the new time step $\Delta t = \min\{\Delta t_a, \Delta t_b\} = 1$ is less restrictive than the classical time step $\Delta t_0 = 0.5$. Figure 1.7(a) shows that the new time step is optimal since when $\Delta t = 0.999 < 1$ the *Low Froude scheme* is stable while when $\Delta t = 1.001$ the *Low Froude scheme* is unstable.



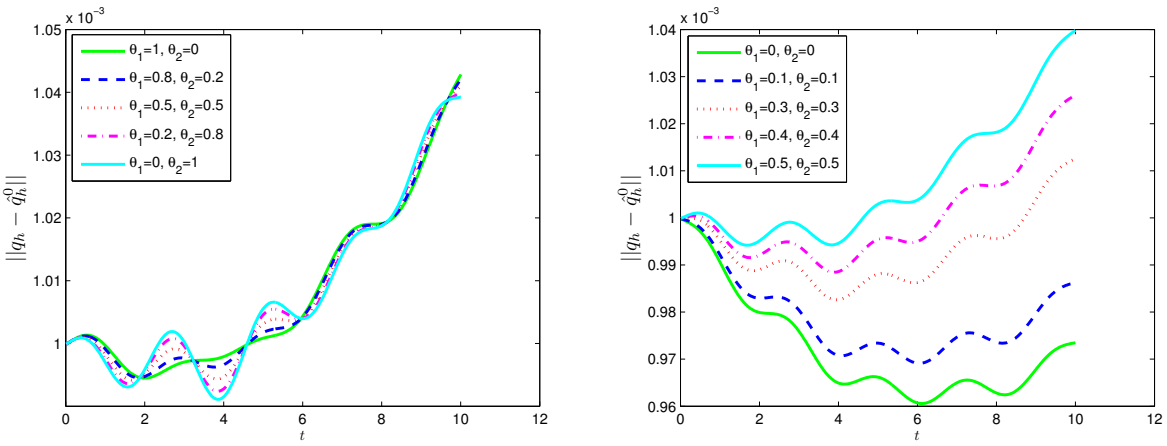
(a) Total deviation ($\Delta x = 2\pi \times 10^{-2}$) **(b)** Total deviation ($\Delta x = 2\pi \times 10^{-3}$)

Figure 1.4: *Comparisons of schemes: proximity to the discrete kernel as time increases.*



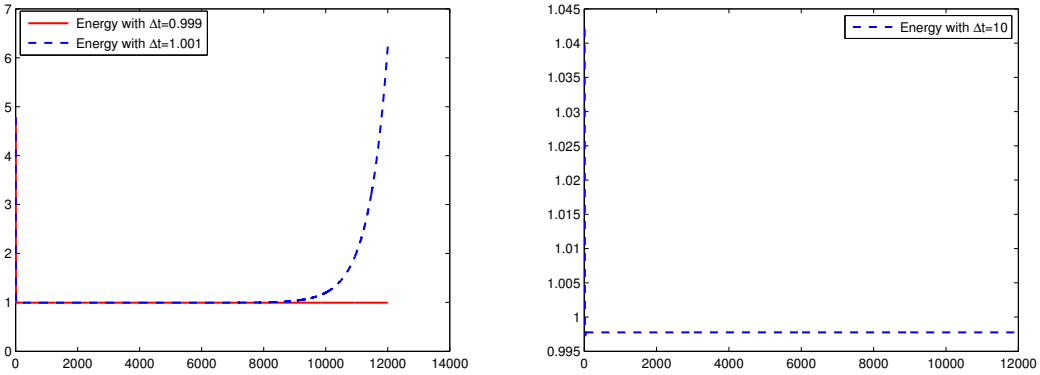
(a) Total deviation with $\theta_1 + \theta_2 = 1$. **(b)** Total deviation with various values of $\theta_1 + \theta_2$.

Figure 1.5: *Comparisons of Low Froude schemes: proximity to the discrete kernel as time increases.*



(a) Total deviation with $\theta_1 + \theta_2 = 1$. **(b)** Total deviation with various values of $\theta_1 + \theta_2$.

Figure 1.6: *Comparisons of All Froude schemes: proximity to the discrete kernel as time increases.*



(a) $\theta_1 = 0.5, \theta_2 = 0$ **(b)** $\theta_1 = 0, \theta_2 = 0$

Figure 1.7: *Influence of the time step upon the Low Froude scheme*

On the other hand, if we take $\theta_1 = 0$ and $\theta_2 = 0$, then $\Theta_1 = 1$ and $\Theta_3 = 1$. Due to the fact that $\frac{\kappa_u^2 a_\star^2}{\omega^2 \Delta x^2} \leq \Theta_3$, the constraint over the time step for the Low Froude scheme is prescribed by Δt_a. However as $a_\star = \omega \Delta x \sqrt{\Theta_1}$, $\Delta t_a = +\infty$ and the Low Froude scheme is always stable without regard to the time step Δt. Figure 1.7(b) confirms this statement by showing that the Low Froude scheme is stable even for $\Delta t = 10$.

In Figure 1.8, we take $\Delta t = \Delta t_0 = 0.5$ and change the value of θ_1 and θ_2. This figure indicates that the behaviour of the energy of the Low Froude scheme depends on the value of θ_1 and θ_2. The choice $\theta_1 = \theta_2 = \frac{1}{2}$ (Crank-Nicolson approximation for the Coriolis term) is able to preserve the energy exaclty although this choice requires a more restrictive constraint upon the time step than for $\theta_1, \theta_2 \leq \frac{1}{2}$.

1.6 Conclusion

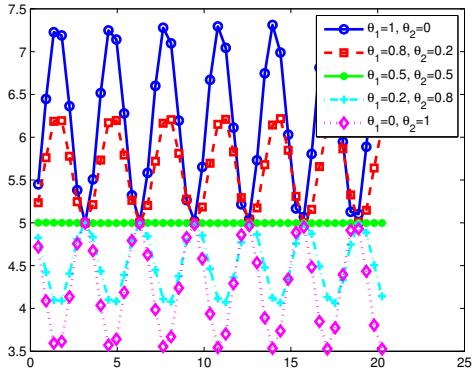
It is well known that the *classical Godunov scheme* applied to the linear wave equation is not accurate at low Froude number on cartesian meshes in dimension 2 [20]. In this work, we have shown that, when a Coriolis source term is involved, the *classical Godunov scheme* is not accurate at low Froude number even in dimension 1.

This is because the stationary space of the *classical Godunov* discrete operator is not a good approximation of the invariant subspace $\mathcal{E}_{\omega \neq 0}$. The loss of invariance of $\mathcal{E}_{\omega \neq 0}$ is explained by studying the associated modified equation. It is strongly related to the numerical diffusion κ_r on the pressure equation. In particular when we set $\kappa_r = 0$, the inaccuracy problem does not occur. As a result, we derived two modified schemes by decreasing the value of the numerical diffusion κ_r on the pressure equation. From this, we deduce that:

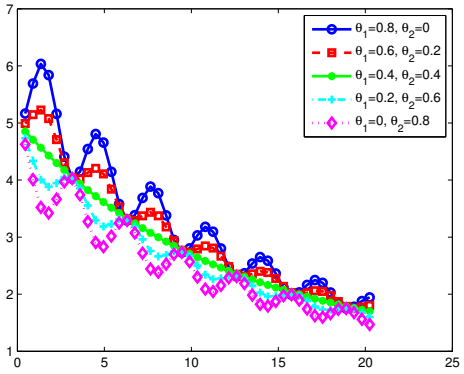
- The *Low Froude Godunov scheme* ($\kappa_r = 0$) is accurate at low Froude number.
- The *All Froude Godunov scheme* ($\kappa_r = M$) is accurate at low Froude number locally in time.

We then proved that both schemes are stable under suitable constraints upon the time step. These stability conditions turn out to be less restrictive than classical ones for a suitable treatment of the Coriolis source term.

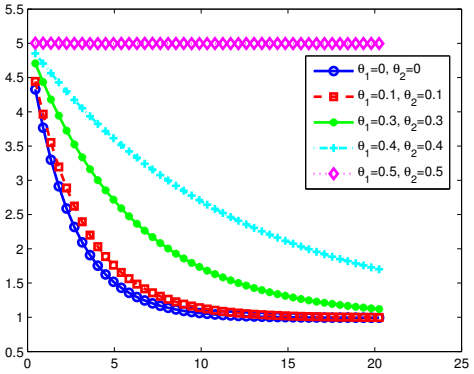
In forthcoming works, we shall extend our analysis to the two-dimensional case and to the nonlinear framework in order to derive and analyse accurate and stable numerical schemes for



(a) Energy with $\theta_1 + \theta_2 = 1$.



(b) Energy with $\theta_1 + \theta_2 = 0.8$.



(c) Energy with various values of $\theta_1 + \theta_2$.

Figure 1.8: *Comparisons of Low Froude schemes depending on the time step*

System (1.1). In particular we aim at comparing our work with previous approaches from [25, 27, 28].

1.A Analysis of splitting scheme

Let us define a two-step Godunov scheme using a splitting strategy to take into account the Coriolis source term. The first step is related to the acoustic term

$$\begin{cases} r_i^\star - r_i^n + a_\star \Delta t \frac{u_{i+1}^n - u_{i-1}^n}{2\Delta x} - \nu_r \Delta t \frac{r_{i+1}^n - 2r_i^n + r_{i-1}^n}{\Delta x^2} = 0, \\ u_i^\star - u_i^n + a_\star \Delta t \frac{r_{i+1}^n - r_{i-1}^n}{2\Delta x} - \nu_u \Delta t \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{\Delta x^2} = 0, \\ v_i^\star - v_i^n = 0, \end{cases} \tag{1.A.1a}$$

and we use a θ-scheme to deal with the Coriolis term in the second step

$$\begin{cases} r_i^{n+1} = r_i^\star, \\ u_i^{n+1} - u_i^\star = \omega \Delta t \left[\theta_1 v_i^\star + (1-\theta_1) v_i^{n+1} \right], \\ v_i^{n+1} - v_i^\star = -\omega \Delta t \left[\theta_2 u_i^\star + (1-\theta_2) u_i^{n+1} \right], \end{cases} \tag{1.A.1b}$$

for $0 \leq \theta_1, \theta_2 \leq 1$.

Lemma 1.7. *With the splitting scheme, we have:*

i. For $\nu_r = 0$, the splitting scheme preserves steady states only if $\theta_2 = 0$.

ii. For $\nu_r \neq 0$, steady states are not preserved without regard to the value of θ_1 and θ_2.

Proof. Let us assume that the numerical solution at time $t^n = n\Delta t$ belongs to the discrete kernel $\mathcal{E}^h_{\omega \neq 0}$

$$\forall i \in \mathbb{Z}, \quad u_i^n = 0 \quad \text{and} \quad a_\star \frac{r_{i+1}^n - r_{i-1}^n}{2\Delta x} = \omega v_i^n.$$

We shall show that at the next time step the numerical solution does not lie in the discrete kernel anymore. After the first step, we easily obtain

$$\begin{cases} r_i^\star = r_i^n + \nu_r \Delta t \frac{r_{i+1}^n - 2r_i^n + r_{i-1}^n}{\Delta x^2}, \\ u_i^\star = u_i^n - a_\star \Delta t \frac{r_{i+1}^n - r_{i-1}^n}{2\Delta x} = -\omega \Delta t v_i^n, \\ v_i^\star = v_i^n. \end{cases}$$

Then the second step leads to

$$\begin{cases} r_i^{n+1} = r_i^n + \nu_r \Delta t \frac{r_{i+1}^n - 2r_i^n + r_{i-1}^n}{\Delta x^2}, \\ u_i^{n+1} = \omega \Delta t (1-\theta_1)(v_i^{n+1} - v_i^n), \\ v_i^{n+1} + \omega \Delta t (1-\theta_2) u_i^{n+1} = \left[1 + (\omega \Delta t)^2 \theta_2 \right] v_i^n. \end{cases}$$

As a result, we have

$$v_i^{n+1} + (\omega \Delta t)^2 (1-\theta_1)(1-\theta_2)(v_i^{n+1} - v_i^n) = \left[1 + (\omega \Delta t)^2 \theta_2 \right] v_i^n$$

from which it follows that

$$\left[1 + (\omega \Delta t)^2 (1-\theta_1)(1-\theta_2) \right] v_i^{n+1} = \left[1 + (\omega \Delta t)^2 (1-\theta_1)(1-\theta_2) + (\omega \Delta t)^2 \theta_2 \right] v_i^n.$$

Therefore, $v_i^{n+1} = v_i^n$ (and $u_i^{n+1} = 0$) if and only if $\theta_2 = 0$. The kernel is recovered if $\nu_r = 0$ as $r_i^{n+1} = r_i^n$. □

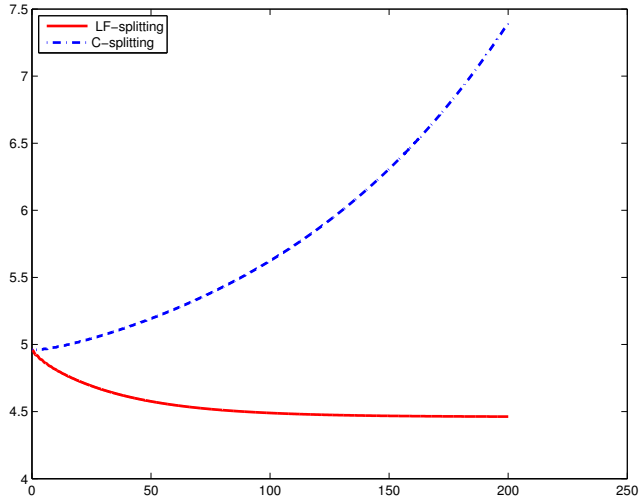


Figure 1.9: *The energy of the Low Froude splitting scheme ($\kappa_r = 0$) and Classical splitting scheme with the initial condition $r^0 = u^0 = 1$, $v^0 = \chi_{[-\frac{1}{2},\frac{1}{2}]}(x)$ on domain $\mathbb{T}_1 = [-1,1]$ and the parameters: $a_\star = \omega = 1$, $\Delta x = 0.02$, $\theta_1 = \theta_2 = \frac{1}{2}$.*

Let us now note that the choice $\theta_2 = 0$ is not really a splitting method since it can be written as a one-step method

$$\begin{cases} r_i^{n+1} - r_i^n + a_\star \Delta t \frac{u_{i+1}^n - u_{i-1}^n}{2\Delta x} - \nu_r \Delta t \frac{r_{i+1}^n - 2r_i^n + r_{i-1}^n}{\Delta x^2} = 0, \\ u_i^{n+1} - u_i^n + a_\star \Delta t \frac{r_{i+1}^n - r_{i-1}^n}{2\Delta x} - \nu_u \Delta t \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{\Delta x^2} = \omega \Delta t [\theta_1 v_i^n + (1-\theta_1) v_i^{n+1}], \\ v_i^{n+1} - v_i^n = -\omega \Delta t u_i^{n+1}. \end{cases}$$

Remark 1.8. *Let us mention about the relation between the Von Neumann condition and the L^2 stability. We begin with the standard Godunov scheme ($\kappa_r = \kappa_u = 1$). In this case, the amplification matrix of the first step (acoustic step) is a normal matrix. We can obtain the L^2 stability for this step due to the fact that in this case the Von Neumann condition is not only necessary but also sufficient. We mention [20] for this stability condition. On the other hand, we have the L^2 stability for the Coriolis step when $\theta_1 = \theta_2 \leq \frac{1}{2}$. As a result, we can ensure L^2 stability for the standard Godunov splitting scheme.*
However, the numerical viscosity on the pressure (κ_r) is a crucial for the inaccuracy of standard Godunov scheme. *To recover the expected accuracy, we have modify this diffusion term in order to obtain* low Froude Godunov scheme *($\kappa_r = 0$) or* All Froude Godunov scheme *($\kappa_r = M$). This makes the amplification matrix for the first step is no more symmetric. As a consequence, the Von Neumann condition is just necessary condition for L^2 stability (see [36] for details). Therefore, in this case, to find the stability condition of the splitting scheme, we have to investigate the Von Neumann analysis for the amplification matrix combined by two steps instead of using the stability condition satisfied for both steps. In fact, from the numerical point of view, the modified splitting scheme is unstable when $\theta_2 > 0$ (see Figure (1.9)).*

1.B Discrete Hodge decomposition

For the *Low Froude Godunov scheme* ($\kappa_r = 0$), the discrete kernel defined at the center of each cell is given by

$$\mathcal{E}^h_{\omega \neq 0} = \left\{ \hat{q}_h = (\hat{r}_h, \hat{u}_h, \hat{v}_h) \in \mathbb{R}^{3N} : u_i = 0, \frac{a_\star}{2\Delta x}(r_{i+1} - r_{i-1}) = \omega v_i \right\} \tag{1.B.2}$$

and then we have the following result

Lemma 1.8. *The orthogonal space of $\mathcal{E}^h_{\omega\neq 0}$ is given by*

$$\mathcal{E}^{h,\perp}_{\omega\neq 0} = \left\{ \tilde{q}_h = (\tilde{r}_h, \tilde{u}_h, \tilde{v}_h) \in \mathbb{R}^{3N} : a_\star \frac{(\tilde{v}_{i+1} - \tilde{v}_{i-1})}{2\Delta x} = \omega \tilde{r}_i \right\}.$$

Moreover, we also have the discrete Hodge decomposition $\mathcal{E}^h_{\omega\neq 0} \oplus \mathcal{E}^{h,\perp}_{\omega\neq 0} = \mathbb{R}^{3N}$.

Proof. Let us denote that

$$\mathbb{A}_h = \left\{ \tilde{q}_h = (\tilde{r}_h, \tilde{u}_h, \tilde{v}_h) \in \mathbb{R}^{3N} : \frac{a_\star}{2\Delta x}(\tilde{v}_{i+1} - \tilde{v}_{i-1}) = \omega \tilde{r}_i \right\}.$$

To begin with, we take an arbitrary $\tilde{q}_h = (\tilde{r}_h, \tilde{u}_h, \tilde{v}_h) \in \mathbb{R}^{3N}$, then $\forall \hat{q}_h = (\hat{r}_h, \hat{u}_h, \hat{v}_h) \in \mathcal{E}^h_{\omega\neq 0}$, by using periodic boundary condition, we obviously get

$$\begin{aligned} \langle \tilde{q}_h, \hat{q}_h \rangle &= \sum_{i=1}^N \Delta x \, (\tilde{r}_i \hat{r}_i + \tilde{v}_i \hat{v}_i) = \sum_{i=1}^N \Delta x \tilde{r}_i \hat{r}_i + \tilde{v}_i \frac{a_\star}{2\omega}(\hat{r}_{i+1} - \hat{r}_{i-1}) \\ &= \sum_{i=1}^N \Delta x \left[\tilde{r}_i - a_\star \frac{(\tilde{v}_{i+1} - \tilde{v}_{i-1})}{2\Delta x \omega} \right] \hat{r}_i. \end{aligned}$$

Therefore, if $\tilde{q}_h \in \mathbb{A}_h$ we clearly obtain $\langle \tilde{q}_h, \hat{q}_h \rangle = 0 \ \forall \hat{q}_h \in \mathcal{E}^h_{\omega\neq 0}$ which leads to $\tilde{q}_h \in \mathcal{E}^{h,\perp}_{\omega\neq 0}$. This means that $\mathbb{A}_h$ is a subset of $\mathcal{E}^{h,\perp}_{\omega\neq 0}$. On the other hand, if $\tilde{q}_h \in \mathcal{E}^{h,\perp}_{\omega\neq 0}$, the equation $\langle \tilde{q}_h, \hat{q}_h \rangle = 0, \quad \forall \hat{q}_h \in \mathcal{E}^h_{\omega\neq 0}$ implies that $\tilde{q}_h \in \mathbb{A}_h$. In other words, $\mathcal{E}^{h,\perp}_{\omega\neq 0}$ is a subset of $\mathbb{A}_h$. For those reasons, we conclude that $\mathcal{E}^{h,\perp}_{\omega\neq 0} = \mathbb{A}_h$.
Since we clearly have $\mathcal{E}^h_{\omega\neq 0} \oplus \mathcal{E}^{h,\perp}_{\omega\neq 0} \subset \mathbb{R}^{3N}$, we only have to prove that each element in $\mathbb{R}^{3N}$ can be written as

$$q_h = \hat{q}_h + \tilde{q}_h \quad \text{where} \hat{q}_h \in \mathcal{E}^h_{\omega\neq 0} \quad \text{and} \quad \tilde{q}_h \in \mathcal{E}^{h,\perp}_{\omega\neq 0}.$$

Now, $\forall (\tilde{p}_h, \tilde{s}_h, \tilde{w}_h) \in \mathcal{E}^{h,\perp}_{\omega\neq 0}$, by using the orthogonality between $\mathcal{E}^h_{\omega\neq 0}$ and $\mathcal{E}^{h,\perp}_{\omega\neq 0}$, we obtain

$$\langle \hat{r}_h, \tilde{p}_h \rangle + \langle \hat{v}_h, \tilde{w}_h \rangle = 0 \Rightarrow \langle \tilde{v}_h, \tilde{w}_h \rangle + \langle \tilde{r}_h, \tilde{p}_h \rangle = \langle r_h, \tilde{p}_h \rangle + \langle v_h, \tilde{w}_h \rangle.$$

We now denote $\alpha = \frac{a_\star}{2\omega\Delta x}$ and by the definition of the discrete kernel and orthogonal subspace, the above equation can be written as

$$\langle \tilde{v}_h, \tilde{w}_h \rangle + \alpha^2 \sum_{i=1}^N \Delta x (\tilde{v}_{i+1} - \tilde{v}_{i-1})(\tilde{w}_{i+1} - \tilde{w}_{i-1}) = \langle v_h, \tilde{w}_h \rangle + \alpha \sum_{i=1}^N \Delta x r_i (\tilde{w}_{i+1} - \tilde{w}_{i-1})$$

In consideration of the periodic boundary condition, this equation is equivalent to

$$\langle \tilde{v}_h, \tilde{w}_h \rangle - \alpha^2 \sum_{i=1}^N \Delta x \tilde{w}_i (\tilde{v}_{i+2} - 2\tilde{v}_i + \tilde{v}_{i-2}) = \langle v_h, \tilde{w}_h \rangle - \alpha \sum_{i=1}^N \Delta x \tilde{w}_i (r_{i+1} - r_{i-1}).$$

We now choose the special $(\tilde{p}_h, \tilde{s}_h, \tilde{w}_h) \in \mathcal{E}^{h,\perp}_{\omega\neq 0}$ such that $\tilde{w}_i = 1$ and $\tilde{w}_{j\neq i} = 0$ to obtain

$$\tilde{v}_i - \alpha^2 (\tilde{v}_{i+2} - 2\tilde{v}_i + \tilde{v}_{i-2}) = v_i - \alpha (r_{i+1} - r_{i-1}). \tag{1.B.3}$$

By doing the same way, we also obtain

$$\hat{r}_i - \alpha^2 (\hat{r}_{i+2} - 2\hat{r}_i + \hat{r}_{i-2}) = r_i - \alpha (v_{i+1} - v_{i-1}). \tag{1.B.4}$$

As a result, we can find $\hat{r}$ and $\tilde{v}$ by solving the following linear systems

$$\mathcal{A}\hat{r} = \mathcal{B}_r \quad \text{and} \quad \mathcal{A}\tilde{v} = \mathcal{B}_v$$

where the matrix A, B_r and B_v are respectively given by

$$\mathcal{A} = \begin{pmatrix} 1+2\alpha^2 & 0 & -\alpha^2 & 0 & 0 & ..... & 0 & 0 & -\alpha^2 & 0 \\ 0 & 1+2\alpha^2 & 0 & -\alpha^2 & 0 & ..... & 0 & 0 & 0 & -\alpha^2 \\ -\alpha^2 & 0 & 1+2\alpha^2 & 0 & -\alpha^2 & ..... & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & & & \cdots & \vdots & \vdots & \vdots & \vdots \\ -\alpha^2 & 0 & 0 & 0 & 0 & \cdots & -\alpha^2 & 0 & 1+2\alpha^2 & 0 \\ 0 & -\alpha^2 & 0 & 0 & 0 & \cdots & 0 & -\alpha^2 & 0 & 1+2\alpha^2 \end{pmatrix},$$

$$\mathcal{B}_r = \begin{pmatrix} r_1 - \alpha(v_2 - v_0) \\ r_2 - \alpha(v_3 - v_1) \\ r_3 - \alpha(v_4 - v_2) \\ \vdots \\ r_{N-1} - \alpha(v_N - v_{N-2}) \\ r_N - \alpha(v_{N+1} - v_{N-1}) \end{pmatrix} \quad \text{and} \quad \mathcal{B}_v = \begin{pmatrix} v_1 - \alpha(r_2 - r_0) \\ v_2 - \alpha(r_3 - r_1) \\ v_3 - \alpha(r_4 - r_2) \\ \vdots \\ v_{N-1} - \alpha(r_N - r_{N-2}) \\ v_N - \alpha(r_{N+1} - r_{N-1}) \end{pmatrix}.$$

We note that the matrix $\mathcal{A}$ is a M-matrix, so it is invertible. Therefore, we have the uniqueness of $\hat{r}_h$ and $\tilde{v}_h$ and then we can easily obtain $\hat{v}_h$ and $\tilde{r}_h$ by using the definition of the discrete kernel and orthogonal subspace. We can also check again $r_h = \hat{r}_h + \tilde{r}_h$ and $v_h = \hat{v}_h + \tilde{v}_h$. □

Remark 1.9. *Let us note that when the numerical solution $q_h \in \mathcal{E}^h_{\omega\neq 0}$, the right hand side of the linear system B_v is equal to 0. From (1.B.3), we obviously have $\tilde{v}_h = 0$ which leads to $\tilde{r}_h = 0$ by using the definition of the orthogonal subspace. On the contrary, when the numerical solution q_h belongs to the orthogonal subspace $\mathcal{E}^{h,\perp}_{\omega\neq 0}$, we will obtain $\hat{r}_h = 0$ from (1.B.4) and $\hat{v}_h = 0$ with the definition of the discrete kernel.*

CHAPTER 2

Analysis of Apparent Topography scheme applied to the linear wave equation with Coriolis source term

The only way of discovering the limits of the possible is to venture a little way past them into the impossible.

Clarke's Second Law.

This work has been done in collaboration with Emmanuel Audusse, Pascal Omnes and Yohan Penel. It has been published in Finite Volumes For Complex Applications VIII, Springer Proceedings in Mathematics.

Abstract

The shallow water equations can be used to model many phenomena in geophysical fluid mechanics. For large scales, the Coriolis force plays an important role and the geostrophic equilibrium which corresponds to the balance between the pressure gradient and the Coriolis force is an important feature. In this communication, we investigate the stability condition and the behavior of the so-called Apparent Topography scheme which is capable of capturing a discrete version of the geostrophic equilibrium.

Chapter content

2.1 Introduction

In order to study the Shallow Water equations with Coriolis source term, we consider the dimensionless formulation on the rotating frame which is given by

$$\begin{cases} \mathsf{St}\partial_t h + \nabla \cdot (h\bar{\mathbf{u}}) = 0, & \text{(2.1a)} \\ \mathsf{St}\partial_t (h\bar{\mathbf{u}}) + \nabla \cdot (h\bar{\mathbf{u}} \otimes \bar{\mathbf{u}}) + \dfrac{1}{\mathsf{Fr}^2} \nabla \left(\dfrac{h^2}{2} \right) = -\dfrac{1}{\mathsf{Fr}^2} h \nabla b - \dfrac{1}{\mathsf{Ro}} h \bar{\mathbf{u}}^{\perp}, & \text{(2.1b)} \end{cases}$$

where unknowns h and $\bar{\mathbf{u}}$ respectively denote the water depth and the average velocity over the water column and function $b(\mathbf{x})$ denotes the topography of the considered oceanic basin and is a given function. Dimensionless numbers St, Fr and Ro respectively stand for the Strouhal, the Froude and the Rossby numbers defined below. In the sequel, we shall focus on cases where

$$\mathsf{St} := \frac{L}{UT} = \mathcal{O}\left(\frac{1}{\mathsf{M}}\right), \quad \mathsf{Fr} := \frac{U}{\sqrt{gH}} = \mathcal{O}(\mathsf{M}), \quad \mathsf{Ro} := \frac{U}{\Omega L} = \mathcal{O}(\mathsf{M}),$$

with M a small parameter. The parameters g and Ω denote the gravity coefficient and the angular velocity of the Earth. Constants U, H, L and T are some characteristic velocity, vertical and horizontal lengths and time. These orders of magnitude correspond to the study of short-time dynamics and standard conditions for large scale oceanic flows.

For data independent of y and with a flat topography, the solution of System (2.1) then satisfies at the leading order the quasi-1d linear wave equation with Coriolis source term (see [37] for the derivation)

$$\begin{cases} \partial_t r + a_\star \partial_x u = 0, \\ \partial_t u + a_\star \partial_x r = \omega v, \\ \partial_t v = -\omega u, \end{cases} \tag{2.2}$$

where $a_\star$ and ω are constants of order $\mathcal{O}(1)$ – respectively related to the wave velocity and to the rotating velocity – r is the first order perturbation of the water depth h and (u, v) is the leading order for the velocity field. The stationary state corresponding to System (2.2) is the 1d version of the so-called *geostrophic equilibrium* and is given by

$$u = 0, \quad a_\star \partial_x r = \omega v. \tag{2.3}$$

A first study of the accuracy of numerical schemes applied to system (2.2) for initial data that are close to the kernel (2.3) was performed in [37]. It was shown that the standard Godunov scheme applied to the linear wave equation with Coriolis source term is inaccurate at low Froude number and the numerical viscosity on the pressure equation is the main reason for this inaccuracy. A modified *low Froude* Godunov scheme was proposed to cure the problem. The scheme was shown to be L^2 stable under a suitable CFL condition. The proofs extend the ideas introduced in [20] for the study of the homogeneous wave equations in *low Mach* number regimes.

In this paper, our objective is to study in the same context the numerical scheme introduced in [13] as a well-balanced (WB) scheme for the Shallow Water equations with Coriolis source term (2.1). In particular we prove the L^2 stability of the scheme under suitable CFL conditions. Moreover we compare this scheme, called *apparent topography* scheme in the following, and the *low Froude* one in terms of dispersion relations and accuracy for some test cases.Note that a high order extension of the *apparent topography* scheme for the non-linear SW equations with Coriolis source term has been studied in [27], where the authors also paid attention to the linear dispersion relation (hence related to (2.2)).

2.2 The numerical schemes

Both *low Froude* and *apparent topography* schemes are colocated finite volume schemes and can be interpreted as a way to modify the numerical diffusion of the classical Godunov scheme on the pressure equation. In the *low Froude* scheme proposed in [37], the numerical diffusion on the pressure equation is simply deleted. In the *apparent topography* scheme introduced in [13], the diffusion term of the classical Godunov scheme remains and an additional consistent term is introduced in the pressure equation such that the numerical diffusion vanishes when applied to an element of the linear kernel (2.3). The name *apparent topography* comes from the fact that the scheme was first developed in the context of WB methods for the shallow water equation with topography, see[5]. The two aforementioned semi-discrete schemes applied to (2.2) read

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t} r_j + a_\star \frac{u_{j+1}-u_{j-1}}{2\Delta x} - \frac{\kappa_r a_\star \Delta x}{2} \frac{r_{j+1}-2r_j+r_{j-1}}{\Delta x^2} + \frac{\kappa_r \omega}{2} \frac{v_{j+1}-v_{j-1}}{2} = 0, \\ \frac{\mathrm{d}}{\mathrm{d}t} u_j + a_\star \frac{r_{j+1}-r_{j-1}}{2\Delta x} - \frac{\kappa_u a_\star \Delta x}{2} \frac{u_{j+1}-2u_j+u_{j-1}}{\Delta x^2} = \omega f(v_{j-1}, v_j, v_{j+1}), \\ \frac{\mathrm{d}}{\mathrm{d}t} v_j = -\omega f(u_{j-1}, u_j, u_{j+1}). \end{cases} \tag{2.4}$$

where *Low Froude* scheme corresponds to

$$\kappa_r = 0 \quad \text{and} \quad f(x,y,z) = y$$

and the *Apparent Topography* scheme to

$$\kappa_r = \kappa_u \quad \text{and} \quad f(x,y,z) = \frac{x+2y+z}{4}.$$

.

2.2.1 Study of the semi-discrete scheme - Dispersion relations

We now study the stability of the semi-discrete Godunov type schemes by means of Fourier modes:

$$r_j(t) = \varphi_r(t) e^{\boldsymbol{i} k x_j}, \quad u_j(t) = \varphi_u(t) e^{\boldsymbol{i} k x_j} \quad \text{and} \quad v_j(t) = \varphi_v(t) e^{\boldsymbol{i} k x_j}.$$

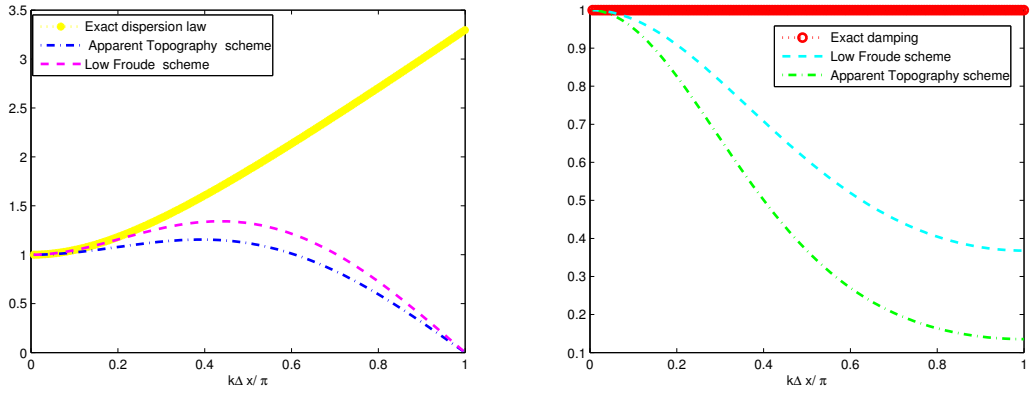
Substituting these expressions into (2.4), we obtain the following linear system of differential equations

$$\begin{pmatrix} \varphi_r'(t) \\ \varphi_u'(t) \\ \varphi_v'(t) \end{pmatrix} + \begin{pmatrix} \kappa_r a_\star \frac{\sin^2(\frac{k\Delta x}{2})}{\frac{\Delta x}{2}} & i a_\star \frac{\sin(k\Delta x)}{\Delta x} & i \frac{\kappa_r \omega \Delta x}{2} \frac{\sin(k\Delta x)}{\Delta x} \\ i a_\star \frac{\sin(k\Delta x)}{\Delta x} & \kappa_u a_\star \frac{\sin^2(k\Delta x/2)}{\Delta x/2} & -\omega\zeta \\ 0 & \omega\zeta & 0 \end{pmatrix} \begin{pmatrix} \varphi_r(t) \\ \varphi_u(t) \\ \varphi_v(t) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \tag{2.5}$$

where $\zeta = 1$ for the *low Froude* scheme and $\zeta = \cos^2(\frac{k\Delta x}{2})$ for the *apparent topography* scheme. The first eigenvalue of the amplification matrix is $\lambda = 0$, corresponding to the discrete stationary state (2.3). The other two, corresponding to the inertia-gravity modes, are given in Table 2.1. Their real part $\Re(\lambda)$ characterizes the decay of Fourier modes k. Since $\Re(\lambda) > 0$, both *low Froude* and *apparent topography* schemes are damping. The damping rate of the *apparent topography* scheme is twice larger than that of the *low Froude* scheme. The imaginary part $\Im(\lambda)$ characterizes the propagation properties of the Fourier modes. Note that for the *low Froude* scheme, the eigenvalues may be real for $k\Delta x$ close to π which means the corresponding modes do not propagate and are only damped. For small $a_* k/\omega$, the dispersion relation $\Im(\lambda)/\omega$ of the *low Froude* scheme is closer to the exact one (for the rotating wave equation (2.2)) whereas the converse holds for large $a_* k/\omega$.

Wave equation	$\pm i\sqrt{a_\star^2 k^2+\omega^2}$
Low Froude	$a_\star \frac{\kappa_u}{2}\frac{\sin^2(\frac{k\Delta x}{2})}{\frac{\Delta x}{2}} \pm i\sqrt{a_\star^2\left(\frac{\sin(k\Delta x/2)}{\Delta x/2}\right)^2\left[\cos^2(\frac{k\Delta x}{2})-(\frac{\kappa_u}{2})^2\sin^2(\frac{k\Delta x}{2})\right]+\omega^2}$
Apparent Topography	$a_\star\kappa_u\frac{\sin^2(\frac{k\Delta x}{2})}{\frac{\Delta x}{2}} \pm i\sqrt{a_\star^2\left(\frac{\sin(k\Delta x)}{\Delta x}\right)^2+\omega^2\left(\frac{1+\cos(k\Delta x)}{2}\right)^2}$

Table 2.1: *The eigenvalues corresponding to the inertia-gravity modes for small $k\Delta x$.*



(a) Dispersion relation $\Im(\lambda)/\omega$ **(b)** Damping error $e^{-\Re(\lambda)}$

Figure 2.1: *Numerical properties of the semi-discrete schemes with the Rossby deformation $R_d := \frac{a_\star}{\omega} = \Delta x$ and $(\kappa_r,\kappa_u) = (0,1)$ for LF, $(\kappa_r,\kappa_u) = (1,1)$ for AT.*

2.3 Study of the fully discrete scheme: kernel and L^2-stability

The fully discrete *apparent topography* scheme applied to (2.2) can be written as

$$\begin{cases} \frac{r_j^{n+1}-r_j^n}{\Delta t} + a_\star\frac{u_{j+1}^n-u_{j-1}^n}{2\Delta x} - \frac{\kappa_r a_*\Delta x}{2}\frac{r_{j+1}^n-2r_j^n+r_{j-1}^n}{\Delta x^2} + \frac{\kappa_r\omega}{2}\frac{(v_{j+1}^n-v_{j-1}^n)}{2} = 0, \\ \frac{u_j^{n+1}-u_j^n}{\Delta t} + a_\star\frac{r_{j+1}^n-r_{j-1}^n}{2\Delta x} - \frac{\kappa_u a_*\Delta x}{2}\frac{u_{j+1}^n-2u_j^n+u_{j-1}^n}{\Delta x^2} \\ \qquad = \omega\left[\theta_1\frac{v_{j+1}^n+2v_j^n+v_{j-1}^n}{4} + (1-\theta_1)\frac{v_{j+1}^{n+1}+2v_j^{n+1}+v_{j-1}^{n+1}}{4}\right], \\ \frac{v_j^{n+1}-v_j^n}{\Delta t} = -\omega\left[\theta_2\frac{u_{j+1}^n+2u_j^n+u_{j-1}^n}{4} + (1-\theta_2)\frac{u_{j+1}^{n+1}+2u_j^{n+1}+u_{j-1}^{n+1}}{4}\right] \end{cases} \tag{2.6}$$

for $j\in\{1,\dots,N\}$ and $0\le\theta_1,\theta_2\le 1$. Setting $q=(r,u,v)$, periodic boundary conditions read

$$q_0^{n+1} = q_N^{n+1}, q_{N+1}^{n+1} = q_1^{n+1}.$$

For practical reasons, we assume that the cell number N is odd.

2.3.1 Analysis of the discrete kernel and orthogonal space

Lemma 2.1. *The kernel of the Apparent Topography scheme (2.6) is given by*

$$\mathcal{E}^h_{\omega\neq 0} = \ker L_{\kappa,h} = \left\{ q=(r,u,v) \;\middle|\; u_j = 0,\ a_\star\frac{r_{j+1}-r_j}{\Delta x} = \omega\frac{v_{j+1}+v_j}{2} \right\}. \tag{2.7}$$

Proof. A stationary state consists in $r_j^{n+1} = r_j^n$, $u_j^{n+1} = u_j^n$ and $v_j^{n+1} = v_j^n$. Therefore, we obtain

$$\begin{cases} a_\star \frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} - \frac{\kappa_r a_\star \Delta x}{2} \frac{r_{j+1}^n - 2r_j^n + r_{j-1}^n}{\Delta x^2} + \frac{\kappa_r \omega}{2} \frac{v_{j+1}^n - v_{j-1}^n}{2} = 0, \\ a_\star \frac{r_{j+1}^n - r_{j-1}^n}{2\Delta x} - \frac{\kappa_u a_\star \Delta x}{2} \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{\Delta x^2} = \omega \frac{v_{j+1}^n + 2v_j^n + v_{j-1}^n}{4}, \\ 0 = -\omega \frac{u_{j+1}^n + 2u_j^n + u_{j-1}^n}{4}. \end{cases} \tag{2.8}$$

As a result, we have the homogeneous linear recurrence $u_{j+1}^n + 2u_j^n + u_{j-1}^n = 0$. This implies that the solution of this recurrence relation is

$$u_j^n = (\alpha + \beta j)(-1)^j, \quad \forall j \in \{0, \dots, N+1\}$$

where α and β are constants.
If we consider N is an even number, by using periodic boundary condition, we will obtain $u_j^n = \alpha(-1)^{-j}$ which is the checkerboard mode. In the other case, when N is an odd number, the periodic boundary condition leads to

$$\alpha = -(\alpha + \beta N) \quad \text{and} \quad -(\alpha + \beta) = \alpha + \beta(N+1).$$

It follows that $\alpha = \beta = 0$. As a result, we have $u_j^n = 0, \quad \forall j \in \{0, \dots, N+1\}$. We then deduce from (2.8)

$$a_\star \frac{r_{j+1}^n - 2r_j^n + r_{j-1}^n}{\Delta x} = \omega \frac{v_{j+1}^n - v_{j-1}^n}{2} \quad \text{and} \quad a_\star \frac{r_{j+1}^n - r_{j-1}^n}{\Delta x} = \omega \frac{v_{j+1}^n + 2v_j^n + v_{j-1}^n}{2}.$$

Summing the two equations yields the discrete kernel (2.7). Conversely, any element satisfying (2.7) is a stationary state of relations (2.6). This discrete kernel is a consistent discretization, defined at the cell interfaces, of the continuous kernel (2.3). □

Remark 2.1. *Let us recall that the discrete kernel of the low Froude scheme is*

$$u_j = 0, \qquad a_\star \frac{r_{j+1} - r_{j-1}}{2\Delta x} = \omega v_j$$

(see [37]) which is another consistent discretization, defined at the cell centers, of the continuous kernel (2.3).

Remark 2.2. *When the number of points is even, checkerboard modes for velocity u may exist in the discrete kernel of the apparent topography scheme. Note that the low Froude scheme suffers the same drawback, but for the pressure r.*

Lemma 2.2. *The orthogonal space of $\mathcal{E}^h_{\omega \neq 0}$ is verified by*

$$\mathcal{E}^{h,\perp}_{\omega \neq 0} = \left\{ \tilde{q} = (\tilde{r}, \tilde{u}, \tilde{v}) \in \mathbb{R}^{3N} : \frac{a_\star}{\Delta x}(\tilde{v}_{i+1} - \tilde{v}_i) = \omega \frac{\tilde{r}_{i+1} + \tilde{r}_i}{2} \right\}.$$

Proof. We denote

$$\mathbb{A}_h = \left\{ \tilde{q} = (\tilde{r}, \tilde{u}, \tilde{v}) \in \mathbb{R}^{3N} : \frac{a_\star}{\Delta x}(\tilde{v}_{i+1} - \tilde{v}_i) = \omega \frac{\tilde{r}_{i+1} + \tilde{r}_i}{2} \right\}.$$

We firstly show that $\mathbb{A}_h$ is a subset of $\mathcal{E}^{h,\perp}_{\omega\neq 0}$. Let $\tilde{q}=(p,s,w)\in\mathbb{A}_h$, then $\forall q=(r,u,v)\in\mathcal{E}^{h}_{\omega\neq 0}$, we have

$$\langle \tilde{q},q\rangle \quad = \quad \sum_{i=1}^{N}\Delta x p_i r_i+\Delta x v_i w_i \tag{2.9}$$

Moreover, by using periodic boundary condition, we obtain

$$\begin{aligned}\sum_{i=1}^{N}\Delta x v_i w_i &= \frac{1}{2}\left(\sum_{i=1}^{N}\Delta x v_i w_i+\sum_{i=1}^{N}\Delta x v_{i+1}w_{i+1}\right)=\frac{1}{2}\sum_{i=1}^{N}\Delta x(v_i+v_{i+1})w_{i+1}-\frac{1}{2}\sum_{i=1}^{N}\Delta x v_i(w_{i+1}-w_i)\\ &=\sum_{i=1}^{N}\frac{a_\star}{\omega}(r_{i+1}-r_i)w_{i+1}-\sum_{i=1}^{N}\Delta x\frac{v_i}{2}(w_{i+1}-w_i)\\ &=-\sum_{i=1}^{N}\frac{a_\star}{\omega}(w_{i+1}-w_i)r_i-\sum_{i=1}^{N}\Delta x\frac{v_i}{2}(w_{i+1}-w_i)=-\sum_{i=1}^{N}\Delta x\left(r_i+\frac{\omega\Delta x}{2a_\star}v_i\right)\frac{a_\star}{\omega}\frac{w_{i+1}-w_i}{\Delta x}.\end{aligned} \tag{2.10}$$

Similarly, we also have

$$\begin{aligned}\sum_{i=1}^{N}\Delta x r_i p_i &= \frac{1}{2}\left(\sum_{i=1}^{N}\Delta x r_i p_i+\sum_{i=1}^{N}\Delta x r_{i+1}p_{i+1}\right)=\sum_{i=1}^{N}\Delta x\frac{(p_i+p_{i+1})}{2}r_i+\sum_{i=1}^{N}\Delta x p_{i+1}\frac{\omega\Delta x}{a_\star}\frac{(v_{i+1}+v_i)}{2}\\ &=\sum_{i=1}^{N}\Delta x\frac{(p_i+p_{i+1})}{2}r_i+\sum_{i=1}^{N}\Delta x\frac{\omega\Delta x}{a_\star}v_i\frac{(p_{i+1}+p_i)}{2}=\sum_{i=1}^{N}\Delta x(r_i+\frac{\omega\Delta x}{a_\star}\frac{v_i}{2})\frac{p_{i+1}+p_i}{2}.\end{aligned} \tag{2.11}$$

Therefore, from (2.9),(2.10) and (2.11), we clearly have

$$\langle\tilde{q},q\rangle=\sum_{i=1}^{N}(r_i-\frac{\Delta x}{2}v_i)\left(\frac{p_i+p_{i-1}}{2}-\frac{w_i-w_{i-1}}{\Delta x}\right)=0\Rightarrow\mathbb{A}_h\subset\mathcal{E}^{h,\perp}_{\omega\neq 0}.$$

Next, let us consider $e^{r,i}$ is a vector with a value 1 in the ith coordinate and 0 elsewhere. Then, corresponding to each $e^{r,i}$, we can construct vector $e^{v,i}$ such that $(e^{r,i},0,e^{v,i})$ belongs to the discrete kernel. Particularly, we can define $e^{v,i}$ with a value $-\frac{2a_\star}{\omega\Delta x}$ in the the $(i-1)th$ coordinate, a value $\frac{2a_\star}{\omega\Delta x}$ in the the $(i+1)th$ coordinate and 0 elsewhere. Hence, we can say that we have N degrees of freedom with the discrete kernel. This implies that $dim(\mathcal{E}^{h}_{\omega\neq 0})=N$. We now apply the same strategy for the subset $\mathbb{A}_h$ of $\mathcal{E}^{h,\perp}_{\omega\neq 0}$ in order to have N degrees of freedom for $\tilde{v}$. Moreover, it is obvious to see that we can have N degrees of freedom for $\tilde{u}$. As a result, we obtain

$$dim(\mathbb{A}_h)=2N\Rightarrow\mathbb{A}_h=\mathcal{E}^{h,\perp}_{\omega\neq 0}.$$

□

Remark 2.3. *We have the discrete Hodge decomposition $\mathbb{R}^{3N}=\mathcal{E}^{h}_{\omega\neq 0}\oplus\mathcal{E}^{h,\perp}_{\omega\neq 0}$ which means that an element q in $\mathbb{R}^{3N}$ can be decomposed into*

$$q=\hat{q}+\tilde{q}\quad \textit{where}\quad \hat{q}\in\mathcal{E}^{h}_{\omega\neq 0}\quad \textit{and}\quad \tilde{q}\in\mathcal{E}^{h,\perp}_{\omega\neq 0}.$$

This allow us to define projection on the discrete kernel by $\mathbb{P}q=\hat{q}$. Moreover, $\hat{q}$ can be written as

$$\hat{q}=\sum_{i=1}^{N}\alpha_i e_i$$

where the coefficients $\alpha_1,\cdots,\alpha_N$ are real numbers and the set $\{e_i : 1 \le i \le N\}$ is one basis of the discrete kernel. By orthogonality between $\mathcal{E}^h_{\omega\neq 0}$ and $\mathcal{E}^{h,\perp}_{\omega\neq 0}$, vector $\alpha = (\alpha_1,\cdots,\alpha_N)$ is simply verified by solving the following linear system

$$\langle q, e_j\rangle = \sum_{i=1}^{N} \alpha_i \langle e_i, e_j\rangle \quad \forall j \in \{1,\cdots,N\}.$$

2.3.2 Stability condition of the fully discrete scheme

We will now investigate the L^2 stability of the *apparent topography* scheme. Let us first mention that when $0 < \theta_1, \theta_2 < 1$, the *apparent topography* scheme requires to solve a linear system at each time step, which leads to an additional computational cost. On the other hand, the case $\theta_1 = \theta_2 = 1$, that corresponds to a fully explicit scheme, is known to be unstable – see [27]. Therefore, we restrict our study to the two cases $\theta_1 = 0, \theta_2 = 1$ and $\theta_1 = 1, \theta_2 = 0$. Note that in [37], the L^2 stability of the *low Froude* scheme was studied for all values of $(\theta_1,\theta_2) \in [0,1]^2$.

Lemma 2.3. *Under the hypothesis*

$$\kappa_r\kappa_u \le 1 + \frac{\omega^2\Delta x^2}{4a_\star^2}, \tag{2.12}$$

the apparent topography scheme is L^2 stable under the CFL condition

$$\Delta t \le \min\{\Delta t_a, \Delta t_b, \Delta t_c\}$$

where

$$\Delta t_b := min\left\{\frac{1}{\kappa_r}, \frac{1}{\kappa_u}\right\}\frac{\Delta x}{|a_\star|}\ , \qquad \Delta t_c := \frac{2}{\omega}$$

and Δt_a is given by the following cases:

i. When $\theta_2 = 0$,$\theta_1 = 1$, we have

$$\Delta t_a := \frac{\kappa_r + \kappa_u}{2}\frac{\Delta x}{|a_\star|}.$$

ii. When $\theta_2 = 1$, $\theta_1 = 0$, we obtain

$$\Delta t_a := \begin{cases} \dfrac{-\frac{|a_\star|}{\Delta x} + \sqrt{\frac{a_\star^2}{\Delta x^2} + (\kappa_r+\kappa_u)\kappa_r\omega^2}}{\kappa_r\omega^2} & \text{if} \quad \kappa_r \neq 0, \\ \dfrac{\kappa_u}{2}\dfrac{\Delta x}{|a_\star|} & \text{otherwise.} \end{cases}$$

Remark 2.4. *Note that the choice $\kappa_r = 0$ is similar to the low Froude scheme, but with a discretisation of the Coriolis term at the interfaces. We then retrieve the same CFL condition as that of the cell-centered low Froude scheme, see [37].*

Remark 2.5. *Hypothesis (2.12) is not restrictive since the low Froude scheme always satisfies this condition and the classical choice for the apparent topography scheme is to take $\kappa_r = \kappa_u = 1$.*

Remark 2.6. *The bound Δt_c is the classical CFL condition for the inertial oscillations phenomenon.*

Remark 2.7. *The bound Δt_b is one of the classical CFL conditions for the problem without rotation. For $\Delta x \ll 1$, the asymptotic expansion of the bound Δt_a leads to the other classical CFL condition for the problem without rotation*

$$\Delta t_a = \frac{\kappa_r + \kappa_u}{2} \frac{\Delta x}{|a_\star|}.$$

Proof. We perform a Von Neumann analysis to investigate the stability condition. Let us denote

$$\sigma = \frac{\Delta t}{\Delta x}, \quad \gamma = \omega \Delta t, \quad s = \sin\left(\frac{k\Delta x}{2}\right), \quad \mu = \cos^2\left(\frac{k\Delta x}{2}\right) = 1 - s^2.$$

By substituting the discrete Fourier modes $r_j^n = \varphi_r^n e^{ikx_j}$, $u_j^n = \varphi_u^n e^{ikx_j}$ and $v_j^n = \varphi_v^n e^{ikx_j}$ into the fully discrete scheme (2.6), we obtain $\mathcal{A}\varphi^{n+1} = \mathcal{B}\varphi^n$ where the matrices $\mathcal{A}$ and $\mathcal{B}$ are given by

$$\mathcal{A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -(1-\theta_1)\gamma\mu \\ 0 & (1-\theta_2)\gamma\mu & 1 \end{pmatrix}$$

and

$$\mathcal{B} = \begin{pmatrix} 1 - 2\kappa_r|a_\star|\sigma s^2 & -a_\star \sigma i \sin(k\Delta x) & -\frac{\kappa_r \omega \Delta t}{2} i \sin(k\Delta x) \\ -a_\star \sigma i \sin(k\Delta x) & 1 - 2\kappa_u |a_\star| \sigma s^2 & \theta_1 \gamma \mu \\ 0 & -\theta_2 \gamma \mu & 1 \end{pmatrix}.$$

We then search for the eigenvalues of the amplification matrix $\mathcal{C} = \mathcal{A}^{-1}\mathcal{B}$, that are the roots of the third order polynomial $\mathcal{P}(\lambda) = \det(\mathcal{B} - \lambda\mathcal{A})$. Easy computations lead to

$$\mathcal{P}(\lambda) = (1-\lambda)(\Lambda\lambda^2 + \xi\lambda + \zeta) \tag{2.13}$$

with

$$\begin{aligned} \Lambda &= 1 + \gamma^2\mu^2(1-\theta_1)(1-\theta_2) > 0 \\ \xi &= -2 + \gamma^2\mu^2(\theta_1 + \theta_2 - 2\theta_1\theta_2) + 2(\kappa_r + \kappa_u)|a_\star|\sigma s^2 + 2\kappa_r|a_\star|\sigma s^2\gamma^2\mu^2(1-\theta_1)(1-\theta_2) \\ \zeta &= 1 + \gamma^2\mu^2\theta_1\theta_2 - 2(\kappa_r+\kappa_u)|a_\star|\sigma s^2 + 4a_\star^2\sigma^2 s^2(1-s^2) + 4\kappa_r\kappa_u a_\star^2\sigma^2 s^4 \\ &\quad + 2\kappa_r|a_\star|\sigma s^2\gamma^2\mu^2\theta_2(1-\theta_1). \end{aligned}$$

The eigenvalue $\lambda_0 = 1$ corresponds to the discrete kernel (2.7). In order to ensure that the other two roots of (2.13) are in the unit circle ($|\lambda_\pm| \le 1$), the coefficients Λ, ξ and ζ have to satisfy $|\zeta| \le \Lambda$ and $|\xi| \le \Lambda + \zeta$. Computations are then similar to the ones in [37]. More precisely, condition $\zeta \le \Lambda$ will lead to the condition involving Δt_a and condition $|\xi| \le \Lambda + \zeta$ will lead to conditions involving Δt_b and Δt_c. Particularly, we consider the following cases:

- Firstly, the condition $\zeta \le \Lambda$ is equivalent to

$$\begin{aligned} \mathcal{F}_1(s^2) &:= -2(\kappa_r+\kappa_u)|a_\star|\sigma s^2 + 4a_\star^2\sigma^2 s^2(1-s^2) + 4\kappa_r\kappa_u a_\star^2\sigma^2 s^4 \\ &- [1-(\theta_1+\theta_2)]\gamma^2(1-s^2)^2 + 2\kappa_r|a_\star|\sigma s^2\gamma^2\mu^2\theta_2(1-\theta_1) \le 0. \end{aligned}$$

  We now consider the special case $\theta_1 = 1, \theta_2 = 0$. In this case, the stability condition reduces to

$$\mathcal{G}_1(s^2) := -(\kappa_r+\kappa_u) + 2|a_\star|\sigma(1-s^2) + 2\kappa_r\kappa_u|a_\star|\sigma s^2 \le 0.$$

  Therefore, we obtain the time step Δt according to the value of $\kappa_r\kappa_u$

$$\Delta t \le \Delta t_\star := \begin{cases} \frac{\kappa_r+\kappa_u}{2}\frac{\Delta x}{|a_\star|} & \text{if} \quad \kappa_r\kappa_u \le 1 \\ \frac{\kappa_r+\kappa_u}{2\kappa_r\kappa_u}\frac{\Delta x}{|a_\star|} & \text{if} \quad \kappa_r\kappa_u > 1. \end{cases}$$

Next, we consider another special case with $\theta_1 = 0, \theta_2 = 1$. The stability condition becomes

$$\mathcal{G}_1(s^2) := -(\kappa_r + \kappa_u) + 2|a_\star|\sigma(1-s^2) + 2\kappa_r\kappa_u|a_\star|\sigma s^2 + \kappa_r\gamma^2(1-s^2)^2 \leq 0.$$

Moreover, we also have

$$\mathcal{G}_1'(s^2) = -2|a_\star|\sigma + 2\kappa_r\kappa_u|a_\star|\sigma - 2\kappa_r\gamma^2(1-s^2).$$

We now notice that $\mathcal{G}_1'(s^2) = 0$ at $\mathcal{S} = 1 + (1-\kappa_r\kappa_u)\frac{|a_\star|\sigma}{\kappa_r\gamma^2}$. Therefore, if we now assume that $\kappa_r\kappa_u \leq 1$, we need the condition $\mathcal{G}_1(0) \leq 0$. This condition can be written as

$$\kappa_r\gamma^2 + 2|a_\star|\sigma - (\kappa_r + \kappa_u) \leq 0 \quad \Rightarrow \Delta t \leq \frac{-\frac{|a_\star|}{\Delta x} + \sqrt{\frac{a_\star^2}{\Delta x^2} + (\kappa_r + \kappa_u)\kappa_r\omega^2}}{\kappa_r\omega^2}$$

When $\kappa_r\kappa_u > 1$, we have to take into account the condition $\mathcal{G}_1(1) \leq 0$ which is equivalent to

$$-(\kappa_r + \kappa_u) + 2\kappa_r\kappa_u|a_\star|\sigma \leq 0 \quad \Rightarrow \Delta t \leq \frac{(\kappa_r + \kappa_u)}{2\kappa_r\kappa_u}\frac{\Delta x}{|a_\star|}.$$

- Next, the condition $\zeta \geq -\Lambda$ can be written as

$$\begin{aligned}\mathcal{F}_2(s^2) &:= 2 - 2(\kappa_r + \kappa_u)|a_\star|\sigma s^2 + 4a_\star^2\sigma^2 s^2(1-s^2) + 4\kappa_r\kappa_u a_\star^2\sigma^2 s^4 \\ &+ \gamma^2\mu^2[1 - (\theta_1 + \theta_2) + 2\theta_1\theta_2] + 2\kappa_r|a_\star|\sigma s^2\gamma^2\mu^2\theta_2(1-\theta_1) \geq 0.\end{aligned}$$

We shall see below that this constraint is weaker that another one ($\mathcal{F}_4(s^2) \geq 0$) and needs not be taken into account.

- Let us now turn to the condition upon ξ. The first case $-\xi \leq \Lambda + \zeta$ reads

$$\mathcal{F}_3(s^2) := -\gamma^2\mu^2 - 4a_\star^2\sigma^2 s^2(1-s^2) - 4\kappa_r\kappa_u a_\star^2\sigma^2 s^4 - 2\kappa_r|a_\star|\sigma s^2\gamma^2\mu^2(1-\theta_1) \leq 0.$$

This inequality always holds and does not imply an additional constraint upon Δt.

- Finally, we consider the case $\xi \leq \Lambda + \zeta$. This leads to

$$\begin{aligned}\mathcal{F}_4(s^2) &:= \gamma^2\mu^2(1-2\theta_1)(1-2\theta_2) + 4(1 - (\kappa_r + \kappa_u)|a_\star|\sigma s^2) + 4a_\star^2\sigma^2 s^2(1-s^2) \\ &+ 4\kappa_r\kappa_u a_\star^2\sigma^2 s^4 + 2\kappa_r|a_\star|\sigma s^2\gamma^2\mu^2(1-\theta_1)(2\theta_2 - 1) \geq 0.\end{aligned}$$

Let us note that $2\mathcal{F}_2(s^2) - \mathcal{F}_4(s^2) \geq 0$ over $[0,1]$. This implies that the condition $\mathcal{F}_2(s^2) \geq 0$ is a consequence of $\mathcal{F}_4(s^2) \geq 0$. It is obvious to see that the condition to ensure that $\mathcal{F}_4(s^2) \geq 0$ of the case the $\theta_1 = 1, \theta_2 = 0$ is more restrictive than that of the case $\theta_1 = 0, \theta_2 = 1$, so we only consider the special case $\theta_1 = 1, \theta_2 = 0$ and in this case, we have

$$\mathcal{F}_4(s^2) := -\gamma^2\mu^2 + 4(1 - (\kappa_r + \kappa_u)|a_\star|\sigma s^2) + 4a_\star^2\sigma^2 s^2(1-s^2) + 4\kappa_r\kappa_u a_\star^2\sigma^2 s^4$$

which leads to

$$\mathcal{F}_4'(s^2) = 2\gamma^2(1-s^2) - 4(\kappa_r + \kappa_u)|a_\star|\sigma + 4a_\star^2\sigma^2 - 8a_\star^2\sigma^2 s^2 + 8\kappa_r\kappa_u a_\star^2\sigma^2 s^2.$$

Therefore, $\mathcal{F}_4'(s^2) = 0$ at $\mathcal{S} = \frac{-4(\kappa_r+\kappa_u)|a_\star|\sigma + 4a_\star^2\sigma^2 + 2\gamma^2}{8(1-\kappa_r\kappa_u)a_\star^2\sigma^2 + 2\gamma^2}$. Now, if we assume that

$$\kappa_r\kappa_u \leq 1 + \frac{\omega^2\Delta x^2}{4a_\star^2}$$

the minimum value of $\mathcal{F}_4(s^2)$ is either $\mathcal{F}_4(0)$ or $\mathcal{F}_4(1)$. Then,

$$\mathcal{F}_4(0) = -\gamma^2 + 4 \geq 0 \Rightarrow \Delta t \leq \frac{2}{\omega}$$

Next, the condition $\mathcal{F}_4(1) = 4 - 4(\kappa_r + \kappa_u)|a_\star|\sigma + 4\kappa_r\kappa_u a_\star^2\sigma^2 \geq 0$ leads to

$$\Delta t \leq \left[\frac{\kappa_r + \kappa_u}{2} - \frac{|\kappa_r - \kappa_u|}{2}\right]\frac{\Delta x}{a_\star\kappa_r\kappa_u} \quad \Rightarrow \Delta t \leq \min\{\frac{1}{\kappa_r}, \frac{1}{\kappa_u}\}\frac{\Delta x}{a_\star}.$$

□

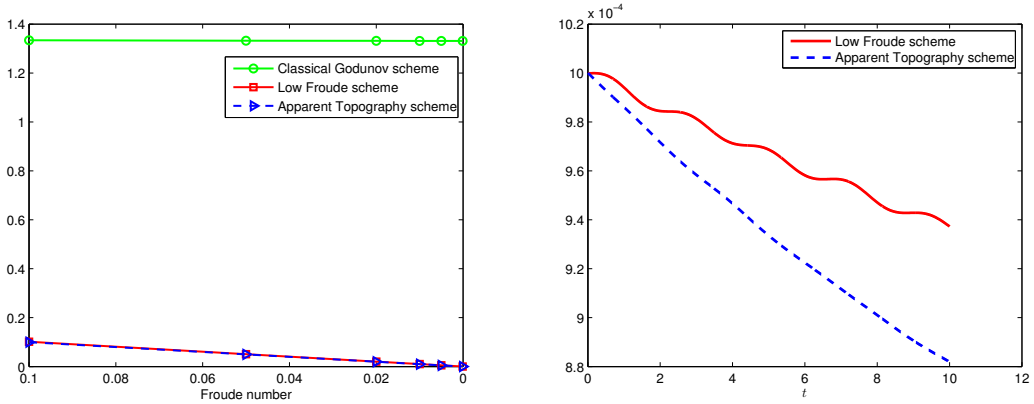
2.4 Numerical results

2.4.1 Accuracy test case

Let us fix the parameters $a_\star = 1$, $\omega = 1$, $\theta_1 = 1, \theta_2 = 0$ and consider the initial condition on the domain $(0, 2\pi)$

$$q_i^0 = \hat{q}_i^0 + M\frac{\tilde{q}_i^0}{\|\tilde{q}_i^0\|} \quad \text{where} \begin{cases} \hat{q}_h^0(x) &= \Big(\sin(\omega x), 0, a_\star\cos(\omega x)\Big) &\in \mathcal{E}^h_{\omega\neq 0}, \\ \tilde{q}_h^0(x) &= \Big(a_\star\cos(\omega x), 1, \sin(\omega x)\Big) &\in \mathcal{E}^{h,\perp}_{\omega\neq 0}, \end{cases}$$

which is close to the kernel $\mathcal{E}^h_{\omega\neq 0}$ up to a perturbation of order M. We solve the 1D linear wave equation (2.2) by means of the *Apparent Topography* scheme (2.6), the *low Froude* scheme and the classical Godunov scheme. We observe on Figure 2.2 (left) that the classical Godunov scheme is inaccurate since the deviation from the kernel does not remain of order M, while the two schemes designed for the geostrophic regime have the correct behaviour as the Froude number goes to 0. We now investigate the accuracy with time at a fixed Froude number. As exhibited for the semi-discrete scheme, we see on Figure 2.2 (right) that the *Apparent Topography* scheme is more diffusive than the *low Froude* scheme for the part of the signal which is in the orthogonal of the kernel.



(a) Maximum in time of the deviation $\|q_h - \hat{q}_h^0\|$ **(b)** Evolution of the deviation $\|q_h - \hat{q}_h^0\|$ for $M = 10^{-3}$ depending on the Froude number

Figure 2.2: *Comparisons of classical and WB schemes*

2.4.2 Stability test case

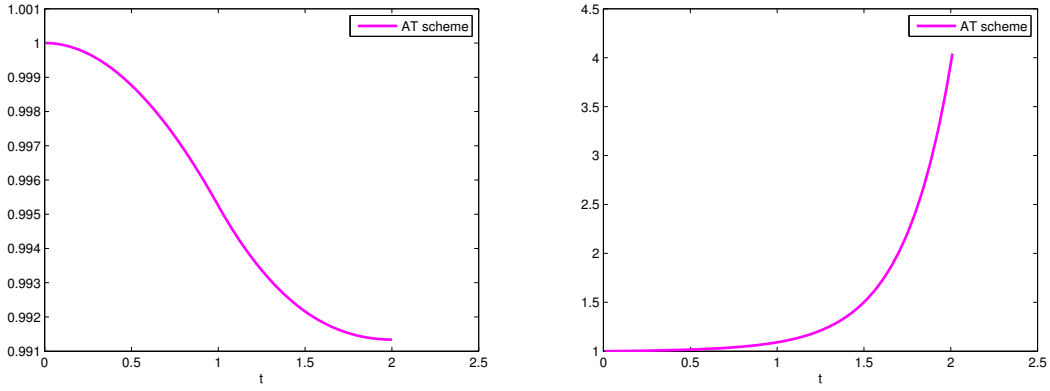
We now turn to another test case when we consider the initial condition given by

$$\begin{cases} r^0 = \chi_{[-\frac{1}{2},\frac{1}{2}]}(x), \quad x \in [-1,1] \\ u^0 = 0, \\ v^0 = 0. \end{cases}$$

In this test case, we choose $\kappa_r = \kappa_u = 1$, $a_\star = \omega = 1$, $\Delta x = 0.01$ and $\theta_1 = 1, \theta_2 = 0$. Therefore, in this case, the the time step of *the Apparent Topography scheme* must satisfies

$$\Delta t \leq \min\{\Delta t_a, \Delta t_b, \Delta t_c\} = \min\{\frac{\Delta x}{a_\star}, \frac{2}{\omega}\} = \Delta x = 0.01.$$

Figure (2.3) present that the the time step is optimal. This is because the Apparent Topography



(a) Energy with $\Delta t = 0.01$ **(b)** Energy with $\Delta t = 0.0101$

Figure 2.3: *Influence of time step upon the Apparent Topography scheme.*

scheme is stable with $\Delta t = 0.01$ while this scheme is unstable with the time step $\Delta t = 0.0101$ which is slightly greater than $\Delta t = 0.01$.

2.5 Conclusion

In this work, we extend the study done in [37] to the *Apparent Topography* scheme proposed in [13]. Particularly, we investigate the stability condition for this scheme as well as we compare their results to the *Low Froude* scheme obtained in [37] in terms of dispersion relations and accuracy. Both schemes are well balanced due to the fact that they can capture the discrete steady state and the kernel of the spatial operator of both schemes correspond to a consistent discretization of the geostrophic equilibrium.

In future works, the authors will apply the two schemes to linear 2D cases before considering nonlinear applications in order to discriminate them.

CHAPTER 3

Analysis of staggered scheme for the linear wave equation with Coriolis force

Our greatest weakness
lies in giving up.
The most certain way to succeed is
always to try just one more time.

Thomas A. Edison.

Abstract

The standard Godunov scheme applied to the linear wave equation with Coriolis source term is inaccurate at low Froude number. By analyzing the kernel of the first order modified equation, the work in [37] shows that the numerical viscosity on the pressure equation is responsible for this inaccuracy problem. One simple correction for this problem is to cancel the numerical diffusion related to the pressure equation to obtain the *Low Froude scheme*. Another correction based on the Apparent Topography method developed in [13] and [25] is to introduce an additional term to this numerical diffusion such that this term will cancel the diffusion at the equilibrium. The present work deals with those strategies to construct well-balanced schemes on staggered meshes. Moreover, a Fourier analysis is performed for a class of well-balanced schemes on staggered and collocated meshes. The numerical dispersion law and damping error are also investigated for the semi-discrete in space and fully discrete staggered type schemes. Finally, stability conditions of the fully discrete scheme are obtained through a Von Neumann analysis.

Chapter content

3.1 Introduction

The shallow water equations (SWE), a simplified 2D model from the 3D incompressible Euler system, is currently used to simulate rivers, coastal flows, dam-break floods and oceans. At large scale, it is important to take into account the Coriolis force coming from the rotation of the Earth and the dimensionless shallow water system in the rotating frame is given by

$$\begin{cases} \partial_t h + \nabla \cdot (h\bar{\mathbf{u}}) = 0, & \text{(3.1a)} \\ St\partial_t(h\bar{\mathbf{u}}) + \nabla \cdot (h\bar{\mathbf{u}} \otimes \bar{\mathbf{u}}) + \dfrac{1}{Fr^2}\nabla\left(\dfrac{h^2}{2}\right) = -\dfrac{1}{Fr^2}h\nabla b - \dfrac{1}{Ro}h\bar{\mathbf{u}}^{\perp} & \text{(3.1b)} \end{cases}$$

In System (3.1) unknowns h and $\bar{\mathbf{u}}$ respectively denote the water depth and the velocity of the water column and function $b(x)$ denotes the topography of the considered oceanic basin and is a given function. Dimensionless numbers St, Fr and Ro respectively stand for the Strouhal, the Froude and the Rossby numbers. In the sequel, we shall focus on cases where

$$Ro = \mathcal{O}(M) \quad \text{and} \quad Fr = \mathcal{O}(M)$$

with M a small parameter. For large scale oceanographic flows, typical values lead to $M \sim 10^{-2}$. Let us now suppose the topography is flat and the solution is independent of the y direction. For a Strouhal number of order $\mathcal{O}(\frac{1}{M})$, i.e for short time, the solution of system (3.1) satisfies at the leading order the quasi-1D linear wave equation with Coriolis term

$$\begin{cases} \partial_t r + a_\star \partial_x u = 0, \\ \partial_t u + a_\star \partial_x r = \omega v, \\ \partial_t v = -\omega u \end{cases} \tag{3.2}$$

where $a_\star$ and ω being constants of order one, respectively related to the wave velocity and to the rotating velocity. We mention [37] for details of this derivation. The stationary state corresponding to Equation (3.2) is the 1D version of the geostrophic equilibrium which is given by

$$u = 0, \quad a_\star \partial_x r = \omega v. \tag{3.3}$$

The ability of a numerical scheme to capture this non trivial steady state currently receives a great attention. In the collocated framework, the Low Froude strategy introduced in [37] and Apparent Topography proposed in [13] by adapting hydrostatic reconstruction [5] are shown to capture well the balance state (3.3). Moreover, the Apparent Topography method applied to the linear wave equation (3.2) is proved to be stable under some CFL conditions in [38] and this strategy is also extended to high order schemes in [27] for the 2D nonlinear shallow water system. However, the dispersion law of these well-balanced schemes on collocated grids is not a monotone function like it is in the continuous system. This property is required to avoid oscillations for the modes with the shortest wavelength $2\Delta x$ and to ensure that the waves move in the correct direction [26].

In the present work, we adapt the Low Froude and Apparent Topography strategy to staggered grids to obtain new schemes which possess all satisfactory properties of the collocated scheme such as preserving the nontrivial steady state and accuracy at low Froude number when the initial condition is around the geostrophic equilibrium. Moreover, the new staggered schemes have much better dispersion relation than the collocated schemes. In section 3.2, we first propose a version of the Low Froude and Apparent Topography schemes on staggered grids. Secondly, we perform the analysis for the semi-discrete staggered scheme with focus on the well-balanced and orthogonality preserving property. The dispersion relation, damping error, group and phase

velocity are also investigated in this section. In Section 3.3, we introduce the time discretization and perform the Fourier analysis for the fully discrete scheme. More importantly, in this section, we show that the proposed staggered schemes are stable under some CFL conditions. Section 3.4 gathers numerical simulations that illustrate our purposes.

3.2 Analysis of the semi-discrete staggered schemes

One of the most common schemes applied to the linear wave equation (3.2) on collocated grid is the leapfrog scheme. This second order accurate scheme can be written as

$$\begin{cases} \frac{r_i^{n+1}-r_i^{n-1}}{2\Delta t} + a_\star \frac{u_{i+1}^n - u_{i-1}^n}{2\Delta x} = 0 \\ \frac{u_i^{n+1}-u_i^{n-1}}{2\Delta t} + a_\star \frac{r_{i+1}^n - r_{i-1}^n}{2\Delta x} = \omega v_i^n \\ \frac{v_i^{n+1}-v_i^{n-1}}{2\Delta t} = -\omega u_i^n. \end{cases} \tag{3.4}$$

It is worth pointing out that in this collocated scheme, all the space derivatives are taken over

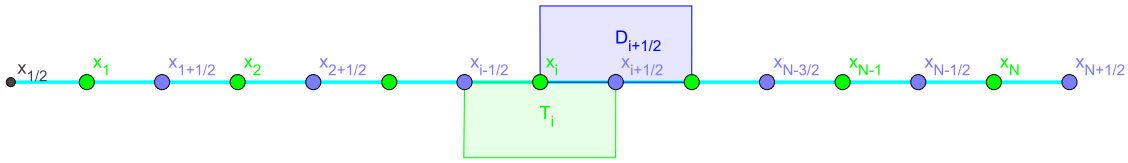


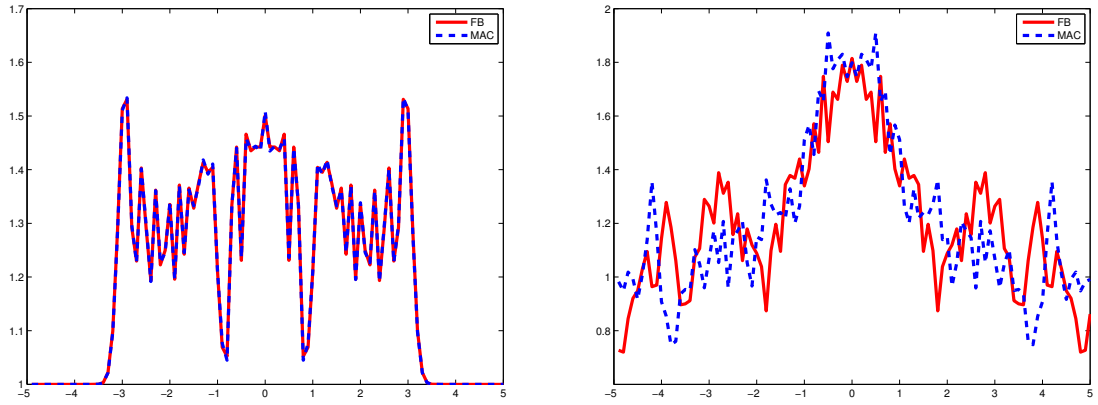
Figure 3.1: *Primary (green) cell and dual (blue) cell.*

the distance $2\Delta x$. However, the distance between adjacent grid nodes is only Δx. Therefore, it is reasonable to use staggered schemes which still ensure second order accuracy in space, but the derivative is performed over the distance Δx. As a result, with a staggered grid (see Fig. 3.1), we obtain a scheme which is more compact than the one defined on a collocated grid. There exist some classical staggered schemes without numerical diffusion than can be applied to the linear wave equation (3.2). For instance, the well known Marker and Cell (MAC) scheme which is staggered not only in space but also in time and the forward-backward scheme obtained by using an explicit discretization in time for one equation and an implicit one for the other equations. The analysis of the MAC and forward-backward is performed in Appendix 3.A. However, the numerical scheme without diffusion terms can introduce some unphysical oscillations for discontinuous initial solutions (see Figure 3.2). Hence, in this section, we will analyze the behavior of the semi discrete staggered scheme with some diffusion terms coming from the Godunov scheme. This staggered scheme can be written as

$$\begin{cases} \frac{d}{dt} r_{i+1/2}(t) + \frac{a_\star}{\Delta x}\left[u_{i+1} - u_i\right] - \frac{\kappa_r a_\star}{2\Delta x}\left[r_{i+3/2} - 2r_{i+1/2} + r_{i-1/2}\right] + \frac{\eta_r \omega}{2}(v_{i+1} - v_i) = 0 \\ \frac{d}{dt} u_i(t) + \frac{a_\star}{\Delta x}\left[r_{i+1/2} - r_{i-1/2}\right] - \frac{\kappa_u a_\star}{2\Delta x}\left[u_{i+1} - 2u_i + u_{i-1}\right] = \omega v_i \\ \frac{d}{dt} v_i(t) = -\omega u_i \end{cases} \tag{3.5}$$

where κ_r, κ_u stand for the parameters of the standard diffusion term and η_r corresponds to the correction term based on the Apparent Topography method introduced in [13]. We also note that the classical staggered scheme (with some additional diffusion) corresponds to $\kappa_r = \kappa_u = 1, \eta_r = 0$, the *Low Froude staggered scheme* corresponds to $\kappa_r = \eta_r = 0$ and the *Apparent Topography staggered scheme* has $\kappa_r = \eta_r > 0$.

Remark 3.1. *Since the stationary state of the* collocated *Apparent Topography scheme is defined as the interface, it is necessary to use more than one cell to approximate the Coriolis source term. We mention the work [38] for more details. However, with the* staggered *Apparent Topography scheme, the Coriolis force is computed by using only one cell.*



(a) $r(x,t)$ at time $t=2$ **(b)** $r(x,t)$ at time $t=400$

Figure 3.2: *The pressure $r(x,t)$ of the forward-backward and MAC scheme with initial fluid at rest ($u^0=v^0=0$) and discontinuous initial height given by $r^0(x)=1+\chi_{[-1,1]}(x)$ on domain $[-5,5]$.*

3.2.1 Discrete operators

To analyze the behavior of the semi-discrete staggered type schemes, it is convenient to construct the discrete version of some differential operators. Let $u_h=(u_i)_{i\in[1,N]}$ be in $\mathbb{R}^N$, we define $\partial_{x,h}u_h$ and $\partial^2_{x^2,h}u_h$ respectively by

$$\forall i\in[1,N]:(\partial_{x,h}u_h)_{i+1/2}:=\frac{u_{i+1}-u_i}{\Delta x}\quad\text{and}\quad\forall i\in[1,N]:(\partial^2_{x^2,h}u_h)_i:=\frac{u_{i+1}-2u_i+u_{i-1}}{\Delta x^2}.$$

Let $r_h=(r_{i+1/2})_{i\in[1,N]}$ be in $\mathbb{R}^N$, we define $\partial_{x,h}r_h$ and $\partial^2_{x^2,h}r_h$ respectively by

$$\forall i\in[1,N]:(\partial_{x,h}r_h)_i:=\frac{r_{i+1/2}-r_{i-1/2}}{\Delta x}\quad\text{and}\quad\forall i\in[1,N]:(\partial^2_{x^2,h}r_h)_{i+1/2}:=\frac{r_{i+3/2}-2r_{i+1/2}+r_{i-1/2}}{\Delta x^2}.$$

In these definitions, we use periodic boundary conditions when needed ($u_{N+1}=u_1$, $u_0=u_N$, $r_{1/2}=r_{N+1/2}$ and $r_{N+3/2}=r_{3/2}$).

Moreover, for $q^1_h=(r^1_h,u^1_h,v^1_h)\in\mathbb{R}^{3N}$ and $q^2_h=(r^2_h,u^2_h,v^2_h)\in\mathbb{R}^{3N}$, let us define the following discrete scalar products

$$\begin{aligned}\langle r^1_h,r^2_h\rangle&=\sum_{i=1}^N\Delta x r^1_{i+1/2}r^2_{i+1/2}\\ \langle u^1_h,u^2_h\rangle&=\sum_{i=1}^N\Delta x u^1_i\cdot u^2_i\\ \langle q^1_h,q^2_h\rangle&=\sum_{i=1}^N\Delta x r^1_{i+1/2}r^2_{i+1/2}+\sum_{i=1}^N\Delta x\boldsymbol{u}^1_i\cdot\boldsymbol{u}^2_i,\end{aligned}$$

where we denote $\boldsymbol{u}^s_i=(u^s_i,v^s_i)$ for $s\in\{1;2\}$ and all i. Although we use the same notations for all three different scalar products, the context in which they will be used in the sequel cannot bring any confusion.

We have the following properties for the above discrete operators. For all $u_h = (u_i)_{i\in[1,N]}$, $v_h = (v_i)_{i\in[1,N]}$ and all $r_h = (r_{i+1/2})_{i\in[1,N]}$, $\phi_h = (\phi_{i+1/2})_{i\in[1,N]}$, it holds that:

Lemma 3.1. *With periodic boundary conditions, the discrete operators satisfy the following discrete integration by part formula:*

$$\langle \partial_{x,h} u_h, r_h\rangle = -\langle \partial_{x,h} r_h, u_h\rangle. \tag{3.6}$$

which also implies that

$$\langle \partial^2_{x^2,h} u_h, v_h\rangle = -\langle \partial_{x,h} u_h, \partial_{x,h} v_h\rangle \quad \text{and} \quad \langle \partial^2_{x^2,h} r_h, \phi_h\rangle = -\langle \partial_{x,h} r_h, \partial_{x,h} \phi_h\rangle.$$

Proof. The result comes from simple direct computations. □

3.2.2 Evolution of the discrete energy

Lemma 3.2. *Let $q_h(t) = (r_h(t), u_h(t), v_h(t))$ be the solution of system (3.5). With $\eta_r = 0$ and the discrete energy defined as follows*

$$E_h(t) = \langle q_h(t), q_h(t)\rangle, \tag{3.7}$$

we have the dissipation of the discrete energy

$$\frac{d}{dt} E_h(t) \le 0.$$

Proof. We take the discrete scalar product of the semi-discrete staggered scheme (3.5) with $q_h(t)$ and we obtain

$$\begin{aligned}\frac{1}{2}\frac{d}{dt} E_h(t) = &-a_\star \langle \partial_{x,h} u_h, r_h\rangle - a_\star \langle \partial_{x,h} r_h, u_h\rangle - \omega \langle \boldsymbol{u}_h^\perp, \boldsymbol{u}_h\rangle \\ &+ \frac{\kappa_r a_\star \Delta x}{2} \langle \partial^2_{x^2,h} r_h, r_h\rangle + \frac{\kappa_u a_\star \Delta x}{2} \langle \partial^2_{x^2,h} u_h, u_h\rangle,\end{aligned}$$

where we denote $\boldsymbol{u}_h^\perp = (-v_h, u_h)$. By using the properties of the discrete operators stated in Lemma 3.1, the above equation implies that

$$\frac{d}{dt} E_h(t) = -\kappa_r a_\star \Delta x \left[\sum_{i=1}^N \Delta x \left(\frac{r_{i+3/2} - r_{i+1/2}}{\Delta x}\right)^2\right] - \kappa_u a_\star \Delta x \left[\sum_{i=1}^N \Delta x \left(\frac{u_{i+1} - u_i}{\Delta x}\right)^2\right] \le 0$$

□

3.2.3 Analysis of the discrete kernel and orthogonal subspace

In this section, we will carry out an analysis of the well-balanced properties of the staggered type schemes. To begin with, let us define the discrete kernel as

$$\mathcal{E}^h_{\omega\neq 0} = \left\{ q_h = (r_h, u_h, v_h) \in \mathbb{R}^{3N} : u_i = 0, \quad a_\star \frac{r_{i+1/2} - r_{i-1/2}}{\Delta x} = \omega v_i \right\}$$

which is exactly the steady state of the staggered scheme without diffusion term, as obviously seen in (3.5) when $\kappa_r = \nu_r = 0$ and $\kappa_u = 0$. In order to figure out whether one numerical scheme can capture this discrete steady state or not, we analyze the discrete kernels of the staggered type schemes. In particular, we have the following results:

Lemma 3.3. *The discrete kernel of the staggered scheme strongly depends on the numerical viscosity in the pressure equation. Particularly, the discrete kernel of the standard scheme ($\kappa_r > 0$, $\eta_r = 0$, and $\kappa_u \geq 0$)*

$$\ker L^h_{\kappa_r\neq 0,\eta_r=0} = \left\{ q_h = (r_h, u_h, v_h) \in \mathbb{R}^{3N} : r_{i+1/2} = \text{const}, u_i = 0, v_i = 0, \forall i \in [1, N] \right\}. \tag{3.8}$$

Moreover, the Low Froude ($\kappa_r = \eta_r = 0$) and the Apparent Topography scheme ($\kappa_r = \eta_r > 0$) have the same discrete kernel which is given by

$$\ker L^h_{\kappa_r=\eta_r} = \left\{ q_h = (r_h, u_h, v_h) \in \mathbb{R}^{3N} : u_i = 0, \quad a_\star \frac{r_{i+1/2} - r_{i-1/2}}{\Delta x} = \omega v_i \right\} = \mathcal{E}^h_{\omega\neq 0}. \tag{3.9}$$

Proof. The semi-discrete staggered scheme (3.5) can be written as

$$\frac{d}{dt} q_h + L^h_{\kappa,\eta} q_h = 0 \tag{3.10}$$

where $q_h = (r_h, u_h, v_h) \in \mathbb{R}^{3N}$, $L^h_{\kappa,\eta} = (L^1_{\kappa,\eta}, .., L^i_{\kappa,\eta}, .., L^N_{\kappa,\eta})^T$ and

$$L^i_{\kappa,\eta} q_h = \begin{pmatrix} \frac{a_\star}{\Delta x}\left[u_{i+1} - u_i\right] - \frac{\kappa_r a_\star}{2\Delta x}\left[r_{i+3/2} - 2r_{i+1/2} + r_{i-1/2}\right] + \frac{\eta_r \omega}{2}(v_{i+1} - v_i) \\ \frac{a_\star}{\Delta x}\left[r_{i+1/2} - r_{i-1/2}\right] - \frac{\kappa_u a_\star}{2\Delta x}\left[u_{i+1} - 2u_i + u_{i-1}\right] - \omega v_i \\ \omega u_i \end{pmatrix}.$$

- In any case, we always have $u_i = 0$ for all $i \in [1, N]$, as soon as $\omega > 0$.
- We then consider the case $\eta_r = 0$ and $\kappa_r \neq 0$. Any element in the kernel verifies $L^h_{\kappa,\eta} q_h = 0$ and thus $\langle L^h_{\kappa,\eta} q_h, q_h \rangle = 0$. The computation performed in Lemma 3.2 leads to

  $$-\kappa_r a_\star \sum_{i=1}^{N} (r_{i+3/2} - r_{i+1/2})^2 - \kappa_u a_\star \sum_{i=1}^{N} (u_{i+1} - u_i)^2 = 0$$

  which implies that $r_h = \text{const}$ when the numerical diffusions are such that $\kappa_r > 0$ and $\kappa_u \geq 0$. Therefore, in this case, the discrete kernel is given by (3.8).
- We now turn to the other cases when $\kappa_r = \eta_r$. With $u_i = 0$, the second condition in $L^h_{\kappa,\eta} q_h = 0$ obviously leads to the fact that $\mathcal{E}^h_{\omega\neq 0} \subset \ker L^h_{\kappa_r=\eta_r}$, and then this is sufficient to obtain the first condition in $L^h_{\kappa,\eta} q_h = 0$, so that (3.9) is verified.

□

Remark 3.2. *For the* collocated *scheme, the kernel of the low Froude scheme introduced in [37] and the Apparent Topography scheme discussed in [38] are different one from the other since the first one is defined at the cell center and the second one is defined at the interface. On the contrary, we have exactly the same discrete kernel with those strategies on a staggered grid.*

Remark 3.3. *The relation (3.8) indicates that* $\ker L^h_{\kappa_r\neq 0,\eta_r=0}$ *is a poor subspace of* $\mathcal{E}^h_{\omega\neq 0}$*, so it is unable to approximate the continuous kernel* $\mathcal{E}_{\omega\neq 0}$*. On the contrary, with relation (3.9), we can say that* $\ker L^h_{\kappa_r=\eta_r}$ *is accurate enough to approximate* $\mathcal{E}_{\omega\neq 0}$ *correctly.*

Lemma 3.4. *The orthogonal space of $\mathcal{E}^h_{\omega\neq 0}$ is given by*

$$\mathcal{E}^{h,\perp}_{\omega\neq 0} = \left\{ q_h = (r_h, u_h, v_h) \in \mathbb{R}^{3N} : \quad a_\star \frac{v_{i+1} - v_i}{\Delta x} = \omega r_{i+1/2} \right\}.$$

This leads to the following discrete Hodge decomposition $\mathbb{R}^{3N} = \mathcal{E}^h_{\omega\neq 0} \oplus \mathcal{E}^{h,\perp}_{\omega\neq 0}$, which means that an element $q_h \in \mathbb{R}^{3N}$ can be decomposed into

$$q_h = \hat{q}_h + \tilde{q}_h \quad \textit{with} \quad \hat{q}_h \in \mathcal{E}^h_{\omega\neq 0} \quad \textit{and} \quad \tilde{q}_h \in \mathcal{E}^{h,\perp}_{\omega\neq 0}.$$

Proof. Let us denote $\mathbb{A}_h$ the set

$$\mathbb{A}_h = \left\{ q_h = (r_h, u_h, v_h) \in \mathbb{R}^{3N} : a_\star \frac{v_{i+1} - v_i}{\Delta x} = \omega r_{i+1/2} \right\}.$$

Next, we consider an arbitrary $\tilde{q}_h = (\tilde{r}_h, \tilde{u}_u, \tilde{v}_h)$, then for all $\hat{q}_h = (\hat{r}_h, \hat{u}_h, \hat{v}_h) \in \mathcal{E}^h_{\omega\neq 0}$, using the properties of $\mathcal{E}^h_{\omega\neq 0}$ and the discrete integration by part formula of Lemma 3.1 we obtain

$$\begin{aligned}
\langle \tilde{q}_h, \hat{q}_h \rangle &= \sum_{i=1}^N \tilde{r}_{i+1/2} \hat{r}_{i+1/2} + \sum_{i=1}^N \tilde{v}_i \hat{v}_i = \sum_{i=1}^N \tilde{r}_{i+1/2} \hat{r}_{i+1/2} + \frac{a_\star}{\omega} \sum_{i=1}^N \tilde{v}_i \frac{\hat{r}_{i+1/2} - \hat{r}_{i-1/2}}{\Delta x} \\
&= \sum_{i=1}^N \hat{r}_{i+1/2} \left(\tilde{r}_{i+1/2} - \frac{a_\star}{\omega} \frac{\tilde{v}_{i+1} - \tilde{v}_i}{\Delta x} \right).
\end{aligned}$$

Therefore, if $\tilde{q}_h$ belongs to $\mathbb{A}_h$, we obviously have $\langle \tilde{q}_h, \hat{q}_h \rangle = 0$ which leads to $\mathbb{A}_h \subset \mathcal{E}^{h,\perp}_{\omega\neq 0}$. Now we consider $\tilde{q}_h \in \mathcal{E}^{h,\perp}_{\omega\neq 0}$. Since $\hat{r}_h$ can be arbitrary in $\mathbb{R}^N$ (as can be seen from the definition of $\mathcal{E}^h_{\omega\neq 0}$), then for each $i \in \{1, \cdots, N\}$ we can choose one special $\hat{q}_h \in \mathcal{E}^h_{\omega\neq 0}$ such that r_h has value 1 in $i+1/2$ and 0 elsewhere to ensure that $\tilde{r}_{i+1/2} - \frac{a_\star}{\omega} \frac{\tilde{v}_{i+1} - \tilde{v}_i}{\Delta x} = 0$. This implies that $\mathcal{E}^{h,\perp}_{\omega\neq 0} \subset \mathbb{A}_h$. To sum up, we have $\mathcal{E}^{h,\perp}_{\omega\neq 0} = \mathbb{A}_h$. □

Remark 3.4. *The discrete Hodge decomposition allows us to define the discrete orthogonal projection*

$$\mathbb{P}_h : \left\{ \begin{array}{ccc} \mathbb{R}^{3N} & \longrightarrow & \mathcal{E}^h_{\omega\neq 0} \\ q_h & \longmapsto & \hat{q}_h \end{array} \right. \tag{3.11}$$

which is computed in the following way: Consider $q_h = (r_h, u_h, v_h) \in \mathbb{R}^{3N}$, and an arbitrary $(\tilde{p}_h, \tilde{s}_h, \tilde{w}_h) \in \mathcal{E}^{h,\perp}_{\omega\neq 0}$. By using the fact that $\hat{u}_h = 0$ and the orthogonality property, we easily obtain

$$\langle \hat{r}_h, \tilde{p}_h \rangle + \langle \hat{v}_h, \tilde{w}_h \rangle = 0 \Rightarrow \langle \tilde{r}_h, \tilde{p}_h \rangle + \langle \tilde{v}_h, \tilde{w}_h \rangle = \langle r_h, \tilde{p}_h \rangle + \langle v_h, \tilde{w}_h \rangle$$

which implies, using the properties of elements of $\mathcal{E}^{h,\perp}_{\omega\neq 0}$

$$\left(\frac{a_\star}{\omega} \right)^2 \langle \partial_{x,h} \tilde{v}_h, \partial_{x,h} \tilde{w}_h \rangle + \langle \tilde{v}_h, \tilde{w}_h \rangle = \frac{a_\star}{\omega} \langle r_h, \partial_{x,h} \tilde{w}_h \rangle + \langle v_h, \tilde{w}_h \rangle.$$

Using the discrete integration by part formula of Lemma 3.1, we obtain

$$\langle \tilde{v}_h, \tilde{w}_h \rangle - \left(\frac{a_\star}{\omega \Delta x} \right)^2 \sum_{i=1}^N \tilde{w}_i (\tilde{v}_{i+1} - 2\tilde{v}_i + \tilde{v}_{i-1}) = \langle v_h, \tilde{w}_h \rangle - \frac{a_\star}{\omega \Delta x} \sum_{i=1}^N \tilde{w}_i (r_{i+1/2} - r_{i-1/2}).$$

We now choose the special $(\tilde{p}_h, \tilde{s}_h, \tilde{w}_h) \in \mathcal{E}^{h,\perp}_{\omega\neq 0}$ *such that* $\tilde{w}_i = 1$ *and* $\tilde{w}_{j\neq i} = 0$ *to obtain*

$$\tilde{v}_i - \left(\frac{a_\star}{\omega\Delta x}\right)^2 (\tilde{v}_{i+1} - 2\tilde{v}_i + \tilde{v}_{i-1}) = v_i - \frac{a_\star}{\omega\Delta x}(r_{i+1/2} - r_{i-1/2}). \tag{3.12}$$

Similarly, we can obtain

$$\hat{r}_{i+1/2} - \left(\frac{a_\star}{\omega\Delta x}\right)^2 (\hat{r}_{i+3/2} - 2\hat{r}_{i+1/2} + \hat{r}_{i-1/2}) = r_{i+1/2} - \frac{a_\star}{\omega\Delta x}(v_{i+1} - v_i). \tag{3.13}$$

As a result, we can find $\tilde{v}_h$ *and* $\hat{r}_h$ *by solving the linear systems* (3.12) *and* (3.13)*. We also note that the matrix of the above linear systems are the same and is an M-matrix, so it is invertible. This implies that* $\hat{r}_h$ *and* $\tilde{v}_h$ *are well defined by* (3.12) *and* (3.13)*. Then, we can easily construct* $\hat{v}_h$ *and* $\tilde{r}_h$ *by using the definition of the discrete kernel and orthogonal subspace.*

3.2.4 Orthogonality preserving property

In the previous section, the well-balanced properties of the staggered type schemes were investigated. We can say that both Low Froude and Apparent Topography schemes can capture the discrete geostrophic equilibrium: they both have a discrete kernel which is a consistent discretization of the continuous kernel (3.3). In other words, it is possible to accurately discretize an initial continuous geostrophic equilibrium and the numerical solution of those well balanced schemes will remain constant at any time.

In this section, we will consider another important aspect which is named "the orthogonality preserving property". At the continuous level, when the initial condition is in the subspace orthogonal to the kernel, the solution of the linear wave equation remains in this orthogonal subspace. Since this is one property of the continuous model, we want to investigate whether this property is also verified at the discrete level with the numerical schemes. If one scheme satisfies this property, we say that it is an *orthogonality preserving scheme.*

Lemma 3.5. *For the staggered type scheme, we have:*

i. The standard staggered scheme and the Apparent Topography scheme ($\kappa_r = \eta_r > 0$) *are not orthogonality preserving schemes.*

ii. The Low Froude staggered scheme ($\kappa_r = \eta_r = 0$) *is an orthogonality preserving scheme.*

Proof. Consider the solution of (3.5) with initial condition $q_h(0) = q_h^0 \in \mathcal{E}^{h,\perp}_{\omega\neq 0}$. and an arbitrary element $\hat{q}_h \in \mathcal{E}^h_{\omega\neq 0}$. If $\frac{d}{dt}\langle q_h(t), \hat{q}_h\rangle = 0$, then $\langle q_h(t), \hat{q}_h\rangle = \langle q_h(0), \hat{q}_h\rangle = 0$.

For all $\hat{q}_h \in \mathcal{E}^h_{\omega\neq 0}$, we have

$$\hat{u}_h = 0 \quad \text{and} \quad a_\star \partial_{x,h}\hat{r}_h = \omega\hat{v}_h.$$

Then, using (3.5) and the discrete integration by part, we obtain

$$\left\langle \frac{d}{dt}q_h(t), \hat{q}_h \right\rangle = -a_\star\langle \partial_{x,h}u_h, \hat{r}_h\rangle - \omega\langle u_h, \hat{v}_h\rangle + \frac{\kappa_r a_\star \Delta x}{2}\langle \partial^2_{x^2,h} r_h, \hat{r}_h\rangle - \frac{\eta_r\omega\Delta x}{2}\langle \partial_{x,h}v_h, \hat{r}_h\rangle$$

$$= \frac{\kappa_r a_\star \Delta x}{2}\langle \partial^2_{x^2,h} r_h, \hat{r}_h\rangle - \frac{\eta_r\omega\Delta x}{2}\langle \partial_{x,h}v_h, \hat{r}_h\rangle.$$

It is important to see that with the standard scheme or the *Apparent Topography staggered scheme*, the right-hand side of the above equation does not vanish. Therefore, Point *(i)* is proved. However, the *Low Froude staggered scheme* imposes $\kappa_r = \eta_r = 0$, and the above right-hand side vanishes. Therefore, we conclude that $q_h(t) \in \mathcal{E}^{h,\perp}_{\omega\neq 0}$ for the solution of the Low Froude staggered scheme. This proves Point *(ii)*. □

3.2.5 Behavior of the solution of the staggered scheme

In this section, we will show some properties of the numerical solution of the staggered type schemes. Particularly, we focus on the accuracy around the discrete geostrophic equilibrium. The analysis in the previous sections helps us understand which kind of scheme can preserve discrete geostrophic equilibria, or their orthogonal states. We will go further with the analysis and consider the case when the initial condition is close to a discrete geostrophic equilibrium and we want to figure out if the numerical solution of the staggered type schemes remains close to the it at any time or at least in short time, which is a property of the continuous system. We will see that the classical scheme does not fulfill this requirement. We mention the works [20–22] for such kind of analysis on collocated meshes using the first order modified equation associated to the homogeneous linear wave equation and [39] for the same equation taking into account porosity effects.

In order to perform the analysis conveniently, we shall denote the parameters of the standard diffusion terms by

$$\nu_r = \frac{\kappa_r a_\star \Delta x}{2} \quad \text{and} \quad \nu_u = \frac{\kappa_u a_\star \Delta x}{2}.$$

and the parameter corresponding to the correction term by

$$\gamma_r = \frac{\eta_r a_\star \Delta x}{2}.$$

Theorem 3.1. *Let $q_{\kappa,h}(t)$ be the solution of (3.5). Then*

i. When $\kappa_r = \eta_r = 0$, if $\|q_h^0 - \mathbb{P}_h(q_h^0)\| = C_1 M$ with $C_1 \in \mathbb{R}^+$, then we have $\forall t \geq 0$, $\|q_{\kappa,h}(t) - \mathbb{P}_h(q_h^0)\| \leq C_1 M$.

ii. Let $\kappa_r = \mathcal{O}(M)$, $\eta_r = 0$, and $\|q_h^0 - \mathbb{P}_h(q_h^0)\| = C_1 M$ with $C_1 > 0$. Then, $\forall C_2 = \mathcal{O}(1) > 0$, there exists $C_3 = \mathcal{O}(1) \in \mathbb{R}^+$ not depending on M such that $\forall t \leq C_2$, we have $\|q_{\kappa,h}(t) - \mathbb{P}_h(q_h^0)\| \leq C_3 M$.

iii. Let $\kappa_r = \mathcal{O}(1)$, $\eta_r = 0$, and $\Delta x = \mathcal{O}(M)$. Let $\|q_h^0 - \mathbb{P}_h(q_h^0)\| = C_1 M$ with $C_1 > 0$. Then, $\forall C_2 = \mathcal{O}(1) > 0$, there exists $C_3 = \mathcal{O}(1) \in \mathbb{R}^+$ not depending on M such that $\forall t \leq C_2$, we have $\|q_{\kappa,h}(t) - \mathbb{P}_h(q_h^0)\| \leq C_3 M$.

Proof. By linearity, the solution of semi-discrete staggered scheme $q_h(t)$ can be written as

$$q_{\kappa,h}(t) = q^a_{\kappa,h}(t) + q^b_{\kappa,h}(t)$$

where $q^a_{\kappa,h}(t)$ is the solution of (3.10) with the initial condition

$$q^a_{\kappa,h}(0) = \mathbb{P}_h(q_h^0)$$

and $q^b_{\kappa,h}(t)$ is the solution of (3.10) with the initial condition

$$q^b_{\kappa,h}(0) = q_h^0 - \mathbb{P}_h(q_h^0).$$

Then, we have

$$\|q_{\kappa,h}(t) - \mathbb{P}_h(q_h^0)\| = \|q^a_{\kappa,h}(t) + q^b_{\kappa,h}(t) - \mathbb{P}_h(q_h^0)\| \leq \|q^a_{\kappa,h}(t) - \mathbb{P}_h(q_h^0)\| + \|q^b_{\kappa,h}(t)\|$$

Moreover, Lemma 3.2 about the dissipation of the semi-discrete staggered scheme when $\eta_r = 0$ leads to the conclusion that $\|q^b_{\kappa,h}(t)\| \leq \|q^b_{\kappa,h}(0)\| = C_1 M$. For this reason, the accuracy of the

scheme is linked to the behavior of $q^a_{\kappa,h}(t)$. When $\kappa_r = \eta_r = 0$, we obviously have $q^a_{\kappa,h}(t) = \mathbb{P}_h(q^0_h)$ which leads to Point *(i)*.
To be convenient, we denote the discrete projection of q^0_h on the discrete kernel $\mathcal{E}^h_{\omega\neq 0}$ by $\hat{q}^0_h = \mathbb{P}_h(q^0_h)$. By using the fact that $\hat{q}^0_h$ is in the discrete kernel $\mathcal{E}^h_{\omega\neq 0}$, and thus verifies $\hat{u}^0_h = 0$ and $a_\star \partial_{x,h}\hat{r}^0_h = \omega \hat{v}^0_h$, we obtain $\forall i \in \{1, \cdots, N\}$

$$\begin{cases} \partial_t(r^a_{\kappa,i+1/2} - \hat{r}^0_{i+1/2}) + a_\star \partial_{x,h}(u^a_{\kappa,h} - \hat{u}^0_h)_{i+1/2} - \nu_r \partial^2_{x^2,h}(r^a_{\kappa,h} - \hat{r}^0_h)_{i+1/2} - \nu_r \partial^2_{x^2,h}(\hat{r}^0_h)_{i+1/2} = 0 \\ \partial_t(u^a_{\kappa,i} - \hat{u}^0_i) + a_\star \partial_{x,h}(r^a_{\kappa,h} - \hat{r}^0_h)_i - \nu_u \partial^2_{x^2,h}(u^a_{\kappa,h} - \hat{u}^0_h)_i = \omega(v^a_{\kappa,i} - \hat{v}^0_i) \\ \partial_t(v^a_{\kappa,i} - \hat{v}^0_i) = -\omega(u^a_{\kappa,i} - \hat{u}^0_i) \end{cases} \tag{3.14}$$

Therefore, by multiplying (3.14) with $q^a_{\kappa,i} - \hat{q}^0_i$, performing discrete integrations by part and using the periodic boundary condition, we deduce that

$$\frac{1}{2}\frac{d}{dt}\|q^a_{\kappa,h} - \hat{q}^0_h\|^2 = -\nu_r ||\partial_{x,h}(r^a_{\kappa,h} - \hat{r}^0_h)||^2 - \nu_u ||\partial_{x,h}(u^a_{\kappa,h} - \hat{u}^0_h)||^2 + \nu_r \langle \partial^2_{x^2,h}(\hat{r}^0_h), (r^a_{\kappa,h} - \hat{r}^0_h)\rangle$$

which leads to

$$\frac{1}{2}\frac{d}{dt}\|q^a_{\kappa,h} - \hat{q}^0_h\|^2 \quad \leq \quad \nu_r \|\partial^2_{x^2,h}(\hat{r}^0_h)\| \, \|q^a_{\kappa,h} - \hat{q}^0_h\|$$

and to

$$\frac{d}{dt}\|q^a_{\kappa,h} - \hat{q}^0_h\| \leq \|\partial^2_{x^2,h}(\hat{r}^0_h)\|.$$

This inequality leads to

$$\|q^a_{\kappa,h} - \hat{q}^0_h\| \leq t\,\nu_r \|\partial^2_{x^2,h}(\hat{r}^0_h)\|.$$

As a result, we get

$$\|q_{\kappa,h}(t) - \mathbb{P}_h(q^0_h)\| \leq \|q^0_h - \mathbb{P}_h(q^0_h)\| + t\,\nu_r \|\partial^2_{x^2,h}(\hat{r}^0_h)\|.$$

We then deduce Points *(ii)* and *(iii)* respectively for $\kappa_r = \mathcal{O}(M)$ and $\Delta x = \mathcal{O}(M)$.

□

3.2.6 Fourier analysis for the semi-discrete staggered schemes

We now construct the plane wave solutions for the 3×3 system of equation in (3.2). Let us denote q the column vector $q = (r, u, v)^T$, then the linear wave equation (3.2) can be written in the general form

$$\partial_t q + A\partial_x q + Bq = 0 \tag{3.15}$$

where $A = \begin{pmatrix} 0 & a_\star & 0 \\ a_\star & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$ and $B = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -\omega \\ 0 & \omega & 0 \end{pmatrix}$ are constant-coefficient matrices. We look for the plane wave solutions of (3.15) of the form

$$q = e^{i(kx+\tau t)}\hat{q} \tag{3.16}$$

where k is the wave number and τ is the frequency of the wave. These functions can be solutions to the linear wave equation only under a *dispersion relation* between τ and k which is commonly written as $\tau = \tau(k)$. In general, this relation lies in the complex set: the real part $\Re(\tau)$ and the imaginary part $\Im(\tau)$ indicate respectively propagation and decay of Fourier modes. In order to find $\hat{q}$ as well as $\tau(k)$, we substitute the plane wave solutions (3.16) into (3.15) to obtain

$$i\tau(k)\hat{q} + \mathcal{A}(k)\hat{q} = 0$$

and conclude that $-i\tau(k)$ and $\hat{q}$ are the right eigenvalue and eigenvector of the matrix $\mathcal{A}(k)$ which is given by

$$\mathcal{A}(k) = ikA + B = \begin{pmatrix} 0 & a_\star ik & 0 \\ a_\star ik & 0 & -\omega \\ 0 & \omega & 0 \end{pmatrix}$$

then, the eigenvalues of this matrix are given by

$$\lambda = 0, \pm i\sqrt{a_\star^2 k^2 + \omega^2}.$$

Therefore, the dispersion relations of the linear wave equation are given by

$$\tau(k) = 0, \pm\sqrt{a_\star^2 k^2 + \omega^2},$$

corresponding to the steady wave (geostrophic mode) and Poincaré wave (inertia-gravity modes). Phase (C) and group (G) velocities are given by

$$C = \frac{\tau}{k} \text{ and } G = \frac{\partial \tau}{\partial k}.$$

It is important to see that the phase velocity depends on the wave number, so wave components of different wavelengths travel at different speeds. On the other hand, $\tau(k)$ is a monotone function which indicates the free spurious oscillations of the shortest wave, namely $2\Delta x$ [26].
The numerical dispersion relation of the semi-discrete staggered scheme is found by using the discrete Fourier modes and we only consider $k\Delta x$ on the range $(0, \pi]$. It is expected that in the region of interest, i.e $k\Delta x \leq 0.1$, the dispersion relation of the numerical scheme has the same behavior as that of the continuous modes and it preserves the amplitude of the waves which means that there is no damping error with the numerical scheme. Let us emphasize that most of energy transfer occurs in this long wave region. On the other hand, since the phase error of the waves with the shortest wavelength $2\Delta x$ might produce some oscillations, numerical damping is really useful for high frequencies. However, numerical diffusion also affects the wave speed, and waves might therefore transfer with incorrect speed. Hence, we have to control numerical viscosity such that it does not destroy the steady state, ensures the stability of the scheme and has a minimal impact on the wave speed.
In order to obtain the dispersion relation of the semi-discrete Godunov scheme, we look for the solution of the semi-discrete Godunov type scheme under discrete Fourier modes:

$$r_{i+1/2}(t) = \varphi_r(t) e^{\boldsymbol{i}k(x_i + \frac{\Delta x}{2})}, \quad u_i(t) = \varphi_u(t) e^{\boldsymbol{i}kx_i} \quad \text{and} \quad v_i(t) = \varphi_v(t) e^{\boldsymbol{i}kx_i}.$$

Substituting these expressions in the semi-discrete scheme and setting $\eta_r = \kappa_r$ (because both Low Froude and Apparent Topography schemes use this equality), we obtain the following linear system of differential equations:

$$\begin{pmatrix} \varphi_r'(t) \\ \varphi_u'(t) \\ \varphi_v'(t) \end{pmatrix} + \begin{pmatrix} \kappa_r a_\star \frac{2\sin^2(\frac{k\Delta x}{2})}{\Delta x} & i a_\star \frac{2\sin(\frac{k\Delta x}{2})}{\Delta x} & i\frac{\kappa_r \omega \Delta x}{2} \frac{2\sin(\frac{k\Delta x}{2})}{\Delta x} \\ i a_\star \frac{2\sin(\frac{k\Delta x}{2})}{\Delta x} & \kappa_u a_\star \frac{2\sin^2(\frac{k\Delta x}{2})}{\Delta x} & -\omega \\ 0 & \omega & 0 \end{pmatrix} \begin{pmatrix} \varphi_r(t) \\ \varphi_u(t) \\ \varphi_v(t) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}. \tag{3.17}$$

We shall denote by $\mathcal{A}(k, \Delta x)$ the matrix of the linear differential equation (3.17). The first eigenvalue of the matrix $\mathcal{A}(k, \Delta x)$ is $\lambda = 0$, corresponding to the discrete stationary state. The other two eigenvalues, corresponding to the inertia-gravity modes, are given by

$$\lambda_\pm = \frac{\nu_r + \nu_u}{2} \frac{\alpha^2}{\Delta x^2} \pm i\sqrt{\omega^2 + a_\star^2 \frac{\alpha^2}{\Delta x^2} - \left(\frac{\nu_r - \nu_u}{2}\right)^2 \left(\frac{\alpha^2}{\Delta x^2}\right)^2} \quad \text{with} \quad \alpha = 2\sin\left(\frac{k\Delta x}{2}\right). \tag{3.18}$$

The real part of eigenvalues $\Re(\lambda_\pm)$ indicate the decay of Fourier modes k. Due to the fact that $\Re(\lambda_\pm) > 0$, both Low Froude and Apparent Topography schemes are damping. More importantly, the damping rate of the *Apparent Topography staggered scheme* is larger than that of the *Low Froude staggered scheme* (which uses $\eta_r = 0$) especially for the shortest wavelength $2\Delta x$ for which $\frac{k\Delta x}{2} = \frac{\pi}{2}$ and thus α is maximal. Therefore, from the numerical point of view, the *Apparent Topography scheme* will probably present fewer oscillations than the *Low Froude scheme*. Moreover, the imaginary parts $\Im(\lambda_\pm)$ are of different signs, which means that the inertia-gravity modes move in different directions, which is the same behavior as the continuous system.

Remark 3.5. *On a collocated mesh, the eigenvalues $\lambda_\pm$ of the Godunov type scheme can be written as*

$$\lambda_\pm = \frac{\nu_r + \nu_u}{2}\frac{\alpha^2}{\Delta x^2} \pm i\sqrt{\omega^2\vartheta^2 + a_\star^2\frac{\sin(k\Delta x)^2}{\Delta x^2} - \left(\frac{\nu_r - \nu_u}{2}\right)^2\left(\frac{\alpha^2}{\Delta x^2}\right)^2} \tag{3.19}$$

where

$$\vartheta = \begin{cases} 1 & \textit{Low Froude scheme,} \\ \cos^2(\frac{k\Delta x}{2}) & \textit{Apparent Topography scheme.} \end{cases}$$

We mention [38] for more details.

Remark 3.6. *It is highly recommended to use the same numerical diffusion term on the pressure and velocity equation for the Apparent Topography scheme on both collocated and staggered mesh, since the final term in the eigenvalue of the inertia-gravity modes disappears with $\nu_r = \nu_u$.*

We have the following comments based on the Fourier analysis of the semi-discrete scheme:

- We see in Figure 3.3a that the staggered schemes give better dispersion laws than the corresponding collocated schemes. Since the term $\sin(\frac{k\Delta x}{2})$ of $\Im(\lambda_\pm)$ in (3.18) is a monotone function until $k = \frac{\pi}{\Delta x}$, we get monotonic curves for the dispersion laws of the staggered schemes. However, this important property does not hold with the collocated schemes. We can observe that when $k\Delta x \approx 0.4\pi$, the dispersion laws of the collocated schemes start to decrease. This problem can be explained by the non-monotonic function $\sin(k\Delta x)$ in (3.19). Therefore, we do not have the spurious $2\Delta x$ oscillations with the staggered schemes but we might have them with the collocated schemes.

- The error phase velocity error is shown in Figure 3.3b where the following quantity is plotted:

$$M_C = \frac{C_h - C_0}{C_0}$$

  where C_h and C_0 respectively stand for the numerical and exact phase velocity. This figure shows us another important property of the staggered schemes: We can notice that the phase velocity error is much smaller with the staggered scheme than with the collocated scheme.

- Figure 3.3c indicates that the group velocity of the staggered schemes is always positive while it becomes negative with the collocated schemes when $k\Delta x > 0.4\pi$. Hence, in the short wave region, i.e for high frequencies, the energy of the collocated schemes moves in the wrong direction.

- Considering the staggered type schemes, from Figure 3.3, we can see that the *Apparent Topography staggered scheme* curves are closer to the analytical ones than those of the *Low Froude staggered scheme* in terms of dispersion law, phase and group velocity.

- From Figure 3.3d, the damping rate of the Apparent Topography schemes is twice larger than that of the Low Froude schemes (the damping rates do not depend on the type of mesh, collocated or staggered). As a result, on a collocated mesh, we may expect fewer oscillations with the Apparent Topography schemes, especially for the waves with shortest wavelength $2\Delta x$.

- Figure 3.3 also indicates that in order to make sure that the properties of the numerical solution are similar to those of the exact solution, with the collocated schemes, we have to consider meshes which satisfy $\frac{k\Delta x}{\pi} \leq 0.2$, while the staggered schemes require $\frac{k\Delta x}{\pi} \leq 0.4$. It means that we can reduce the mesh size by a factor of two with the staggered schemes. Hence, the staggered schemes are preferred rather than the collocated schemes in terms of efficiency.

- Figures 3.4 and 3.6 indicate that the behaviors of the dispersion law and group velocity strongly depend on the Rossby deformation radius. When $R_d < \Delta x$, the *Low Froude collocated scheme* produces much better dispersion law and group velocity than the *Apparent Topography collocated scheme* in the region $k\Delta x < 0.4$. More importantly, the group velocity of the *Apparent Topography collocated scheme* is negative even in the region of interest which indicates the wrong moving of the energy. Moreover, Figure 3.5 shows that the error of phase velocity of the *Low Froude collocated scheme* is smaller than that of the *Apparent Topography collocated scheme*. For example, with the shortest wavelength $2\Delta x$, the error of phase velocity is 50% with the low Froude scheme and 100% with the Apparent Topography scheme. However, in the regime $R_d > \Delta x$, the Apparent Topography collocated scheme is preferable. This is because with the *Low Froude collocated scheme*, waves with short wavelengths ($k\Delta x > 0.8$) do not propagate at all since the corresponding eigenvalues are real numbers (their imaginary part vanishes) and thus the error in the phase velocity for short wavelengths is larger with the *Low Froude Collocated scheme* and their energy does not propagate.

3.3 Analysis of fully discrete staggered scheme

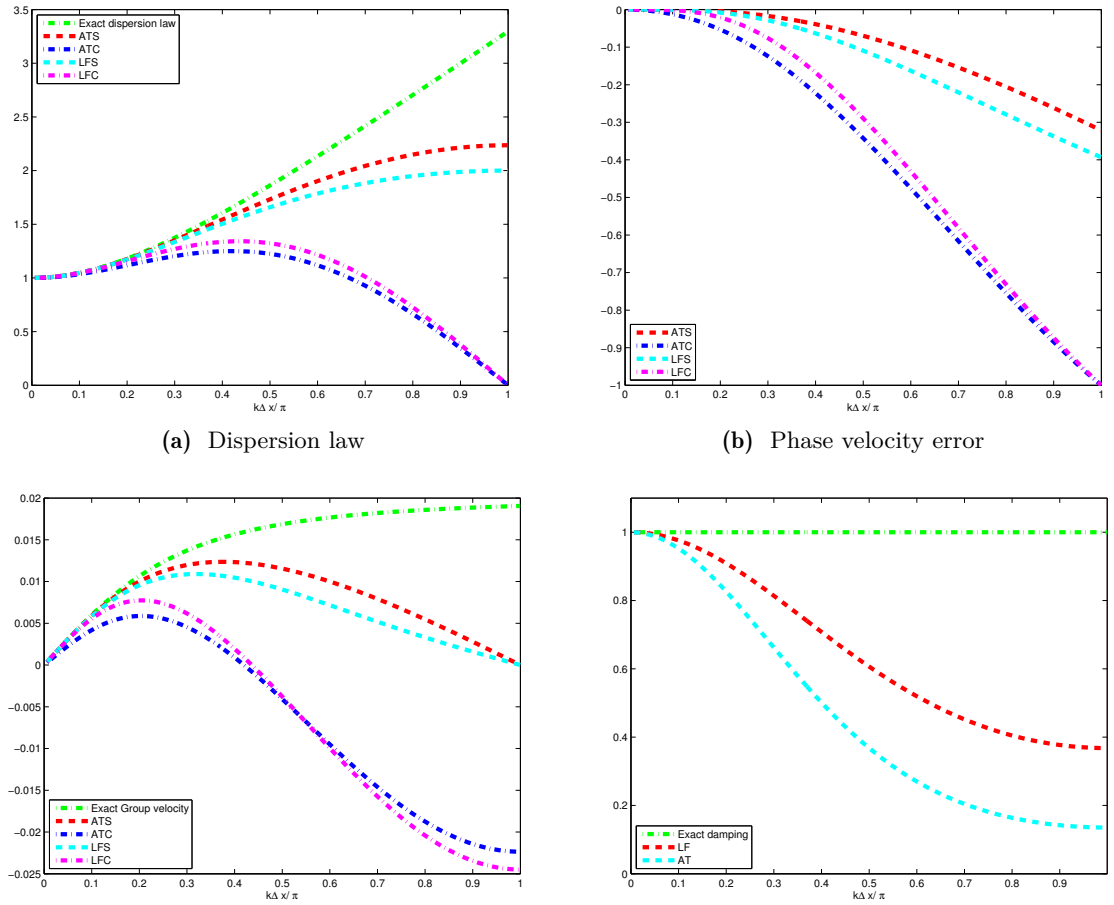
3.3.1 Fourier analysis of fully discrete scheme

The numerical properties of the semi-discrete scheme may change a lot when we take into account the time discretization. In this section, we investigate the numerical dispersion law and damping error (amplification) of the fully discrete schemes. We mention the work of Manuel J. Castro et al. in [27] for the study of these properties with high order schemes in space an in time.
In general, when a θ-scheme is applied to the Coriolis term, the first order time discretization of the staggered schemes can be written as

$$\mathcal{T}_\theta q_i^{n+1} = q_i^n - \Delta t \mathcal{L}_{\kappa,\eta}^{i,\theta}(q^n) \tag{3.20}$$

where the matrix $\mathcal{T}_\theta$ is given by

$$\mathcal{T}_\theta = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -(1-\theta_1)\omega\Delta t \\ 0 & (1-\theta_2)\omega\Delta t & 1 \end{pmatrix} \tag{3.21}$$



(a) Dispersion law **(b)** Phase velocity error

(c) Group velocity **(d)** Damping

Figure 3.3: *Numerical properties of the semi-discrete Godunov type schemes with Rossby deformation* $R_d = \Delta x$.

and $\mathcal{L}^{i,\theta}_{\kappa,\eta}$ is the operator corresponding to the spatial discretization defined by

$$\mathcal{L}^{i,\theta}_{\kappa,\eta} q^n = \begin{pmatrix} \frac{a_\star}{\Delta x}\left[u^n_{i+1} - u^n_i\right] - \frac{\kappa_r a_\star}{2\Delta x}\left[r^n_{i+3/2} - 2r^n_{i+1/2} + r^n_{i-1/2}\right] + \frac{\eta_r \omega}{2}(v^n_{i+1} - v^n_i) \\ \frac{a_\star}{\Delta x}\left[r^n_{i+1/2} - r^n_{i-1/2}\right] - \frac{\kappa_u a_\star}{2\Delta x}\left[u^n_{i+1} - 2u^n_i + u^n_{i-1}\right] - \theta_1 \omega v^n_i \\ \theta_2 \omega u^n_i \end{pmatrix}.$$

It is useful to see that equation (3.20) can be rewritten under the following form

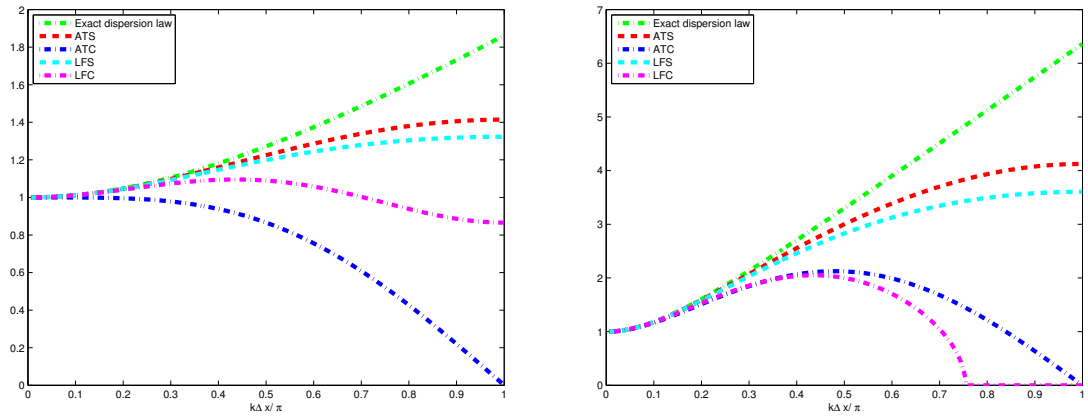
$$q^{n+1}_i = \mathcal{T}_\theta^{-1}[q^n_i - \Delta t \mathcal{L}^{i,\theta}_{\kappa,\eta} q^n], \tag{3.22}$$

where the matrix $\mathcal{T}_\theta^{-1}$ is given by

$$\mathcal{T}_\theta^{-1} = \frac{1}{\Lambda(\theta_1,\theta_2)} \begin{pmatrix} \Lambda(\theta_1,\theta_2) & 0 & 0 \\ 0 & 1 & \omega\Delta t(1-\theta_1) \\ 0 & -\omega\Delta t(1-\theta_2) & 1 \end{pmatrix} \quad \text{with} \quad \Lambda(\theta_1,\theta_2) = 1 + (\omega\Delta t)^2(1-\theta_1)(1-\theta_2). \tag{3.23}$$

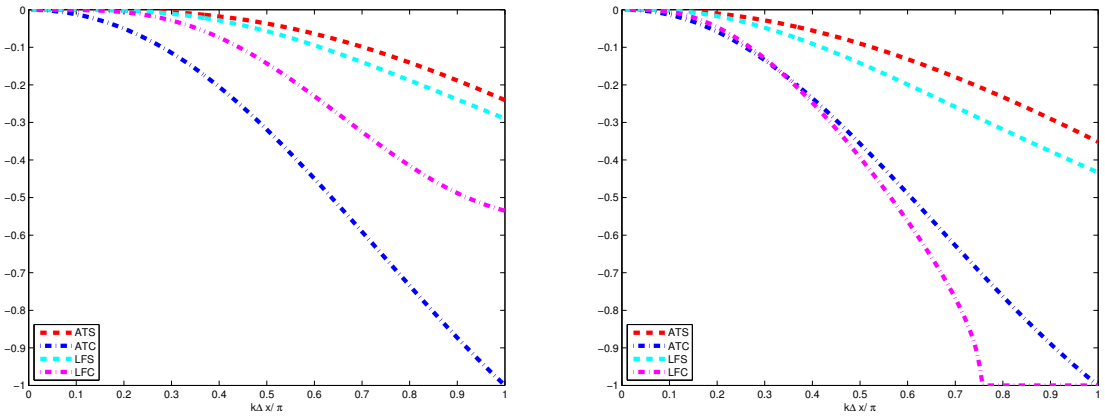
We now conduct a Fourier analysis for the fully discrete scheme by substituting the discrete Fourier modes

$$r^n_i = \varphi^n_r e^{\boldsymbol{i}kx_i}, \quad u^n_i = \varphi^n_u e^{\boldsymbol{i}kx_i} \quad \text{and} \quad v^n_i = \varphi^n_v e^{\boldsymbol{i}kx_i} \tag{3.24}$$



(a) $R_d = \frac{1}{2}\Delta x$ **(b)** $R_d = 2\Delta x$

Figure 3.4: *Dispersion laws of semi-discrete Godunov type schemes with different values of Rossby deformation.*



(a) $R_d = \frac{1}{2}\Delta x$ **(b)** $R_d = 2\Delta x$

Figure 3.5: *The errors of phase velocity of semi-discrete Godunov type schemes with different values of Rossby deformation.*

into the fully discrete scheme (3.22) in order to obtain

$$\varphi^{n+1} = \mathcal{M}_\theta \varphi^n \tag{3.25}$$

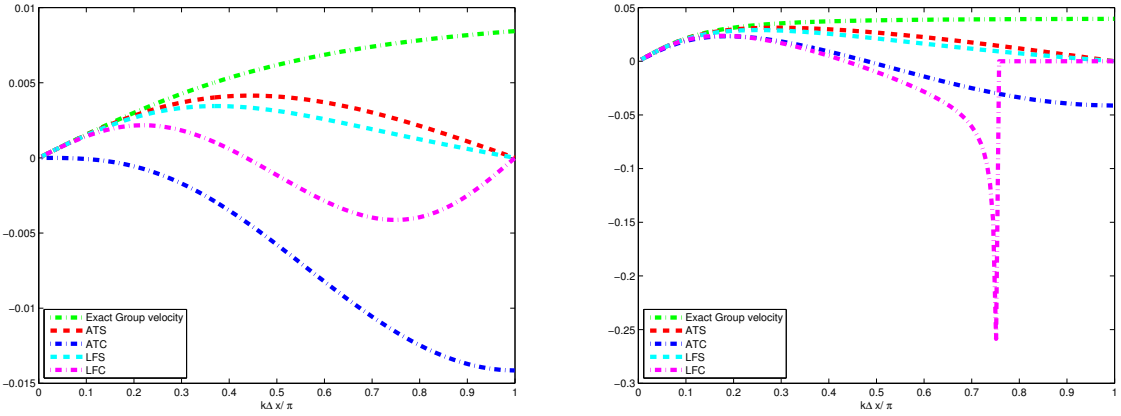
where the matrix $\mathcal{M}_\theta$ is given by

$$\mathcal{M}_\theta = \mathcal{T}_\theta^{-1}[\mathcal{I} - \Delta t \mathcal{A}_\theta] \tag{3.26}$$

with

$$\mathcal{A}_\theta = \begin{pmatrix} \kappa_r a_\star \frac{\sin^2(\frac{k\Delta x}{2})}{\frac{\Delta x}{2}} & i a_\star \frac{\sin(\frac{k\Delta x}{2})}{\frac{\Delta x}{2}} & i\frac{\kappa_r \omega \Delta x}{2}\frac{\sin(\frac{k\Delta x}{2})}{\frac{\Delta x}{2}} \\ i a_\star \frac{\sin(\frac{k\Delta x}{2})}{\frac{\Delta x}{2}} & \kappa_u a_\star \frac{\sin^2(\frac{k\Delta x}{2})}{\frac{\Delta x}{2}} & -\theta_1 \omega \\ 0 & \theta_2 \omega & 0 \end{pmatrix}.$$

Remark 3.7. *We now consider one special case when $\theta_1 = 1$ and $\theta_2 = 0$. In this case, let us note that the one step scheme is the same as the following splitting scheme, in which the first order*



(a) $R_d = \frac{1}{2}\Delta x$ **(b)** $R_d = 2\Delta x$

Figure 3.6: *Group velocity of semi-discrete Godunov type schemes with different values of Rossby deformation.*

time discretization of the fully discrete scheme is composed of two steps:

$$\begin{aligned} q_i^{(1)} &= q_i^n - \Delta t \mathcal{B} \mathcal{L}_{\kappa,\eta}^i q^n \\ q_i^{n+1} &= q_i^{(1)} - \Delta t \mathcal{C} \mathcal{L}_{\kappa,\eta}^i q^{(1)} \end{aligned} \tag{3.27}$$

where

$$\mathcal{B} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \textit{and} \quad \mathcal{C} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Therefore, we obtain

$$\mathcal{M} = (\mathcal{I} - \Delta t \mathcal{C}\mathcal{A}).(\mathcal{I} - \Delta t \mathcal{B}\mathcal{A}). \tag{3.28}$$

For every eigenvalue α of matrix $\mathcal{M}$ (or $\mathcal{M}_\theta$), the discrete frequency of the wave is defined by

$$\Re(\tau) = \frac{arg(\alpha)}{\Delta t} \tag{3.29}$$

where $arg(\alpha)$ represents the argument of the complex number α.
The amplification factor after a time $\frac{\Delta x}{a_\star}$ is given by

$$\varrho = |\alpha|^{\frac{\Delta x}{a_\star \Delta t}}. \tag{3.30}$$

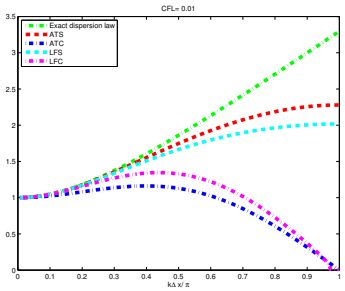
Now if we denote $CFL = \frac{a_\star \Delta t}{\Delta x}$, the amplification factor can be written as $\varrho = |\alpha|^{1/CFL}$.
Figures 3.7 and 3.8 show the numerical dispersion laws and damping of the fully discrete schemes with first order time discretization and $\theta_1 = \theta_2 = 0.5$. Those figures indicate the effects of the time step through the CFL parameter and of the value of the Rossby deformation radius.

- We observe that the properties of the fully discrete scheme are nearly the same as those of the semi-discrete scheme when the time step is small ($CFL = 0.01$). More importantly, we still have monotonic curves for the dispersion laws of the fully staggered schemes, like in the semi-discrete case. However, when the time step is large, there appears a drawback with the *Apparent Topography staggered scheme*: This scheme produces numerical waves that are faster than the exact ones. Hence, we can say that this scheme is very sensitive to the time step Δt. On the contrary, the *Low Froude staggered scheme* is not sensitive to the CFL parameter.

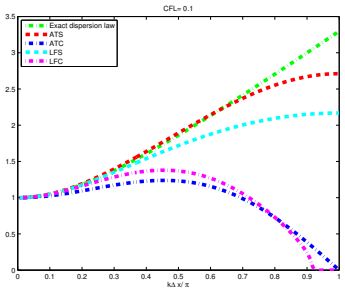
- In consideration of the *Apparent Topography collocated scheme*, the numerical dispersion law is closer to the exact one when the CFL value is larger (i.e. for large time steps).

- About the damping error, when the time step is small, there is no difference between the collocated and staggered schemes. The damping rate of the Apparent Topography scheme is twice larger than that of the Low Froude scheme. However, the damping rate of those schemes are different one from the other when the time step gets larger. We can observe that the greater the CFL value, the less damping error there is with the staggered schemes.

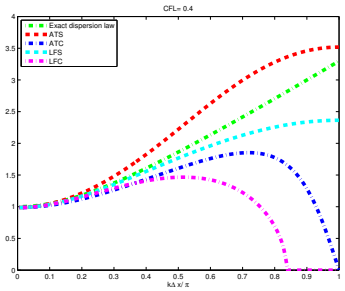
- The properties of the *Low Froude collocated scheme* also depend on the ratio between the Rossby deformation radius and the space step Δx. The dispersion law goes faster to zero in the regime $R_d > \Delta x$ than in the regime $R_d \leq \Delta x$.



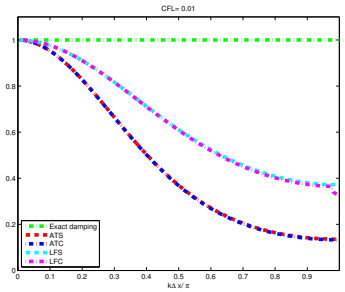
(a) Dispersion law with $CFL = 0.01$



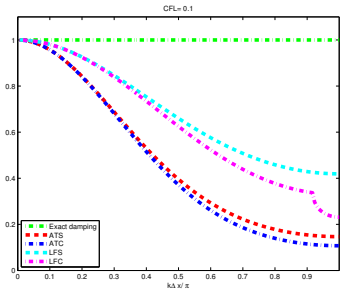
(b) Dispersion law with $CFL = 0.1$



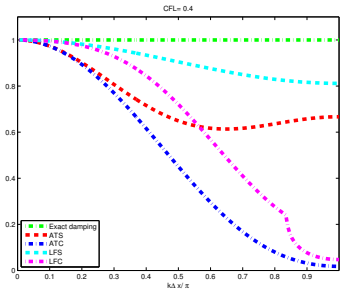
(c) Dispersion law with $CFL = 0.4$



(d) Damping error with $CFL = 0.01$

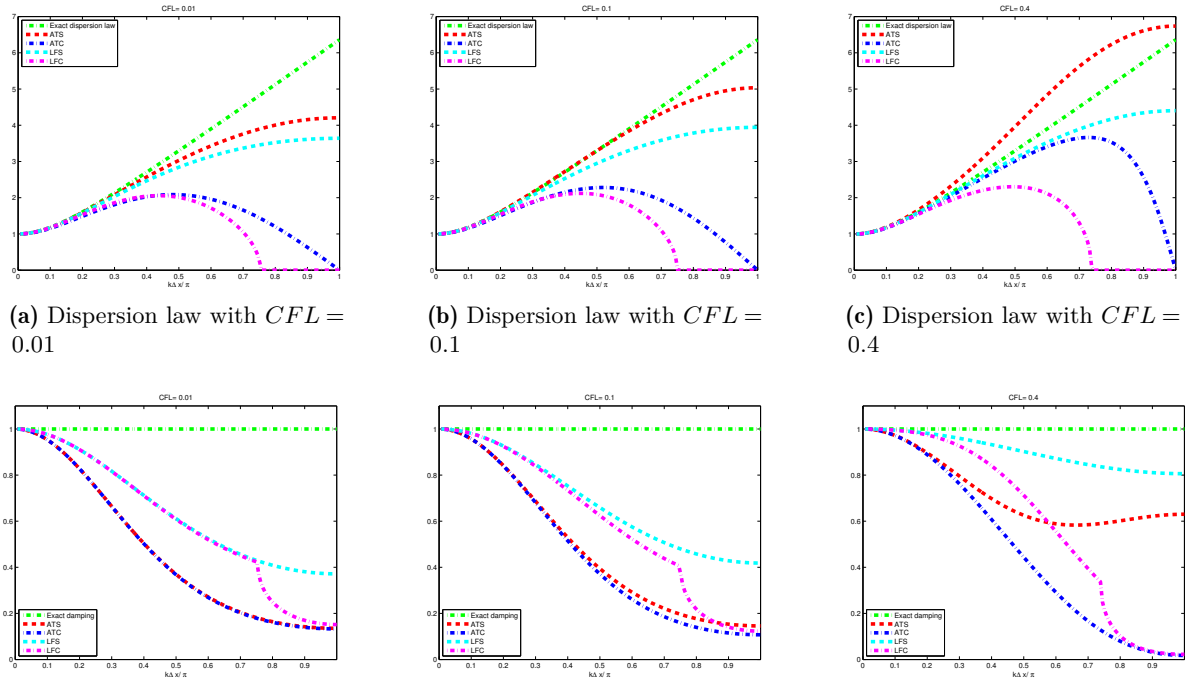


(e) Damping error with $CFL = 0.1$



(f) Damping error with $CFL = 0.4$

Figure 3.7: *Numerical properties of Godunov type schemes with first order time discretization when* $\theta_1 = \theta_2 = \frac{1}{2}$*, with Rossby deformation radius* $R_d = \Delta x$ *and* $\kappa_r = \kappa_u = 1$ *for the staggered and collocated Apparent Topography schemes, while* $\kappa_r = 0, \kappa_u = 1$ *for the staggered and collocated Low Froude schemes.*



(a) Dispersion law with $CFL = 0.01$ **(b)** Dispersion law with $CFL = 0.1$ **(c)** Dispersion law with $CFL = 0.4$

(d) Damping error with $CFL = 0.01$ **(e)** Damping error with $CFL = 0.1$ **(f)** Damping error with $CFL = 0.4$

Figure 3.8: *Numerical properties of Godunov type schemes with first order time discretization when* $\theta_1 = \theta_2 = \frac{1}{2}$*, with Rossby deformation radius* $R_d = 2\Delta x$ *and* $\kappa_r = \kappa_u = 1$ *for the staggered and collocated Apparent Topography schemes, while* $\kappa_r = 0, \kappa_u = 1$ *for the staggered and collocated Low Froude schemes.*

3.3.2 Stability condition of the staggered type schemes

We consider a homogeneous cartesian mesh. The one step fully discrete staggered scheme is given by

$$\begin{cases} \dfrac{r_{i+1/2}^{n+1} - r_{i+1/2}^{n}}{\Delta t} + a_\star \left(\dfrac{u_{i+1}^n - u_i^n}{\Delta x} \right) - \dfrac{\kappa_r a_\star \Delta x}{2} \left(\dfrac{r_{i+3/2}^n - 2r_{i+1/2}^n + r_{i-1/2}^n}{\Delta x^2} \right) + \dfrac{\kappa_r \omega \Delta x}{2} \left(\dfrac{v_{i+1}^n - v_i^n}{\Delta x} \right) = 0, \\ \dfrac{u_i^{n+1} - u_i^n}{\Delta t} + a_\star \left(\dfrac{r_{i+1/2}^n - r_{i-1/2}^n}{\Delta x} \right) - \dfrac{\kappa_u a_\star \Delta x}{2} \left(\dfrac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{\Delta x^2} \right) = \omega \left[\theta_1 v_i^n + (1-\theta_1) v_i^{n+1} \right], \\ \dfrac{v_i^{n+1} - v_i^n}{\Delta t} = -\omega \left[\theta_2 u_i^n + (1-\theta_2) u_i^{n+1} \right] \end{cases} \tag{3.31}$$

for $i \in \{1, \ldots, N\}$ and $0 \leq \theta_1, \theta_2 \leq 1$.
To be convenient, let us denote:

$$\Theta_1 = 1 - \theta_1 - \theta_2, \quad \Theta_2 = \theta_1\theta_2 + (1-\theta_1)(1-\theta_2) \in [0,1], \quad \Theta_3 = (1-2\theta_1)(1-2\theta_2) \in [-1,1].$$

Theorem 3.2. *We have:*

i. When $\theta_1 + \theta_2 > 1$*, the staggered scheme* (3.31) *is unstable.*

ii. For the Low Froude staggered scheme ($\kappa_r = 0$ *and* $\kappa_u > 0$)*, we consider the following two cases:*

(a) If $\frac{\kappa_u^2 a_\star^2}{\omega^2\Delta x^2} \le \frac{4a_\star^2}{\omega^2\Delta x^2} + \Theta_3$, *the Low Froude staggered scheme is stable under the sufficient condition:*

$$\Delta t \le \Delta t_a := \frac{\kappa_u \Delta x}{2\left(|a_\star| - \frac{\omega^2\Delta x^2}{4|a_\star|}\Theta_1\right)_+}; \tag{3.32a}$$

(b) If $\frac{\kappa_u^2 a_\star^2}{\omega^2\Delta x^2} > \frac{4a_\star^2}{\omega^2\Delta x^2} + \Theta_3$, *the Low Froude staggered scheme is stable under the sufficient condition:*

i. When $\Theta_3 \ge 0$,

$$\Delta t \le \min\{\Delta t_a, \Delta t_b\} \tag{3.32b}$$

where $\Delta t_b := \begin{cases} \frac{2\kappa_u|a_\star|}{\frac{4a_\star^2}{\Delta x} + \omega^2\Theta_3\Delta x}\left[1 - \sqrt{1 - \frac{4a_\star^2 + \omega^2\Theta_3\Delta x^2}{\kappa_u^2 a_\star^2}}\right], & \text{if } 4a_\star^2 + \omega^2\Theta_3\Delta x^2 \ne 0, \\ \frac{\Delta x}{\kappa_u|a_\star|}, & \text{otherwise.} \end{cases}$

ii. When $\Theta_3 < 0$,

$$\Delta t \le \min\{\Delta t_a, \Delta t_b, \Delta t_c\} \quad \text{where} \quad \Delta t_c := \frac{2}{\omega\sqrt{|\Theta_3|}}. \tag{3.32c}$$

iii. For the Apparent Topography staggered scheme ($\kappa_r > 0$), *we have the following sufficient CFL condition:*

$$\Delta t \le \min\{\Delta t_a, \Delta t_b\}$$

where

$$\Delta t_a := \begin{cases} \frac{-\frac{|a_\star|}{\Delta x}(1+\kappa_r\kappa_u) + \sqrt{(\frac{a_\star}{\Delta x})^2(1+\kappa_r\kappa_u)^2 + \kappa_r(\kappa_r+\kappa_u)\omega^2\theta_2(1-\theta_1)}}{\kappa_r\omega^2\theta_2(1-\theta_1)}, & \text{if } \theta_2(1-\theta_1) \ne 0 \\ \frac{\kappa_r+\kappa_u}{2(1+\kappa_r\kappa_u)}\frac{\Delta x}{|a_\star|} & \text{otherwise.} \end{cases} \tag{3.32d}$$

and

$$\Delta t_b := \begin{cases} \min\left\{\frac{1}{\kappa_r}, \frac{1}{\kappa_u}\right\}\frac{\Delta x}{a_\star}, & \text{if } \theta_2 = \frac{1}{2}, \\ \min\left\{\frac{1}{2(\kappa_r+\kappa_u)}\frac{\Delta x}{a_\star}, \frac{1}{\omega}\right\}, & \text{otherwise.} \end{cases} \tag{3.32e}$$

Remark 3.8. *The time step of the Low Froude staggered scheme is more restrictive upon* Δt_a *than the one of the Low Froude collocated scheme. However, it is less restrictive in consideration of* Δt_b. *Moreover, in the general case, we normally have* $\kappa_u \le 1$ *and if we only consider the case* $0 \le \theta_1, \theta_2 \le \frac{1}{2}$, *we will obtain* $\Theta_3 \ge 0$. *As a result, the CFL condition in this case is only restricted by* Δt_a *and we can conclude that the CFL condition of the Low Froude staggered scheme does not depend on the Coriolis parameter* ω.

Proof. We perform a Von Neumann analysis to investigate the stability condition for the staggered scheme (3.31). To begin with, let us denote

$$\sigma = \frac{\Delta t}{\Delta x},\ \gamma = \omega\Delta t \ \text{ and } \ s = \sin\left(\frac{k\Delta x}{2}\right).$$

Next, we substitute

$$q_j^n = \begin{pmatrix} r_j^n \\ u_j^n \\ v_j^n \end{pmatrix} = \begin{pmatrix} R_n \\ U_n \\ V_n \end{pmatrix} e^{ikj\Delta x}$$

into (3.31) in order to obtain

$$\mathcal{T}_\theta q_j^{n+1} = B q_j^n \tag{3.33}$$

where the matrix $\mathcal{T}_\theta$ is given by (3.21) and B is given by

$$B = \begin{pmatrix} 1-2\kappa_r|a_\star|\sigma s^2 & -2a_\star\sigma i s & -i\kappa_r\omega\Delta t s \\ -2a_\star\sigma i s & 1-2\kappa_u|a_\star|\sigma s^2 & \theta_1\gamma \\ 0 & -\theta_2\gamma & 1 \end{pmatrix}.$$

In addition, $\mathcal{T}_\theta^{-1}$ is given by (3.23). Therefore, we can rewrite (3.33) as $q_j^{n+1} = C q_j^n$ where the amplification matrix C is given by

$$C = \mathcal{T}_\theta^{-1} B = \frac{1}{\Lambda(\theta_1,\theta_2)} \begin{pmatrix} (1-2\kappa_r|a_\star|\sigma s^2)\Lambda(\theta_1,\theta_2) & -2a_\star\sigma i s\Lambda(\theta_1,\theta_2) & -i\kappa_r\omega\Delta t s\Lambda(\theta_1,\theta_2) \\ -2a_\star\sigma i s & 1-\gamma^2\theta_2(1-\theta_1)-2\kappa_u|a_\star|\sigma s^2 & \gamma \\ 2\gamma(1-\theta_2)a_\star\sigma i s & -\gamma[1-(1-\theta_2)2\kappa_u|a_\star|\sigma s^2] & 1-\gamma^2\theta_1(1-\theta_2) \end{pmatrix}.$$

The characteristic polynomial $\mathcal{P}(\lambda)$ of this amplification matrix has one solution $\lambda_0 = 1$ and the other two roots $\lambda_\pm$ are the solutions of a second degree equation

$$\lambda^2 + \xi\lambda + \zeta = 0 \tag{3.34}$$

where the coefficients ξ and ζ are given by

$$\xi = -\frac{2-\gamma^2(\theta_1+\theta_2-2\theta_1\theta_2)-2\kappa_u|a_\star|\sigma s^2}{\Lambda(\theta_1,\theta_2)} + 2\kappa_r|a_\star|\sigma s^2$$

and

$$\zeta = \frac{1+\gamma^2\theta_1\theta_2-2\kappa_u|a_\star|\sigma s^2+4a_\star^2\sigma^2 s^2}{\Lambda(\theta_1,\theta_2)} - 2\kappa_r|a_\star|\sigma s^2\frac{1-2\kappa_u|a_\star|\sigma s^2}{\Lambda(\theta_1,\theta_2)} + 2\kappa_r|a_\star|\sigma s^2\frac{\gamma^2\theta_2(1-\theta_1)}{\Lambda(\theta_1,\theta_2)}.$$

In order to ensure that the roots of (3.34) are in the unit circle ($|\lambda_\pm| \le 1$), the coefficients ξ and ζ must satisfy

$$|\zeta| \le 1 \quad \text{and} \quad |\xi| \le 1+\zeta. \tag{3.35}$$

First of all, we consider the modes which are constant in space ($k = 0$, so $s = 0$). In this case the condition $|\zeta| \le 1$ reduces to

$$1+\gamma^2\theta_1\theta_2 \le 1+\gamma^2(1-\theta_1)(1-\theta_2).$$

This condition is fulfilled if and only if

$$\gamma^2[(\theta_1+\theta_2)-1] \le 0.$$

Therefore, the parameters θ_1 and θ_2 must satisfy the stability condition $\theta_1+\theta_2 \le 1$. This proves Point 1. Therefore, in what follows, we consider the case $\theta_1+\theta_2 \le 1$, which implies in particular that $\Theta_1 \ge 0$.
Next, we consider the Low Froude staggered scheme by setting the numerical viscosity $\kappa_r = 0$. In this case, condition (3.35) can be expressed by the following procedures:

- Firstly, the condition $\zeta \leq 1$ is equivalent to

$$\frac{1+\gamma^2\theta_1\theta_2 - 2\kappa_u|a_\star|\sigma s^2 + 4a_\star^2\sigma^2 s^2}{\Lambda(\theta_1,\theta_2)} \leq 1$$

which leads to

$$f_1(s^2) := -\gamma^2\Theta_1 - 2\kappa_u|a_\star|\sigma s^2 + 4a_\star^2\sigma^2 s^2 \leq 0.$$

With s varying in $[-1,1]$, the previous condition holds provided $\max_{[0,1]} f_1 \leq 0$. We notice that

$$\max_{[0,1]} f_1 = \begin{cases} f_1(0), & \text{if } |a_\star|\sigma \leq \frac{\kappa_u}{2}, \\ f_1(1), & \text{otherwise.} \end{cases}$$

We first note that with $|a_\star|\sigma \leq \frac{\kappa_u}{2}$, the condition $f_1(0) \leq 0$ is always satisfied. So the condition $\Delta t \leq \frac{\kappa_u \Delta x}{2|a_\star|}$ is sufficient to have $\zeta \leq 1$.

On the other hand, if $\Delta t > \frac{\kappa_u \Delta x}{2|a_\star|}$, which means $|a_\star|\sigma > \frac{\kappa_u}{2}$, the condition $f_1(1) \leq 0$ reads

$$-\gamma^2\Theta_1 - 2\kappa_u|a_\star|\sigma + 4a_\star^2\sigma^2 \leq 0 \iff \left(\frac{(2|a_\star|)^2}{\Delta x^2} - \omega^2\Theta_1\right)\Delta t \leq \frac{2|a_\star|}{\Delta x}\kappa_u,$$

which means $\Delta t \leq \Delta t_a$ defined in (3.32a). So by gathering the two cases, the condition $\zeta \leq 1$ will be satisfied if and only if $\Delta t \leq \Delta t_a$.

- Next, the condition $\zeta \geq -1$ can be written as

$$f_2(s^2) := \gamma^2\Theta_2 + 2(1 - \kappa_u|a_\star|\sigma s^2) + 4a_\star^2\sigma^2 s^2 \geq 0.$$

We shall see below that this constraint is weaker than another one ($f_3(s^2) \geq 0$) and needs not be taken into account.

- Let us now turn to the condition upon ξ. The first case $-\xi \leq 1 + \zeta$ reads

$$2 - \gamma^2(\theta_1 + \theta_2 - 2\theta_1\theta_2) - 2\kappa_u|a_\star|\sigma s^2 \leq 2 + \gamma^2[1 - (\theta_1 + \theta_2) + 2\theta_1\theta_2] - 2\kappa_u|a_\star|\sigma s^2 + 4a_\star^2\sigma^2 s^2$$

which comes down to

$$-\gamma^2 - 4a_\star^2 s^2 \leq 0.$$

The latter inequality always holds and does not imply any additional constraint upon Δt.

- Finally, we consider the case $\xi \leq 1 + \zeta$. This leads to

$$-2 + \gamma^2(\theta_1 + \theta_2 - 2\theta_1\theta_2) + 2\kappa_u|a_\star|\sigma s^2 \leq 2 + \gamma^2[1 - (\theta_1 + \theta_2) + 2\theta_1\theta_2] - 2\kappa_u|a_\star|\sigma s^2 + 4a_\star^2\sigma^2 s^2.$$

It follows that

$$f_3(s^2) := \gamma^2\Theta_3 + 4(1 - \kappa_u|a_\star|\sigma s^2) + 4a_\star^2\sigma^2 s^2 \geq 0.$$

From $\Theta_3 = 2\Theta_2 - 1$, we infer that $2f_2(s^2) - f_3(s^2) \geq 0$ over $[0,1]$. This implies that the condition $f_2(s^2) \geq 0$ is a consequence of $f_3(s^2) \geq 0$.

With s varying in $[-1,1]$, the previous condition holds provided $\min_{[0,1]} f_3 \geq 0$. Moreover, we have

$$\min_{[0,1]} f_3 = \begin{cases} f_3(0), & \text{if } |a_\star|\sigma > \kappa_u, \\ f_3(1), & \text{otherwise.} \end{cases}$$

When $|a_\star|\sigma > \kappa_u$, the condition $f_3(0) \geq 0$ is given by

$$f_3(0) = \gamma^2\Theta_3 + 4 \geq 0$$

which is always satisfied when $\Theta_3 \geq 0$ and in case $\Theta_3 < 0$, this condition leads to

$$\Delta t \leq \frac{2}{\omega\sqrt{|\Theta_3|}}. \tag{3.36}$$

When $|a_\star|\sigma \leq \kappa_u$, the condition $f_3(1) \geq 0$ is equivalent to

$$Q_3(\Delta t) = \left(\omega^2\Theta_3 + \frac{4a_\star^2}{\Delta x^2}\right)\Delta t^2 - 4\kappa_u\frac{|a_\star|}{\Delta x}\Delta t + 4 \geq 0$$

When $\frac{\kappa_u^2 a_\star^2}{\omega^2\Delta x^2} \leq \frac{4a_\star^2}{\omega^2\Delta x^2} + \Theta_3$, this condition is always satisfied. If not, then the solutions of the second order equation $Q_3(\Delta t) = 0$ are given by

$$\Delta t = \frac{2\kappa_u|a_\star|}{\frac{4a_\star^2}{\Delta x} + \omega^2\Theta_3\Delta x}\left[1 \pm \sqrt{1 - \frac{4a_\star^2 + \omega^2\Theta_3\Delta x^2}{\kappa_u^2 a_\star^2}}\right] \tag{3.37}$$

which leads to the stability condition

$$\Delta t \leq \Delta t_b = \begin{cases} \frac{2\kappa_u|a_\star|}{\frac{4a_\star^2}{\Delta x}+\omega^2\Theta_3\Delta x}\left[1 - \sqrt{1 - \frac{4a_\star^2+\omega^2\Theta_3\Delta x^2}{\kappa_u^2 a_\star^2}}\right], & \text{if } 4a_\star^2 + \omega^2\Theta_3\Delta x^2 \neq 0, \\ \frac{\Delta x}{\kappa_u|a_\star|}, & \text{otherwise.} \end{cases} \tag{3.38}$$

This concludes Point 2.

We now consider the Apparent Topography scheme with $\kappa_r > 0$. In this case, in order to find a sufficient CFL condition based on (3.35), we combine the following conditions:

- First, the condition $\zeta \leq 1$ is equivalent to

$$f_1(s^2) := -2(\kappa_r+\kappa_u)|a_\star|\sigma s^2 + 4a_\star^2\sigma^2 s^2 + 4\kappa_r\kappa_u a_\star^2\sigma^2 s^4 + 2\kappa_r|a_\star|\sigma s^2\gamma^2\theta_2(1-\theta_1) - \gamma^2[1-(\theta_1+\theta_2)] \leq 0.$$

With s varying in $[-1,1]$, the maximum value of $f_1(s^2)$ is either $f_1(0)$ or $f_1(1)$ due to the the fact that $4\kappa_r\kappa_u a_\star^2\sigma^2 > 0$. Moreover, we have $f_1(0) = -\gamma^2[1-(\theta_1+\theta_2)]$, so the condition $f_1(0) \leq 0$ is always satisfied since we only consider the case $\theta_1+\theta_2 \leq 1$. We now focus on the condition $f_1(1) \leq 0$ which is given by

$$-2(\kappa_r+\kappa_u)|a_\star|\sigma + 4a_\star^2\sigma^2 + 4\kappa_r\kappa_u a_\star^2\sigma^2 + 2\kappa_r|a_\star|\sigma\gamma^2\theta_2(1-\theta_1) - \gamma^2[1-(\theta_1+\theta_2)] \leq 0. \tag{3.39}$$

We note that a sufficient condition for (3.39) to hold is when we have

$$-2(\kappa_r+\kappa_u)|a_\star|\sigma + 4a_\star^2\sigma^2 + 4\kappa_r\kappa_u a_\star^2\sigma^2 + 2\kappa_r|a_\star|\sigma\gamma^2\theta_2(1-\theta_1) \leq 0. \tag{3.40}$$

We now notice that when $\theta_2 = 0$ or $\theta_1 = 1$, condition (3.40) reduces to

$$-2(\kappa_r+\kappa_u)|a_\star|\sigma + 4a_\star^2\sigma^2 + 4\kappa_r\kappa_u a_\star^2\sigma^2 \leq 0 \quad \Rightarrow \quad \Delta t \leq \frac{\kappa_r+\kappa_u}{2(1+\kappa_r\kappa_u)}\frac{\Delta x}{a_\star}.$$

On the other hand, when $\theta_2(1-\theta_1) \neq 0$, (3.40) leads to the condition

$$\Delta t \leq \Delta t_a := \frac{-\frac{|a_\star|}{\Delta x}(1+\kappa_r\kappa_u) + \sqrt{(\frac{a_\star}{\Delta x})^2(1+\kappa_r\kappa_u)^2 + \kappa_r(\kappa_r+\kappa_u)\omega^2\theta_2(1-\theta_1)}}{\kappa_r\omega^2\theta_2(1-\theta_1)}. \tag{3.41}$$

- Next, the condition $\zeta \geq -1$ can be written as

$$\begin{aligned} f_2(s^2) := \quad & 2 - 2(\kappa_r + \kappa_u)|a_\star|\sigma s^2 + 4a_\star^2\sigma^2 s^2 + 4\kappa_r\kappa_u a_\star^2\sigma^2 s^4 \\ & + 2\kappa_r|a_\star|\sigma s^2\gamma^2\theta_2(1-\theta_1) + \gamma^2[1-(\theta_1+\theta_2)+2\theta_1\theta_2] \geq 0. \end{aligned}$$

  We shall see below that this constraint is weaker that another one ($f_3(s^2) \geq 0$) and needs not be taken into account.

- The condition $-\xi \leq 1 + \zeta$ reads

$$\begin{aligned} 2 - \gamma^2(\theta_1+\theta_2-2\theta_1\theta_2) - 2(\kappa_r+\kappa_u)|a_\star|\sigma s^2 \quad &\leq \quad 2 - 2(\kappa_r+\kappa_u)|a_\star|\sigma s^2 + 4a_\star^2\sigma^2 s^2 + 4\kappa_r\kappa_u a_\star^2\sigma^2 s^4 \\ &+ \quad 2\kappa_r|a_\star|\sigma s^2\gamma^2(1-\theta_1) + \gamma^2[1-(\theta_1+\theta_2)+2\theta_1\theta_2] \end{aligned}$$

  which comes down to

$$-\gamma^2 - 4a_\star^2\sigma^2 s^2 - 4\kappa_r\kappa_u a_\star^2\sigma^2 s^4 - 2\kappa_r|a_\star|\sigma s^2\gamma^2(1-\theta_1) \leq 0.$$

  This inequality always holds, so we do not need any additional constraint upon Δt.

- Finally, we consider the case $\xi \leq 1 + \zeta$. This leads to

$$\begin{aligned} -2 + \gamma^2(\theta_1+\theta_2-2\theta_1\theta_2) + 2(\kappa_r+\kappa_u)|a_\star|\sigma s^2 \quad &\leq \quad 2 - 2(\kappa_r+\kappa_u)|a_\star|\sigma s^2 + 4a_\star^2\sigma^2 s^2 + 4\kappa_r\kappa_u a_\star^2\sigma^2 s^4 \\ &+ \quad 2\kappa_r|a_\star|\sigma s^2\gamma^2(1-\theta_1)(2\theta_2-1) \\ &+ \quad \gamma^2[1-(\theta_1+\theta_2)+2\theta_1\theta_2] \end{aligned}$$

  It follows that

$$f_3(s^2) := \gamma^2\Theta_3 + 2\kappa_r|a_\star|\sigma\gamma^2 s^2(1-\theta_1)(2\theta_2-1) + 4 - 4(\kappa_r+\kappa_u)|a_\star|\sigma s^2 + 4a_\star^2\sigma^2 s^2 + 4\kappa_r\kappa_u a_\star^2\sigma^2 s^4 \geq 0. \tag{3.42}$$

  We note that $2f_2 \geq f_3$ so that indeed $f_2 \geq 0$ is a consequence of $f_3 \geq 0$.
  Since $\Theta_3 \geq -1$ and $(1-\theta_1)(2\theta_2-1) \geq -1$, this condition is always satisfied when

$$g_3(s^2) := -\gamma^2 - 2\kappa_r|a_\star|\sigma\gamma^2 s^2 + 4 - 4(\kappa_r+\kappa_u)|a_\star|\sigma s^2 + 4a_\star^2\sigma^2 s^2 + 4\kappa_r\kappa_u a_\star^2\sigma^2 s^4 \geq 0. \tag{3.43}$$

  The minimum of g_3 is reached in $s^2 = X_3$ with

$$X_3 = \frac{\frac{\kappa_r}{2}\gamma^2 + (\kappa_r+\kappa_u) - |a_\star|\sigma}{2\kappa_r\kappa_u|a_\star|\sigma}.$$

  We realize that under the condition $\Delta t \leq \frac{\kappa_r+\kappa_u}{2(1+\kappa_r\kappa_u)}\frac{\Delta x}{|a_\star|}$ (which implies $(\kappa_r+\kappa_u) \geq 2|a_\star|\sigma(1+\kappa_r\kappa_u)$) we have

$$X_3 \geq 1 + \frac{1}{2\kappa_r\kappa_u} > 1.$$

  Therefore, the stability condition is satisfied when

$$g_3(s^2=1) = -\gamma^2 - 2\kappa_r|a_\star|\sigma\gamma^2 + 4 - 4(\kappa_r+\kappa_u)|a_\star|\sigma + 4a_\star^2\sigma^2 + 4\kappa_r\kappa_u a_\star^2\sigma^2 \geq 0. \tag{3.44}$$

  One sufficient condition for (3.44) is given by

$$\Delta t \leq \Delta t_b := \min\left\{\frac{1}{2(\kappa_r+\kappa_u)}\frac{\Delta x}{|a_\star|}, \frac{1}{\omega}\right\}.$$

  because then $\gamma \leq 1$, moreover $2\kappa_r|a_\star|\sigma\gamma^2 \leq \frac{\kappa_r}{\kappa_r+\kappa_u}$ and $4(\kappa_r+\kappa_u)|a_\star|\sigma \leq 2$.

As a special case, going back to (3.42), we now notice that when $\theta_2 = \frac{1}{2}$ the first two terms of (3.42) vanish and the condition $f_3(s^2) \geq 0$ leads to

$$4 - 4(\kappa_r + \kappa_u)|a_\star|\sigma + 4a_\star^2\sigma^2 + 4\kappa_r\kappa_u a_\star^2\sigma^2 \geq 0.$$

This condition is fulfilled when

$$1 - (\kappa_r + \kappa_u)|a_\star|\sigma + \kappa_r\kappa_u a_\star^2\sigma^2 \geq 0. \tag{3.45}$$

Solving (3.45) is easy and leads to the following sufficient condition

$$\Delta t \leq \Delta t_b := \min\left\{\frac{1}{\kappa_r}, \frac{1}{\kappa_u}\right\}\frac{\Delta x}{|a_\star|}.$$

□

3.4 Numerical results

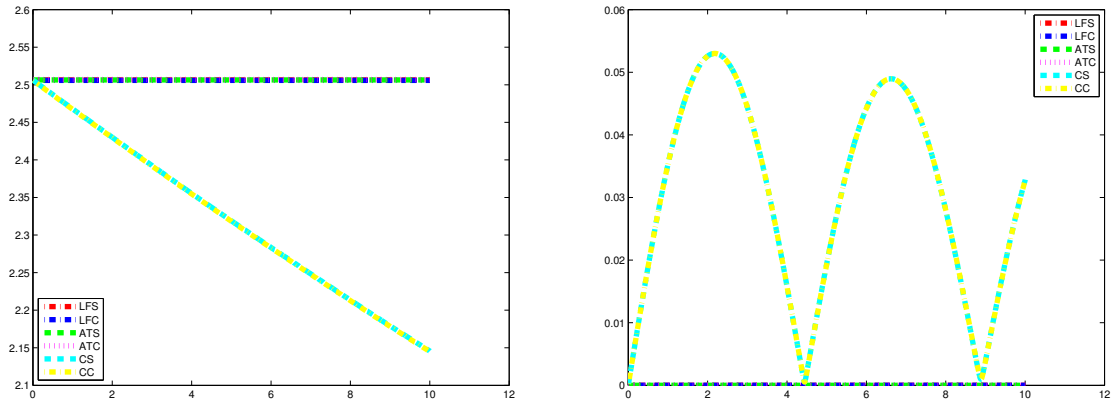
3.4.1 Well balanced test case

Let us fix the parameters $a_\star = 1$, $\omega = 1$, $\theta_1 = \frac{1}{2}, \theta_2 = \frac{1}{2}$ and consider the initial condition on the periodic domain $(0, 2\pi)$

$$q^0 = \Big(\sin(\omega x), 0, a_\star \cos(\omega x)\Big) \tag{3.46}$$

which is in the kernel $\mathcal{E}_{\omega\neq 0}$. At the discrete level, in order to ensure that q_h^0 is in the discrete kernel, we first interpolate r^0 to where it is located according to the scheme ($r_{i+1/2}$ for staggered schemes and r_i for collocated schemes), and then use the definitions of the discrete kernels to compute v_i from $r_{i+1/2}$ and $r_{i-1/2}$ for the staggered schemes and from r_{i+1} and r_{i-1} for the collocated schemes.

Figure 3.9 indicates that the classical scheme is unable to capture the discrete kernel since it introduces spurious waves in the orthogonal subspace (Figure 3.9b). Moreover, the damping of the kernel part in Figure 3.9a is another evidence to show the incorrect behaviour of the classical scheme. On the contrary, the Low Froude and Apparent Topography strategies are well-balanced schemes. This is because they preserve the kernel part and do not create any wave in the orthogonal part as well.



(a) $\|\hat{q}\|(t)$ **(b)** $\|\tilde{q}\|(t)$

Figure 3.9: *Well-balanced test case: the evolution of the kernel and orthogonal components of the fully discrete scheme.*

3.4.2 Orthogonality preserving test case

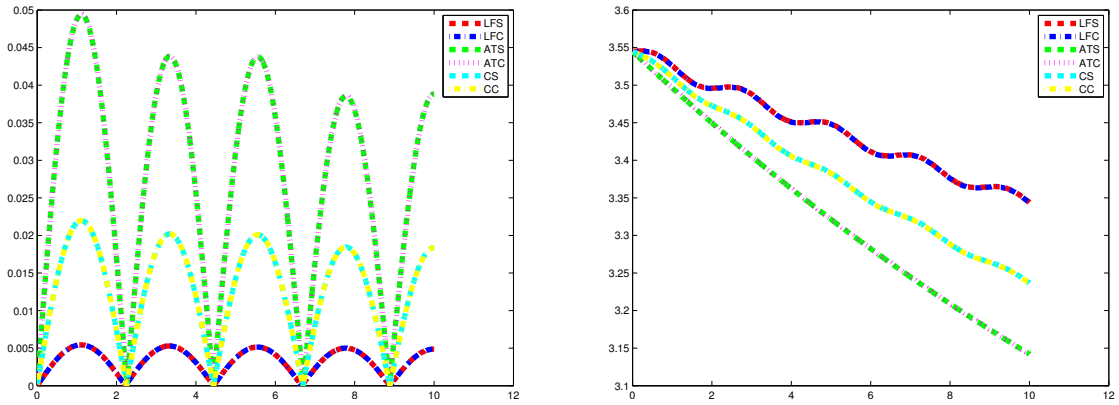
In this test case, we investigate the behavior of the numerical scheme with one initial condition in the orthogonal subspace given by

$$q^0 = \Big(a_\star \cos(\omega x), 1, \sin(\omega x)\Big) \tag{3.47}$$

At the discrete level, in order to ensure that q_h^0 is in the discrete kernel, we first interpolate v^0 to define v_i, and then use the definitions of the orthogonal of the discrete kernels to compute $r_{i+1/2}$ from v_{i+1} and v_i for the staggered schemes and r_i from v_{i+1} and v_{i-1} for the collocated schemes.

Figure 3.10b shows that the orthogonal part is damped much faster with the Apparent Topography schemes than with the classical and Low Froude schemes. As a result, the Apparent Topography schemes create waves with larger amplitudes in the kernel than the other schemes, which is shown in Figure 3.10a. As can be seen, although the classical Low Froude is not orthogonality preserving, it introduces waves with smaller amplitudes in the kernel than the other schemes.

In Figure 3.11, we make the analysis for some different Low Froude schemes. The parameter τ stands for the time discretization of the velocity u in the pressure equation (we replace u^n by $\tau u^n + (1-\tau)u^{n+1}$. So the explicit choice corresponds to the case $\tau = 1$ (note that the scheme remains explicit even when $\tau < 1$ because u^{n+1} can be computed independently of r^{n+1}). From Figure 3.11, we can observe that only the Low Froude scheme with $\tau = \frac{1}{2}$ can preserve the orthogonal subspace. On the other hand, this Figure also shows that the damping rate in the orthogonal part also depends on the parameter τ. The more implicit, the larger the damping rate.



(a) $\|\hat{q}\|(t)$ **(b)** $\|\tilde{q}\|(t)$

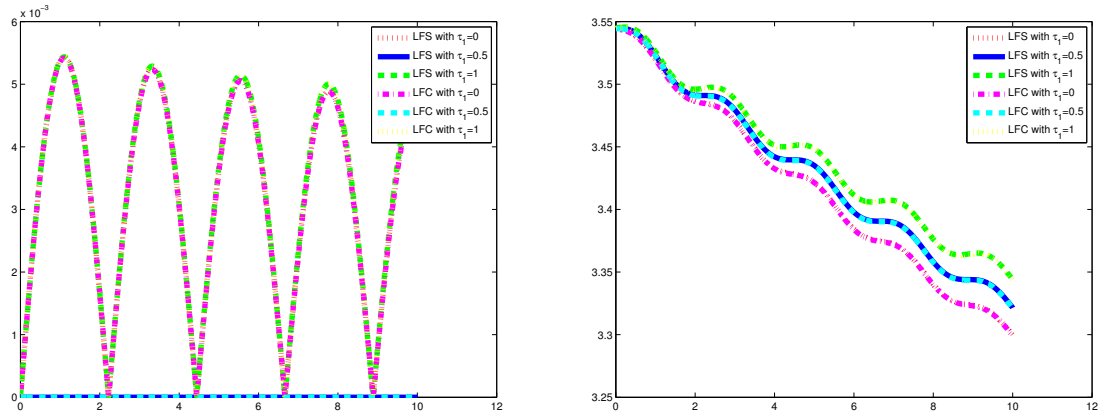
Figure 3.10: *Orthogonality preserving test case: the evolution of the kernel and orthogonal part of the fully discrete scheme.*

3.4.3 Accuracy at low Froude number test case

We now consider the following condition

$$q_i^0 = \hat{q}_i^0 + M\frac{\tilde{q}_i^0}{\|\tilde{q}_i^0\|} \quad \text{where} \begin{cases} \hat{q}^0(x) &= \Big(\sin(\omega x), 0, a_\star \cos(\omega x)\Big) &\in \mathcal{E}_{\omega\neq 0}, \\ \tilde{q}^0(x) &= \Big(a_\star \cos(\omega x), 1, \sin(\omega x)\Big) &\in \mathcal{E}_{\omega\neq 0}^{\perp}, \end{cases}$$

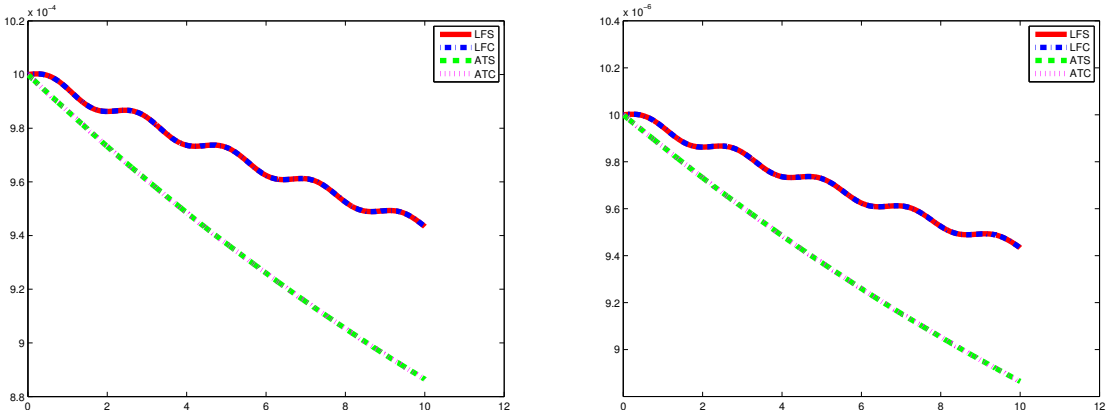
which is close to the kernel $\mathcal{E}_{\omega\neq 0}$ up to a perturbation of order M, and where the discrete components $\hat{q}_i^0$ and $\tilde{q}_i^0$ are computed as in the previous two test cases.



(a) $\|\hat{q}\|(t)$ **(b)** $\|\tilde{q}\|(t)$

Figure 3.11: *Orthogonal preserving test case: the evolution of the kernel and orthogonal components of the Low Froude type schemes.*

We can observe from Figure 3.12 that all the presented well-balanced schemes are accurate at low Froude number, since, for those schemes, the deviations of the solution from the projection of the initial condition into the kernel are of size $\mathcal{O}(M)$ when the initial condition is close to the discrete kernel ($\|q^0 - \mathbb{P}q^0\| = \mathcal{O}(M)$). On the other hand, the total deviation for the Low Froude scheme remains with a norm much larger than with the Apparent Topography scheme. One explanation for this is that we have more damping with the Apparent Topography scheme in the orthogonal subspace. This also implies that the Low Froude scheme tends to the steady state slower than the Apparent Topography scheme.



(a) $\|q - \mathbb{P}q^0\|(t)$ with $M = 10^{-3}$ **(b)** $\|q - \mathbb{P}q^0\|(t)$ with $M = 10^{-5}$

Figure 3.12: *Accuracy at low Froude number test case: deviation of the solution from the initial projection in the kernel for various fully discrete schemes.*

3.4.4 Water column test case and geostrophic adjustment

In this test, we consider the initial condition given by

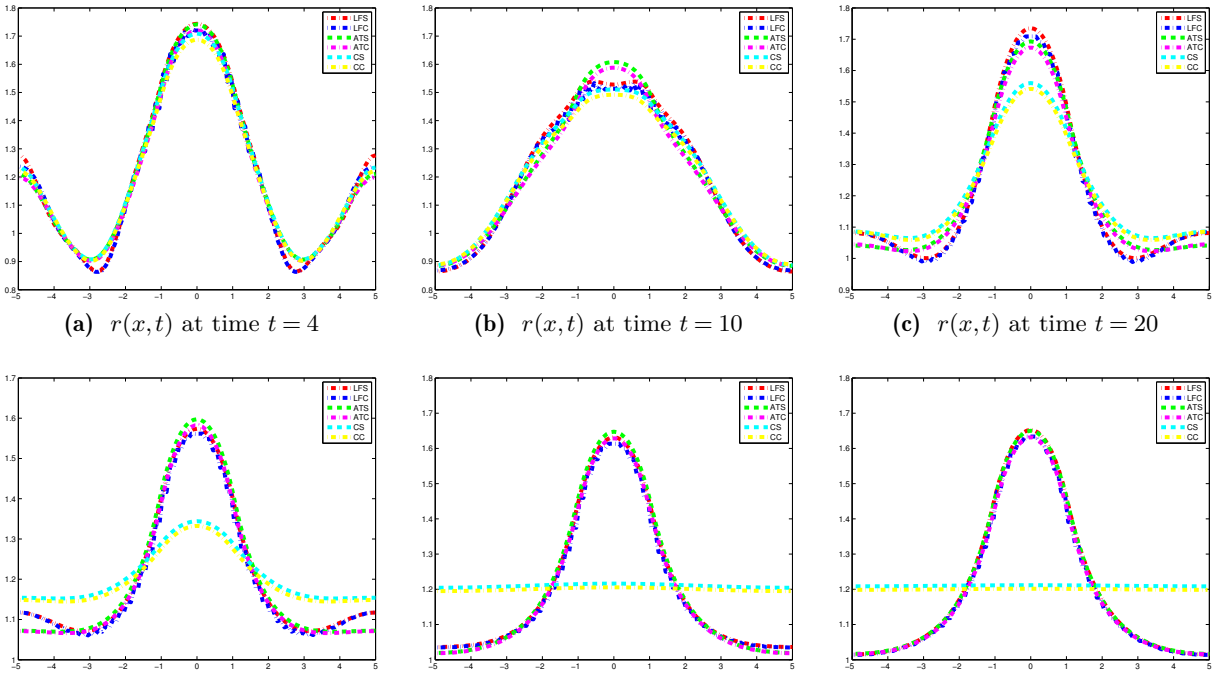
$$\begin{cases} r(x,t=0) = \begin{cases} 1+A_0, & \text{if} \quad |x| \leq R_0 \\ 1, & \text{if} \quad |x| > R_0. \end{cases} \\ u(x,t=0) = 0, \\ v(x,t=0) = 0. \end{cases} \tag{3.48}$$

with periodic boundary condition on the domain $[-5,5]$, $\theta_1 = 1$, $\theta_2 = 0$, $a_\star = \omega = 1$ and 100 grid cells. We note that this initial condition is far from the geostrophic equilibrium.

Figure 3.13 and 3.14 respectively present the evolution of the pressure and vertical velocity. We can observe that with the classical scheme, the pressure tends to a constant and the vertical velocity has a tendency to zero. This corresponds to the discrete kernel (3.8). However, the Low Froude and Apparent Topography schemes tend to another steady state corresponding to the correct discrete kernel (3.9). On the other hand, as can be seen, there are small oscillations during the evolution of the LFC scheme. One possible reason for this problem comes from the fact that the dispersion relation of the collocated schemes is not monotone. Therefore, we really need the damping effect for the waves with shortest wavelengths to avoid this unwanted behaviour of the collocated scheme. Since the damping of the ATC is twice larger than that of the LFC scheme, the oscillation problem is avoided with this scheme.

The adjustment process of the staggered schemes is shown in Figure 3.15 and 3.16. We can see that the ATS scheme tend to the geostrophic equilibrium faster than the LFS scheme. These figures are another evidence to confirm that the Low Froude and Apparent Topography strategies have a correct discrete steady state.

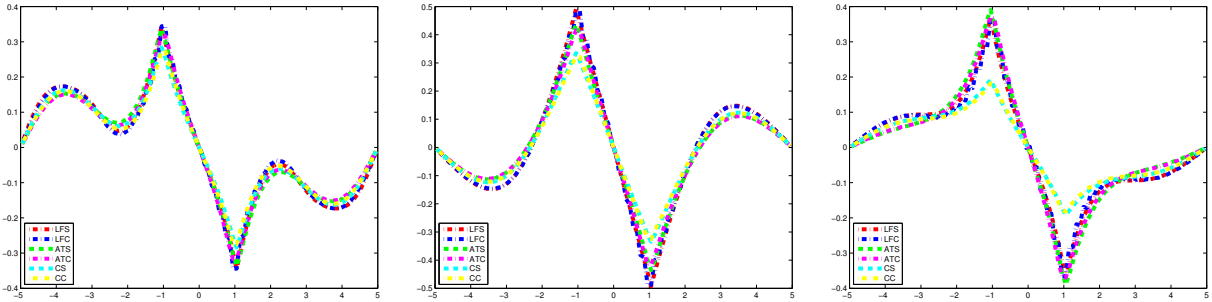
The final state of a well balanced scheme also depends on the parameter A_0 and R_0. This property is presented in Figure 3.17. The positive value of A_0 stands for the unbalanced height elevation and the negative value of A_0 corresponds to the height depression.



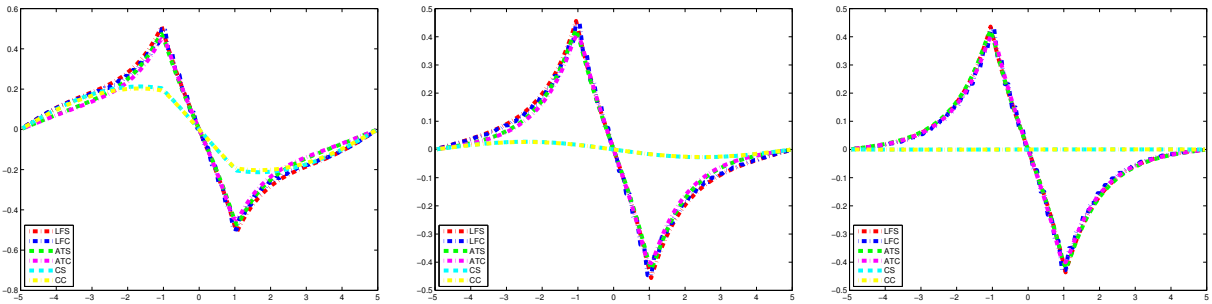
(a) $r(x,t)$ at time $t = 4$ **(b)** $r(x,t)$ at time $t = 10$ **(c)** $r(x,t)$ at time $t = 20$

(d) $r(x,t)$ at time $t = 40$ **(e)** $r(x,t)$ at time $t = 200$ **(f)** $r(x,t)$ at time $t = 400$

Figure 3.13: *Water column test case: the evolution of the pressure with* $A_0 = R_0 = 1$

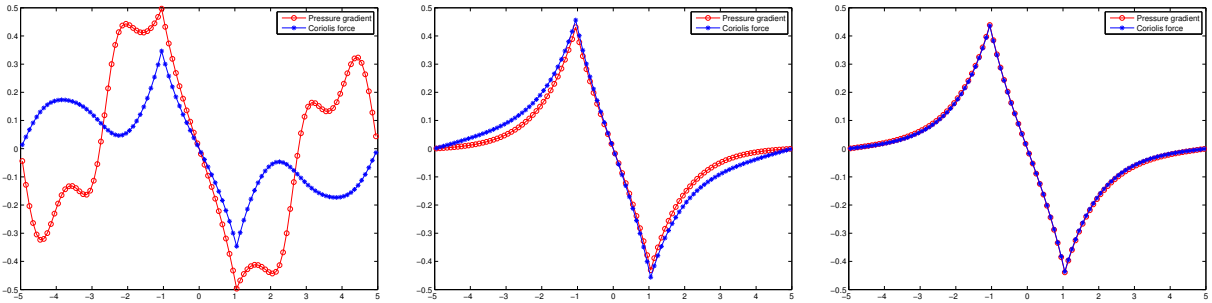


(a) $v(x,t)$ at time $t = 4$ **(b)** $v(x,t)$ at time $t = 10$ **(c)** $v(x,t)$ at time $t = 20$



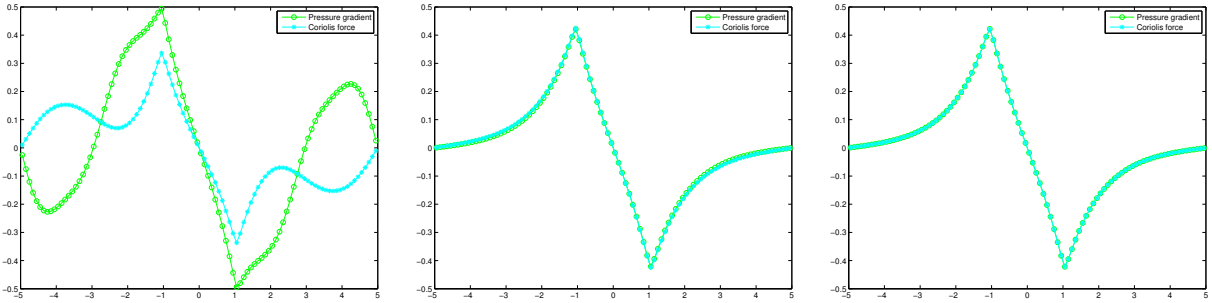
(d) $v(x,t)$ at time $t = 40$ **(e)** $v(x,t)$ at time $t = 200$ **(f)** $v(x,t)$ at time $t = 400$

Figure 3.14: *Water column test case: the evolution of the vertical velocity with* $A_0 = R_0 = 1$



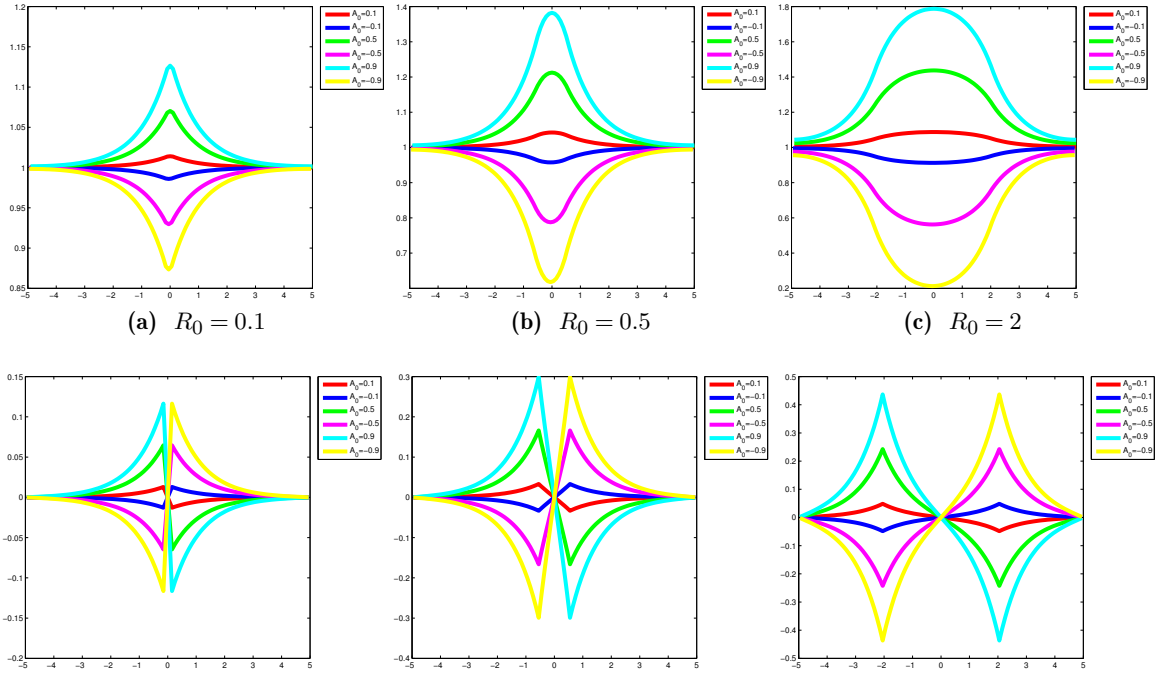
(a) $t = 4$ **(b)** $t = 200$ **(c)** $t = 400$

Figure 3.15: *The pressure gradient and Coriolis force of the LFS scheme with* $A_0 = R_0 = 1$



(a) $t = 4$ **(b)** $t = 200$ **(c)** $t = 400$

Figure 3.16: *The pressure gradient and Coriolis force of the ATS scheme with* $A_0 = R_0 = 1$



(a) $R_0 = 0.1$ (b) $R_0 = 0.5$ (c) $R_0 = 2$

(d) $R_0 = 0.1$ (e) $R_0 = 0.5$ (f) $R_0 = 2$

Figure 3.17: *Water column test case with different values of A_0 and R_0 at time $t = 400$: pressure r (top row) and vertical velocity v (bottom row).*

3.5 Conclusion

In the present work, we adapt the Low Froude and Apparent Topography strategies, developed for collocated grids in [13, 37], to staggered grids in order to obtain the Low Froude staggered scheme and the Apparent Topography staggered scheme, which are able to capture correctly discrete steady states. The new schemes on staggered grids have better dispersion laws than those on collocated grids. More importantly, the Low Froude staggered scheme is robust with respect to the time step as well as to the relation between the Rossby deformation and space step. On the other hand, the Low Froude staggered scheme is accurate at low Froude number when the initial condition is close to the kernel, and this scheme may also be tuned to preserve the orthogonal subspace. Besides the properties of the numerical schemes, we provide CFL conditions to ensure the stability of the staggered schemes in Theorem 3.2. We now aim to extend these results to the 2D case with Arakawa A-E grids [40, 41].

3.A Analysis of staggered type schemes without diffusion term

3.A.1 MAC type schemes

The are some kinds of MAC schemes which can be applied to the linear wave equation (3.2). One of them is simply given by

$$\begin{cases} \frac{u_i^{n+1}-u_i^n}{\Delta t} + a_\star \frac{r_{i+1/2}^{n+1/2}-r_{i-1/2}^{n+1/2}}{\Delta x} = \omega \frac{v_i^{n+1}+v_i^n}{2}, \\ \frac{v_i^{n+1}-v_i^n}{\Delta t} = -\omega \frac{u_i^{n+1}+u_i^n}{2}, \\ \frac{r_{i+1/2}^{n+3/2}-r_{i+1/2}^{n+1/2}}{\Delta t} + a_\star \frac{u_{i+1}^{n+1}-u_i^{n+1}}{\Delta x} = 0 \end{cases} \tag{3.A.1}$$

which computes the velocities at the cell centers and the pressures at the interfaces. In this scheme, we use a staggered discretization in time, in which the unknowns are not defied at the same time. Particularly, the pressure field is defined at $t^{n+1/2} = (n + \frac{1}{2})\Delta t$ and the velocity field is defined at $t^n = n\Delta t$. Let us note that since we do not have the velocity field at time $t^{n+1/2} = (n + \frac{1}{2})\Delta t$, we use a semi-implicit discretization in time for the Coriolis source term. One may think that we can also use a staggered grid in time for the velocity field by defining the discrete horizontal and vertical velocities respectively at t^n and $t^{n+1/2}$ in order to get another kind of MAC scheme, namely MAC-ω. This scheme can be written as

$$\begin{cases} \frac{u_i^{n+1}-u_i^n}{\Delta t} + a_\star \frac{r_{i+1/2}^{n+1/2}-r_{i-1/2}^{n+1/2}}{\Delta x} = \omega v_i^{n+1/2}, \\ \frac{v_i^{n+3/2}-v_i^{n+1/2}}{\Delta t} = -\omega u_i^{n+1}, \\ \frac{r_{i+1/2}^{n+3/2}-r_{i+1/2}^{n+1/2}}{\Delta t} + a_\star \frac{u_{i+1}^{n+1}-u_i^{n+1}}{\Delta x} = 0. \end{cases} \tag{3.A.2}$$

Remark 3.9. *The MAC type schemes* (3.A.1) *and* (3.A.2) *are second order accurate in space and in time.*

Stability condition

Lemma 3.6. *[(i)]*

i. The MAC scheme (3.A.1) *is stable when the following CFL condition is satisfied*

$$\Delta t \left(\frac{a_\star}{\Delta x} \right) \leq 1. \tag{3.A.3}$$

ii. The MAC-ω scheme (3.A.2) *is stable under the following condition*

$$\Delta t \left(\frac{a_\star}{\Delta x} + \frac{\omega}{2} \right) \leq 1. \tag{3.A.4}$$

Proof. To begin with, we multiply the first equation of (3.A.1) with $u^{n+1} + u^n$ in order to obtain

$$\frac{\|u^{n+1}\|^2 - \|u^n\|^2}{\Delta t} + a_\star \langle \partial_{x,h} r^{n+1/2}, u^{n+1} + u^n \rangle = \frac{\omega}{2} \langle v^{n+1} + v^n, u^{n+1} + u^n \rangle. \tag{3.A.5}$$

Next, by multiplying the second equation with $v^{n+1} + v^n$, we easily get

$$\frac{\|v^{n+1}\|^2 - \|v^n\|^2}{\Delta t} = -\frac{\omega}{2} \langle u^{n+1} + u^n, v^{n+1} + v^n \rangle. \tag{3.A.6}$$

On the other hand, we now multiply the final equation with $r^{n+3/2} + r^{n+1/2}$ to get the following relation

$$\frac{\|r^{n+3/2}\|^2 - \|r^{n+1/2}\|^2}{\Delta t} + a_\star \langle \partial_{x,h} u^{n+1}, r^{n+3/2} + r^{n+1/2} \rangle = 0. \tag{3.A.7}$$

By using the discrete integration by part and taking the sum of all equations from (3.A.5) to (3.A.7), we obtain one kind of conservation of energy

$$\mathbb{E}_h = \|r^{n+3/2}\|^2 + \|u^{n+1}\|^2 + \|v^{n+1}\|^2 + a_\star \Delta t \langle \partial_{x,h} u^{n+1}, r^{n+3/2} \rangle = \|r^{n+1/2}\|^2 + \|u^n\|^2 + \|v^n\|^2 + a_\star \Delta t \langle \partial_{x,h} u^n, r^{n+1/2} \rangle$$

As a result, we have

$$\|r^{n+3/2}\|^2 + \|u^{n+1}\|^2 + \|v^{n+1}\|^2 - a_\star \Delta t \|r^{n+3/2}\| . \|u^{n+1}\| \leq \mathbb{E}_h \tag{3.A.8}$$

Due to the fact that $\|\partial_{x,h}u^{n+1}\| \leq \frac{2}{\Delta x}\|u^{n+1}\|$, inequality (3.A.8) leads to

$$\|r^{n+3/2}\|^2 - 2\frac{a_\star \Delta t}{\Delta x}\|u^{n+1}\|.\|r^{3+1/2}\| + \|u^{n+1}\|^2 + \|v^{n+1}\|^2 \leq \mathbb{E}_h.$$

By the inequality $(1-\alpha)(x^2+y^2) \leq x^2 - 2\alpha xy + y^2$, this leads to

$$\left(1 - \frac{a_\star \Delta t}{\Delta x}\right)\left(\|r^{n+3/2}\|^2 + \|u^{n+1}\|^2 + \|v^{n+1}\|^2\right) \leq \mathbb{E}_h.$$

Therefore, under the stability condition given by (3.A.3), it follows that the energy is always bounded a constant

$$\|r^{n+3/2}\|^2 + \|u^{n+1}\|^2 + \|v^{n+1}\|^2 \leq \frac{\mathbb{E}_h}{1 - a_\star\left(\frac{\Delta t}{\Delta x}\right)}.$$

This proves Point (i).
For Point (ii), we follow the same way and easily obtain the conservation of the following quantity

$$\|r^{n+3/2}\|^2 + \|u^{n+1}\|^2 + \|v^{n+3/2}\|^2 + a_\star \Delta t\langle \partial_{x,h}u^{n+1}, r^{n+3/2}\rangle + \omega\Delta t\langle u^{n+1}, v^{n+3/2}\rangle = \mathbb{E}_h$$

This relation implies that

$$\left(1 - \frac{a_\star \Delta t}{\Delta x}\right)\|r^{n+3/2}\|^2 + \left(1 - \frac{a_\star \Delta t}{\Delta x} - \frac{\omega \Delta t}{2}\right)\|u^{n+1}\|^2 + \left(1 - \frac{\omega \Delta t}{2}\right)\|v^{n+3/2}\|^2 \leq \mathbb{E}_h$$

Therefore, the condition given by (3.A.4) is sufficient for stability. □

The dispersion law

We look for the solution of the MAC scheme (3.A.2) under the following discrete Fourier modes

$$r_{j+1/2}^{n+1/2} = \hat{r}e^{i(k(j+1/2)\Delta x - \ell(n+1/2)\Delta t)}, \quad u_j^n = \hat{u}e^{i(kj\Delta x - \ell n\Delta t)} \quad \text{and} \quad v_j^{n+1/2} = \hat{v}e^{i(kj\Delta x - \ell(n+1/2)\Delta t)}. \tag{3.A.9}$$

Introducing these expressions in (3.A.2), we obtain the following system

$$\begin{pmatrix} a_\star\left(\frac{e^{i\frac{k\Delta x}{2}} - e^{-i\frac{k\Delta x}{2}}}{\Delta x}\right) & \frac{e^{-\frac{i\ell\Delta t}{2}} - e^{\frac{i\ell\Delta t}{2}}}{\Delta t} & -\omega \\ 0 & \omega & \frac{e^{-\frac{i\ell\Delta t}{2}} - e^{\frac{i\ell\Delta t}{2}}}{\Delta t} \\ \frac{e^{-\frac{i\ell\Delta t}{2}} - e^{\frac{i\ell\Delta t}{2}}}{\Delta t} & a_\star\left(\frac{e^{i\frac{k\Delta x}{2}} - e^{-i\frac{k\Delta x}{2}}}{\Delta x}\right) & 0 \end{pmatrix} \begin{pmatrix} \hat{r} \\ \hat{u} \\ \hat{v} \end{pmatrix} = 0. \tag{3.A.10}$$

It is clear that (3.A.10) have nontrivial solutions when the determinant is zero. It leads to the following dispersion relation

$$\sin\left(\frac{k\Delta x}{2}\right) = 0 \ \text{(steady states)} \quad \text{or} \quad \frac{2\sin(\frac{\ell\Delta t}{2})}{\Delta t} = \pm\sqrt{\omega^2 + a_\star^2\left(\frac{2\sin(\frac{k\Delta x}{2})}{\Delta x}\right)^2} \tag{3.A.11}$$

Let us note that from the dispersion relation (3.A.11), we can say that the necessary condition for stability is

$$\Delta t\sqrt{\left(\frac{a_\star}{\Delta x}\right)^2 + \left(\frac{\omega}{2}\right)^2} \leq 1, \tag{3.A.12}$$

which is actually better than (3.A.4), which was obtained by energy estimates.

3.A.2 The forward-backward type schemes

The forward-backward scheme applied to the linear wave equation (3.2) is given by

$$\begin{cases} \frac{u_i^{n+1}-u_i^n}{\Delta t} + a_\star \frac{r_{i+1/2}^n - r_{i-1/2}^n}{\Delta x} = \omega \left[\theta_1 v_i^n + (1-\theta_1) v_i^{n+1}\right], \\ \frac{v_i^{n+1}-v_i^n}{\Delta t} = -\omega \left[\theta_2 u_i^n + (1-\theta_2) u_i^{n+1}\right], \\ \frac{r_{i+1/2}^{n+1}-r_{i+1/2}^n}{\Delta t} + a_\star \frac{u_{i+1}^{n+1}-u_i^{n+1}}{\Delta x} = 0. \end{cases} \tag{3.A.13}$$

Remark 3.10. *The forward-backward scheme (3.A.13) is second-order accurate in space, but only first-order accurate in time.*

Remark 3.11. *Although the term u^{n+1} appear in the second and the third equation, the forward-backward scheme is totally explicit in the sense that the quantity u^{n+1} is calculated in the first equation and then it is used in the other equations without having to invert any matrix.*

Lemma 3.7. *With $\theta_1 + \theta_2 \leq 1$ and $\Theta = (1-2\theta_1)(1-2\theta_2)$, we consider the following cases:*

i. If $\Theta \geq \frac{4a_\star^2}{\omega^2 \Delta x^2}$, the forward-backward scheme (3.A.13) is always stable.

ii. If $\Theta < \frac{4a_\star^2}{\omega^2 \Delta x^2}$, the stability condition of the the forward-backward scheme (3.A.13) is given by

$$\Delta t \leq \frac{1}{\sqrt{\left(\frac{a_\star}{\Delta x}\right)^2 - \left(\frac{\omega}{2}\right)^2 \Theta}} \tag{3.A.14}$$

Proof. The characteristic polynomial of the amplification matrix associated to the forward-backward scheme (3.A.13) has one solution $\lambda = 0$. The other roots are the solutions of the following equation

$$\lambda^2 + \xi \lambda + \zeta = 0$$

where

$$\xi = -\frac{2 - \gamma^2(\theta_1 + \theta_2 - 2\theta_1\theta_2) - 4a_\star^2 \sigma^2 s^2}{\Lambda(\theta_1, \theta_2)} \quad \text{and} \quad \zeta = \frac{1 + \gamma^2 \theta_1 \theta_2}{\Lambda(\theta_1, \theta_2)}.$$

Let us note that the parameters σ, γ and $\Lambda(\theta_1, \theta_2)$ are defined in the proof of theorem 3.2. When the parameters θ_1 and θ_2 satisfy the stability relation $\theta_1 + \theta_2 \leq 1$, it is obvious to see that the conditions $|\zeta| \leq 1$ and $-\xi \leq 1 + \zeta$ are always true. Hence, the stability of this scheme is obtained when $\xi \leq 1 + \zeta$, which leads to

$$\gamma^2 \Theta + 4 - 4a_\star^2 \sigma^2 \geq 0.$$

Therefore, we deduce Point (i) and (ii). □

Part II

Analysis of numerical schemes for the linear equation with Coriolis source term in 2D

CHAPTER 4

Analysis of modified Godunov type schemes for the two-dimensional linear wave equation with Coriolis source term on cartesian meshes

It always seems impossible
until it's done.

Nelson Mandela.

This work has been done in collaboration with Emmanuel Audusse, Pascal Omnes and Yohan Penel. It has been submitted.

Abstract

The study deals with colocated Godunov type finite volume schemes applied to the two-dimensional linear wave equation with Coriolis source term. The purpose is to explain the wrong behaviour of the classical scheme and to modify it in order to avoid accuracy issues around the geostrophic equilibrium and in geostrophic adjustment processes. To do so, a Hodge-like decomposition is introduced. Then three different well-balanced strategies are introduced. Some properties of the associated modified equation are proven and then extended to the semi-discrete case. Stability of fully discrete schemes under a classical CFL condition is established thanks to a Von Neumann analysis. Some numerical results reinforce the purpose.

Chapter content

4.1 Introduction

The primitive equations of the ocean are widely used to model oceanographic flows at global or regional scales, see [42] and [43, 44] for a mathematical study. The presence of specific source terms, in particular those accounting for Coriolis force and non trivial bathymetry, plays an important role and is not obvious to deal with in numerical simulations. A good model to study the impact of the discretisation of these source terms on the quality of numerical solutions is the shallow water system (4.1) presented below. It is simpler than the primitive equations, in particular due to the reduction of dimension from 3D to 2D, but still contains most of the issues raised by the presence of source terms. In this context, the discretisation of the topographic source term in a collocated finite volume framework has been dealt with in many works over the last two decades, see the reference book [25] or [45] for a more recent review. In the present work, we focus on the collocated finite volume discretisation of the Coriolis source term.

At large scales, many oceanographic flows are perturbations of the so-called geostrophic equilibrium (4.3) that corresponds to a balance between the pressure gradient and the Coriolis force. It follows that the accuracy of numerical methods is strongly related in many situations to their ability to maintain this balance. In the collocated finite volume framework, very few works were devoted to this question. In [13], see also [14, 25], the authors propose to extend a technique originally developed for the topographic term in [5]. The resulting scheme is named the *Apparent Topography* method. It turns out to yield good results for one dimensional experiments. In [27] the technique is extended to higher order schemes and assessed on two-dimensional problems. One of the main results in the present work is to prove that the Apparent Topography method alone is not accurate around the two-dimensional geostrophic equilibrium and has to be supplemented by other developments. In [28] the authors propose an alternative method, namely the Finite Volume Evolution Galerkin (FVEG) method, but they also mainly consider the one-dimensional geostrophic equilibrium, *i.e.*when the two-dimensional velocity is function of only one space variable. Very recently [46], an RS-IMEX scheme (for Reference Solution IMplicit EXplicit scheme) was designed for shallow water equations with Coriolis force and proven to be asymptotically consistent with the Quasi-Geostrophic Equations. Here we only consider time discretisations that lead to explicit computations, *i.e.*with no linear systems to solve. As previously mentioned, the present work is also restricted to the collocated finite volume framework, we refer to [12, 47] for other approaches.

The velocity field associated with the geostrophic equilibrium (4.3) is obviously divergence free. This implies that our study will share important properties with works devoted to the study of low Mach number (for Euler equations) or low Froude number (for shallow water equations) regimes. The reader is referred to [15, 20–22, 48–50] where some accurate numerical schemes are proposed. In particular, we shall often refer to the framework introduced in [20].

In the present work we investigate the preservation of the geostrophic equilibrium in the context of the two-dimensional wave equation with Coriolis force (4.2), that is the linearised version of the shallow water equations. It generalises a study initiated in [37, 38] in the one dimensional context. In Section 4.2 we recall the main characteristics of the wave equation with Coriolis force. In Section 4.3 we show that the classical collocated finite volume Godunov scheme is not accurate in the vicinity of the geostrophic equilibrium. This inaccuracy is mainly related to the numerical diffusion terms that make the stationary states of the scheme not consistent with those of the continuous model. In Section 4.4, we show that the Apparent Topography (AT) method proposed in [13] and a Divergence Penalisation (DP) method mentioned in [20] can be used to cure the problem, provided they are combined together or to other Low Froude strategies inspired from [20–22]. For that, we analyze the modified equations related to the aforementioned corrections. In Section 4.5 we turn to the related semi-discrete (in space) analysis. In particular we construct discrete operators that possess mimetic properties that are proven to be necessary

for accuracy, see also [47]. We also propose two consistent discretisations of the continuous steady states and we design the corresponding numerical schemes. Finally we exhibit the dispersion relations associated to the different numerical schemes and we prove that one of the proposed scheme is energy dissipative. In Section 4.6 we introduce the time discretisation and we prove the main result of this work which is that some of the proposed schemes are accurate and linearly stable under some CFL conditions. In Section 4.7 we illustrate the previous properties through some two dimensional numerical results.

4.2 Properties of the linear wave equation with Coriolis source term in 2D

In order to study the dimensionless shallow water equation in the rotating frame

$$\begin{cases} \mathsf{St}\,\partial_t h + \nabla\cdot(h\bar{\mathbf{u}}) = 0, \\ \mathsf{St}\,\partial_t(h\bar{\mathbf{u}}) + \nabla\cdot(h\bar{\mathbf{u}}\otimes\bar{\mathbf{u}}) + \dfrac{1}{\mathsf{Fr}^2}\nabla\left(\dfrac{h^2}{2}\right) = -\dfrac{1}{\mathsf{Fr}^2}h\nabla b - \dfrac{1}{\mathsf{Ro}}h\bar{\mathbf{u}}^\perp, \end{cases} \tag{4.1}$$

we focus on the linear wave equation with Coriolis source term

$$\begin{cases} \partial_t r + a_\star \nabla\cdot\boldsymbol{u} = 0 \\ \partial_t \boldsymbol{u} + a_\star\nabla r = -\omega\boldsymbol{u}^\perp \end{cases} \iff \partial_t q + L_w q = 0 \tag{4.2}$$

with $\boldsymbol{u} = (u,v)^T$, $\boldsymbol{u}^\perp = (-v,u)^T$, $q = (r,u,v)$ and

$$L_\omega q = \begin{pmatrix} a_\star\nabla\cdot\boldsymbol{u} \\ a_\star\nabla r + \omega\boldsymbol{u}^\perp \end{pmatrix}.$$

In the sequel, we assume that $a_* > 0$ is a constant.

System (4.2) is obtained from the shallow water equation (4.1) when the Froude number (Fr) and the Rossby number (Ro) are of order $\mathcal{O}(M)$ and the Strouhal number (St) is of order $\mathcal{O}(M^{-1})$, *i.e.* for short times, for a small parameter $M \ll 1$, and $b \equiv cte$.

To begin with, let us introduce the Hilbert space

$$\left(L^2(\mathbb{T}^2)\right)^3 = \left\{ q = (r,u,v) \;\middle|\; \int_{\mathbb{T}^2} r^2 \;\mathrm{d}\boldsymbol{x} + \int_{\mathbb{T}^2}\left(u^2+v^2\right)\;\mathrm{d}\boldsymbol{x} < \infty \right\}$$

equipped with the scalar product

$$\langle q_1, q_2\rangle = \int_{\mathbb{T}^2} r_1 r_2 \;\mathrm{d}\boldsymbol{x} + \int_{\mathbb{T}^2}(u_1u_2 + v_1v_2)\;\mathrm{d}\boldsymbol{x}.$$

4.2.1 Structure of the kernel of the original model

Since the preservation of steady-states of (4.2) is crucial in the design of accurate numerical schemes, especially in the limit $M \to 0$, we recall some well-known results about those steady-states. Let us define the kernel of the linear operator L_ω as

$$\mathcal{E}_{\omega\neq 0} := \ker L_{\omega\neq 0} = \left\{ (r,\boldsymbol{u}) \in H^1(\mathbb{T}^2)\times\left(L^2(\mathbb{T}^2)\right)^2 \;\middle|\; a_\star\nabla r = -\omega\boldsymbol{u}^\perp \right\}. \tag{4.3}$$

Since $\omega\,\boldsymbol{u} = a_\star(\nabla r)^\perp$ implies that $\nabla\cdot\boldsymbol{u} = 0$, being a steady-state of (4.2) is equivalent to belonging to $\mathcal{E}_{\omega\neq 0}$. Let us mention that $\mathcal{E}_{\omega=0}$ is named the incompressible subspace (see [20] for more details). We shall keep the same terminology in the present work. As we shall see later on, the orthogonal space of the kernel plays an important role in the analysis of the behaviour of numerical schemes. Hence the following statement:

Proposition 4.1. *The orthogonal space of $\mathcal{E}_{\omega\neq 0}$ is given by*

$$\mathcal{E}_{\omega\neq 0}^{\perp} = \left\{ (p, \boldsymbol{v}) \in L^2(\mathbb{T}^2) \times \boldsymbol{H}(\mathrm{curl}, \mathbb{T}^2) \;\middle|\; \omega p = a_\star \nabla \times \boldsymbol{v} \right\}, \tag{4.4}$$

where $\nabla \times \boldsymbol{u} := \partial_x u_y - \partial_y u_x$ and $\boldsymbol{H}(\mathrm{curl}, \mathbb{T}^2) := \left\{ \boldsymbol{u} \in \left(L^2(\mathbb{T}^2)\right)^2 \;\middle|\; \nabla \times \boldsymbol{u} \in L^2(\mathbb{T}^2) \right\}$.

Moreover, we have $\mathcal{E}_{\omega\neq 0} \oplus \mathcal{E}_{\omega\neq 0}^{\perp} = \left(L^2(\mathbb{T}^2)\right)^3$. In other words, any $q \in \left(L^2(\mathbb{T}^2)\right)^3$ can be decomposed into

$$q = \hat{q} + \tilde{q}$$

where $\hat{q} \in \mathcal{E}_{\omega\neq 0}$ and $\tilde{q} \in \mathcal{E}_{\omega\neq 0}^{\perp}$ and this decomposition is unique.

The proof of this proposition can be found in Appendix 4.A.

Remark 4.1. *The periodic boundary condition implies that for all elements $\tilde{q} = (p, \boldsymbol{v}) \in \mathcal{E}_{\omega\neq 0}^{\perp}$, we have*

$$p \in L_0^2(\mathbb{T}^2) := \left\{ f \in L^2(\mathbb{T}^2) \;\middle|\; \int_{\mathbb{T}^2} f \, \mathrm{d}\boldsymbol{x} = 0 \right\}.$$

4.2.2 Energy conservation

Let us define the energy as $E = \langle q, q \rangle$.

Proposition 4.2. *Let q be a solution of System (4.2) on $\mathbb{T}^2$. Then, the energy is preserved*

$$E(t > 0) = E(t = 0).$$

Proof. To compute the energy estimate associated to System (4.2), we directly obtain

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\langle q, q \rangle = -\langle L_\omega q, q \rangle = 0$$

since L_ω is antisymmetric. □

Remark 4.2. *Energy conservation and linearity imply uniqueness of the solution of System (4.2).*

4.2.3 Behaviour of solutions

Proposition 4.3. *Let q be the solution of System (4.2) with initial condition $q^0(\boldsymbol{x})$. Then:*

i. $\forall\, q^0(\boldsymbol{x}) \in \mathcal{E}_{\omega\neq 0}$, *we have* $q(t > 0, \boldsymbol{x}) = q^0(\boldsymbol{x}) \in \mathcal{E}_{\omega\neq 0}$.

ii. $\forall\, q^0(\boldsymbol{x}) \in \mathcal{E}_{\omega\neq 0}^{\perp}$, *we have* $q(t > 0, \boldsymbol{x}) \in \mathcal{E}_{\omega\neq 0}^{\perp}$.

Proof. Any initial condition $q^0 = (r^0, u^0, v^0) \in \mathcal{E}_{\omega\neq 0}$ is obviously a solution of (4.2); by the uniqueness property mentioned above, Point *i.* is proven.

As far as Point *ii.* is concerned, we consider $q^0 \in \mathcal{E}_{\omega\neq 0}^{\perp}$. For all $\hat{q} = (\hat{r}, \hat{\boldsymbol{u}})$ belonging to the kernel $\mathcal{E}_{\omega\neq 0}$, due to periodic boundary conditions, we obtain

$$\begin{aligned}
\left\langle \frac{\mathrm{d}}{\mathrm{d}t} q, \hat{q} \right\rangle &= -a_\star \int_{\mathbb{T}^2} \hat{r} \nabla \cdot \boldsymbol{u} \, \mathrm{d}\boldsymbol{x} - a_\star \int_{\mathbb{T}^2} \nabla r \cdot \hat{\boldsymbol{u}} \, \mathrm{d}\boldsymbol{x} - \omega \int_{\mathbb{T}^2} \boldsymbol{u}^\perp \cdot \hat{\boldsymbol{u}} \, \mathrm{d}\boldsymbol{x} \\
&= \int_{\mathbb{T}^2} \boldsymbol{u} \cdot \left(a_\star \nabla \hat{r} + \omega \hat{\boldsymbol{u}}^\perp \right) \mathrm{d}\boldsymbol{x} + a_\star \int_{\mathbb{T}^2} r \nabla \cdot \hat{\boldsymbol{u}} \, \mathrm{d}\boldsymbol{x} = 0
\end{aligned}$$

which implies that

$$\forall\, \hat{q} \in \mathcal{E}_{\omega\neq 0},\ \frac{\mathrm{d}}{\mathrm{d}t}\langle q, \hat{q}\rangle = 0 \implies \langle q(t,\cdot), \hat{q}\rangle = \langle q(t=0,\cdot), \hat{q}\rangle = 0$$

that is to say

$$q(t,\cdot) \in \mathcal{E}^{\perp}_{\omega\neq 0}.$$

This proves Point *ii.* □

Corollary 4.1. *Let q be the solution of System (4.2) with initial condition q^0. Let $\mathbb{P}q^0$ be the orthogonal projection of q^0 onto the incompressible subspace $\mathcal{E}_{\omega\neq 0}$. Then, q can be decomposed into*

$$q(t,\cdot) = \mathbb{P}q^0 + \tilde{q}(t,\cdot) \in \mathcal{E}_{\omega\neq 0} \oplus \mathcal{E}^{\perp}_{\omega\neq 0}$$

where $\tilde{q}$ is the solution of System (4.2) with initial condition $(q^0 - \mathbb{P}q^0)$.

Moreover, the conservation of the energy for $\tilde{q}$ implies that for all times $t > 0$, $\|q(t,\cdot) - \mathbb{P}q^0\| = \|q^0 - \mathbb{P}q^0\|$ which allows to say that

$$\|q^0 - \mathbb{P}q^0\| = \mathcal{O}(M) \implies \forall t > 0, \|q - \mathbb{P}q^0\|(t) = \mathcal{O}(M). \tag{4.5}$$

In other words, when the initial condition q^0 is close to the incompressible subspace $\mathcal{E}_{\omega\neq 0}$, the solution of the linear wave equation (4.2) is still close to the projection of the initial condition onto $\mathcal{E}_{\omega\neq 0}$.

One of the problems encountered with the Godunov scheme applied to (4.2) is that it does not reproduce this closeness to the projection of the initial condition on $\mathcal{E}_{\omega\neq 0}$ for values of $M \ll 1$. This inaccurate behaviour is explained in the next section. A numerical scheme for which the solution satisfies relation (4.5) will be said *accurate at low Froude number at any time*, as defined in [37].

4.3 Inaccuracy of the classical Godunov scheme

4.3.1 Numerical highlighting

We consider a cartesian mesh with mesh sizes Δx (*resp.* Δy) in the x (*resp.* y) direction. The semi-discrete Godunov scheme applied to the linear wave equation (4.2) can be written

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t} r_{i,j} + a_\star \left(\frac{u_{i+1,j} - u_{i-1,j}}{2\Delta x} + \frac{v_{i,j+1} - v_{i,j-1}}{2\Delta y} \right) - \frac{\kappa_r a_\star}{2} \left(\frac{r_{i+1,j} - 2r_{i,j} + r_{i-1,j}}{\Delta x} + \frac{r_{i,j+1} - 2r_{i,j} + r_{i,j-1}}{\Delta y} \right) = 0, \\ \frac{\mathrm{d}}{\mathrm{d}t} u_{i,j} + a_\star \frac{r_{i+1,j} - r_{i-1,j}}{2\Delta x} - \frac{\kappa_u a_\star}{2} \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{\Delta x} = \omega v_{i,j}, \\ \frac{\mathrm{d}}{\mathrm{d}t} v_{i,j} + a_\star \frac{r_{i,j+1} - r_{i,j-1}}{2\Delta y} - \frac{\kappa_v a_\star}{2} \frac{v_{i,j+1} - 2v_{i,j} + v_{i,j-1}}{\Delta y} = -\omega u_{i,j}, \end{cases} \tag{4.6}$$

where parameters κ_r, κ_u and κ_v lie in $[0,1]$ and represent the standard numerical diffusion of the Godunov type schemes. The classical Godunov scheme corresponds to $\kappa_r = \kappa_u = \kappa_v = 1$. The following facts are now well-known:

- In the 1D case, the classical Godunov scheme applied to the homogeneous system (*i.e.*with no Coriolis source term) is accurate for low M. It is no more the case when the Coriolis force is involved, see [37, 38]. Indeed, in that case the diffusion on the pressure equation is shown to be responsible for the inaccuracy. A simple correction consists in setting $\kappa_r = 0$ and is proven to be a stable strategy in [37]. This scheme is referred to in the sequel as the LF-C strategy, since we adapt the diffusion in the pressure equation to the Low-Froude (LF) case and we keep the classical (C) diffusion on the velocity equation.

- In the 2D case, the classical Godunov scheme applied to the homogeneous system (*i.e.*with no Coriolis source term) is inaccurate for low M on cartesian meshes. This is due to the numerical viscosity on the velocity equation, see [20, 21] for more details. It is corrected in [20] by setting $\kappa_u = \kappa_v = 0$ to obtain a stable and accurate scheme. This scheme is referred to in the sequel as the C-LF strategy, since we keep the classical (C) diffusion on the pressure equation and we adapt the diffusion in the velocity equation to the Low-Froude (LF) case.

The purpose here is to show that none of these corrections (neither the LF-C nor the C-LF strategies) cures the inaccuracy of the Godunov scheme applied to the 2D wave equation with a Coriolis source term for low values of M. To do so, we consider the classical Godunov scheme and the modified versions proposed in [20, 37] when the initial condition is at the geostrophic equilibrium (see Fig. 4.1)

$$\begin{cases} r(t=0,x,y) = 1 - \exp\left[-\left(\frac{3x}{0.5}\right)^2 - \left(\frac{3y}{0.5}\right)^2\right] \\ u(t=0,x,y) = -\frac{6y}{0.5}\exp\left[-\left(\frac{3x}{0.5}\right)^2 - \left(\frac{3y}{0.5}\right)^2\right] \\ v(t=0,x,y) = \frac{6x}{0.5}\exp\left[-\left(\frac{3x}{0.5}\right)^2 - \left(\frac{3y}{0.5}\right)^2\right]. \end{cases} \tag{4.7}$$

in the periodic domain $\mathbb{T}^2 = [-0.5, 0.5] \times [-0.5, 0.5]$. This initial condition is obviously a steady state of System (4.2).

In Figures 4.2 and 4.3, we present the numerical results for a 50×50 grid at time $t = 10$. It indicates that the Godunov type schemes with standard diffusion (Fig. 4.2(b)), and both corrected LF-C (Fig. 4.2(c)) and C-LF (Fig. 4.2(d)) schemes are unable to capture the steady state. At the same time, it is not possible to use a LF-LF strategy and completely delete both diffusion terms (*i.e.*using $\kappa_r = \kappa_u = \kappa_v = 0$), because the resulting fully discrete explicit scheme would then obviously be unstable. Note that for this test, the modification on the diffusion in the velocity equation provides more substantial improvements than the modification on the diffusion in the pressure equation: in the first case, the 2D structure of the solution is more or less preserved, see Figure 4.2, and the solution remains not so far from the exact one, see Figure 4.3, whereas in the second case, the solution is quite close to the one of the classical scheme. Nevertheless, this behaviour is related to this particular test case and we need more investigations to obtain numerical schemes which are able to exactly preserve steady states and then be accurate in any situation.

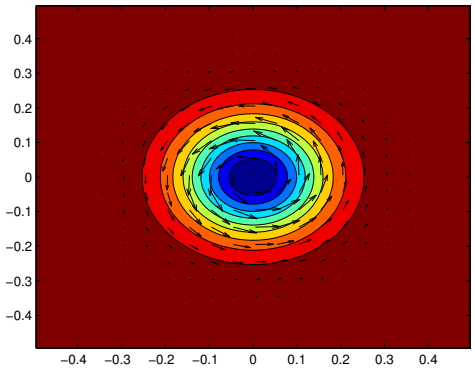
Before doing that, let us analyze the discrete kernel of the semi-discrete Godunov scheme in order to point out the main reason of the inaccuracy problem.

4.3.2 Analysis of the discrete kernel

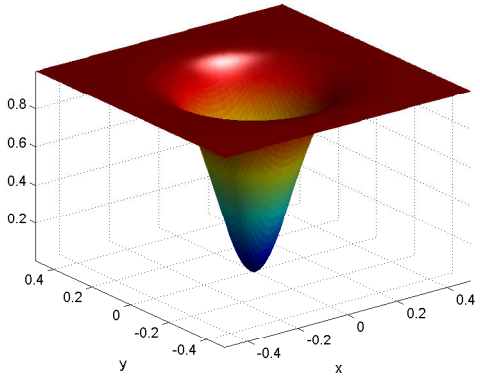
Let us denote by $L_{\omega,\kappa,h}$ the spatial operator in the semi-discrete scheme (4.6), so that (4.6) reads $q'_{i,j} + L_{\omega,\kappa,h} q_{i,j} = 0$.

Lemma 4.1. *Let us define the discrete energy of System (4.6) by*

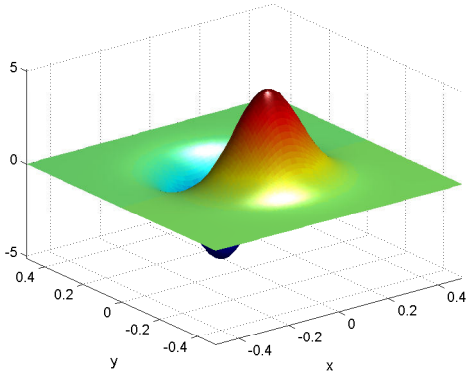
$$E_h(t) = \Delta x \Delta y \left[\sum_{i,j} r_{i,j}(t)^2 + \sum_{i,j} u_{i,j}(t)^2 + \sum_{i,j} v_{i,j}(t)^2\right].$$



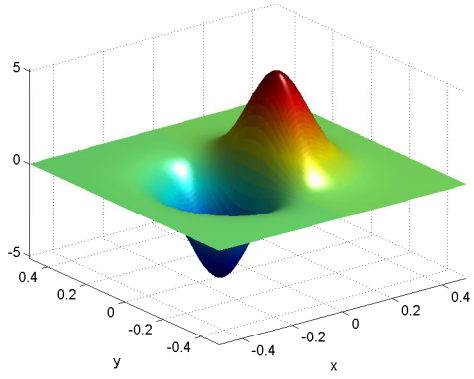
(a) Contour of pressure and vector field



(b) Pressure r

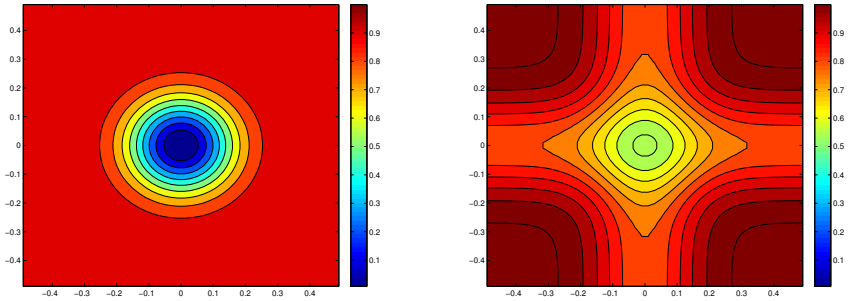


(c) Horizontal velocity u

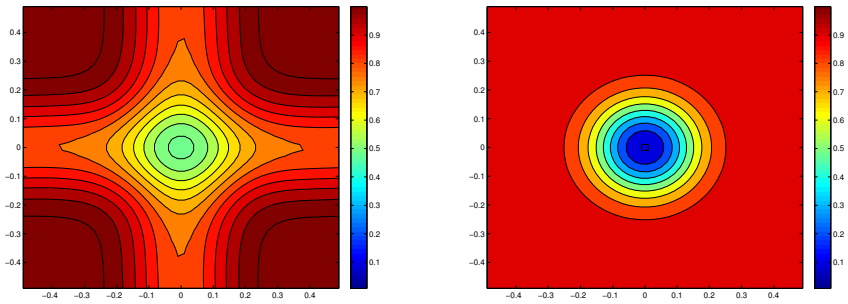


(d) Vertical velocity v

Figure 4.1: *Initial condition: stationary vortex.*



(a) Initial condition **(b)** Classical Godunov scheme



(c) LF-C scheme **(d)** C-LF scheme

Figure 4.2: *Contours of the pressure r.*

Then for any $\kappa_{r,u,v} \in [0,1]$

$$\frac{\mathrm{d}}{\mathrm{d}t} E_h(t) \leq 0,$$

which means that the energy associated to Godunov type schemes is dissipated.

Proof. Let us multiply the semi-discrete scheme (4.6) by $q_{i,j}\Delta x\Delta y$ and sum over all cells (i,j). Due to periodic boundary conditions, we obtain by standard calculations

$$\left\langle \frac{\mathrm{d}q}{\mathrm{d}t}, q \right\rangle = -\frac{\kappa_r a_\star \Delta y}{2} \sum_{i,j} (r_{i+1,j} - r_{i,j})^2 - \frac{\kappa_r a_\star \Delta x}{2} \sum_{i,j} (r_{i,j+1} - r_{i,j})^2 - \frac{\kappa_u a_\star \Delta y}{2} \sum_{i,j} (u_{i+1,j} - u_{i,j})^2 - \frac{\kappa_v a_\star \Delta x}{2} \sum_{i,j} (v_{i,j+1} - v_{i,j})^2 \leq 0. \quad (4.8)$$

□

Although the semi-discrete scheme is energy-dissipative, the fact still remains that it is not consistent with the incompressible space.

Lemma 4.2.

- *For $\kappa_r \neq 0$, $\kappa_u \neq 0$ and $\kappa_v \neq 0$, i.e. for the classical Godunov scheme, we have*

$$\ker L_{\omega,\kappa,h} = \left\{ q = (r,u,v) \in \mathbb{R}^{3N} \,\middle|\, \exists\, c \in \mathbb{R},\, r = c, u = 0, v = 0 \right\} \quad (4.9a)$$

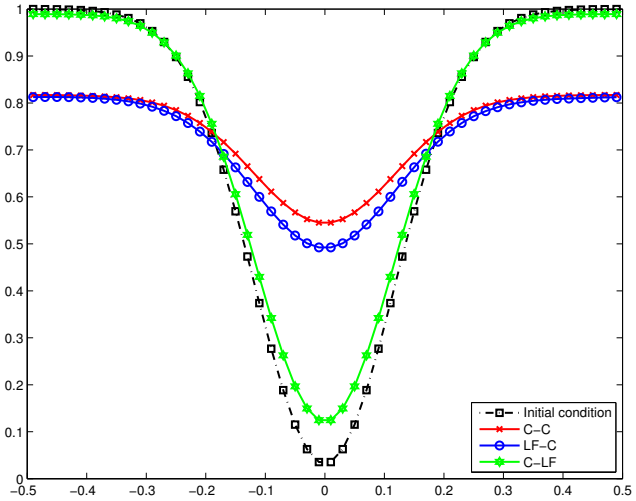


Figure 4.3: *Cross-section ($y=0$) of the pressure r.*

- *For $\kappa_r \neq 0$, $\kappa_u = \kappa_v = 0$, i.e. for the C-LF strategy, we have*

$$\ker L_{\omega,\kappa,h} = \left\{ q=(r,u,v)\in\mathbb{R}^{3N} \;\middle|\; \exists\, c\in\mathbb{R},\ r=c, u=0, v=0 \right\} \tag{4.9b}$$

- *For $\kappa_r = 0$, $\kappa_u \neq 0$ and $\kappa_v \neq 0$, i.e. for the LF-C strategy, we have*

$$\ker L_{\omega,\kappa,h} = \left\{ q=(r,u,v)\in\mathbb{R}^{3N} \;\middle|\; \exists (u_j,v_i)\in\mathbb{R}^N\times\mathbb{R}^N, \forall\,(i,j),\ u_{i,j}=u_j,\ v_{i,j}=v_i, \right.$$
$$\left. a_\star \begin{pmatrix} \frac{r_{i+1,j}-r_{i-1,j}}{2\Delta x} \\ \frac{r_{i,j+1}-r_{i,j-1}}{2\Delta y} \end{pmatrix} = -\omega \begin{pmatrix} -v_i \\ u_j \end{pmatrix} \right\} \tag{4.9c}$$

Remark 4.3. *Although all kernels above correspond to discrete versions of the exact relation $a_*\nabla r = -\omega \boldsymbol{u}^\perp$, constraints upon the velocity field are too strong, so that those kernels do not match with the exact one $\mathcal{E}_{\omega\neq 0}$ defined in (4.3).*

Proof. Since any steady state of (4.6) satisfy $\frac{\mathrm{d}q_{i,j}}{\mathrm{d}t}=0$, Equation (4.8) implies

$$\sum_{i,j}\left[\kappa_r\left(\Delta y(r_{i+1,j}-r_{i,j})^2+\Delta x(r_{i,j+1}-r_{i,j})^2\right)+\kappa_u\Delta y(u_{i+1,j}-u_{i,j})^2+\kappa_v\Delta x(v_{i,j+1}-v_{i,j})^2\right]=0. \tag{4.10}$$

- When $\kappa_r \neq 0$, we easily get from (4.10)

$$\forall\,(i,j)\in[1,N_x]\times[1,N_y],\ r_{i+1,j}=r_{i,j} \text{ and } r_{i,j+1}=r_{i,j} \implies r=\text{const}. \tag{4.11}$$

  When $\kappa_u \neq 0$ and $\kappa_v \neq 0$, we also have $u_{i+1,j}=u_{i,j}$ and $v_{i,j+1}=v_{i,j}$ for all (i,j), which implies that there exist $(u_j,v_i)\in\mathbb{R}^N\times\mathbb{R}^N$ such that

$$u_{i,j}=u_j \text{ and } v_{i,j}=v_i \ \forall (i,j).$$

  Going back to (4.6), $r=const$ implies that $u=0$ and $v=0$. Therefore, this leads to (4.9a).

- Likewise, when $\kappa_r \neq 0$ but $\kappa_u=\kappa_v=0$, (4.11) still holds. Then we deduce from (4.6) that $u_{i,j}=v_{i,j}=0$ and consequently (4.9b).

- Now, we consider the case $\kappa_r = 0$ (and $\kappa_u \neq 0$, $\kappa_v \neq 0$). We deduce from (4.10) that $u_{i,j} = u_j$ and $v_{i,j} = v_i$. Hence, the steady version of (4.6) now reads

$$\frac{a_\star}{2\Delta x}(r_{i+1,j} - r_{i-1,j}) = \omega v_i \quad \text{and} \quad \frac{a_\star}{2\Delta y}(r_{i,j+1} - r_{i,j-1}) = -\omega u_j$$

  which is nothing but (4.9c).

□

4.4 Properties of the first order modified equation with correction terms

In the previous section, we have shown that the classical Godunov scheme is inaccurate near the geostrophic equilibrium and that simple corrections consisting in deleting diffusion terms (LF-C or C-LF strategies) are not enough to ensure the accuracy. As it is not possible to delete all diffusion terms at the same time for stability reasons, it is essential to introduce some correction terms for the standard diffusion.

We aim at deriving a numerical scheme which is able to preserve steady states or to be accurate around steady states. It is worth pointing out that we not only have to deal with the balance between the pressure gradient and the Coriolis force but also to take into account the divergence free condition.

4.4.1 Definition of the schemes

We mention below two strategies to deal with diffusion terms. Each strategy leads to a different numerical scheme which will be referred to as X-Y scheme where X is related to the diffusion on the pressure equation and Y to the diffusion on the velocity equation.

i. The *Apparent Topography* scheme was introduced in [13], see also [27], to deal with the geostrophic equilibrium in the 1D nonlinear shallow water system. This strategy was proven to be stable in [38] for the 1D linear wave equation. Here we extend the procedure to the 2D linear wave equation. We notice that the steady state defined by $a_\star \nabla r = -\omega \boldsymbol{u}^\perp$ also satisfies

$$\nabla \cdot \left(\nabla r + \frac{\omega}{a_\star} \boldsymbol{u}^\perp \right) = 0.$$

   It suggests that the numerical diffusion on the pressure equation can be modified into $\nabla \cdot (\nabla r + \frac{\omega}{a_\star} \boldsymbol{u}^\perp)$ instead of the classical operator Δr – see (4.6). As for the velocity equations, either we keep the classical diffusion to obtain the *Apparent Topography-Classical scheme* (AT-C) or we delete diffusion terms which leads to the *Apparent Topography-Low Froude scheme* (AT-LF). In 1D, they are shown to be both stable and accurate [38].

ii. The *Divergence Penalisation* method consists in a modification on the diffusion on the velocity equation that is based on the operator $\nabla(\nabla \cdot \boldsymbol{u})$ instead of the classical diffusion in (4.6) since the equilibrium states satisfy $\nabla \cdot \boldsymbol{u} = 0$. This idea was mentioned in [20, § 5.6] to be applied to the homogeneous linear wave equation, but not studied. We propose to extend it to the present case and to analyze its properties. As for the pressure equation, we can choose the classical diffusion to obtain the *Classical-Divergence Penalisation* scheme (C-DP) or delete this diffusion term to get the *Low Froude-Divergence Penalisation* scheme (LF-DP).

iii. Finally, we can combine both strategies to get the *Apparent Topography-Divergence Penalisation* method (AT-DP). It comes down to considering $\nabla \cdot (\nabla r + \frac{\omega}{a_\star} \boldsymbol{u}^\perp)$ for the diffusion on the pressure equation and $\nabla(\nabla \cdot \boldsymbol{u})$ for the velocity equation.

Schemes	κ_r	κ_u	κ_v	η_r	η_u	η_v
AT-LF	$\mathcal{O}(1)$	0	0	κ_r	0	0
AT-C	$\mathcal{O}(1)$	$\mathcal{O}(1)$	$\mathcal{O}(1)$	κ_r	0	0
LF-DP	0	$\mathcal{O}(1)$	$\mathcal{O}(1)$	0	κ_u	κ_v
C-DP	$\mathcal{O}(1)$	$\mathcal{O}(1)$	$\mathcal{O}(1)$	0	κ_u	κ_v
AT-DP	$\mathcal{O}(1)$	$\mathcal{O}(1)$	$\mathcal{O}(1)$	κ_r	κ_u	κ_v

Table 4.1: *Parameters of Godunov type schemes with corrections.*

We shall prove below that the AT-LF, LF-DP and AT-DP approaches are accurate and stable. The AT-C and C-DP strategies, like the LF-C and C-LF ones previously mentioned in paragraph4.3.1, will not be able to cure the problem since only one issue is taken into account.

To carry out the accuracy analysis, we shall analyze the first order modified equation which is the common tool to study stability and accuracy of finite difference schemes. We refer to [51] for more details about this method. The first order modified equation corresponding to the aforementioned strategies is given by

$$\begin{cases} \partial_t r + a_\star(\partial_x u + \partial_y v) - \dfrac{\kappa_r^x a_\star \Delta x}{2}\dfrac{\partial^2 r}{\partial x^2} - \dfrac{\kappa_r^y a_\star \Delta y}{2}\dfrac{\partial^2 r}{\partial y^2} + \dfrac{\eta_r^x a_\star \Delta x}{2}\dfrac{\omega}{a_\star}\dfrac{\partial v}{\partial x} - \dfrac{\eta_r^y a_\star \Delta y}{2}\dfrac{\omega}{a_\star}\dfrac{\partial u}{\partial y} = 0, \\ \partial_t u + a_\star \partial_x r - \dfrac{\kappa_u a_\star \Delta x}{2}\dfrac{\partial^2 u}{\partial x^2} - \dfrac{\eta_u a_\star \Delta x}{2}\dfrac{\partial^2 v}{\partial x \partial y} = \omega v, \\ \partial_t v + a_\star \partial_y r - \dfrac{\kappa_v a_\star \Delta y}{2}\dfrac{\partial^2 v}{\partial y^2} - \dfrac{\eta_v a_\star \Delta y}{2}\dfrac{\partial^2 u}{\partial y \partial x} = -\omega u, \end{cases} \tag{4.12}$$

where parameters $\eta_r^x \geq 0$, $\eta_r^y \geq 0$, $\eta_u \geq 0$ and $\eta_v \geq 0$ stand for the correction terms. We recall that $\kappa_{r,u,v}^{x,y} \in [0,1]$. The modified equation reads in a compact form

$$\begin{cases} \partial_t q + \mathcal{L} q = 0, & \text{(4.13a)} \\ q(t=0,\boldsymbol{x}) = q^0(\boldsymbol{x}). & \text{(4.13b)} \end{cases}$$

The spatial operator is defined by $\mathcal{L} = L_\omega - \mathcal{B}_{\kappa,\eta}$, with L_ω as in (4.2) and

$$\mathcal{B}_{\kappa,\eta} q = \begin{pmatrix} \frac{\kappa_r^x a_\star \Delta x}{2}\frac{\partial^2 r}{\partial x^2} + \frac{\kappa_r^y a_\star \Delta y}{2}\frac{\partial^2 r}{\partial y^2} \\ \frac{\kappa_u a_\star \Delta x}{2}\frac{\partial^2 u}{\partial x^2} \\ \frac{\kappa_v a_\star \Delta y}{2}\frac{\partial^2 v}{\partial y^2} \end{pmatrix} + \begin{pmatrix} -\frac{\eta_r^x a_\star \Delta x}{2}\frac{\omega}{a_\star}\frac{\partial v}{\partial x} + \frac{\eta_r^y a_\star \Delta y}{2}\frac{\omega}{a_\star}\frac{\partial u}{\partial y} \\ \frac{\eta_u a_\star \Delta x}{2}\frac{\partial^2 v}{\partial x \partial y} \\ \frac{\eta_v a_\star \Delta y}{2}\frac{\partial^2 u}{\partial y \partial x} \end{pmatrix}.$$

The choices of parameters in (4.12-4.13) corresponding to the aforementioned strategies are summarised in Table 4.1.

Remark 4.4. *For the numerical diffusion on the velocity equation to be consistent with $\nabla(\nabla\cdot\boldsymbol{u})$, we have to take $\kappa_u \Delta x = \kappa_v \Delta y$ and $\eta_u = \kappa_u$, $\eta_v = \kappa_v$. Likewise, to recover the operator $\nabla\cdot(\nabla r + \frac{\omega}{a_\star}\boldsymbol{u}^\perp)$, we must consider the case $\kappa_r^x \Delta x = \kappa_r^y \Delta y$, $\eta_r^x = \kappa_r^x$ and $\eta_r^y = \kappa_r^y$.*

From now on, we shall denote and assume that

$$\begin{aligned} \nu_r &= \frac{\kappa_r^x a_\star \Delta x}{2} = \frac{\kappa_r^y a_\star \Delta y}{2}, \quad \nu_u = \frac{\kappa_u a_\star \Delta x}{2} = \frac{\kappa_v a_\star \Delta y}{2}, \\ \gamma_r &= \frac{\eta_r^x a_\star \Delta x}{2} = \frac{\eta_r^y a_\star \Delta y}{2}, \quad \gamma_u = \frac{\eta_u a_\star \Delta x}{2} = \frac{\eta_v a_\star \Delta y}{2} \end{aligned} \tag{4.14}$$

in the correction terms. With such assumptions, the action of the diffusion operator may be rewritten as follows

$$B_{\kappa,\eta}q = B_{\nu,\gamma}q = \begin{pmatrix} \nabla\cdot(\nu_r\nabla r + \gamma_r\frac{\omega}{a_\star}\boldsymbol{u}^\perp) \\ \frac{\partial}{\partial x}(\nu_u\frac{\partial u}{\partial x} + \gamma_u\frac{\partial v}{\partial y}) \\ \frac{\partial}{\partial y}(\gamma_u\frac{\partial u}{\partial x} + \nu_u\frac{\partial v}{\partial y}) \end{pmatrix}.$$

Then we can study the behaviour of the schemes for an initial solution in the incompressible space $\mathcal{E}_{\omega\neq 0}$ or in its orthogonal $\mathcal{E}^\perp_{\omega\neq 0}$, see Prop. 4.3.

Proposition 4.4.

i. A solution in the incompressible space $\mathcal{E}_{\omega\neq 0}$ is preserved for all time by the LF-DP, AT-LF and AT-DP schemes.

ii. The orthogonal subspace $\mathcal{E}^\perp_{\omega\neq 0}$ is preserved by the LF-DP scheme.

Proof. All schemes such that $\gamma_r = \nu_r$ and $\gamma_u = \nu_u$ (namely LF-DP, AT-LF and AT-DP) satisfy $q \in \mathcal{E}_{\omega\neq 0} \implies \mathcal{B}_{\nu,\gamma}q = 0$ (Point *i*).
The proof for Point *ii* is very similar to the one in Prop. 4.3 up to the term

$$\nu_u \int_{\mathbb{T}^2} \hat{\boldsymbol{u}}\cdot\nabla(\nabla\cdot\boldsymbol{u})\,\mathrm{d}\boldsymbol{x} = -\nu_u\int_{\mathbb{T}^2}(\nabla\cdot\hat{\boldsymbol{u}})(\nabla\cdot\boldsymbol{u})\,\mathrm{d}\boldsymbol{x} = 0$$

since $\hat{\boldsymbol{u}}$ is in the incompressible subspace $\mathcal{E}_{\omega\neq 0}$. □

For the LF-DP strategy, we can also study the evolution of the energy, see Prop 4.2.

Proposition 4.5. *The LF-DP and C-DP schemes are energy-dissipative.*

Proof. Due to the fact that $\langle L_\omega q, q\rangle = 0$ as L_ω is antisymmetric, when $\gamma_r = 0$ and $\nu_u = \gamma_u$, we have

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|q\|^2 = \langle\mathcal{B}_{\nu,\gamma}q, q\rangle = -\nu_r\|\nabla r\|^2 - \nu_u\|\nabla\cdot\boldsymbol{u}\|^2 \leq 0.$$

This means that the modified equation associated to the LF-DP and C-DP schemes is dissipative. □

Hence the LF-DP strategy enables to mimic Corollary 4.1.

Corollary 4.2. *The solution $q_{\nu,\gamma}$ of the modified equation for the LF-DP parameters satisfies the inequality*

$$\forall t \geq 0,\ \|q_{\nu,\gamma} - \mathbb{P}q^0\|(t) \leq \|q^0 - \mathbb{P}q^0\|,$$

which means the solution is accurate at low Froude number at any time, as defined at the end of Section 4.2.

Proof. Let us first notice that $\mathbb{P}q^0$ is the solution of Eq. (4.13a) for the LF-DP parameters ($\nu_r = \gamma_r = 0$ and $\nu_u = \gamma_u$) with initial condition $\mathbb{P}q^0$ due to Prop. 4.4.*i*. We deduce by linearity that any solution of (4.13) reads $q_{\nu,\gamma}(t,\boldsymbol{x}) = \mathbb{P}q^0(\boldsymbol{x}) + \tilde{q}(t,\boldsymbol{x})$ where $\tilde{q}$ satisfies

$$\begin{cases} \partial_t\tilde{q} + \mathcal{L}\tilde{q} = 0, \\ \tilde{q}(t=0,\boldsymbol{x}) = q^0(\boldsymbol{x}) - \mathbb{P}q^0(\boldsymbol{x}). \end{cases}$$

As the energy is decreasing (Prop. 4.4.*ii*), we have

$$\|q_{\nu,\gamma} - \mathbb{P}q^0\|(t) = \|\tilde{q}\|(t) \leq \|\tilde{q}^0\| = \|q^0 - \mathbb{P}q^0\|.$$

□

4.4.2 Stability properties

For the AT strategy, we are not able to establish energy estimates as for the LF-DP strategy. So we investigate the stability of this approach by studying the behaviour of the Fourier modes of the solution.

Lemma 4.3. *Fourier modes associated to the LF-DP, C-DP, AT-LF and AT-DP schemes are damped.*

Proof. We now look for plane wave solutions of the modified equation (4.12) under the form

$$q(t,\boldsymbol{x}) = \exp\left[\imath(\tau t + \boldsymbol{k}\cdot\boldsymbol{x})\right]\hat{q} \tag{4.15}$$

where $\boldsymbol{k} = (k_x, k_y)$ is the wave number and τ is the wave frequency. To ensure that these waves are captured by the scheme, we assume

$$|\boldsymbol{k}| \leq \frac{\pi}{\Delta x}. \tag{4.16}$$

Such functions are generally solutions of the modified equation under a dispersion relation, *i.e.*a relation between τ and $\boldsymbol{k}$ commonly written as $\tau = \tau(\boldsymbol{k})$. In the present case, the Fourier modes (4.15) are some solutions provided

$$\mathcal{A}\hat{q} = -\imath\tau\hat{q} \tag{4.17}$$

and the matrix $\mathcal{A}$ is given by

$$\mathcal{A} = \begin{pmatrix} \nu_r k_x^2 + \nu_r k_y^2 & a_\star \imath k_x - \gamma_r \frac{\omega}{a_\star}\imath k_y & a_\star \imath k_y + \gamma_r \frac{\omega}{a_\star}\imath k_x \\ a_\star \imath k_x & \nu_u k_x^2 & \gamma_u k_x k_y - \omega \\ a_\star \imath k_y & \gamma_u k_x k_y + \omega & \nu_u k_y^2. \end{pmatrix}$$

The statement of the lemma is equivalent to saying that the real part of all eigenvalues are positive. Indeed, $-\imath\tau$ is an eigenvalue due to (4.17). The decrease for long times in (4.15) requires a negative coefficient for t.

The characteristic polynomial $\mathcal{P}(\lambda)$ reads

$$\begin{aligned}\mathcal{P}(\lambda) = \lambda^3 - (\nu_r + \nu_u)|\boldsymbol{k}|^2\lambda^2 + \left[a_\star^2|\boldsymbol{k}|^2 + \omega^2 + \nu_r\nu_u|\boldsymbol{k}|^4 + (\nu_u^2 - \gamma_u^2)k_x^2k_y^2\right]\lambda \\ - (\nu_r - \gamma_r)\omega^2|\boldsymbol{k}|^2 - (\nu_u^2 - \gamma_u^2)\nu_r k_x^2 k_y^2|\boldsymbol{k}|^2 - (\nu_u - \gamma_u)2a_\star^2 k_x^2 k_y^2 - \gamma_r(\nu_u - \gamma_u)k_x k_y \omega(k_x^2 - k_y^2).\end{aligned}$$

Let us mention that Prop. 4.4.*i* corresponds to the fact that $\lambda = 0$ is a root of $\mathcal{P}$ for the LF-DP, AT-LF and AT-DP schemes.

- For the *LF-DP* scheme ($\nu_r = \gamma_r = 0$ and $\nu_u = \gamma_u$), $\lambda_0 = 0$ is an eigenvalue while the other two are given by

  $$\lambda_c = \frac{\nu_u|\boldsymbol{k}|^2}{2} \pm \imath\sqrt{\omega^2 + a_\star^2|\boldsymbol{k}|^2 - \left(\frac{\nu_u}{2}\right)^2|\boldsymbol{k}|^4}.$$

  Hypothesis (4.16) and $\kappa_u \in [0,1]$ ensure that the term in the square root is positive. Hence $\Re(\lambda_c) > 0$.

- For the *C-DP* scheme ($\gamma_r = 0$ and $\nu_u = \gamma_u$), the linear system $\mathcal{A}q = \lambda q$ reads

$$\nu_r|\boldsymbol{k}|^2 r + \imath a_\star k_x u + \imath a_\star k_y v = \lambda r, \tag{4.18a}$$

$$\imath a_\star k_x r + \nu_u k_x^2 u + (\nu_u k_x k_y - \omega) v = \lambda u, \tag{4.18b}$$

$$\imath a_\star k_y r + (\nu_u k_x k_y + \omega) u + \nu_u k_y^2 v = \lambda v. \tag{4.18c}$$

  We now multiply (4.18a) by $\bar{r}$, (4.18b) by $\bar{u}$ and (4.18c) by $\bar{v}$ in order to obtain

$$\lambda\left(|r|^2 + |u|^2 + |v|^2\right) = \nu_r|\boldsymbol{k}|^2|r|^2 + \nu_u(k_x^2|u|^2 + k_y^2|v|^2) + 2\nu_u k_x k_y \Re(u\bar{v}) + 2\imath\left[a_\star \Re\left(r(k_x\bar{u} + k_y\bar{v})\right) + \omega\Im(u\bar{v})\right]$$

  which implies that $\Re(\lambda) > 0$ by using the fact that $k_x^2|u|^2 + k_y^2|v|^2 \geq 2|k_x k_y||uv|$ and $\Re(u\bar{v}) \geq -|uv|$.

- For the *AT-LF* scheme ($\nu_r = \gamma_r$ and $\nu_u = \gamma_u = 0$), $\lambda_0 = 0$ is an eigenvalue. The other two are given by

$$\lambda_c = \frac{\nu_r|\boldsymbol{k}|^2}{2} \pm \imath\sqrt{\omega^2 + a_\star^2|\boldsymbol{k}|^2 - \left(\frac{\nu_r}{2}\right)^2|\boldsymbol{k}|^4} \implies \Re(\lambda_c) \geq 0.$$

- Finally, we consider the *AT-DP* scheme ($\nu_r = \gamma_r$ and $\nu_u = \gamma_u$). It is obvious that $\lambda = 0$ is a solution. The other solutions satisfy the following equation

$$\lambda^2 - (\nu_r + \nu_u)|\boldsymbol{k}|^2\lambda + \nu_r\nu_u|\boldsymbol{k}|^4 + \left[\omega^2 + a_\star^2|\boldsymbol{k}|^2\right] = 0.$$

  The solution of the above equation is given by

$$\lambda = \frac{\nu_r + \nu_u}{2}|\boldsymbol{k}|^2 \pm \imath\sqrt{\omega^2 + a_\star^2|\boldsymbol{k}|^2 - \left(\frac{\nu_r - \nu_u}{2}\right)^2|\boldsymbol{k}|^4}.$$

  which means that the Fourier modes are damped with speed $\frac{\nu_r+\nu_u}{2}|\boldsymbol{k}|^2$.

□

Remark 4.5. *The exact Fourier modes of the linear wave equation (4.2) are such that*

$$\lambda_{wave} = \pm\imath\sqrt{\omega^2 + a_*^2|\boldsymbol{k}|^2}. \tag{4.19}$$

Consequently, we notice that the AT-DP scheme is the only one that can recover the exact imaginary part by taking $\nu_r = \nu_u$. This choice will be done in the sequel.

Remark 4.6. *For the* AT-C *scheme ($\nu_r = \gamma_r$ and $\gamma_u = 0$), we are not able to prove the Fourier modes are damped. Nevertheless the Fourier analysis provides some information on the behaviour of the solution when the diffusion on the velocity equation is small. In that case, the characteristic polynomial becomes*

$$\chi(\lambda, \nu_u) = \lambda^3 - (\nu_r + \nu_u)|\boldsymbol{k}|^2\lambda^2 + \left[\omega^2 + a_\star^2|\boldsymbol{k}|^2 + \nu_r\nu_u|\boldsymbol{k}|^4 + \nu_u^2 k_x^2 k_y^2\right]\lambda - \nu_r|\boldsymbol{k}|^2\nu_u^2 k_x^2 k_y^2 - 2\nu_u a_\star^2 k_x^2 k_y^2 - \nu_r\nu_u\omega k_x k_y(k_x^2 - k_y^2).$$

We note that under (4.16) and due to $\kappa_{r,u} \in [0,1]$, $\partial_\lambda\chi(\lambda, \nu_u)$ does not vanish which means there is a single real root. Therefore, by the implicit function theorem, we can define for ν_u small enough a function $\nu_u \longmapsto \lambda_0(\nu_u)$ corresponding to the unique root of the polynomial. We have

$$\lambda_0(\nu_u) \underset{\nu_u \to 0}{\sim} \lambda_0'(\nu_u = 0)\nu_u = -\frac{\partial_{\nu_u}\chi(0,0)}{\partial_\lambda\chi(0,0)}\nu_u = \frac{2a_\star^2 k_x^2 k_y^2 + \nu_r\omega k_x k_y(k_x^2 - k_y^2)}{\left(k_x^2 + k_y^2\right)a_\star^2 + \omega^2}\nu_u.$$

As a result, we deduce that when $\kappa_u = \mathcal{O}(M)$, then $\lambda_0(\nu_u) = \mathcal{O}(M)$. Note that the sign of the eigenvalue remains undetermined.

4.5 Analysis of the semi-discrete Godunov type schemes

In this section, we investigate some ways to construct "well-balanced" schemes, *i.e.*that preserve a discrete version of the incompressible subspace. The study of the modified equation leads us to focus on the numerical viscosity on both pressure and velocity equations. As a result, it is essential to consider suitable diffusion terms for the Godunov scheme. More specifically, we proposed to introduce the diffusion operators $\nabla\cdot(\nabla r+\omega\boldsymbol{u}^\perp)$ and $\nabla(\nabla\cdot\boldsymbol{u})$ for pressure and velocity equations respectively.

We now turn to the space discretisation of the aforementioned strategies. We consider a collocated finite volume framework. To ensure that the resulting schemes satisfy properties similar to those proved at the continuous level, one first has to construct some discrete differential operators and corresponding discrete kernels consistent with $\mathcal{E}_{\omega\neq 0}$. This requires to choose the location where the differential relation defining the kernels holds: either at the cell centers or at the vertices. One finally has to verify that the resulting schemes still satisfy stability properties.

4.5.1 Cell-centered scheme

A first possibility consists in locating the kernels at the same place as the unknowns, namely at the cell centers. Let us denote $r_h=(r_{i,j})$, $u_h=(u_{i,j})$ and $v_h=(v_{i,j})$ be in $\mathbb{R}^N$ where $N=N_x\times N_y$.

Discrete operators

We define the gradient ∇^c_{2h}, the divergence $\nabla^c_{2h}\cdot$ and the curl $\nabla^c_{2h}\times$ as

$$
\begin{aligned}
[\nabla^c_{2h}r_h]_{i,j} &= \begin{pmatrix} \frac{r_{i+1,j}-r_{i-1,j}}{2\Delta x} \\ \frac{r_{i,j+1}-r_{i,j-1}}{2\Delta y}\end{pmatrix}\\
[\nabla^c_{2h}\cdot\boldsymbol{u}_h]_{i,j} &= \frac{u_{i+1,j}-u_{i-1,j}}{2\Delta x}+\frac{v_{i,j+1}-v_{i,j-1}}{2\Delta y},\\
[\nabla^c_{2h}\times\boldsymbol{u}_h]_{i,j} &= -\nabla^c_{2h}\cdot\boldsymbol{u}_h^\perp=\frac{v_{i+1,j}-v_{i-1,j}}{2\Delta x}-\frac{u_{i,j+1}-u_{i,j-1}}{2\Delta y}.
\end{aligned}
$$

Lemma 4.4. *These operators satisfy the following mimetic properties:*

i. $\langle\nabla^c_{2h}r_h,\boldsymbol{u}_h\rangle=-\langle r_h,\nabla^c_{2h}\cdot\boldsymbol{u}_h\rangle$ *which implies that* $\langle r_h,\nabla^c_{2h}\times\boldsymbol{u}_h\rangle=-\left\langle(\nabla^c_{2h}r_h)^\perp,\boldsymbol{u}_h\right\rangle$;

ii. $\nabla^c_{2h}\times\nabla^c_{2h}r_h=0$.

Such properties turn out to be crucial for stability purposes as claimed in [47, 52].

Discrete kernel

We now define the discrete kernel at the cell centers as the natural equivalent to $\mathcal{E}_{\omega\neq 0}$ defined in (4.3)

$$
\mathcal{E}^c_{\omega\neq 0,h}=\left\{\hat{q}_h=(\hat{r}_h,\hat{u}_h,\hat{v}_h)\in\mathbb{R}^{3N}\;\middle|\;a_\star\nabla^c_{2h}\hat{r}_h=-\omega\hat{\boldsymbol{u}}_h^\perp\right\}. \tag{4.20}
$$

In particular, we prove the following lemma which is the semi-discrete counterpart to Proposition 4.1.

Lemma 4.5. *The orthogonal space of* $\mathcal{E}^c_{\omega\neq 0,h}$ *is*

$$
\mathcal{E}^{c,\perp}_{\omega\neq 0,h}=\left\{\tilde{q}_h=(\tilde{r}_h,\tilde{u}_h,\tilde{v}_h)\in\mathbb{R}^{3N}\;\middle|\;a_\star\nabla^c_{2h}\times\tilde{\boldsymbol{u}}_h=\omega\tilde{r}_h\right\}. \tag{4.21}
$$

This implies the following discrete Hodge decomposition

$$\mathbb{R}^{3N} = \mathcal{E}^c_{\omega\neq 0,h} \oplus \mathcal{E}^{c,\perp}_{\omega\neq 0,h}.$$

Proof. By definition, an element $\tilde{q}_h = (\tilde{r}_h, \tilde{\boldsymbol{u}}_h)$ of the orthogonal of $\mathcal{E}^c_{\omega\neq 0,h}$ verifies, for all $\hat{q}_h = (\hat{r}_h, \hat{\boldsymbol{u}}_h)$ in $\mathcal{E}^c_{\omega\neq 0,h}$:

$$\langle \tilde{r}_h, \hat{r}_h\rangle + \langle \tilde{\boldsymbol{u}}_h, \hat{\boldsymbol{u}}_h\rangle = \langle \tilde{r}_h, \hat{r}_h\rangle + \langle \tilde{\boldsymbol{u}}_h^\perp, \hat{\boldsymbol{u}}_h^\perp\rangle = 0.$$

Using the definition of $\mathcal{E}^c_{\omega\neq 0,h}$ and Lemma 4.4 this implies

$$\langle \tilde{r}_h, \hat{r}_h\rangle - \frac{a_\star}{\omega}\langle \tilde{\boldsymbol{u}}_h^\perp, \nabla^c_{2h}\hat{r}_h\rangle = \langle \tilde{r}_h, \hat{r}_h\rangle + \frac{a_\star}{\omega}\langle \nabla^c_{2h}\cdot \tilde{\boldsymbol{u}}_h^\perp, \hat{r}_h\rangle = \langle \tilde{r}_h - \frac{a_\star}{\omega}\nabla^c_{2h}\times \tilde{\boldsymbol{u}}_h, \hat{r}_h\rangle = 0.$$

Since $\hat{r}_h$ can be arbitrary in $\mathbb{R}^N$, this is exactly equivalent to $\omega\tilde{r}_h - a_\star\nabla^c_{2h}\times\tilde{\boldsymbol{u}}_h$. □

Remark 4.7. *For any $q_h \in \mathbb{R}^{3N}$, the unique decomposition*

$$q_h = \hat{q}_h + \tilde{q}_h \quad \text{with} \quad \hat{q}_h = (\hat{r}_h, \hat{u}_h, \hat{v}_h) \in \mathcal{E}^c_{\omega\neq 0,h} \quad \text{and} \quad \tilde{q}_h = (\tilde{r}_h, \tilde{u}_h, \tilde{v}_h) \in \mathcal{E}^{c,\perp}_{\omega\neq 0,h}$$

may be found by the following process: Let $\hat{r}_h$ satisfy the equation

$$\hat{r}_h - \frac{a_*^2}{\omega^2}\nabla^c_{2h}\cdot(\nabla^c_{2h}\hat{r}_h) = r_h - \frac{a_\star}{\omega}\nabla^c_{2h}\times\boldsymbol{u}_h. \tag{4.22}$$

It can be shown that (4.22) has a unique solution since it amounts to solving a linear system involving an M-matrix. Then, let us define $\hat{\boldsymbol{u}}_h$ by

$$\hat{\boldsymbol{u}}_h = \frac{a_\star}{\omega}(\nabla^c_{2h}\hat{r}_h)^\perp \tag{4.23}$$

so that $\hat{q}_h = (\hat{r}_h, \hat{\boldsymbol{u}}_h) \in \mathcal{E}^c_{\omega\neq 0,h}$. Finally, we set $\tilde{q}_h = q_h - \hat{q}_h$ and it remains to prove that $\tilde{q}_h \in \mathcal{E}^{c,\perp}_{\omega\neq 0,h}$. It suffices to notice that

$$\begin{aligned} a_\star\nabla^c_{2h}\times\tilde{\boldsymbol{u}}_h = a_\star\left(\nabla^c_{2h}\times\boldsymbol{u}_h + \nabla^c_{2h}\cdot\hat{\boldsymbol{u}}_h^\perp\right) &\overset{(4.23)}{=} a_\star\left(\nabla^c_{2h}\times\boldsymbol{u}_h - \frac{a_\star}{\omega}\nabla^c_{2h}\cdot(\nabla^c_{2h}\hat{r}_h)\right) \\ &\overset{(4.22)}{=} a_\star\nabla^c_{2h}\times\boldsymbol{u}_h - \omega\left(\hat{r}_h - r_h + \frac{a_\star}{\omega}\nabla^c_{2h}\times\boldsymbol{u}_h\right) = \omega\tilde{r}_h. \end{aligned}$$

Semi-discrete scheme

The cell-centered semi-discrete scheme reads

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t}r_{i,j}(t) + a_\star[\nabla^c_{2h}\cdot\boldsymbol{u}_h]_{i,j} - \nu_r\left[\nabla^c_{2h}\cdot\left(\nabla^c_{2h}r_h + \dfrac{\omega}{a_\star}\boldsymbol{u}_h^\perp\right)\right]_{i,j} = 0, & \text{(4.24a)} \\ \dfrac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{u}_{i,j}(t) + a_\star[\nabla^c_{2h}r_h]_{i,j} - \nu_u\left[\nabla^c_{2h}(\nabla^c_{2h}\cdot\boldsymbol{u}_h)\right]_{i,j} = -\omega\boldsymbol{u}_{i,j}^\perp. & \text{(4.24b)} \end{cases}$$

The modified equation associated to the scheme (4.24) is (4.12) for coefficients chosen as in Remark 4.4. The stencil associated to the scheme (4.24) is a 13-point stencil: it involves the 8 points around the considered one (*i.e.*at a distance Δx or Δy) and 4 points to a distance $2\Delta x$ (or $2\Delta y$) in the definition of both diffusion terms. Moreover the definition of the diffusion terms induces no relation between odd and even cells. This may be the reason for checkerboard type oscillations. The interface scheme (4.26) we propose in the sequel will not be affected by this drawback.

Proposition 4.6.

i. Steady states of the semi-discrete scheme (4.24) are the discrete geostrophic equilibria from (4.20).

ii. The pressure gradient and Coriolis forces are energy conservative.

iii. The discrete energy of the $LF-DP$ scheme ($\nu_r = 0$) is decreasing.

Proof. On the one hand, by construction, discrete geostrophic equilibria (4.20) are steady states of (4.24). On the other hand, let us consider steady states of (4.24). Applying the operator $\nabla^c_{2h}\times$ to (4.24b), we obtain $\nabla^c_{2h}\cdot\boldsymbol{u}_h = 0$ due to Lemma 4.4.*ii*. This proves Point *i*.

Point *ii* is a straightforward consequence of Lemma 4.4 *i*. Moreover, when $\nu_r = 0$, the scalar product with q_h leads to

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}E_h(t) = -a_\star\langle\nabla^c_{2h}\cdot\boldsymbol{u}_h, r_h\rangle - a_\star\langle\nabla^c_{2h}r_h, \boldsymbol{u}_h\rangle + \nu_u\langle\nabla^c_{2h}[\nabla^c_{2h}\cdot\boldsymbol{u}_h], \boldsymbol{u}_h\rangle = -\nu_u\|\nabla^c_{2h}\cdot\boldsymbol{u}_h\|^2 \leq 0$$

thanks to Lemma 4.4.*i*. This proves Point *iii*. □

4.5.2 Vertex-based scheme

The original *Apparent Topography* scheme [13] was designed in 1D so that equilibrium states are located at the interfaces while the unknowns are still at the cell centers. That is why we are interested in this part in investigating another version of the scheme.

Discrete kernel

Let us define the discrete kernel by imposing the geostrophic equilibrium at the interfaces of each cell

$$\mathcal{E}^v_{\omega\neq 0,h} = \left\{ \hat{q}_h = (\hat{r}_h, \hat{u}_h, \hat{v}_h) \in \mathbb{R}^{3N} \;\middle|\; a_\star \begin{pmatrix} \frac{\hat{r}_{i+1,j}-\hat{r}_{i,j}}{\Delta x} \\ \frac{\hat{r}_{i,j+1}-\hat{r}_{i,j}}{\Delta y} \end{pmatrix} = -\omega \begin{pmatrix} -\frac{\hat{v}_{i+1,j}+\hat{v}_{i,j}}{2} \\ \frac{\hat{u}_{i,j+1}+\hat{u}_{i,j}}{2} \end{pmatrix} \right\}. \tag{4.25}$$

Discrete operators

To design the numerical scheme, we first define the discrete operators at the vertices of each cell (i,j) – see Figure 4.4

$$\begin{aligned}
[\nabla^{\mathrm{v}}_h r_h]_{i+1/2,j+1/2} &= \begin{pmatrix} \frac{(r_{i+1,j+1}+r_{i+1,j})-(r_{i,j+1}+r_{i,j})}{2\Delta x} \\ \frac{(r_{i+1,j+1}+r_{i,j+1})-(r_{i+1,j}+r_{i,j})}{2\Delta y} \end{pmatrix} \\
[\nabla^{\mathrm{v}}_h\cdot\boldsymbol{u}_h]_{i+1/2,j+1/2} &= \frac{(u_{i+1,j+1}+u_{i+1,j})-(u_{i,j+1}+u_{i,j})}{2\Delta x} + \frac{(v_{i+1,j+1}+v_{i,j+1})-(v_{i+1,j}+v_{i,j})}{2\Delta y} \\
[\nabla^{\mathrm{v}}_h\times\boldsymbol{u}_h]_{i+1/2,j+1/2} &= -\nabla^{\mathrm{v}}_h\cdot\boldsymbol{u}^\perp_h, \\
[f^{\mathrm{v}}_h(u_h)]_{i+1/2,j+1/2} &= \frac{u_{i+1,j+1}+u_{i,j+1}+u_{i+1,j}+u_{i,j}}{4}.
\end{aligned}$$

We shall also need dual operators that enable to switch from the vertex grid to the center grid. For $\boldsymbol{\varphi}_h = (\varphi_h, \psi_h)$ defined at the vertices, we define the following operators

$$
\begin{aligned}
[\nabla_h^{\mathrm{c}} \varphi_h]_{i,j} &= \frac{1}{2}\begin{pmatrix} \frac{\varphi_{i+1/2,j+1/2}-\varphi_{i-1/2,j+1/2}}{\Delta x} \\ \frac{\varphi_{i+1/2,j+1/2}-\varphi_{i+1/2,j-1/2}}{\Delta y} \end{pmatrix} + \frac{1}{2}\begin{pmatrix} \frac{\varphi_{i+1/2,j-1/2}-\varphi_{i-1/2,j-1/2}}{\Delta x} \\ \frac{\varphi_{i-1/2,j+1/2}-\varphi_{i-1/2,j-1/2}}{\Delta y} \end{pmatrix} \\
[\nabla_h^{\mathrm{c}} \cdot \boldsymbol{\varphi}_h]_{i,j} &= \frac{(\varphi_{i+1/2,j+1/2}+\varphi_{i+1/2,j-1/2}) - (\varphi_{i-1/2,j+1/2}+\varphi_{i-1/2,j-1/2})}{2\Delta x} \\
&\quad + \frac{(\psi_{i+1/2,j+1/2}+\psi_{i-1/2,j+1/2}) - (\psi_{i+1/2,j-1/2}+\psi_{i-1/2,j-1/2})}{2\Delta y} \\
[f_h^{\mathrm{c}}(\varphi_h)]_{i,j} &= \frac{\varphi_{i+1/2,j+1/2}+\varphi_{i-1/2,j+1/2}+\varphi_{i+1/2,j-1/2}+\varphi_{i-1/2,j-1/2}}{4}.
\end{aligned}
$$

With such operators, we have the following compatibility property:

Lemma 4.6. *Any $\hat{q}_h \in \mathcal{E}^v_{\omega\neq 0,h}$ satisfies the geostrophic equilibrium and the divergence free condition at the vertices:*

$$
\begin{cases} [\nabla_h^v \hat{r}_h]_{i+1/2,j+1/2} = -\omega [f_h^v(\hat{\boldsymbol{u}}_h^\perp)]_{i+1/2,j+1/2}, \\ [\nabla_h^v \cdot \hat{\boldsymbol{u}}_h]_{i+1/2,j+1/2} = 0. \end{cases}
$$

Moreover, they satisfy mimetic properties:

Lemma 4.7.

i. $\nabla_h^v \times \nabla_h^c \left[f_h^v(r_h)\right] = \nabla_h^v \times f_h^c \left[\nabla_h^v r_h\right] = 0$;

ii. $\left\langle f_h^c \left[\nabla_h^v r_h\right], \boldsymbol{u}_h \right\rangle = -\left\langle r_h, f_h^c \left[\nabla_h^v \cdot \boldsymbol{u}_h\right] \right\rangle$ *and* $\left\langle f_h^c \left[f_h^v(u_h)\right], v_h \right\rangle = \left\langle u_h, f_h^c \left[f_h^v(v_h)\right] \right\rangle$.

Proof. Each property results from direct computations. For instance:

$$
\begin{aligned}
&\langle f_h^{\mathrm{c}} [\nabla_h^{\mathrm{v}} r_h], \boldsymbol{u}_h \rangle \\
&= \sum_{i,j} \left[\frac{1}{4}\left(\frac{r_{i+1,j+1}-r_{i-1,j+1}}{2\Delta x}\right) + \frac{1}{2}\left(\frac{r_{i+1,j}-r_{i-1,j}}{2\Delta x}\right) + \frac{1}{4}\left(\frac{r_{i+1,j-1}-r_{i-1,j-1}}{2\Delta x}\right)\right] u_{i,j} \\
&\quad + \sum_{i,j} \left[\frac{1}{4}\left(\frac{r_{i+1,j+1}-r_{i+1,j-1}}{2\Delta y}\right) + \frac{1}{2}\left(\frac{r_{i,j+1}-r_{i,j-1}}{2\Delta y}\right) + \frac{1}{4}\left(\frac{r_{i-1,j+1}-r_{i-1,j-1}}{2\Delta y}\right)\right] v_{i,j} \\
&= \sum_{i,j} \left(\frac{r_{i+1,j}-r_{i-1,j}}{2\Delta x}\right)\left(\frac{u_{i,j+1}+2u_{i,j}+u_{i,j-1}}{4}\right) + \sum_{i,j} \left(\frac{r_{i,j+1}-r_{i,j-1}}{2\Delta y}\right)\left(\frac{v_{i+1,j}+2v_{i,j}+v_{i-1,j}}{4}\right) \\
&= -\sum_{i,j} r_{i,j} \left[\frac{1}{4}\left(\frac{u_{i+1,j+1}-u_{i-1,j+1}}{2\Delta x}\right) + \frac{1}{2}\left(\frac{u_{i+1,j}-u_{i-1,j}}{2\Delta x}\right) + \frac{1}{4}\left(\frac{u_{i+1,j-1}-u_{i-1,j-1}}{2\Delta x}\right)\right] \\
&\quad - \sum_{i,j} r_{i,j} \left[\frac{1}{4}\left(\frac{v_{i+1,j+1}-v_{i+1,j-1}}{2\Delta y}\right) + \frac{1}{2}\left(\frac{v_{i,j+1}-v_{i,j-1}}{2\Delta y}\right) + \frac{1}{4}\left(\frac{v_{i-1,j+1}-v_{i-1,j-1}}{2\Delta y}\right)\right] \\
&= -\langle f_h^{\mathrm{c}} [\nabla_h^{\mathrm{v}} \cdot \boldsymbol{u}_h], r_h \rangle .
\end{aligned}
$$

□

Semi-discrete scheme

The semi-discrete scheme with the kernel at the interface is given by

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t} r_{i,j}(t) + a_\star f_h^{\mathrm{c}} [\nabla_h^{\mathrm{v}} \cdot \boldsymbol{u}_h]_{i,j} - \nu_r \nabla_h^{\mathrm{c}} \cdot \left[\nabla_h^{\mathrm{v}} r_h + \omega f_h^{\mathrm{v}}(\boldsymbol{u}_h^\perp) \right]_{i,j} = 0, \\ \dfrac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{u}_{i,j}(t) + a_\star f_h^{\mathrm{c}} [\nabla_h^{\mathrm{v}} r_h]_{i,j} - \nu_u \nabla_h^{\mathrm{c}} [\nabla_h^{\mathrm{v}} \cdot \boldsymbol{u}_h]_{i,j} = -\omega f_h^{\mathrm{c}} \left[f_h^{\mathrm{v}}(\boldsymbol{u}_h^\perp) \right]_{i,j}. \end{cases} \tag{4.26}$$

The modified equation associated to the scheme (4.26) is still (4.12) for coefficients chosen as in Remark 4.4. The stencil associated to this second scheme (4.26) is a classical 9-point stencil since it only involves the 8 points around the considered one. It is then more compact than the one of the cell-centered scheme (4.24).

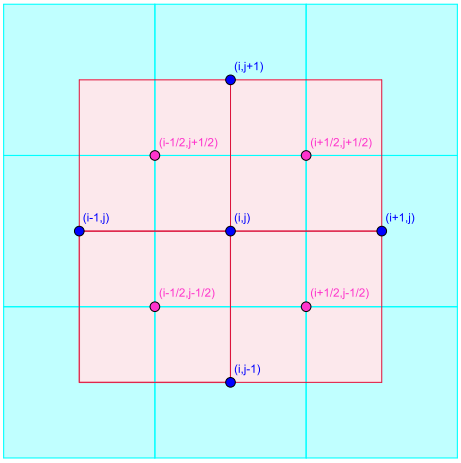


Figure 4.4: *Cell centers* (i,j) *and vertices* $(i+1/2, j+1/2)$.

Proposition 4.7.

i. Steady states of the semi-discrete scheme (4.26) *are the geostrophic equilibria from* (4.25).

ii. The pressure gradient and Coriolis forces are energy conservative.

iii. The energy of the LF-DP scheme ($\nu_r = 0$) *is decreasing.*

Proof. Point *i.* results from Lemma 4.6 and from Lemma 4.7.*i.* Moreover, according to Lemma 4.7.*ii*, we have

$$\langle f_h^{\mathrm{c}} [\nabla_h^{\mathrm{v}} r_h], \boldsymbol{u}_h \rangle + \langle f_h^{\mathrm{c}} [\nabla_h^{\mathrm{v}} \cdot \boldsymbol{u}_h], r_h \rangle = 0 \quad \text{and} \quad \left\langle f_h^{\mathrm{c}} \left[f_h^{\mathrm{v}}(\boldsymbol{u}_h^\perp) \right], \boldsymbol{u}_h \right\rangle = 0$$

which proves Point *ii.*

After some computations, we have

$$\langle \nabla_h^{\mathrm{c}} [\nabla_h^{\mathrm{v}} \cdot \boldsymbol{u}_h], \boldsymbol{u}_h \rangle = -\sum_{i,j} \left[\frac{u_{i+1,j+1} - u_{i,j+1}}{2\Delta x} + \frac{u_{i+1,j} - u_{i,j}}{2\Delta x} + \frac{v_{i+1,j+1} - v_{i+1,j}}{2\Delta y} + \frac{v_{i,j+1} - v_{i,j}}{2\Delta y} \right]^2.$$

Therefore, when $\nu_r = 0$, we deduce that

$$\frac{1}{2} \frac{\mathrm{d}}{\mathrm{d}t} E_h(t) = \nu_u \langle \nabla_h^{\mathrm{c}} [\nabla_h^{\mathrm{v}} \cdot \boldsymbol{u}_h], \boldsymbol{u}_h \rangle = -\nu_u \| \nabla_h^{\mathrm{v}} \cdot \boldsymbol{u}_h \|^2$$

which means that the semi-discrete LF-DP scheme is dissipative. This proves Point *iii.* □

Godunov type scheme	α	β	η
Cell-centered	$\sin(k_x\Delta x)$	$\sin(k_y\Delta y)$	1
Vertex-based	$2\sin(\frac{k_x\Delta x}{2})\cos(\frac{k_y\Delta y}{2})$	$2\sin(\frac{k_y\Delta y}{2})\cos(\frac{k_x\Delta x}{2})$	$\cos(\frac{k_x\Delta x}{2})\cos(\frac{k_y\Delta y}{2})$

Table 4.2: *Parameters α, β an η in the Fourier analysis of the semi-discrete schemes.*

4.5.3 Fourier analysis

Let us carry out a Fourier analysis of the semi-discrete schemes by considering the discrete Fourier modes

$$r_{i,j}(t)=\varphi_r(t)e^{\imath(k_x x_i+k_y y_j)},\quad u_{i,j}(t)=\varphi_u(t)e^{\imath(k_x x_i+k_y y_j)}\quad\text{and}\quad v_{i,j}(t)=\varphi_v(t)e^{\imath(k_x x_i+k_y y_j)}.$$

that are substituted in the cell-centered scheme (4.24) and in the vertex-based scheme (4.26) to obtain the differential system

$$\begin{pmatrix}\varphi_r'(t)\\ \varphi_u'(t)\\ \varphi_v'(t)\end{pmatrix}=\begin{pmatrix}-\nu_r\left(\frac{\alpha^2}{\Delta x^2}+\frac{\beta^2}{\Delta y^2}\right) & -\imath\eta\left(a_\star\frac{\alpha}{\Delta x}-\nu_r\frac{\omega}{a_\star}\frac{\beta}{\Delta y}\right) & -\imath\eta\left(a_\star\frac{\beta}{\Delta y}+\nu_r\frac{\omega}{a_\star}\frac{\alpha}{\Delta x}\right)\\ -\imath a_\star\frac{\alpha}{\Delta x}\eta & -\nu_u\frac{\alpha^2}{\Delta x^2} & -\nu_u\frac{\alpha}{\Delta x}\frac{\beta}{\Delta y}+\omega\eta^2\\ -\imath a_\star\frac{\beta}{\Delta y}\eta & -\nu_u\frac{\alpha}{\Delta x}\frac{\beta}{\Delta y}-\omega\eta^2 & -\nu_u\frac{\beta^2}{\Delta y^2}\end{pmatrix}\begin{pmatrix}\varphi_r(t)\\ \varphi_u(t)\\ \varphi_v(t)\end{pmatrix}\tag{4.27}$$

where parameters α, β an η are specified in Table 4.2 depending on the scheme under study. One eigenvalue of the amplification matrix in (4.27) is $\lambda_0=0$ which corresponds to the stationary state. The other eigenvalues are given by

$$\lambda_c=\frac{\nu_r+\nu_u}{2}\left(\frac{\alpha^2}{\Delta x^2}+\frac{\beta^2}{\Delta y^2}\right)\pm\imath\sqrt{\omega^2\eta^4+a_\star^2\eta^2\left(\frac{\alpha^2}{\Delta x^2}+\frac{\beta^2}{\Delta y^2}\right)-\left(\frac{\nu_r-\nu_u}{2}\right)^2\left(\frac{\alpha^2}{\Delta x^2}+\frac{\beta^2}{\Delta y^2}\right)^2}.$$

As mentioned above, it is essential with the $AT-DP$ scheme to take $\nu_r=\nu_u$ in order to be as close as possible to the exact dispersion relation (4.19), see Figure 4.5.

Remark 4.8. *We notice that the damping rate $\Re(\lambda)$ of the $AT-DP$ scheme is larger than those of the $AT-LF$ and $LF-DP$ schemes.*

4.6 Analysis of the fully discrete Godunov type schemes

We consider an explicit discretisation for the advection term. Nevertheless it is well known that a fully explicit discretisation of the Coriolis term leads in that case to unstable schemes, see [27]. Then let us set

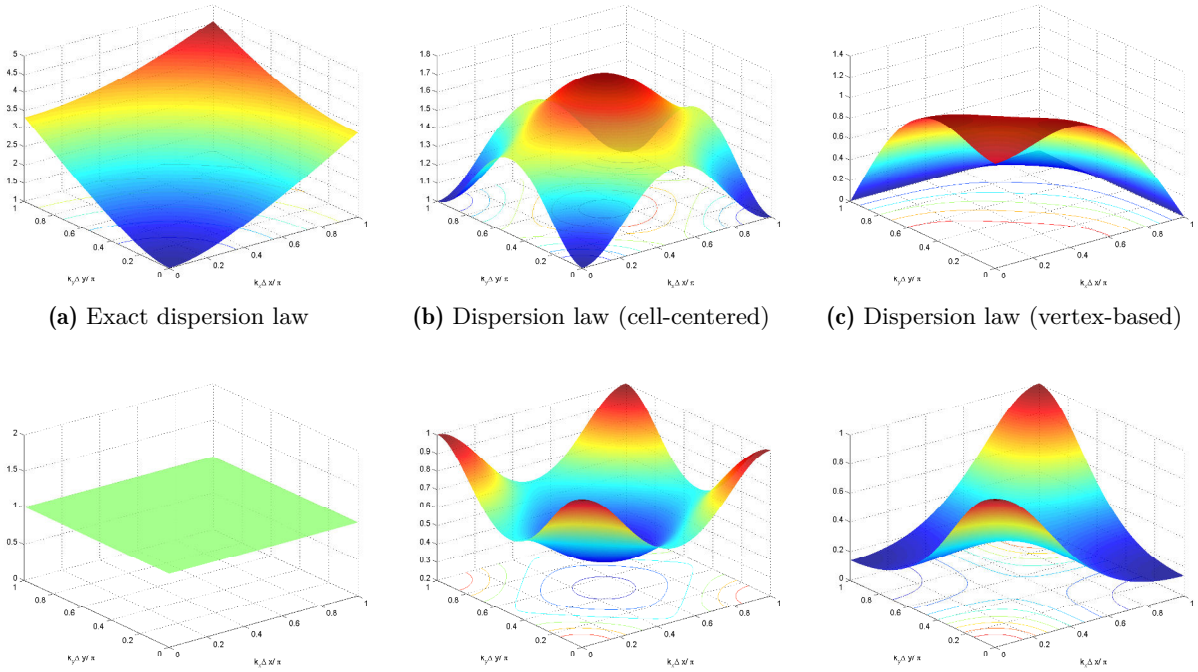
$$\boldsymbol{u}^\theta=\begin{pmatrix}\theta_1u^n+(1-\theta_1)u^{n+1}\\ \theta_2v^n+(1-\theta_2)v^{n+1}\end{pmatrix}$$

for some θ_1, $\theta_2\in[0,1]$. In particular, for $\theta_1=\theta_2=\theta$, then $\boldsymbol{u}^\theta=\theta\boldsymbol{u}^n+(1-\theta)\boldsymbol{u}^{n+1}$.

4.6.1 Stability condition

For the sake of clarity, we shall assume in the sequel that

$$\Delta x=\Delta y=h.$$



(a) Exact dispersion law **(b)** Dispersion law (cell-centered) **(c)** Dispersion law (vertex-based)

(d) Exact damping **(e)** Damping error (cell-centered) **(f)** Damping error (vertex-based)

Figure 4.5: *Dispersion relation and damping for the AT-DP scheme with* $a_* = \omega \Delta x$.

As a consequence, due to (4.14), we have

$$\kappa_r := \kappa_r^x = \kappa_r^y = \eta_r^x = \eta_r^y, \qquad \kappa_u = \kappa_v = \eta_u = \eta_v \qquad \text{and} \qquad \nu_\# = \frac{\kappa_\# a_* \Delta x}{2}.$$

We propose the following time discretisation for the cell-centered scheme

$$\begin{cases} \dfrac{r_{i,j}^{n+1} - r_{i,j}^n}{\Delta t} + a_\star [\nabla_{2h}^c \cdot \boldsymbol{u}_h^n]_{i,j} - \nu_r \left[\nabla_{2h}^c \cdot \left(\nabla_{2h}^c r_h^n + \dfrac{\omega}{a_*} \boldsymbol{u}_h^{n,\perp} \right) \right]_{i,j} = 0, \\ \dfrac{\boldsymbol{u}_{i,j}^{n+1} - \boldsymbol{u}_{i,j}^n}{\Delta t} + a_\star [\nabla_{2h}^c r_h^n]_{i,j} - \nu_u \left[\nabla_{2h}^c (\nabla_{2h}^c \cdot \boldsymbol{u}_h^n) \right]_{i,j} = -\omega \boldsymbol{u}_{i,j}^{\theta,\perp}, \end{cases} \tag{4.28}$$

and for the vertex-based scheme

$$\begin{cases} \dfrac{r_{i,j}^{n+1} - r_{i,j}^n}{\Delta t} + a_\star f_h^{\mathrm{c}} \left[\nabla_h^{\mathrm{v}} \cdot \boldsymbol{u}_h^n \right]_{i,j} - \nu_r \nabla_h^{\mathrm{c}} \cdot \left[\nabla_h^{\mathrm{v}} r_h^n + \omega f_h^{\mathrm{v}} (\boldsymbol{u}_h^n)^\perp \right]_{i,j} = 0, \\ \dfrac{\boldsymbol{u}_{i,j}^{n+1} - \boldsymbol{u}_{i,j}^n}{\Delta t} + a_\star f_h^{\mathrm{c}} \left[\nabla_h^{\mathrm{v}} r_h^n \right]_{i,j} - \nu_u \nabla_h^{\mathrm{c}} \left[\nabla_h^{\mathrm{v}} \cdot \boldsymbol{u}_h^n \right]_{i,j} = -\omega f_h^{\mathrm{c}} \left[f_h^{\mathrm{v}} (\boldsymbol{u}_h^\theta) \right]_{i,j}^\perp. \end{cases} \tag{4.29}$$

In order to avoid inverting a matrix with a large stencil in the computation of the scheme, the vertex-based scheme is restricted to the cases $(\theta_1 = 1, \theta_2 = 0)$ and $(\theta_1 = 0, \theta_2 = 1)$.

Lemma 4.8. *Any choice such that* $\theta_1 + \theta_2 > 1$ *makes schemes* (4.28) *and* (4.29) *unstable. In particular, the explicit case* $\theta_1 = \theta_2 = 1$ *is unstable, as mentioned before.*

The proof of this lemma is embedded in the proof of Theorem 4.1 below.

Theorem 4.1. *For a uniform mesh $\Delta x = \Delta y = h$, the LF-DP schemes (*i.e.(4.28) *and* (4.29) *with $\nu_r = 0$) are stable under the following conditions*

$$\Delta t \leq \min\{\Delta t_a, \Delta t_b\} \quad \text{where} \quad \Delta t_a := \frac{\kappa_u h}{2a_*} \text{ and } \Delta t_b := \frac{2}{\omega|\theta_2 - \theta_1|}.$$

Remark 4.9. *The restriction on the time step Δt_a (resp. Δt_b) is the classical CFL condition for advection (resp. rotation) phenomena. Note that the choice $\theta_2 = \theta_1$ makes the CFL condition independent from the Coriolis parameter.*

Proof. Let us denote

$$\varpi = \omega\Delta t, \quad \sigma = a_\star \frac{\Delta t}{h}.$$

We now perform the Fourier analysis for fully discrete Godunov type schemes by substituting the fully discrete Fourier mode

$$r_{i,j}^n = \varphi_r^n e^{\imath(k_x x_i + k_y y_j)}, \quad u_{i,j}^n = \varphi_u^n e^{\imath(k_x x_i + k_y y_j)} \quad \text{and} \quad v_{i,j}^n = \varphi_v^n e^{\imath(k_x x_i + k_y y_j)}$$

into the fully discrete scheme to obtain

$$\mathcal{T}_\theta \varphi^{n+1} = \mathcal{M}_\theta \varphi^n$$

where

$$\mathcal{T}_\theta = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -(1-\theta_2)\varpi\eta^2 \\ 0 & (1-\theta_1)\varpi\eta^2 & 1 \end{pmatrix}$$

and

$$\mathcal{M}_\theta = \begin{pmatrix} 1 - \frac{\kappa_r\sigma}{2}(\alpha^2+\beta^2) & -\imath\eta\left(\sigma\alpha - \frac{\kappa_r}{2}\varpi\beta\right) & -\imath\eta\left(\sigma\beta + \frac{\kappa_r}{2}\varpi\alpha\right) \\ -\imath\sigma\alpha\eta & 1 - \frac{\kappa\sigma}{2}\alpha^2 & -\frac{\kappa\sigma}{2}\alpha\beta + \theta_2\varpi\eta^2 \\ -\imath\sigma\beta\eta & -\frac{\kappa\sigma}{2}\alpha\beta - \theta_1\varpi\eta^2 & 1 - \frac{\kappa\sigma}{2}\beta^2 \end{pmatrix}.$$

Let us set $\Lambda(\theta_1,\theta_2) = 1 + \varpi^2\eta^4(1-\theta_1)(1-\theta_2) = \det\mathcal{T}_\theta$. The characteristic polynomial of this amplification matrix $\mathcal{T}_\theta^{-1}\mathcal{M}_\theta$ has one root $\lambda = 1$ and the other roots are also roots of the second-order polynomial

$$P(\lambda) := \Lambda\lambda^2 + \xi\lambda + \zeta \tag{4.30}$$

where

$$\xi = -2 + \varpi^2\eta^4(\theta_1 + \theta_2 - 2\theta_1\theta_2) + \frac{\kappa_u\sigma}{2}\left[\alpha^2 + \beta^2 - \varpi\eta^2\alpha\beta(\theta_2 - \theta_1)\right] + \frac{\kappa_r\sigma}{2}\left(\alpha^2+\beta^2\right)\Lambda$$

and

$$\zeta = 1 + \varpi^2\eta^4\theta_1\theta_2 + \frac{\kappa_r\sigma}{2}\varpi^2\eta^4\left[\alpha^2\theta_1(1-\theta_2) + \beta^2\theta_2(1-\theta_1)\right] - \frac{\kappa_r\sigma}{2}\left(\alpha^2+\beta^2\right) + \sigma\left(\sigma\eta^2 - \frac{\kappa_u}{2} + \sigma\frac{\kappa_r\kappa_u}{4}(\alpha^2+\beta^2)\right)\left[\alpha^2 + \beta^2 - \varpi\eta^2\alpha\beta(\theta_2-\theta_1)\right].$$

Let us first prove Lemma 4.8 and consider for that the stationary state, $k_x = k_y = 0$, which implies $\alpha = \beta = 0$. The characteristic polynomial then reduces to

$$P(\lambda) = \Lambda\lambda^2 + \left[-2 + \varpi^2\eta^4(\theta_2 + \theta_1 - 2\theta_2\theta_1)\right]\lambda + 1 + \varpi^2\eta^4\theta_2\theta_1.$$

For the scheme to be stable, all eigenvalues must satisfy $|\lambda| \leq 1$. In this simple case, a necessary condition is $|\lambda_1 \lambda_2| \leq 1$, which is equivalent to

$$\frac{\zeta}{\Lambda} \leq 1 \iff \varpi^2(1-\theta_2-\theta_1) \geq 0.$$

This proves Lemma 4.8. Let us now turn to the proof of Theorem 4.1.

We now consider the fully discrete LF-DP cell-centered scheme:

$$\kappa_r = 0, \quad \eta = 1 \quad \text{and} \quad -1 \leq \alpha, \beta \leq 1.$$

Then parameters ξ and ζ involved in (4.30) reduce to

$$\xi = -2 + \varpi^2(\theta_2 + \theta_1 - 2\theta_2\theta_1) + \frac{\kappa_u \sigma}{2}\left[\alpha^2 + \beta^2 - \varpi\alpha\beta(\theta_2 - \theta_1)\right]$$

and

$$\zeta = 1 + \varpi^2\theta_2\theta_1 + \sigma\left(\sigma - \frac{\kappa_u}{2}\right)\left[\alpha^2 + \beta^2 - \varpi\alpha\beta(\theta_2 - \theta_1)\right].$$

Imposing $|\lambda| \leq 1$ is equivalent to

$$|\zeta| \leq \Lambda \quad \text{and} \quad |\xi| \leq \Lambda + \zeta.$$

- Firstly, the condition $\zeta \leq \Lambda$ can be written as

  $$f_1(\alpha, \beta) = \varpi^2\left[1 - (\theta_2 + \theta_1)\right] + \sigma\left(\frac{\kappa_u}{2} - \sigma\right)\left[\alpha^2 + \beta^2 - \varpi\alpha\beta(\theta_2 - \theta_1)\right] \geq 0$$

  which in particular holds when

  $$\sigma \leq \frac{\kappa_u}{2} \quad \text{and} \quad \varpi|\theta_2 - \theta_1| \leq 2. \tag{4.31}$$

  Indeed, the latter constraint implies that $\alpha^2 + \beta^2 - \varpi\alpha\beta(\theta_2 - \theta_1) \in [0, 4]$ since $\alpha, \beta \in [-1, 1]$.

- The condition $\zeta \geq -\Lambda$ is equivalent to

  $$f_2(\alpha, \beta) = 2 + \varpi^2\left[1 - (\theta_2 + \theta_1) + 2\theta_2\theta_1\right] - \sigma\left(\frac{\kappa_u}{2} - \sigma\right)\left[\alpha^2 + \beta^2 - \varpi\alpha\beta(\theta_2 - \theta_1)\right] \geq 0.$$

  Under (4.31) and due to the fact that $\kappa_u \in [0, 1]$, we have

  $$2 - \sigma\left(\frac{\kappa}{2} - \sigma\right)4 = 4\left(\sigma - \frac{\kappa_u}{2}\right)^2 + 2 - \frac{\kappa_u^2}{4} \geq 0$$

  which ensures that the requirement $f_2 \geq 0$ is always satisfied.

- The case $-\xi \leq \Lambda + \zeta$ reads

  $$f_3(\alpha, \beta) = \varpi^2 + \sigma^2\left[\alpha^2 + \beta^2 - \varpi\alpha\beta(\theta_2 - \theta_1)\right] \geq 0$$

  which always holds under (4.31).

- Finally, the condition $\xi \leq \Lambda + \zeta$ reads

  $$f_4(\alpha, \beta) = 4 + \varpi^2(1 - 2\theta_2)(1 - 2\theta_1) + \sigma(\sigma - \kappa_u)\left[\alpha^2 + \beta^2 - \varpi\alpha\beta(\theta_2 - \theta_1)\right] \geq 0.$$

Let us notice that due to (4.31) and θ_2, $\theta_1 \in [0,1]$, we have

$$f_4(\alpha,\beta) \geq p_4 := 4\left[\sigma^2 - \sigma\kappa_u - \frac{\varpi^2}{4} + 1\right].$$

Either $\omega^2 h^2 < (4-\kappa_u)a_*^2$ and p_4 is a second-order polynomial with respect to Δt that is always positive: there is no additional constraint upon the time step. Or $\omega^2 h^2 \geq (4-\kappa_u)a_*^2$ and Δt must be small enough to ensure that $p_4 \geq 0$, *i.e.*

$$\Delta t \leq \frac{h}{a_*\kappa_u} \times 2\frac{1-\sqrt{1-\frac{4a_*^2-\omega^2h^2}{a_*^2\kappa_u^2}}}{\frac{4a_*^2-\omega^2h^2}{a_*^2\kappa_u^2}}. \tag{4.32}$$

The convexity of the function $x \mapsto 1-\sqrt{1-x}$ shows that when $\omega^2h^2 \leq 4a_*^2$, the bound in (4.32) is greater than $\frac{h}{a_*\kappa_u} \geq \frac{h}{2a_*}$. Hence in that case (4.32) is less restrictive than (4.31). The study of the monotonicity of the bound with respect to κ_u shows that it is also the case when $\omega^2h^2 > 4a_*^2$. Consequently, the only constraint upon the time step is (4.31) which ends the proof of Theorem 4.1.

In the vertex-based case, the only difference is that $\eta \in [0,1]$. It can be shown that $\eta = 1$ is always the most restrictive constraint and the same stability conditions hold. □

Proposition 4.8. *Let us set* $\varphi(x) = \frac{\sqrt{1+x^2}-1}{x^2}$. *The cell-centered/vertex-based AT-LF and AT-DP schemes are stable provided that the time step is smaller than*

Scheme	$(\theta_1=0,\theta_2=0)$	$(\theta_1=1,\theta_2=0)$ *or* $(\theta_1=0,\theta_2=1)$	$(\theta_1=1/2,\theta_2=1/2)$
AT-LF	$\frac{\kappa_r}{2}\frac{h}{a_*}$	$\min\left\{\frac{2}{\omega}, \frac{\kappa_r h}{4a_*}\varphi\left(\frac{\kappa_r\omega h}{4a_*}\right), \frac{4h}{\kappa_r a_*}\varphi\left(\frac{2\omega h}{\kappa_r a_*}\right)\right\}$	$\frac{\kappa_r h}{a_*}\varphi\left(\frac{\kappa_r\omega h}{2a_*}\right)$
AT-DP	$\frac{2\kappa}{2+\kappa^2}\frac{h}{a_*}$	$\min\left\{\frac{\kappa}{2+\kappa^2}\frac{h}{a_*}, \frac{1}{\omega}\right\}$	$\min\left\{\frac{\kappa}{2+\kappa^2}\frac{h}{a_*}, \frac{2}{\omega}\right\}$

Proof. The proof relies on same kind of computations than Theorem 4.1. □

Remark 4.10. *Contrary to the result in Theorem 4.1, for the choice* $\theta_1 = \theta_2 = 1/2$, *the CFL conditions in Prop. 4.8 still depend on the Coriolis parameter* ω. *The only choice for which the CFL condition does not depend on the Coriolis parameter is a fully implicit discretisation of the Coriolis term, i.e.* $\theta_1 = \theta_2 = 0$.

We also notice that the stability conditions associated to the AT-LF scheme are more restrictive than the conditions for the LF-DP scheme.

4.6.2 Orthogonality-preserving property

We now turn to another major aspect of the linear wave equation which is the preservation of the orthogonal subspace, see Prop. 4.3. It means that when the initial condition is in the orthogonal subspace, the numerical solution remains in this subspace at any time. If the numerical scheme satisfies such a property, we say that this scheme is an *orthogonality-preserving scheme*.

As we shall see below, the original schemes (4.28) and (4.29) are not orthogonality-preserving schemes. That is why we have to modify them. To do so, let us change the time discretisation of

the velocity divergence on the pressure equation in the cell-centered scheme as

$$\begin{cases} \dfrac{r_{i,j}^{n+1}-r_{i,j}^{n}}{\Delta t}+a_\star[\nabla_{2h}^c\cdot\boldsymbol{u}_h^\tau]_{i,j}-\nu_r\left[\nabla_{2h}^c\cdot\left(\nabla_{2h}^c r_h^n+\dfrac{\omega}{a_*}(\boldsymbol{u}_h^n)^\perp\right)\right]_{i,j}=0,\\ \dfrac{\boldsymbol{u}_{i,j}^{n+1}-\boldsymbol{u}_{i,j}^{n}}{\Delta t}+a_\star[\nabla_{2h}^c r_h^n]_{i,j}-\nu_u\left[\nabla_{2h}^c(\nabla_{2h}^c\cdot\boldsymbol{u}_h^n)\right]_{i,j}=-\omega\boldsymbol{u}_{i,j}^{\theta,\perp}, \end{cases} \tag{4.33}$$

and in the vertex-based scheme as

$$\begin{cases} \dfrac{r_{i,j}^{n+1}-r_{i,j}^{n}}{\Delta t}+a_\star f_h^{\mathrm{c}}\left[\nabla_h^{\mathrm{v}}\cdot\boldsymbol{u}_h^\tau\right]_{i,j}-\nu_r\nabla_h^{\mathrm{c}}\cdot\left[\nabla_h^{\mathrm{v}} r_h^n+\omega f_h^{\mathrm{v}}\left((\boldsymbol{u}_h^n)^\perp\right)\right]_{i,j}=0,\\ \dfrac{\boldsymbol{u}_{i,j}^{n+1}-\boldsymbol{u}_{i,j}^{n}}{\Delta t}+a_\star f_h^{\mathrm{c}}\left[\nabla_h^{\mathrm{v}} r_h^n\right]_{i,j}-\nu_u\nabla_h^{\mathrm{c}}\left[\nabla_h^{\mathrm{v}}\cdot\boldsymbol{u}_h^n\right]_{i,j}=-\omega f_h^{\mathrm{c}}\Big[f_h^{\mathrm{v}}(\boldsymbol{u}_h^\theta)\Big]_{i,j}^\perp, \end{cases} \tag{4.34}$$

for $\boldsymbol{u}_h^\tau=(\tau_1u^n+(1-\tau_1)u^{n+1},\tau_2v^n+(1-\tau_2)v^{n+1})^T$ with τ_1, $\tau_2\in[0,1]$. Note that these modified schemes are still explicit since the updated velocity can be computed first and then introduced in the pressure equation.

It is straightforward to prove that these modified schemes still preserve the corresponding discrete kernels (4.20) and (4.25). They also preserve the orthogonal subspace:

Proposition 4.9. *The fully discrete cell-centered (4.33) and vertex-based (4.34) schemes are orthogonality-preserving schemes provided that*

$$\kappa_r=0 \quad \textit{and} \quad \tau_1=\theta_1, \quad \tau_2=\theta_2. \tag{4.35}$$

Proof. Let us assume that $q_h^n\in\mathcal{E}_{\omega\neq0,h}^{c,\perp}$ and show that $q_h^{n+1}\in\mathcal{E}_{\omega\neq0,h}^{c,\perp}$.

Taking the discrete scalar product of (4.33) with $\hat{q}_h\in\mathcal{E}_{\omega\neq0,h}^{c}$, we obtain

$$\begin{aligned}\langle q_h^{n+1},\hat{q}_h\rangle=\langle q_h^n,\hat{q}_h\rangle&-a_\star\Delta t\left(\langle\nabla_{2h}^c\cdot\boldsymbol{u}_h^\tau,\hat{r}_h\rangle+\langle\nabla_{2h}^c r_h^n,\hat{\boldsymbol{u}}_h\rangle\right)\\&+\nu_u\Delta t\langle\nabla_{2h}^c[\nabla_{2h}^c\cdot\boldsymbol{u}_h^n],\hat{\boldsymbol{u}}_h\rangle+\nu_r\Delta t\left\langle\nabla_{2h}^c\cdot\left[\nabla_{2h}^c r_h^n+\frac{\omega}{a_\star}\boldsymbol{u}_h^{n,\perp}\right],\hat{r}_h\right\rangle-\omega\Delta t\langle\boldsymbol{u}_h^{\theta,\perp},\hat{\boldsymbol{u}}_h\rangle.\end{aligned}$$

Because of $\nabla_{2h}^c\cdot\hat{\boldsymbol{u}}_h=0$ and due to Lemma 4.4, we have

$$\begin{aligned}\langle\nabla_{2h}^c r_h^n,\hat{\boldsymbol{u}}_h\rangle&=-\langle r_h^n,\nabla_{2h}^c\cdot\hat{\boldsymbol{u}}_h\rangle=0,\\ \langle\nabla_{2h}^c[\nabla_{2h}^c\cdot\boldsymbol{u}_h],\hat{\boldsymbol{u}}_h\rangle&=-\langle\nabla_{2h}^c\cdot\boldsymbol{u}_h,\nabla_{2h}^c\cdot\hat{\boldsymbol{u}}_h\rangle=0.\end{aligned}$$

Moreover $\langle q_h^n,\hat{q}_h\rangle=0$ and

$$-a_\star\Delta t\,\langle\nabla_{2h}^c\cdot\boldsymbol{u}_h^\tau,\hat{r}_h\rangle=a_\star\Delta t\,\langle\boldsymbol{u}_h^\tau,\nabla_{2h}^c\hat{r}_h\rangle=-\omega\Delta t\langle\boldsymbol{u}_h^\tau,\hat{\boldsymbol{u}}_h^\perp\rangle=\omega\Delta t\langle\boldsymbol{u}_h^{\tau,\perp},\hat{\boldsymbol{u}}_h\rangle.$$

As a result, we obtain

$$\langle q_h^{n+1},\hat{q}_h\rangle=\omega\Delta t\left\langle(\boldsymbol{u}_h^\tau-\boldsymbol{u}_h^\theta)^\perp,\hat{\boldsymbol{u}}_h\right\rangle+\nu_r\Delta t\left\langle\nabla_{2h}^c\cdot\left[\nabla_{2h}^c r_h^n+\frac{\omega}{a_\star}\boldsymbol{u}_h^{n,\perp}\right],\hat{r}_h\right\rangle.$$

Therefore, in order to ensure that $\forall\,\hat{q}_h\in\mathcal{E}_{\omega\neq0,h}^{c}$, $\langle q_h^{n+1},\hat{q}_h\rangle=0$, we need $\nu_r=0$ and $\tau_1=\theta_1$, $\tau_2=\theta_2$.

Similarly, for the vertex-based scheme (4.34), we have

$$\begin{aligned}\langle q_h^{n+1},\hat{q}_h\rangle=\langle q_h^n,\hat{q}_h\rangle-a_\star\Delta t\left(\langle f_h^{\mathrm{c}}\left[\nabla_h^{\mathrm{v}} r_h^n\right],\hat{\boldsymbol{u}}_h\rangle+\langle f_h^{\mathrm{c}}\left[\nabla_h^{\mathrm{v}}\cdot\boldsymbol{u}_h^\tau\right],\hat{r}_h\rangle\right)+\nu_u\Delta t\,\langle\nabla_h^{\mathrm{c}}\left[\nabla_h^{\mathrm{v}}\cdot\boldsymbol{u}_h^n\right],\hat{\boldsymbol{u}}_h\rangle\\+\nu_r\Delta t\left\langle\nabla_h^{\mathrm{c}}\cdot\left[\nabla_h^{\mathrm{v}} r_h^n+\omega f_h^{\mathrm{v}}(\boldsymbol{u}_h^{n,\perp})\right],\hat{r}_h\right\rangle-\omega\Delta t\left\langle f_h^{\mathrm{c}}\left[f_h^{\mathrm{v}}(\boldsymbol{u}_h^{\theta,\perp})\right],\hat{\boldsymbol{u}}\right\rangle\end{aligned}$$

and due to Lemma 4.7

$$\langle q_h^{n+1}, \hat{q}_h \rangle = \nu_r \Delta t \left\langle \nabla_h^{\mathrm{c}} \cdot \left[\nabla_h^{\mathrm{v}} r_h^n + \omega f_h^{\mathrm{v}}(\boldsymbol{u}_h^{n,\perp}) \right], \hat{r}_h \right\rangle + \omega \Delta t \left\langle f_h^{\mathrm{c}} \left[f_h^{\mathrm{v}}(\hat{\boldsymbol{u}}_h^{\perp}) \right], \boldsymbol{u}_h^{\theta} - \boldsymbol{u}_h^{\tau} \right\rangle.$$

Therefore, under (4.35), we have $\langle q_h^{n+1}, \hat{q}_h \rangle = 0$ for any $\hat{q}_h \in \mathcal{E}^v_{\omega \neq 0, h}$. □

4.7 Numerical results

4.7.1 Well-balanced test case with initial condition in the kernel

We first come back to the test case presented in Section 4.3 to explain the wrong behaviour of the classical scheme and of the naive corrections referred to as LF-C and C-LF strategies. In practice we define the initial discrete pressure r by using relation (4.7) applied at the cell centers and the initial discrete velocity by using the definition of the discrete kernel (4.20). The initial state is then a discrete stationary solution when we use the scheme defined by (4.28) or (4.33). As expected, the AT-DP, AT-LF and LF-DP strategies exactly maintain the stationary state, whereas the results obtained with the AT-C and C-DP strategies are very similar to the ones obtained with the LF-C and C-LF strategies and are not able to preserve the stationary state, compare Fig. 4.6 and 4.3.

In Fig. 4.6 we present the results for two different grid sizes and two different final times. It is clear that the error decreases when the mesh is refined and increases with time, that is not surprising. As it has already been noticed, it clearly appears that, for this test case, the correction on the diffusion for the velocity equation, *i.e.*C-DP strategy, has a much larger impact than the correction on the diffusion for the pressure equation, *i.e.*AT-C strategy, but is not enough to preserve the stationary state. This behaviour will be investigated in more details in Section 4.7.3.

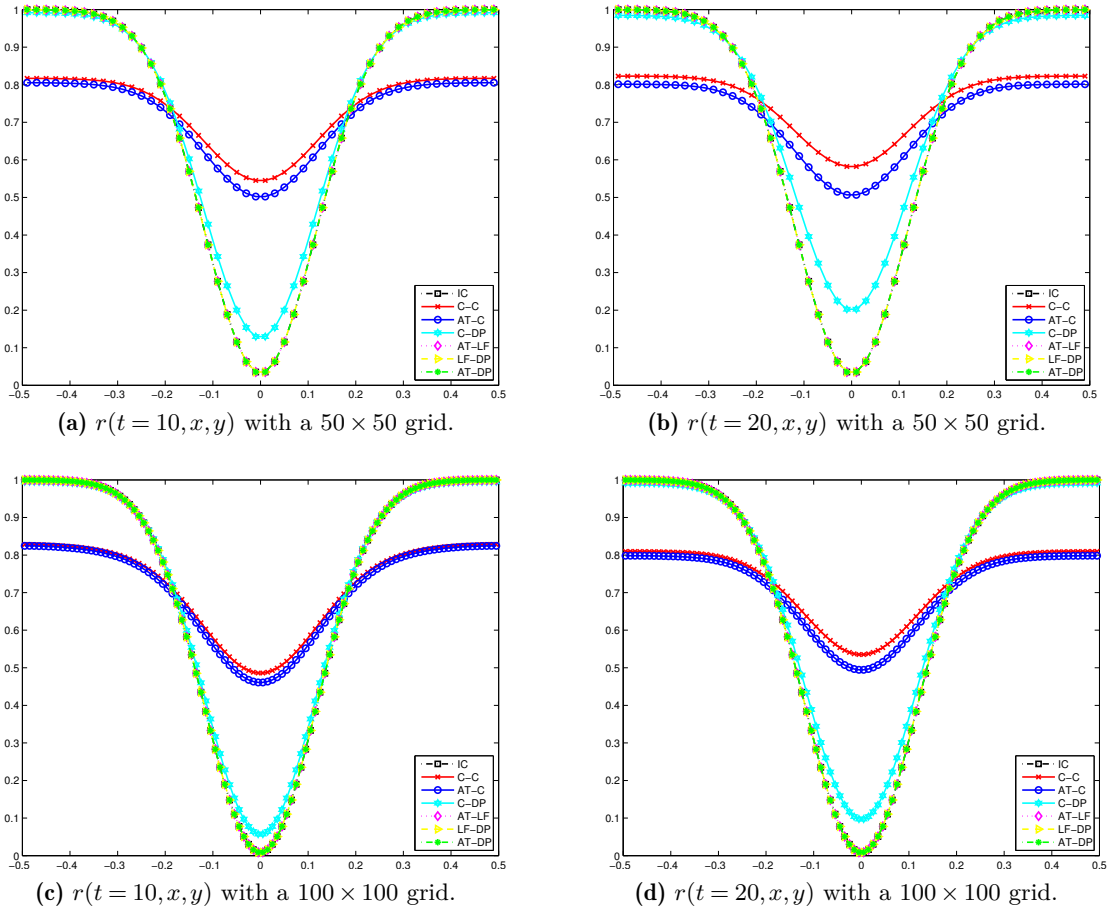
4.7.2 Orthogonality-preserving test case with initial condition in the orthogonal subspace

In this test case, we consider periodic boundary conditions and an initial velocity field given by

$$\begin{cases} u(t=0,x,y) = \frac{1}{2} \exp\left[-\left(\frac{4x}{0.4}\right)^2 - \left(\frac{4y}{0.8}\right)^2 \right] \\ v(t=0,x,y) = \frac{1}{2} \exp\left[-\left(\frac{4x}{0.8}\right)^2 - \left(\frac{4y}{0.4}\right)^2 \right]. \end{cases}$$

in the domain $\mathbb{T}^2 = [-0.5, 0.5] \times [-0.5, 0.5]$. Then the initial pressure $r(t=0,x,y)$ is constructed by using the definition of the discrete orthogonal subspace (4.21). Note that for this test case, we only present results for the cell-centered scheme (4.28) for which we can compute explicitly the orthogonal of the kernel. Note that for the vertex-based scheme (4.29), we can also prove that the LF-DP scheme with the appropriate u^τ velocity preserves the orthogonal, but we cannot provide an explicit expression for this subspace. The time discretisation parameter for Coriolis term is $\theta_1 = \theta_2 = 1/2$ for all the numerical results. A 50×50 grid is used.

As expected, Figure 4.7(a) indicates that, for the choice $\tau = 1$, no scheme is orthogonality-preserving (if ti were the case, the curves would remain exactly zero). Nevertheless it clearly appears that the projection $\hat{q}$ onto the kernel depends on the numerical strategy and is much larger for the C-C and the AT-C schemes than for the other ones. Figure 4.7(b) shows that the orthogonal part of the solution is less damped when the LF strategy is used, i.e. for AT-LF and LF-DP schemes, since the numerical diffusion is canceled for one equation. In Fig. 4.7(c) and Fig. 4.7(d), we present the same results, but focusing on the LF-DP scheme for different values of the parameter τ used for the time discretisation of the velocity in the pressure equation. It appears on Fig. 4.7(c) that the case with $\tau = \theta$ is the only one for which the projection of the



(a) $r(t=10,x,y)$ with a 50×50 grid. **(b)** $r(t=20,x,y)$ with a 50×50 grid.

(c) $r(t=10,x,y)$ with a 100×100 grid. **(d)** $r(t=20,x,y)$ with a 100×100 grid.

Figure 4.6: *Cross-section of pressure.*

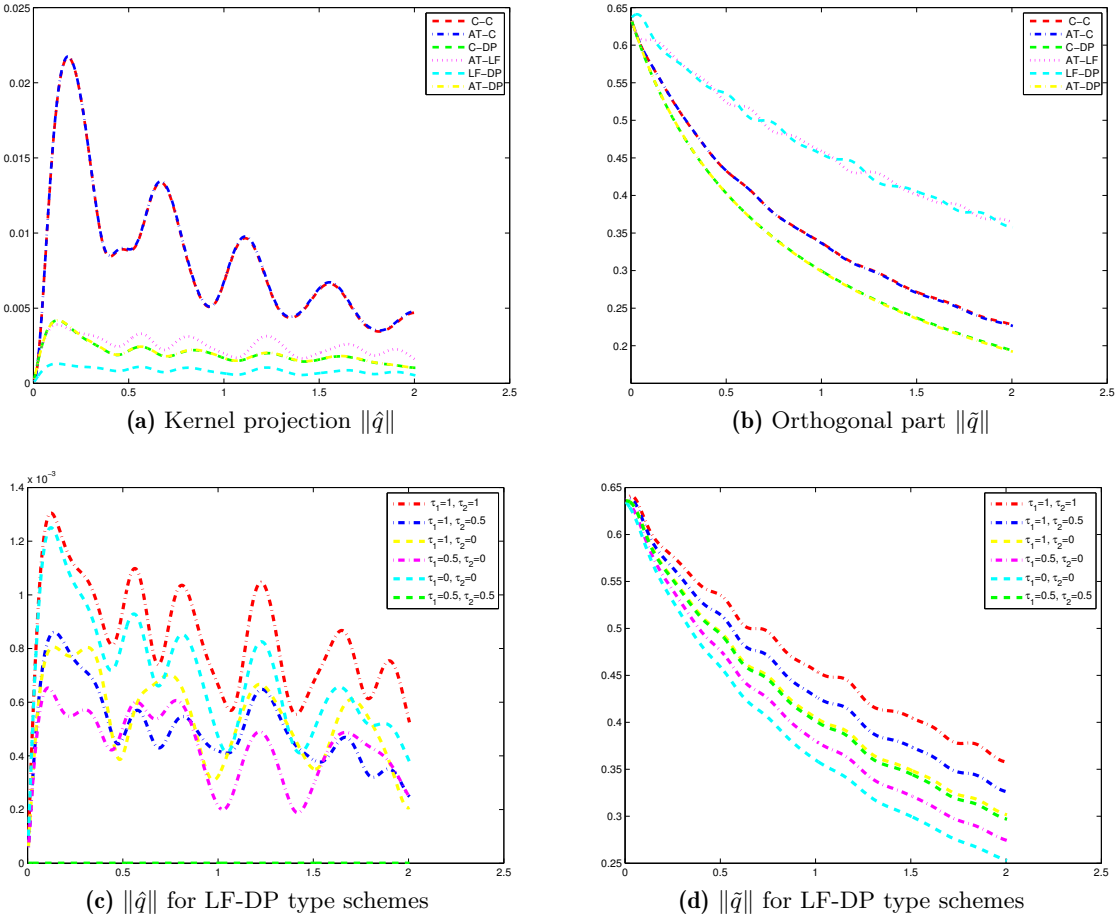
solution in the kernel remains zero for all time, which means that the orthogonal subspace is stable for the scheme. In Figure 4.7(d), it appears that the damping increases when the time discretisation becomes more and more implicit, *i.e.*the parameter τ becomes smaller and smaller. Note that the choice $\tau=\theta=1/2$ for which the orthogonal is a stable subspace corresponds to a mean damping: contrary to the previous test case, the solution evolves but remains in the orthogonal.

4.7.3 Behaviour of the solution with initial condition close to the kernel

We now consider an initial condition close to the discrete kernel up to a perturbation of size $M\ll 1$

$$q_h^0=\hat{q}_h^0+M\frac{\tilde{q}_h^0}{\|\tilde{q}_h^0\|},$$

where $\hat{q}_h^0$ stands for the projection onto the kernel given in Section 4.7.1 and $\tilde{q}_h^0$ is the orthogonal part considered in Section 4.7.2. Here the Froude number M is set equal to 10^{-3} and a 50×50 grid is used. In Figure 4.8(a) we present the evolution in time of the deviation from the initial projection $\hat{q}_h^0$. It appears that for the C-C, AT-C and C-DP schemes, that are not able to maintain steady states, the deviation increases regularly with time. Nevertheless it increases much faster for C-C and AT-C schemes than for C-DP schemes, which reinforces the conclusions of the first numerical example, see Section 4.7.1. For C-C and AT-C schemes, the deviation becomes almost constant when the discrete solution reaches a stationary state of the scheme,



(a) Kernel projection $\|\hat{q}\|$

(b) Orthogonal part $\|\tilde{q}\|$

(c) $\|\hat{q}\|$ for LF-DP type schemes

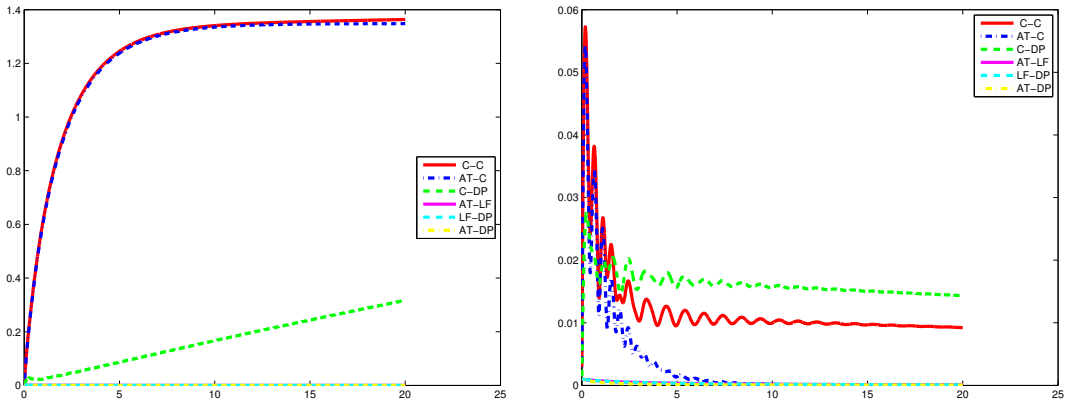
(d) $\|\tilde{q}\|$ for LF-DP type schemes

Figure 4.7: *Evolution of the kernel and orthogonal part for* $\theta_1 = \theta_2 = \frac{1}{2}$.

which is very different from the initial one since the kernels of those scheme are inaccurate approximations of the continuous ones, see Lemma 4.2. The same phenomenon should occur for the C-DP scheme but since the deviation increases slowly, one needs to wait for a long time.

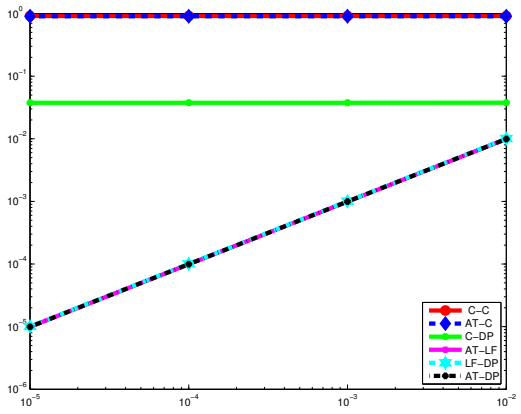
In Fig. 4.8(b) we present the norm of the part of the solution that belongs to the orthogonal subspace. It appears that for each scheme, it is mostly decreasing in time, despite some oscillations, meaning that, for each scheme, the solution tends to a stationary state that belongs to the kernel of the considered scheme. Note that the solution of the AT-C scheme tends quite quickly to a stationary state in its kernel since the orthogonal part vanishes. For C-C and C-DP schemes, the decreasing of the orthogonal part is slower, which explains that the deviation is still increasing in Fig. 4.8(a), even if very slowly for large time for the C-C scheme.

In Fig. 4.9, we present for different values of M, the maximum value, over the time interval, of the deviation from the initial projection $\hat{q}_h^0$. It clearly exhibits that, for the well-balanced LF-DP, AT-LF and AT-DP strategies, the deviation is proportional to M whereas it remains constant for the other strategies, even if the constant is smaller for the C-DP scheme than for the C-C and AT-C schemes. It emphasizes the importance of the well-balanced strategy to ensure the accuracy near the geostrophic equilibrium.



(a) $\|q - \mathbb{P}q^0\|(t)$. **(b)** $\|q - \mathbb{P}q\|(t)$.

Figure 4.8: *Evolution in time of the deviation for an initial condition close to the discrete kernel.*



(a)

Figure 4.9: $\max_{t\in[0,2]} \|q - \mathbb{P}q^0\|(t)$ *as a function of the Froude number (log-log scale)*

4.7.4 Water column test case and geostrophic adjustment

In this test case, we consider a discontinuous initial condition which is given by

$$\begin{cases} r(t=0,x,y) = \begin{cases} 2, & \text{if} \quad x^2+y^2 \leq 1 \\ 1, & \text{if} \quad x^2+y^2 > 1, \end{cases} \\ u(t=0,x,y) = 0, \\ v(t=0,x,y) = 0. \end{cases}$$

with periodic boundary conditions on the domain $[-5,5]\times[-5,5]$. This initial condition corresponds to a circular dam break and is very far from the geostrophic equilibrium (4.3). Hence the solution of the wave equation with Coriolis term (4.2) will contain a travelling wave that should go out of the domain (here due to periodic boundary conditions, the waves remain in the domain but will vanish for long time because of numerical diffusion) and the remaining stationary state will be the geostrophic equilibrium (4.3) corresponding to the initial data. Discrete solutions will exhibit the same behaviour but the remaining state will belong to the discrete kernel of each scheme.

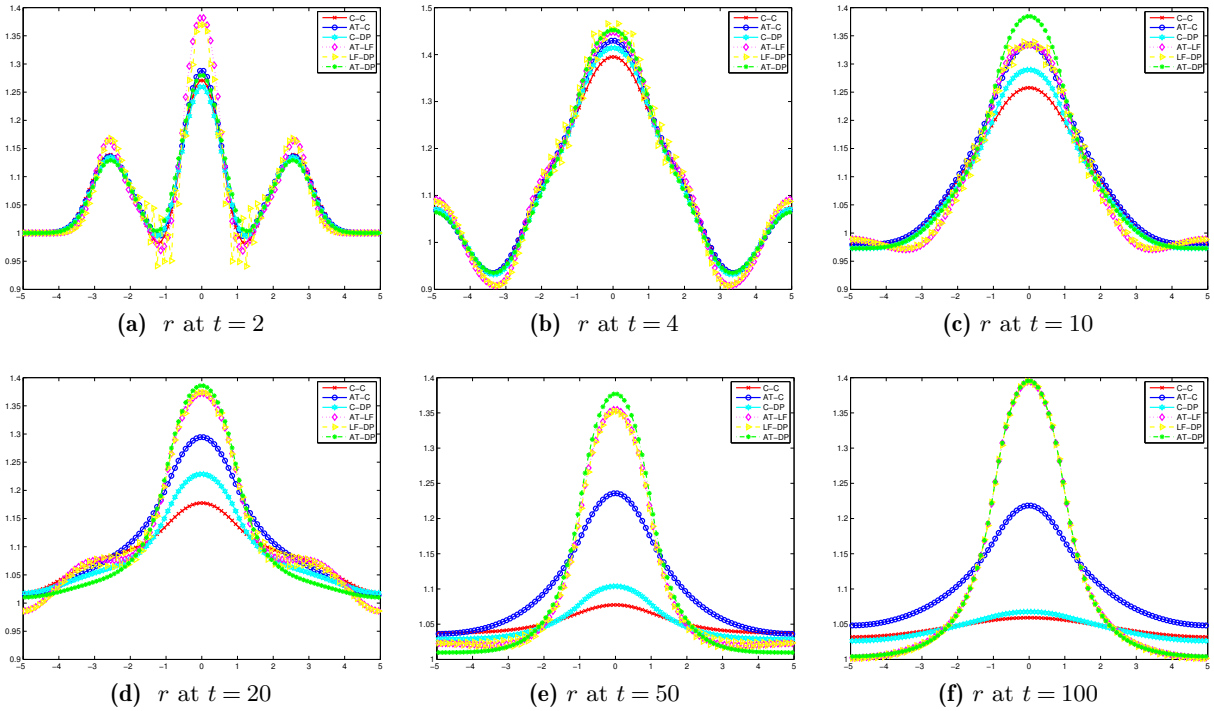
In Fig. 4.10, we present the evolution in time of the pressure r for different schemes. In Fig. 4.10(f), *i.e.*for long time, three groups can be exhibited: the one corresponding to the well-balanced schemes, *i.e.*the LF-DP, AT-LF and AT-DP schemes, the one corresponding to the schemes for which the kernel is given by (4.9a), *i.e.*the C-C and the C-DP schemes, and the AT-C scheme for which the kernel is given by (4.9c). In Fig. 4.12 we present at the final time the results for the quantities r, u and v for three schemes, corresponding to the three groups previously mentioned. Results appear to be very different (note the scale is not the same for the three figures).

On the left column, solutions of the C-C and C-DP schemes are close to a constant state (see the scale on the z-axis) that corresponds to the discrete kernel (4.9a). Note that the discrete kernels (4.9a) and (4.9b) are the same and correspond respectively to the C-C and C-DP schemes. On the center column, u-velocity (*resp.* v-velocity) corresponding to the solution of the AT-C scheme is almost constant in the x-direction (*resp.* in the y-direction). It is in agreement with the definition of the kernel (4.9c), that is neither a constant state, nor a good approximation of the continuous geostrophic equilibrium (4.3).

In the right column, solution for the AT-DP scheme is very similar to the geostrophic equilibrium plotted in Fig. 4.1, which may indicate that the solution is close to the discrete kernel (4.20), that has been proven to be a good approximation of the continuous geostrophic equilibrium (4.3). It is clearly exhibited in Fig. 4.11 where we show that, for long time, the gradient of the pressure along the x-axis balances exactly the x-component of the Coriolis force, which characterizes the geostrophic equilibrium (the result would be the same for any cross-section in any direction). Among the C-C, AT-C and C-DP schemes, that are not well-balanced, note that, whereas the C-DP scheme appeared to be preferable in the previous test cases since the deviation from the discrete geostrophic equilibrium remained relatively small, here, the solution of the C-DP scheme is very similar to the one of the classical C-C scheme and is totally inaccurate. It allows to conclude that the well-balanced property is absolutely necessary to obtain accurate solutions for a large range of test cases.

In Fig. 4.10(a) to 4.10(e) we present the transient part of this geostrophic adjustment. It appears that the time evolution of the solutions of the three well-balanced schemes, even if they converge to the same state, is not completely similar. In particular the solution for the AT-DP scheme is different from a group composed by the solutions corresponding to the AT-LF and LF-DP schemes. Note also that for short time, the solution of the LF-DP scheme presents some oscillations, that are due to the discontinuity of the initial solution. This difference is highlighted

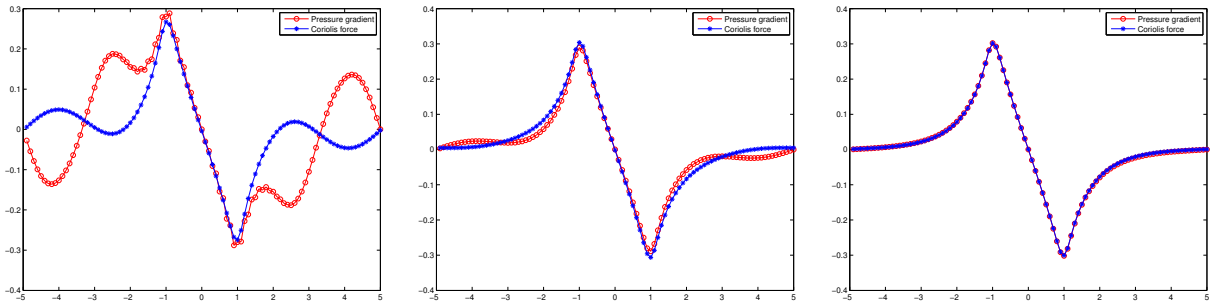
in Fig. 4.13 where we present the time evolution of the energy. Indeed, even if the final state is the same for the three well-balanced schemes, the time evolution is different for, on the one hand, the AT-DP scheme, and, on the other hand, the AT-LF and the LF-DP schemes, for which the energy decreases more slowly. Nevertheless note that, as expected from Th. 4.1, the energy is globally decreasing for all schemes, even if we consider a discontinuous initial condition.



(a) r at $t = 2$ **(b)** r at $t = 4$ **(c)** r at $t = 10$

(d) r at $t = 20$ **(e)** r at $t = 50$ **(f)** r at $t = 100$

Figure 4.10: *Cross section of the pressure r at $y = 0$ at different times.*

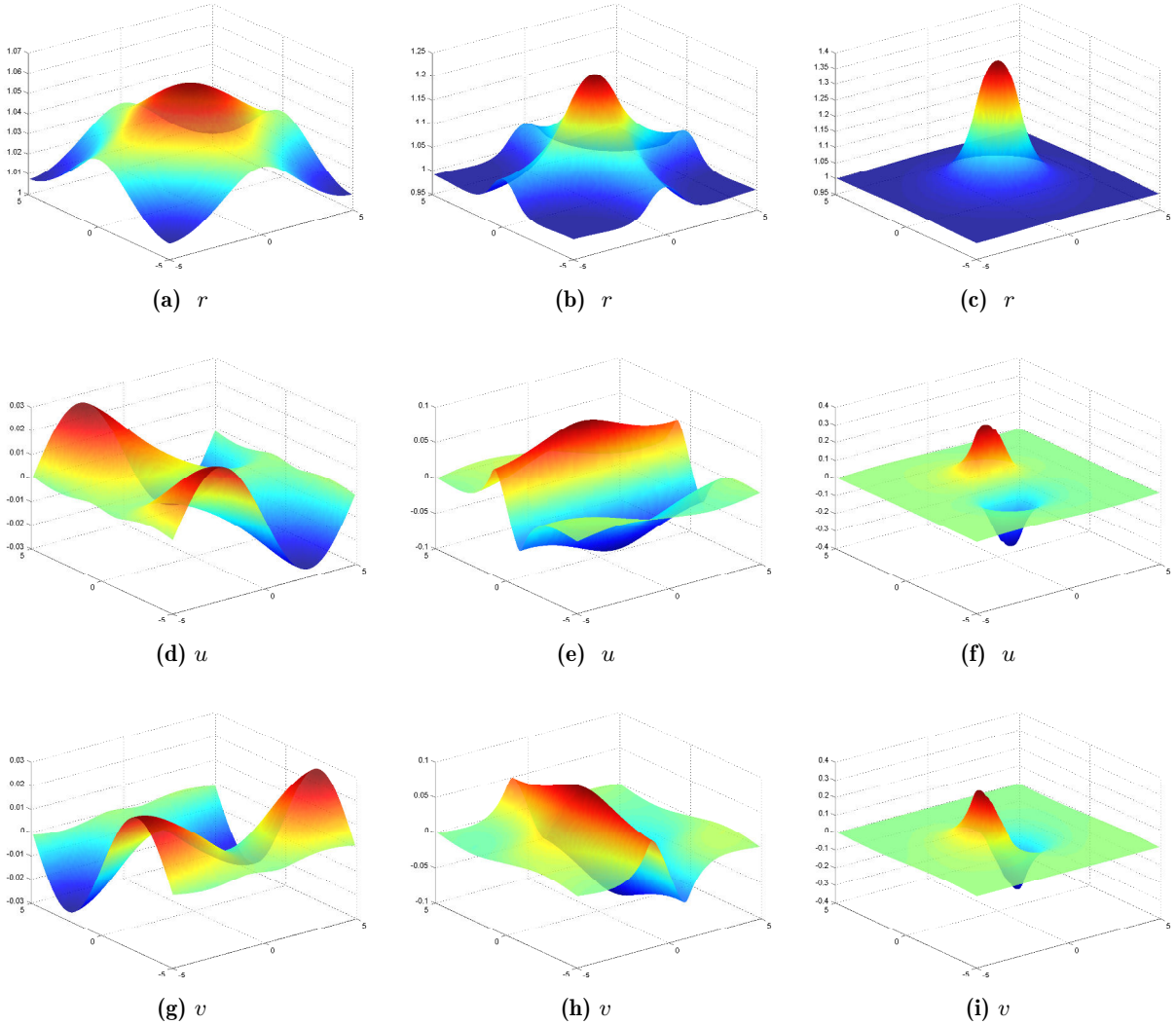


(a) $t = 4$ **(b)** $t = 20$ **(c)** $t = 100$

Figure 4.11: *Cross section of the pressure gradient and Coriolis force at $y = 0$ for AT-DP scheme.*

4.8 Conclusion

In this work we propose new collocated finite volume Godunov type schemes to compute accurate approximate solutions of the wave equation with Coriolis term. The main ingredient of the method is to modify the numerical diffusion of the scheme to make the discrete kernel compatible with the so-called geostrophic equilibrium. It extends techniques proposed in [13] and [20]. We propose three different well-balanced schemes, namely the AT-LF (Apparent Topography & Low



(a) r **(b)** r **(c)** r

(d) u **(e)** u **(f)** u

(g) v **(h)** v **(i)** v

Figure 4.12: *Comparison between C-C (left), AT-C (middle) and AT-DP (right) schemes at time* $t = 100$.

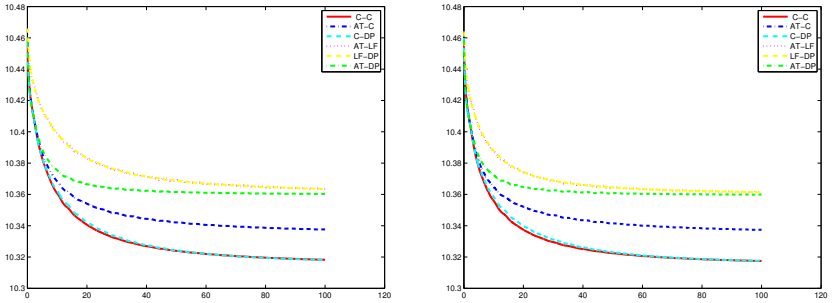
Froude) scheme, the LF-DP (Low Froude & Divergence Penalisation) scheme and the AT-DP (Apparent Topography & Divergence Penalization) scheme, and two different ways to discretise the geostrophic equilibrium, namely at the centers of the cell or at the interfaces.

The main result of the paper is the proof of stability, under classical CFL conditions, of all these modified schemes, see Th. 4.1. Moreover some numerical test cases allow us to investigate the behaviour of the schemes for different kinds of initial solutions, including discontinuous ones, and conclude that the well-balanced property is essential to ensure an accurate geostrophic adjustment. Future works will be dedicated to the extension of these results to the fully nonlinear two-dimensional shallow water equations with Coriolis term (4.1).

4.A Proof of the Hodge decomposition in the continuous case (Prop. 4.1)

Proof. In order to prove (4.4), let us denote by $\mathbb{A}$ the space

$$\mathbb{A} := \left\{ (p, \boldsymbol{v}) \in \left(L^2(\mathbb{T}^2) \right)^3 \;\middle|\; \forall \varphi \in C_c^\infty(\mathbb{T}^2),\ \int_{\mathbb{T}^2} a_\star \boldsymbol{v}^\perp \cdot \nabla \varphi \ \mathrm{d}\boldsymbol{x} = \int_{\mathbb{T}^2} \omega p \varphi \ \mathrm{d}\boldsymbol{x} \right\}.$$



(a) Energy of cell center scheme **(b)** Energy of interface scheme scheme

Figure 4.13: *Evolution in time of the energy*

We first show that $\mathbb{A}$ is a subset of $\mathcal{E}^{\perp}_{\omega\neq 0}$. Let us take $\tilde{q}=(p,\boldsymbol{v})\in\mathbb{A}$. Then for all $q=(r,\boldsymbol{u})\in\mathcal{E}_{\omega\neq 0}$, we have

$$\begin{aligned}\langle\tilde{q},q\rangle &= \int_{\mathbb{T}^2} rp\ \mathrm{d}\boldsymbol{x}+\int_{\mathbb{T}^2}\boldsymbol{u}\cdot\boldsymbol{v}\ \mathrm{d}\boldsymbol{x}=\int_{\mathbb{T}^2} rp\ \mathrm{d}\boldsymbol{x}+\frac{a_\star}{\omega}\int_{\mathbb{T}^2}\boldsymbol{v}\cdot\nabla^{\perp}r\ \mathrm{d}\boldsymbol{x}\\ &= \int_{\mathbb{T}^2}\frac{\omega}{a_\star}pr\ \mathrm{d}\boldsymbol{x}-\int_{\mathbb{T}^2}\boldsymbol{v}^{\perp}\cdot\nabla r\ \mathrm{d}\boldsymbol{x}=0.\end{aligned}$$

By density of $C^{\infty}_c(\mathbb{T}^2)$ in $H^1(\mathbb{T}^2)$, it follows that $\tilde{q}=(p,\boldsymbol{v})\in\mathcal{E}^{\perp}_{\omega\neq 0}$. Therefore, we conclude that $\mathbb{A}\subset\mathcal{E}^{\perp}_{\omega\neq 0}$.

On the other hand, let us take $\tilde{q}=(p,\boldsymbol{v})\in\mathcal{E}^{\perp}_{\omega\neq 0}$. For any $\phi\in H^1(\mathbb{T}^2)$ we have $\hat{q}:=(\frac{\omega}{a_\star}\phi,\nabla^{\perp}\phi)\in\mathcal{E}_{\omega\neq 0}$. This provides

$$\langle\tilde{q},q\rangle=0\implies\int_{\mathbb{T}^2}\frac{\omega}{a_\star}\phi p\ \mathrm{d}\boldsymbol{x}-\int_{\mathbb{T}^2}\boldsymbol{v}^{\perp}\cdot\nabla\phi\ \mathrm{d}\boldsymbol{x}=0.$$

As a result, we have

$$\forall\,\phi\in H^1(\mathbb{T}^2),\ \int_{\mathbb{T}^2}\frac{\omega}{a_\star}\phi p\ \mathrm{d}\boldsymbol{x}=\int_{\mathbb{T}^2}\boldsymbol{v}^{\perp}\cdot\nabla\phi\ \mathrm{d}\boldsymbol{x},$$

which leads to

$$\forall\,\phi\in C^{\infty}_c(\mathbb{T}^2),\ \int_{\mathbb{T}^2}\frac{\omega}{a_\star}\phi p\ \mathrm{d}\boldsymbol{x}=\int_{\mathbb{T}^2}\boldsymbol{v}^{\perp}\cdot\nabla\phi\ \mathrm{d}\boldsymbol{x}.$$

It implies that $\tilde{q}\in\mathbb{A}$, that is to say $\mathcal{E}^{\perp}_{\omega\neq 0}$ is a subset of $\mathbb{A}$. In conclusion, we have

$$\mathcal{E}^{\perp}_{\omega\neq 0}=\mathbb{A}=\left\{(p,\boldsymbol{v})\in\left(L^2(\mathbb{T}^2)\right)^3\,\middle|\,\forall\varphi\in C^{\infty}_c(\mathbb{T}^2),\ \int_{\mathbb{T}^2}a_\star\boldsymbol{v}^{\perp}\cdot\nabla\varphi\ \mathrm{d}\boldsymbol{x}=\int_{\mathbb{T}^2}\omega p\varphi\ \mathrm{d}\boldsymbol{x}\right\}.$$

We eventually have to prove that

$$\mathcal{E}_{\omega\neq 0}\oplus\mathcal{E}^{\perp}_{\omega\neq 0}=\left(L^2(\mathbb{T}^2)\right)^3.$$

By the fact that $\mathcal{E}_{\omega\neq 0}\oplus\mathcal{E}^{\perp}_{\omega\neq 0}\subset\left(L^2(\mathbb{T}^2)\right)^3$ is trivial, we only have to check $\left(L^2(\mathbb{T}^2)\right)^3\subset\mathcal{E}_{\omega\neq 0}\oplus\mathcal{E}^{\perp}_{\omega\neq 0}$. We suppose $q\in(L^2(\mathbb{T}^2))^3$, we shall find $\hat{q}\in\mathcal{E}_{\omega\neq 0}$ and $\tilde{q}\in\mathcal{E}^{\perp}_{\omega\neq 0}$ such that $q=\hat{q}+\tilde{q}$. For $q=(r,u,v)\in\left(L^2(\mathbb{T}^2)\right)^3$, let us denote $\mu(r)=\frac{1}{|\mathbb{T}^2|}\int_{\mathbb{T}^2}r\ \mathrm{d}\boldsymbol{x}$ and consider the following variational form :

Find $h\in H^1(\mathbb{T}^2)$ *such that*: $\forall\ \varphi\in H^1(\mathbb{T}^2)$, $a(h,\varphi)=F(\varphi)$, where

$$a(h,\varphi):=\int_{\mathbb{T}^2}\nabla h\cdot\nabla\varphi\ \mathrm{d}\boldsymbol{x}+\left(\frac{\omega}{a_\star}\right)^2\int_{\mathbb{T}^2}h\varphi\ \mathrm{d}\boldsymbol{x},\ F(\varphi):=\frac{\omega}{a_\star}\int_{\mathbb{T}^2}\boldsymbol{u}^{\perp}\cdot\nabla\varphi\ \mathrm{d}\boldsymbol{x}-\left(\frac{\omega}{a_\star}\right)^2\int_{\mathbb{T}^2}(r-\mu(r))\varphi\ \mathrm{d}\boldsymbol{x}. \tag{4.A.1}$$

The existence and uniqueness of $h \in H^1(\mathbb{T}^2)$ results from the Lax-Milgram theorem for $\omega \neq 0$. We consider the decomposition for r given by

$$r = \hat{r} + \tilde{r} \quad \text{with} \quad \hat{r} = \mu(r) - h \quad \text{and} \quad \tilde{r} = r - \mu(r) + h.$$

For the decomposition of $\boldsymbol{u}$, we simply construct $\hat{\boldsymbol{u}}$ by setting

$$\hat{\boldsymbol{u}} = \frac{a_\star}{\omega} \nabla^\perp \hat{r} \quad \text{and} \quad \tilde{\boldsymbol{u}} = \boldsymbol{u} - \hat{\boldsymbol{u}},$$

which implies $(\hat{r}, \hat{\boldsymbol{u}}) \in \mathcal{E}_{\omega \neq 0}$ and

$$\hat{\boldsymbol{u}}^\perp = -\frac{a_\star}{\omega} \nabla \hat{r} = \frac{a_\star}{\omega} \nabla \hat{h}.$$

Therefore, (4.A.1) implies that for all $\varphi \in H^1(\mathbb{T}^2)$ we have

$$\frac{\omega}{a_\star} \int_{\mathbb{T}^2} (\hat{\boldsymbol{u}} - \boldsymbol{u})^\perp \cdot \nabla \varphi \, \mathrm{d}\boldsymbol{x} + \left(\frac{\omega}{a_\star} \right)^2 \int_{\mathbb{T}^2} \tilde{r} \varphi \, \mathrm{d}\boldsymbol{x} = 0$$

which implies that

$$\forall \varphi \in C_c^\infty(\mathbb{T}^2), \int_{\mathbb{T}^2} a_\star \tilde{\boldsymbol{u}}^\perp \cdot \nabla \varphi \, \mathrm{d}\boldsymbol{x} = \int_{\mathbb{T}^2} \omega \tilde{r} \varphi \, \mathrm{d}\boldsymbol{x}.$$

□

CHAPTER 5

Analysis of staggered type schemes applied to the linear wave equation with Coriolis source term. Part 1: on Cartesian meshes

Problems are not stop signs,
they are guidelines.

Robert H. Schuller.

Abstract

The numerical viscosity on both the pressure and velocity equations are responsible for the inaccuracy problem of the classical Godunov scheme applied to the two dimensional linear wave equation with Coriolis force. To overcome this difficulty, based on the study of the modified equation, the work in [53] proposes corrections to the standard diffusion terms by using a mixture among the *Apparent Topography* method in [13], the *Low Froude* and *Divergence penalization* method mentioned in [20]. In this work, we develop this idea to construct some staggered type schemes on the Arakawa B and D grids, such that those schemes capture well the discrete geostrophic equilibrium which are the stationary states of the system, as well as the subspace which is orthogonal to these stationary states. A Fourier analysis is preformed to compare the staggered type schemes on B and D grids in terms of dispersion laws and damping errors.

Chapter content

5.1 Introduction

The dimensionless shallow water equation on the rotating frame is given by

$$\begin{cases} \partial_t h + \nabla \cdot (h\bar{\mathbf{u}}) = 0, & \text{(5.1a)} \\ St\partial_t(h\bar{\mathbf{u}}) + \nabla \cdot (h\bar{\mathbf{u}} \otimes \bar{\mathbf{u}}) + \dfrac{1}{Fr^2}\nabla\left(\dfrac{h^2}{2}\right) = -\dfrac{1}{Fr^2}h\nabla b - \dfrac{1}{Ro}h\bar{\mathbf{u}}^{\perp} & \text{(5.1b)} \end{cases}$$

In System (5.1) unknowns h and $\bar{\mathbf{u}}$ respectively denote the water depth and the velocity of the water column and function $b(x)$ denotes the topography of the considered oceanic basin and is a given function. Dimensionless numbers St, Fr and Ro respectively stand for the Strouhal, the Froude and the Rossby numbers defined by

$$St = \frac{L}{UT}, \quad Fr = \frac{U}{\sqrt{gH}}, \quad Ro = \frac{U}{\Omega L}$$

where the parameter g and Ω denote the gravity coefficient and the angular velocity of the Earth. Constants U, H, L and T are some characteristic velocity, vertical and horizontal lengths and time. In the sequel, we shall focus on cases where

$$Ro = \mathcal{O}(M) \quad \text{and} \quad Fr = \mathcal{O}(M)$$

with M a small parameter. For large scale oceanographic flows, typical values lead to $M \sim 10^{-2}$. Let us now suppose that the topography is flat. For a Strouhal number of order $\mathcal{O}(\frac{1}{M})$ and for Rossby and Froude numbers of order $\mathcal{O}(M)$, the solution of system (5.1) satisfies at the leading order the linear wave equation with Coriolis source term

$$\begin{cases} \partial_t r + a_\star \nabla . \boldsymbol{u} = 0 \\ \partial_t \boldsymbol{u} + a_\star \nabla r = -\omega \boldsymbol{u}^{\perp} \end{cases} \tag{5.2}$$

where $\boldsymbol{u} = (u, v)^T$, and $\boldsymbol{u}^{\perp} = (-v, u)^T$. The parameters $a_\star$ and ω are constants of order one, respectively related to the wave velocity and to the rotating velocity. The stationary state corresponding to Equation (5.2) is the geostrophic equilibrium which is given by

$$a_\star \nabla r = -\omega \boldsymbol{u}^{\perp}. \tag{5.3}$$

One of the common numerical strategies that can be applied to the linear wave equation (5.2) is the collocated leapfrog scheme (the so called A-grid model). However, this scheme suffers from the problem called "checkerboard mode" that has a pressure state alternating between two constants (see [2] for more details). Hence, it is essential to turn to staggered schemes. Since there is a variety of ways to distribute the variables in the two dimensional case, we have various staggered schemes associated to Arakawa's grids introduced in [40]. A lot of research articles focus on the analysis of the behavior of the dispersion relation for the numerical discretization of (5.2) on Arakawa's grids, e.g. [41, 54]. The main purpose of the present work is to propose staggered schemes with appropriate diffusion terms that avoid oscillating solutions when the initial solution is discontinuous, while at the same time the obtained scheme can capture well the geostrophic equilibrium (5.3).

The outline of this work is the following. In Section 5.2, we perform the analysis for the semi-discrete staggered schemes. In particular, we construct some discrete operators which fulfill mimetic properties. Moreover, we analyze the discrete kernel associated to the semi-discrete staggered scheme to point out the wrong behavior of the classical scheme and we follow the work performed on collocated grids [53] to adapt the Apparent Topography, Low Froude and

Divergence Penalisation methods to staggered grids. This provides schemes that possess discrete steady states which are consistent approximations of the continuous kernel (5.3). On the other hand, based on a Fourier analysis, we exhibit the dispersion relations and damping errors of the semi-discrete (in space) schemes. Next, in Section 5.3, we take into account the time discretization to present fully discrete staggered schemes and we prove some CFL conditions which ensure that the proposed schemes are stable. Besides, we investigate the orthogonality preserving property at the fully discrete level. At last, the analysis is followed by numerical tests in Section 5.4 and the study is completed by some concluding remarks.

5.2 Analysis of the semi-discrete staggered schemes

5.2.1 The semi-discrete staggered scheme on B grids

B grids have velocities discretized at the cell centers, while discrete pressures are located at the vertices of the grid.

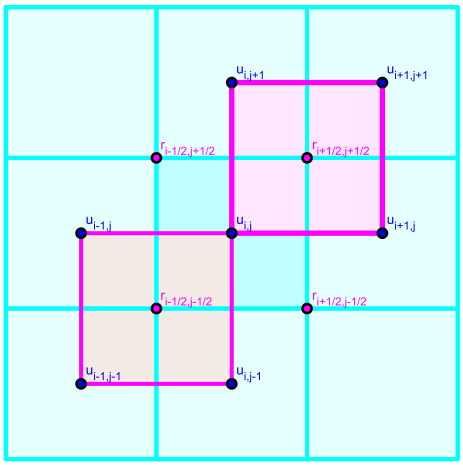


Figure 5.1: *B grid.*

Discrete Operators for B grid type schemes

We first define the discrete version of the gradient and divergence operator. Let $u_h = (u_{i,j})$ and $v_h = (v_{i,j})$ be in $\mathbb{R}^N$ where $N = N_x \times N_y$. We define the discrete divergence $\nabla_h \cdot (\boldsymbol{u}_h)$ by the following formula

$$\begin{aligned}\forall i \in [1, N_x], \forall j \in [1, N_y] : \nabla_h \cdot (\boldsymbol{u}_h)_{i+1/2,j+1/2} = &\frac{(u_{i+1,j+1} + u_{i+1,j}) - (u_{i,j+1} + u_{i,j})}{2\Delta x} \\ &+ \frac{(v_{i+1,j+1} + v_{i,j+1}) - (v_{i+1,j} + v_{i,j})}{2\Delta y}.\end{aligned} \tag{5.4}$$

Moreover, we can define the discrete curl of the vector field by

$$\begin{aligned}\nabla_h \times (\boldsymbol{u}_h)_{i+1/2,j+1/2} &= -\nabla_h \cdot (\boldsymbol{u}_h^{\perp})_{i+1/2,j+1/2} \\ &= \frac{(v_{i+1,j+1} + v_{i+1,j}) - (v_{i,j+1} + v_{i,j})}{2\Delta x} - \frac{(u_{i+1,j+1} + u_{i,j+1}) - (u_{i+1,j} + u_{i,j})}{2\Delta y}.\end{aligned} \tag{5.5}$$

Let $r_h = (r_{i+1/2,j+1/2})$ be a scalar function defined on the dual cells (see Figure 5.1). We can define the discrete gradient by using the following formula

$$\forall i \in [1, N_x], \forall j \in [1, N_y] : \nabla_h(r_h)_{i,j} = \frac{1}{2}\begin{pmatrix} \frac{r_{i+1/2,j+1/2} - r_{i-1/2,j+1/2}}{\Delta x} \\ \frac{r_{i+1/2,j+1/2} - r_{i+1/2,j-1/2}}{\Delta y} \end{pmatrix} + \frac{1}{2}\begin{pmatrix} \frac{r_{i+1/2,j-1/2} - r_{i-1/2,j-1/2}}{\Delta x} \\ \frac{r_{i-1/2,j+1/2} - r_{i-1/2,j-1/2}}{\Delta y} \end{pmatrix}. \tag{5.6}$$

Let us denote the area of the primary cell $\Delta_{i,j} = \Delta x \Delta y$ and the dual cell $\Delta_{i+1/2,j+1/2} = \Delta x \Delta y$. Then, we can define the discrete scalar product between $q_h^1 = (r_h^1, u_h^1, v_h^1)$ and $q_h^2 = (r_h^2, u_h^2, v_h^2)$ by

$$\begin{aligned} \langle q_h^1, q_h^2 \rangle &= \langle r_h^1, r_h^2 \rangle_{\mathcal{D}} + \langle \boldsymbol{u}_h^1, \boldsymbol{u}_h^2 \rangle_{\mathcal{P}} \\ &= \sum_{i,j} \Delta_{i+1/2,j+1/2} r^1_{i+1/2,j+1/2} r^2_{i+1/2,j+1/2} + \sum_{i,j} \Delta_{i,j} \left(u^1_{i,j} u^2_{i,j} + v^1_{i,j} v^2_{i,j} \right). \end{aligned} \tag{5.7}$$

With the help of the discrete operators, the semi-discrete staggered scheme applied to the linear wave equation with Coriolis source term can be written as

$$\begin{cases} \frac{d}{dt} r_{i+1/2,j+1/2}(t) + a_\star \nabla_h \cdot (\boldsymbol{u}_h)_{i+1/2,j+1/2} - \nu_r \nabla_h \cdot \left[\nabla_h(r_h) + \frac{\omega}{a_\star} \boldsymbol{u}_h^\perp \right]_{i+1/2,j+1/2} = 0 \\ \frac{d}{dt} \boldsymbol{u}_{i,j}(t) + a_\star \nabla_h(r_h)_{i,j} - \nu_u \nabla_h [\nabla_h \cdot (\boldsymbol{u}_h)]_{i,j} = -\omega \boldsymbol{u}^\perp_{i,j}. \end{cases} \tag{5.8}$$

where $\nu_r = \frac{\kappa_r^x a_\star \Delta x}{2} = \frac{\kappa_r^y a_\star \Delta y}{2}$ and $\nu_u = \frac{\kappa_u a_\star \Delta x}{2} = \frac{\kappa_v a_\star \Delta y}{2}$ represent the parameters of the diffusion terms.

We also note that the Low Froude – Divergence Penalization (LF-DP) scheme corresponds to $\nu_r = 0, \nu_u > 0$, the Apparent Topography – Low Froude (AT-LF) scheme corresponds to $\nu_r > 0, \nu_u = 0$ and the Apparent Topography – Divergence Penalization (AT-DP) scheme has $\nu_r, \nu_u > 0$.

Remark 5.1. *It is worth pointing out that in the collocated schemes, all the space derivatives are taken over the distance $2h$ where the space step $h = \Delta x$ (resp. $h = \Delta y$) in the x (resp.y) direction. However, the distance between adjacent grid nodes is only h. Therefore, it is reasonable to use staggered schemes which allow us to perform the derivatives over the distance h. As a result, staggered schemes are more compact than the collocated schemes proposed in [53].*

Properties of the discrete operators

Proposition 5.1. *With the discrete divergence, curl, gradient operators and the discrete scalar product defined respectively by (5.4), (5.5), (5.6) and (5.7), we have the following properties for the semi-discrete staggered scheme (5.8):*

i. Energy conservation for the pressure gradient force (discrete integration by part)

$$\langle \nabla_h \cdot (\boldsymbol{u}_h), r_h \rangle_{\mathcal{D}} = -\langle \nabla_h(r_h), \boldsymbol{u}_h \rangle_{\mathcal{P}} \tag{5.9}$$

ii. Energy conservation for the Coriolis force

$$\langle \boldsymbol{u}_h^\perp, \boldsymbol{u}_h \rangle_{\mathcal{P}} = 0. \tag{5.10}$$

iii. No vorticity production for the pressure gradient force

$$\nabla_h \times (\nabla_h(r_h)) = 0. \tag{5.11}$$

Proof. By using periodic boundary condition, we obtain

$$\begin{aligned}
\langle \nabla_h \cdot (\boldsymbol{u}_h), r_h \rangle_{\mathcal{D}} &= \sum_{i,j} \Delta_{i+1/2,j+1/2} \nabla_h \cdot (\boldsymbol{u}_h)_{i+1/2,j+1/2} r_{i+1/2,j+1/2} \\
&= \sum_{i,j} \Delta_{i+1/2,j+1/2} \frac{(u_{i+1,j+1} + u_{i+1,j}) - (u_{i,j+1} + u_{i,j})}{2\Delta x} r_{i+1/2,j+1/2} \\
&+ \sum_{i,j} \Delta_{i+1/2,j+1/2} \frac{(v_{i+1,j+1} + v_{i,j+1}) - (v_{i+1,j} + v_{i,j})}{2\Delta y} r_{i+1/2,j+1/2} \\
&= \sum_{i,j} \Delta_{i,j} u_{i,j} \left(\frac{r_{i-1/2,j-1/2} - r_{i+1/2,j-1/2}}{2\Delta x} + \frac{r_{i-1/2,j+1/2} - r_{i+1/2,j+1/2}}{2\Delta x} \right) \\
&+ \sum_{i,j} \Delta_{i,j} v_{i,j} \left(\frac{r_{i-1/2,j-1/2} - r_{i-1/2,j+1/2}}{2\Delta y} + \frac{r_{i+1/2,j-1/2} - r_{i+1/2,j+1/2}}{2\Delta y} \right) \\
&= -\sum_{i,j} \Delta_{i,j} \nabla_h (r_h)_{i,j} \cdot \boldsymbol{u}_{i,j} = -\langle \nabla_h (r_h), \boldsymbol{u}_h \rangle_{\mathcal{P}},
\end{aligned}$$

which proves Point *(i)*.
Point *(ii)* is obvious and we now turn to Point *(iii)*; we get, after some simplifications

$$\begin{aligned}
\nabla_h \times (\nabla_h (r_h))_{i+1/2,j+1/2} &= \frac{1}{2\Delta y} \left[\left(\frac{r_{i+3/2,j+3/2} - r_{i-1/2,j+3/2}}{2\Delta x} \right) - \left(\frac{r_{i+3/2,j-1/2} - r_{i-1/2,j-1/2}}{2\Delta x} \right) \right] \\
&- \frac{1}{2\Delta x} \left[\left(\frac{r_{i+3/2,j+3/2} - r_{i+3/2,j-1/2}}{2\Delta y} \right) - \left(\frac{r_{i-1/2,j+3/2} - r_{i-1/2,j-1/2}}{2\Delta y} \right) \right] \\
&= 0.
\end{aligned}$$

□

Let us emphasize that the new diffusion term on the velocity equation $\nabla_h(\nabla_h \cdot \boldsymbol{u}_h)$ is a crucial point to ensure a discrete vorticity-divergence relation written on the dual mesh for the staggered scheme: If we apply the operator $\nabla_h \times$ to the velocity equation of the staggered scheme (5.8), we obtain

$$\frac{d}{dt} [\nabla_h \times (\boldsymbol{u}_h)]_{i+1/2,j+1/2} + \omega \nabla_h \cdot (\boldsymbol{u}_h)_{i+1/2,j+1/2} = 0. \tag{5.12}$$

This is because we have no vorticity production of the gradient term (see (5.11)). Of course, with the standard diffusion term $(\partial^2_{xx,h} u_h, \partial^2_{yy,h} v_h)^T$, we generally have

$$\nabla_h \times \left[(\partial^2_{xx,h} u_h, \partial^2_{yy,h} v_h)^T \right] \neq 0.$$

As a consequence, we are unable to obtain the vorticity-divergence relation (5.12) with the standard scheme.

Evolution of the discrete energy

Lemma 5.1. *With $\nu_r = 0$ and the discrete energy defined as follows:*

$$E_h(t) = \sum_{i,j} \Delta_{i+1/2,j+1/2} r^2_{i+1/2,j+1/2}(t) + \sum_{i,j} \Delta_{i,j} \left(u^2_{i,j}(t) + v^2_{i,j}(t) \right), \tag{5.13}$$

we have the dissipation of the discrete energy for the LF-DP scheme:

$$\frac{d}{dt}E_h(t) \leq 0.$$

Proof. We take the scalar product of the staggered scheme (5.8) with $q_h = (r_h, u_h, v_h)$ to obtain

$$\frac{1}{2}\frac{d}{dt}E_h(t) + a_\star\langle\nabla_h\cdot(\boldsymbol{u}_h), r_h\rangle_{\mathcal{D}} + a_\star\langle\nabla_h(r_h), \boldsymbol{u}_h\rangle_{\mathcal{P}} + \langle\boldsymbol{u}_h^\perp, \boldsymbol{u}_h\rangle_{\mathcal{P}} - \nu_u\langle\nabla_h[\nabla_h\cdot(\boldsymbol{u}_h)], \boldsymbol{u}_h\rangle_{\mathcal{P}} = 0.$$

Moreover, the discrete integration by part formula (5.9) implies that

$$\langle\nabla_h[\nabla_h\cdot(\boldsymbol{u}_h)], \boldsymbol{u}_h\rangle_{\mathcal{P}} = -\langle\nabla_h\cdot(\boldsymbol{u}_h), \nabla_h\cdot(\boldsymbol{u}_h)\rangle_{\mathcal{D}} \tag{5.14}$$

Therefore, using (5.9), (5.10) and (5.14), we get

$$\frac{d}{dt}E_h(t) = -2\nu_u\|\nabla_h.(\boldsymbol{u}_h)\|^2,$$

which means that the energy of the LF-DP scheme is decreasing with time. □

Discretized steady-states and their orthogonal subspace on B grids

We now define a set of discretized steady-states with staggered variables on B grids by the following expression

$$\mathcal{E}^{\square}_{\omega\neq 0,B} = \left\{q_h = (r_h, \boldsymbol{u}_h) \in \mathbb{R}^{3N} : a_\star\nabla_h(r_h)_{i,j} = -\omega\boldsymbol{u}^\perp_{i,j}\right\} \tag{5.15}$$

which is a consistent discretization of the geostrophic equilibrium (5.3). Then we have the following result

Lemma 5.2. *The orthogonal space of $\mathcal{E}^{\square}_{\omega\neq 0,B}$ is given by*

$$\mathcal{E}^{\square,\perp}_{\omega\neq 0,B} = \left\{q_h = (r_h, \boldsymbol{u}_h) \in \mathbb{R}^{3N} : \quad a_\star\nabla_h\times(\boldsymbol{u}_h)_{i+1/2,j+1/2} = \omega r_{i+1/2,j+1/2}\right\}, \tag{5.16}$$

which implies the following discrete Hodge decomposition: $\mathbb{R}^{3N} = \mathcal{E}^{\square}_{\omega\neq 0,B}\oplus\mathcal{E}^{\square,\perp}_{\omega\neq 0,B}$.

Proof. First of all, we define the set $\mathcal{A}^B_h$ by

$$\mathcal{A}^B_h := \left\{q_h = (r_h, \boldsymbol{u}_h) \in \mathbb{R}^{3N} : \quad a_\star\nabla_h\times(\boldsymbol{u}_h)_{i+1/2,j+1/2} = \omega r_{i+1/2,j+1/2}\right\}.$$

For each $\hat{q}_h \in \mathcal{E}^{\square}_{\omega\neq 0,B}$ and arbitrary $\tilde{q}_h \in \mathbb{R}^{3N}$, we use the discrete integration by part formula (5.9) to obtain

$$\begin{aligned}
\langle\hat{q}_h, \tilde{q}_h\rangle &= \langle\hat{r}_h, \tilde{r}_h\rangle_{\mathcal{D}} + \langle\hat{\boldsymbol{u}}_h, \tilde{\boldsymbol{u}}_h\rangle_{\mathcal{P}} = \langle\hat{r}_h, \tilde{r}_h\rangle_{\mathcal{D}} + \langle\hat{\boldsymbol{u}}^\perp_h, \tilde{\boldsymbol{u}}^\perp_h\rangle_{\mathcal{P}} \\
&= \langle\hat{r}_h, \tilde{r}_h\rangle_{\mathcal{D}} - \frac{a_\star}{\omega}\left\langle\nabla_h(\hat{r}_h), \tilde{\boldsymbol{u}}^\perp_h\right\rangle_{\mathcal{P}} = \langle\hat{r}_h, \tilde{r}_h\rangle_{\mathcal{D}} + \frac{a_\star}{\omega}\left\langle\hat{r}_h, \nabla_h\cdot(\tilde{\boldsymbol{u}}^\perp_h)\right\rangle_{\mathcal{D}} \\
&= \left\langle\hat{r}_h, \tilde{r}_h - \frac{a_\star}{\omega}\nabla_h\times(\tilde{\boldsymbol{u}}^\perp_h)\right\rangle_{\mathcal{D}}.
\end{aligned}$$

Hence, if $\tilde{q}_h \in \mathcal{A}^B_h$, we obviously have $\langle\hat{q}_h, \tilde{q}_h\rangle = 0$ which leads to $\mathcal{A}^B_h \subset \mathcal{E}^{\square,\perp}_{\omega\neq 0,B}$. On the other hand, since $\hat{r}_h$ can be arbitrary in $\mathbb{R}^N$ when $\hat{q}_h \in \mathcal{E}^{\square}_{\omega\neq 0,B}$, then the equality $\left\langle\hat{r}_h, \tilde{r}_h - \frac{a_\star}{\omega}\nabla_h\times(\tilde{\boldsymbol{u}}^\perp_h)\right\rangle_{\mathcal{D}} = 0$ for all $\hat{q}_h \in \mathcal{E}^{\square}_{\omega\neq 0,B}$ implies that $\tilde{r}_h - \frac{a_\star}{\omega}\nabla_h\times(\tilde{\boldsymbol{u}}^\perp_h) = 0$ and thus $\tilde{q}_h \in \mathcal{A}^B_h$. It follows that $\mathcal{E}^{\square,\perp}_{\omega\neq 0,B} \subset \mathcal{A}^B_h$. □

Remark 5.2. *The discrete Hodge decomposition allows us to define the discrete orthogonal projection*

$$\mathbb{P}_h : \left\{ \begin{array}{ccc} \mathbb{R}^{3N} & \longrightarrow & \mathcal{E}^{\square}_{\omega\neq 0,B} \\ q_h & \longmapsto & \hat{q}_h \end{array} \right.$$

and we can construct $\hat{q}_h$ by what follows.

Let $q_h = (r_h, \boldsymbol{u}_h)$ be given in $\mathbb{R}^{3N}$. For all $(\hat{p}_h, \hat{\boldsymbol{v}}_h) \in \mathcal{E}^{\square}_{\omega\neq 0,B}$, using orthogonality, we have

$$\langle \hat{r}_h, \hat{p}_h\rangle_{\mathcal{D}} + \langle \hat{\boldsymbol{u}}_h, \hat{\boldsymbol{v}}_h\rangle_{\mathcal{P}} = \langle r_h, \hat{p}_h\rangle_{\mathcal{D}} + \langle \boldsymbol{u}_h, \hat{\boldsymbol{v}}_h\rangle_{\mathcal{P}}.$$

We then use the definition of the discrete steady-states and the discrete integration by part formula to get

$$\langle \hat{r}_h, \hat{p}_h\rangle_{\mathcal{D}} - \left(\frac{a_\star}{\omega}\right)^2 \langle \nabla_h \cdot [\nabla_h(\hat{r}_h)], \hat{p}_h\rangle_{\mathcal{D}} = \langle r_h, \hat{p}_h\rangle_{\mathcal{D}} - \left(\frac{a_\star}{\omega}\right) \langle \nabla_h \times \boldsymbol{u}_h, \hat{p}_h\rangle_{\mathcal{D}}.$$

As a result, it is possible to find $\hat{r}_h$ by solving the following linear system

$$\hat{r}_{i+1/2,j+1/2} - \left(\frac{a_\star}{\omega}\right)^2 \nabla_h \cdot [\nabla_h(\hat{r}_h)]_{i+1/2,j+1/2} = r_{i+1/2,j+1/2} - \left(\frac{a_\star}{\omega}\right) \nabla_h \times (\boldsymbol{u}_h)_{i+1/2,j+1/2}. \quad (5.17)$$

Then, by the definition of the discrete steady-states, the part of the velocity field in $\mathcal{E}^{\square}_{\omega\neq 0,B}$ is given by

$$\hat{\boldsymbol{u}}_{i,j} = \left(\frac{a_\star}{\omega}\right) \nabla_h^{\perp}(\hat{r}_h)_{i,j}.$$

Finally, the orthogonal component is simply given by $\tilde{q}_h = q_h - \hat{q}_h$. Moreover, the linear system (5.17) defines a unique solution by the fact that the matrix of this linear system is an M-matrix.

Well-balanced and orthogonality preserving properties

Definition 5.1. *A semi-discrete scheme is said to be well-balanced if*

$$q_h^0 \in \mathcal{E}^{\square}_{\omega\neq 0,B} \quad \Rightarrow \quad \forall t \geq 0, \quad q_h(t) = q_h^0 \in \mathcal{E}^{\square}_{\omega\neq 0,B}.$$

Definition 5.2. *A semi-discrete scheme is said to be orthogonality preserving if*

$$q_h^0 \in \mathcal{E}^{\square,\perp}_{\omega\neq 0,B} \quad \Rightarrow \quad \forall t \geq 0, \quad q_h(t) \in \mathcal{E}^{\square,\perp}_{\omega\neq 0,B}.$$

Lemma 5.3. *For the semi-discrete staggered type scheme (5.8), we have:*

i. It is a well balanced scheme which can capture the discrete steady state (5.15).

ii. It is an orthogonality preserving scheme if $\nu_r = 0$ (LF-DP scheme).

Proof. By the fact that we have no vorticity production of the gradients (see (5.11)), with discrete steady-states, there holds

$$\nabla_h \cdot (\boldsymbol{u}_h)_{i+1/2,j+1/2} = \left(\frac{a_\star}{\omega}\right) \nabla_h \cdot \left[\nabla_h^{\perp}(r_h)\right]_{i+1/2,j+1/2} = -\left(\frac{a_\star}{\omega}\right) \nabla_h \times [\nabla_h(r_h)]_{i+1/2,j+1/2} = 0 \quad (5.18)$$

which implies the well balanced property of the semi-discrete staggered scheme (5.8). This proves Point *(i)*.

We now turn to the second point. By taking the discrete scalar product of the semi-discrete staggered scheme with the stationary state $\hat{q}_h \in \mathcal{E}^{\square}_{\omega\neq 0,B}$, we obtain

$$\left\langle \frac{d}{dt}q_h(t), \hat{q}_h \right\rangle = -a_\star \langle \nabla_h \cdot (\boldsymbol{u}_h), \hat{r}_h \rangle_{\mathcal{D}} + \nu_r \left\langle \nabla_h \cdot \left[\nabla_h(r_h) + \frac{\omega}{a_\star} \boldsymbol{u}_h^{\perp} \right], \hat{r}_h \right\rangle_{\mathcal{D}} - a_\star \langle \nabla_h(r_h), \hat{\boldsymbol{u}}_h \rangle_{\mathcal{P}} + \nu_u \langle \nabla_h[\nabla_h \cdot (\boldsymbol{u}_h)], \hat{\boldsymbol{u}}_h \rangle_{\mathcal{P}} - \omega \langle \boldsymbol{u}_h^{\perp}, \hat{\boldsymbol{u}}_h \rangle_{\mathcal{P}}.$$

By using the discrete integration by part formula and (5.18), we have

$$\langle \nabla_h(r_h), \hat{\boldsymbol{u}}_h \rangle_{\mathcal{P}} = -\langle r_h, \nabla_h \cdot (\hat{\boldsymbol{u}}_h) \rangle_{\mathcal{D}} = 0 \quad , \quad \langle \nabla_h[\nabla_h \cdot (\boldsymbol{u}_h)], \hat{\boldsymbol{u}}_h \rangle_{\mathcal{P}} = -\langle \nabla_h \cdot (\boldsymbol{u}_h), \nabla_h \cdot (\hat{\boldsymbol{u}}_h) \rangle_{\mathcal{D}} = 0$$

and, using the discrete integration by part formula and the fact that $\hat{q}_h \in \mathcal{E}^{\square}_{\omega\neq 0,B}$, we get

$$-a_\star \langle \nabla_h \cdot (\boldsymbol{u}_h), \hat{r}_h \rangle_{\mathcal{D}} - \omega \langle \boldsymbol{u}_h^{\perp}, \hat{\boldsymbol{u}}_h \rangle_{\mathcal{P}} = \left\langle a_\star \nabla_h \hat{r}_h + \omega \hat{\boldsymbol{u}}_h^{\perp}, \boldsymbol{u}_h \right\rangle_{\mathcal{P}} = 0.$$

As a result, the condition to ensure the orthogonality preserving property of the semi-discrete staggered scheme is given by

$$\forall \hat{q}_h \in \mathcal{E}^{\square}_{\omega\neq 0,B}, \quad \nu_r \left\langle \nabla_h . \left[\nabla_h(r_h) + \frac{\omega}{a_\star} \boldsymbol{u}_h^{\perp} \right], \hat{r}_h \right\rangle_{\mathcal{D}} = 0.$$

Therefore, the semi-discrete staggered scheme is orthogonality preserving when we have no diffusion on the pressure equation $\nu_r = 0$. □

Remark 5.3. *The orthogonality preserving property of the staggered scheme with $\nu_r = 0$ allows to ensure that there is no exchange of energy between the kernel and orthogonal kernel during the computation process. On the contrary, if the numerical diffusion on the pressure equation does not vanish ($\nu_r \neq 0$), at each time step, the component of the numerical solution in the orthogonal of the kernel not only damps out, but also partly moves into the kernel. As a result, the kernel part of the numerical solution may be changed at each time step, until the numerical scheme tends to the steady state.*

5.2.2 The semi-discrete staggered scheme on D grids

D grids have the tangential components of the velocities discretized at the midpoints of the edges, while discrete pressures are located at the cell centers.

Discrete operators

In order to design numerical schemes on D grids, we shall first define the discrete versions of some operators. Let $r_h = (r_{i,j})$, $u_h = (u_{i,j+1/2})$ and $v_h = (v_{i+1/2,j})$ be in $\mathbb{R}^N$ where $N = N_x \times N_y$ (see Figure 5.2). We first denote f_h as an averaging operator that uses 4 points around the location where the average is computed. We define

$$f_h(r_h)_{i+1/2,j+1/2} = \frac{r_{i,j} + r_{i+1,j} + r_{i,j+1} + r_{i+1,j+1}}{4},$$

$$f_h(u_h)_{i+1/2,j} = \frac{u_{i,j-1/2} + u_{i+1,j-1/2} + u_{i,j+1/2} + u_{i+1,j+1/2}}{4}$$

and

$$f_h(v_h)_{i,j+1/2} = \frac{v_{i-1/2,j} + v_{i-1/2,j+1} + v_{i+1/2,j} + v_{i+1/2,j+1}}{4}.$$

Moreover, we also need the discrete version of the divergence at the cell corner:

$$\nabla_h^{\mathrm{v}} \cdot (\mathbf{u}_h)_{i+1/2,j+1/2} = \frac{u_{i+1,j+1/2} - u_{i,j+1/2}}{\Delta x} + \frac{v_{i+1/2,j+1} - v_{i+1/2,j}}{\Delta y}.$$

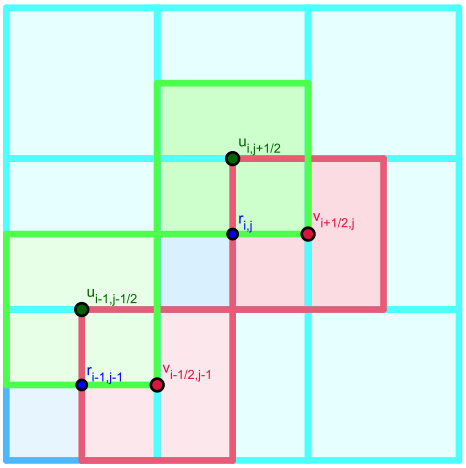


Figure 5.2: *D grid.*

Let $\phi_h = (\phi_{i+1/2,j+1/2})$ be a discrete function defined by its values at the vertices of the cells, we then define the discrete version of some differential operators

$$\partial_{x,h}(\phi_h)_{i,j+1/2} = \frac{\phi_{i+1/2,j+1/2} - \phi_{i-1/2,j+1/2}}{\Delta x} \quad \text{and} \quad \partial_{y,h}(\phi_h)_{i+1/2,j} = \frac{\phi_{i+1/2,j+1/2} - \phi_{i+1/2,j-1/2}}{\Delta y}.$$

Let also $r_h = (r_{i,j})$ be a discrete function defined by its values at the cell centers, we define another discrete version of the same differential operators:

$$\partial_{x,h}(r_h)_{i+1/2,j} = \frac{r_{i+1,j} - r_{i,j}}{\Delta x} \quad \text{and} \quad \partial_{y,h}(r_h)_{i,j+1/2} = \frac{r_{i,j+1} - r_{i,j}}{\Delta y}.$$

Now, for a discrete vector field $\boldsymbol{\varphi}_h = (\varphi_h, \psi_h)$ where $\varphi_h = (\varphi_{i+1/2,j})$ and $\psi_h = (\psi_{i,j+1/2})$, the discrete divergence at the cell center is defined by

$$\nabla_h^{\mathrm{c}} \cdot (\boldsymbol{\varphi}_h)_{i,j} = \frac{\varphi_{i+1/2,j} - \varphi_{i-1/2,j}}{\Delta x} + \frac{\psi_{i,j+1/2} - \psi_{i,j-1/2}}{\Delta y}.$$

Next, we also define the following discrete scalar products

$$\langle r_h^1, r_h^2 \rangle_{\mathcal{P}_r} = \sum_{i,j} \Delta x \Delta y r_{i,j}^1 r_{i,j}^2, \quad \langle u_h^1, u_h^2 \rangle_{\mathcal{D}_u} = \sum_{i,j} \Delta x \Delta y u_{i,j+1/2}^1 u_{i,j+1/2}^2,$$

$$\langle v_h^1, v_h^2 \rangle_{\mathcal{D}_v} = \sum_{i,j} \Delta x \Delta y v_{i+1/2,j}^1 v_{i+1/2,j}^2, \quad \langle \phi_h^1, \phi_h^2 \rangle_{\mathcal{D}_\phi} = \sum_{i,j} \Delta x \Delta y \phi_{i+1/2,j+1/2}^1 \phi_{i+1/2,j+1/2}^2,$$

and

$$\langle q_h^1, q_h^2 \rangle = \langle r_h^1, r_h^2 \rangle_{\mathcal{P}_r} + \langle u_h^1, u_h^2 \rangle_{\mathcal{D}_u} + \langle v_h^1, v_h^2 \rangle_{\mathcal{D}_v}.$$

With the above discrete operators, the semi-discrete staggered type schemes on D grids can be written as

$$\begin{cases} \frac{d}{dt} r_{i,j}(t) + a_\star \nabla_h^{\mathrm{c}} \cdot [f_h(u_h), f_h(v_h)]_{i,j} - \nu_r \nabla_h^{\mathrm{c}} \cdot (\partial_{x,h} r_h - \frac{\omega}{a_\star} v_h, \partial_{y,h} r_h + \frac{\omega}{a_\star} u_h)_{i,j} = 0 \\ \frac{d}{dt} u_{i,j+1/2}(t) + a_\star \partial_{x,h} [f_h(r_h)]_{i,j+1/2} - \nu_u \partial_{x,h} [\nabla_h^{\mathrm{v}} \cdot (\mathbf{u}_h)]_{i,j+1/2} = \omega f_h(v_h)_{i,j+1/2} \\ \frac{d}{dt} v_{i+1/2,j}(t) + a_\star \partial_{y,h} [f_h(r_h)]_{i+1/2,j} - \nu_u \partial_{y,h} [\nabla_h^{\mathrm{v}} \cdot (\mathbf{u}_h)]_{i+1/2,j} = -\omega f_h(u_h)_{i+1/2,j} \end{cases} \tag{5.19}$$

Properties of the discrete operators

Proposition 5.2. *For discrete fields with periodic boundary conditions, we have*

i. The discrete integration by part formula

$$\langle \nabla_h^v \cdot (\mathbf{u}_h), \phi_h \rangle_{\mathcal{D}_\phi} = -\langle \partial_{x,h}(\phi_h), u_h \rangle_{\mathcal{D}_u} - \langle \partial_{y,h}(\phi_h), v_h \rangle_{\mathcal{D}_v}. \tag{5.20}$$

ii. Energy conservation of the Coriolis force

$$\langle f_h(v_h), u_h \rangle_{\mathcal{D}_u} - \langle f_h(u_h), v_h \rangle_{\mathcal{D}_v} = 0. \tag{5.21}$$

iii. No vorticity production for pressure gradient term

$$\nabla_h^c \times [\partial_{x,h}\phi_h, \partial_{y,h}\phi_h] = -\nabla_h^c \cdot [-\partial_{y,h}\phi_h, \partial_{x,h}\phi_h] = 0.$$

Proof. These properties are proved by direct computations. □

Like with the staggered scheme on B grids, we also obtain a vorticity-divergence relation with the staggered scheme on D grids. In particular, we have

$$\frac{d}{dt}\nabla_h^c \times [u_h, v_h]_{i,j} + \omega \nabla_h^c \cdot [f_h(u_h), f_h(v_h)]_{i,j} = 0.$$

Let us emphasize that this relation on D grids is defined at the cell centers.

Evolution of the discrete energy

Lemma 5.4. *With $\nu_r = 0$ and the discrete energy defined by the following expression*

$$E_h^D(t) = \sum_{i,j} \Delta x \Delta y \left[r_{i,j}^2(t) + u_{i,j+1/2}^2(t) + v_{i+1/2,j}^2(t) \right],$$

we have the dissipation of the discrete energy of the LF-DP staggered scheme on D-grids

$$\frac{d}{dt} E_h^D(t) \leq 0.$$

Proof. By taking the discrete scalar product of (5.19) with $q_h = (r_h, u_h, v_h)$, we obtain

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}E_h^D(t) = &-a_\star \left(\langle \nabla_h^c \cdot [f_h(u_h), f_h(v_h)], r_h \rangle_{\mathcal{P}_r} + \langle \partial_{x,h}[f_h(r_h)], u_h \rangle_{\mathcal{D}_u} + \langle \partial_{y,h}[f_h(r_h)], v_h \rangle_{\mathcal{D}_v} \right) \\ &+ \omega \langle f_h(v_h), u_h \rangle_{\mathcal{D}_u} - \omega \langle f_h(u_h), v_h \rangle_{\mathcal{D}_v} + \nu_u \langle \partial_{x,h}[\nabla_h^v \cdot (\mathbf{u}_h)], u_h \rangle_{\mathcal{D}_u} + \nu_u \langle \partial_{y,h}[\nabla_h^v \cdot (\mathbf{u}_h)], v_h \rangle_{\mathcal{D}_v}. \end{aligned} \tag{5.22}$$

Using periodic boundary conditions, it may be proved that

$$\langle \nabla_h^c \cdot [f_h(u_h), f_h(v_h)], r_h \rangle_{\mathcal{P}_r} = \langle \nabla_h^v \cdot (\mathbf{u}_h), f_h(r_h) \rangle_{\mathcal{D}_\phi}.$$

Hence, we apply the discrete integration by part formula (5.20) for $\phi_h = f_h(r_h)$ and then for $\phi_h = \nabla_h^v \cdot (\mathbf{u}_h)$ to respectively obtain

$$\langle \nabla_h^c \cdot [f_h(u_h), f_h(v_h)], r_h \rangle_{\mathcal{P}_r} = -\langle \partial_{x,h}[f_h(r_h)], u_h \rangle_{\mathcal{D}_u} - \langle \partial_{y,h}[f_h(r_h)], v_h \rangle_{\mathcal{D}_v} \tag{5.23}$$

and

$$\nu_u \langle \partial_{x,h}[\nabla_h^v \cdot (\mathbf{u}_h)], u_h \rangle_{\mathcal{D}_u} + \nu_u \langle \partial_{y,h}[\nabla_h^v \cdot (\mathbf{u}_h)], v_h \rangle_{\mathcal{D}_v} = -\nu_u \| \nabla_h^v \cdot (\mathbf{u}_h) \|_{\mathcal{D}_\phi}^2 \tag{5.24}$$

Therefore, from (5.22), (5.23), (5.21) and (5.24), we conclude that

$$\frac{d}{dt}E_h^D(t) = -2\nu_u \|\nabla_h^{\mathrm{v}} \cdot (\mathbf{u}_h)\|^2_{\mathcal{D}_\phi} \leq 0.$$

□

Discretized steady-states and their orthogonal subspace on D grids

We now define a set of discretized stead-states with staggered variables on D-grids by

$$\mathcal{E}^{\square}_{\omega\neq 0,D} = \left\{ \hat{q}_h = (\hat{r}_h, \hat{u}_h, \hat{v}_h) \in \mathbb{R}^{3N} \;\middle|\; a_\star \begin{pmatrix} \frac{\hat{r}_{i+1,j}-\hat{r}_{i,j}}{\Delta x} \\ \frac{\hat{r}_{i,j+1}-\hat{r}_{i,j}}{\Delta y} \end{pmatrix} = -\omega \begin{pmatrix} -\hat{v}_{i+1/2,j} \\ \hat{u}_{i,j+1/2} \end{pmatrix} \right\}, \tag{5.25}$$

which is a consistent discretization of the geostrophic equilibrium (5.3). Then we have the following result

Lemma 5.5. *The orthogonal space of $\mathcal{E}^{\square}_{\omega\neq 0,D}$ is given by*

$$\mathcal{E}^{\square,\perp}_{\omega\neq 0,D} = \left\{ \tilde{q}_h = (\tilde{r}_h, \tilde{u}_h, \tilde{v}_h) \in \mathbb{R}^{3N} \;\middle|\; a_\star \left(\frac{\tilde{v}_{i+1/2,j} - \tilde{v}_{i-1/2,j}}{\Delta x} - \frac{\tilde{u}_{i,j+1/2} - u_{i,j-1/2}}{\Delta y} \right) = \omega \tilde{r}_{i,j} \right\}, \tag{5.26}$$

which implies the following discrete Hodge decomposition $\mathbb{R}^{3N} = \mathcal{E}^{\square}_{\omega\neq 0,D} \oplus \mathcal{E}^{\square,\perp}_{\omega\neq 0,D}$.

Proof. For all $\hat{q}_h = (\hat{r}_h, \hat{u}_h, \hat{v}_h) \in \mathcal{E}^{\square}_{\omega\neq 0,D}$ and an arbitrary $\tilde{q}_h = (\tilde{r}_h, \tilde{u}_h, \tilde{v}_h) \in \mathbb{R}^{3N}$, by using periodic boundary condition, we obtain

$$\begin{aligned}
\langle \hat{q}_h, \tilde{q}_h \rangle &= \sum_{i,j} \Delta x \Delta y \left[\hat{r}_{i,j}\tilde{r}_{i,j} - \frac{a_\star}{\omega}\left(\frac{\hat{r}_{i,j+1}-\hat{r}_{i,j}}{\Delta y}\right)\tilde{u}_{i,j+1/2} + \frac{a_\star}{\omega}\left(\frac{\hat{r}_{i+1,j}-\hat{r}_{i,j}}{\Delta x}\right)\tilde{v}_{i+1/2,j} \right] \\
&= \sum_{i,j} \Delta x \Delta y \hat{r}_{i,j} \left[\tilde{r}_{i,j} - \frac{a_\star}{\omega}\left(\frac{\tilde{v}_{i+1/2,j} - \tilde{v}_{i-1/2,j}}{\Delta x}\right) + \frac{a_\star}{\omega}\left(\frac{\tilde{u}_{i,j+1/2} - \tilde{u}_{i,j-1/2}}{\Delta y}\right) \right].
\end{aligned}$$

This equation implies that the subspace which is in the right-hand side of (5.26) is included in the orthogonal of $\mathcal{E}^{\square}_{\omega\neq 0,D}$. On the other hand, since $\hat{r}_h$ can be arbitrary in $\mathcal{E}^{\square}_{\omega\neq 0,D}$, the above equation implies that any element of the orthogonal of $\mathcal{E}^{\square}_{\omega\neq 0,D}$ verifies

$$\tilde{r}_{i,j} - \frac{a_\star}{\omega}\left(\frac{\tilde{v}_{i+1/2,j} - \tilde{v}_{i-1/2,j}}{\Delta x}\right) + \frac{a_\star}{\omega}\left(\frac{\tilde{u}_{i,j+1/2} - \tilde{u}_{i,j-1/2}}{\Delta y}\right).$$

□

Remark 5.4. *With D-grid type schemes, an element $q_h \in \mathbb{R}^{3N}$ can be decomposed into*

$$q_h = \hat{q}_h + \tilde{q}_h \quad \textit{with} \quad \hat{q}_h \in \mathcal{E}^{\square}_{\omega\neq 0,D} \quad \textit{and} \quad \tilde{q}_h \in \mathcal{E}^{\square,\perp}_{\omega\neq 0,D}$$

and this decomposition can be obtained by solving the following linear system

$$\begin{aligned}
\hat{r}_{i,j} - \left(\frac{a_\star}{\omega}\right)^2 \left[\frac{\hat{r}_{i+1,j} - 2\hat{r}_{i,j} + \hat{r}_{i-1,j}}{\Delta x^2} + \frac{\hat{r}_{i,j+1} - 2\hat{r}_{i,j} + \hat{r}_{i,j-1}}{\Delta y^2} \right] = \\
r_{i,j} - \frac{a_\star}{\omega}\left(\frac{v_{i+1/2,j} - v_{i-1/2,j}}{\Delta x} - \frac{u_{i,j+1/2} - v_{i,j-1/2}}{\Delta y} \right).
\end{aligned}$$

We also note that this system has unique solution since the matrix on the left-hand side is an M-matrix.

Well-balanced and orthogonality preserving properties

Lemma 5.6. *For the semi-discrete staggered type scheme on D grids (5.19), we have:*

i. Discretized steady-states (5.25) are steady-states of (5.19),

ii. It is an orthogonality preserving scheme if $\nu_r = 0$.

Proof. Any element of (5.25) verifies

$$\begin{aligned}\nabla_h^{\mathrm{v}} \cdot (\hat{\mathbf{u}}_h)_{i+1/2,j+1/2} &= \frac{\hat{u}_{i+1,j+1/2} - \hat{u}_{i,j+1/2}}{\Delta x} + \frac{\hat{v}_{i+1/2,j+1} - \hat{v}_{i+1/2,j}}{\Delta y} \\ &= \frac{a_\star}{\omega \Delta x \Delta y}\left[-(\hat{r}_{i+1,j+1} - \hat{r}_{i+1,j}) + (\hat{r}_{i,j+1} - \hat{r}_{i,j}) + (\hat{r}_{i+1,j+1} - \hat{r}_{i,j+1}) - (\hat{r}_{i+1,j} - \hat{r}_{i,j})\right] \\ &= 0,\end{aligned}$$

which means that the velocities of (5.25) are divergence free at the cell vertices. Moreover, we also notice that

$$\nabla_h^{\mathrm{c}} \cdot [f_h(\hat{u}_h), f_h(\hat{v}_h)]_{i,j} = f_h(\nabla_h^{\mathrm{v}} \cdot (\hat{\mathbf{u}}_h))_{i,j} = 0.$$

By using the definition of the discrete kernel, we also obtain

$$\omega f_h(\hat{v}_h)_{i,j+1/2} = \left(\frac{\hat{r}_{i+1,j} - \hat{r}_{i-1,j}}{4\Delta x} + \frac{\hat{r}_{i+1,j+1} - \hat{r}_{i-1,j+1}}{4\Delta x}\right) = a_\star \partial_{x,h}[f_h(r_h)]_{i,j+1/2}$$

and

$$\omega f_h(\hat{u}_h)_{i+1/2,j} = -\left(\frac{\hat{r}_{i,j+1} - \hat{r}_{i,j-1}}{4\Delta y} + \frac{\hat{r}_{i+1,j+1} - \hat{r}_{i+1,j-1}}{4\Delta y}\right) = -a_\star \partial_{y,h}[f_h(r_h)]_{i+1/2,j}.$$

On the other hand, with the steady state, we obviously have $\nabla_h^{\mathrm{c}} \cdot (\partial_{x,h} r_h - \frac{\omega}{a_\star} v_h, \partial_{y,h} r_h + \frac{\omega}{a_\star} u_h) = 0$. Therefore, the semi-discrete scheme (5.19) captures well the discrete geostrophic equilibrium (5.25). This proves Point *(i)*.
To investigate the second point, we take the discrete scalar product of (5.19) with $\hat{q}_h \in \mathcal{E}^{\square}_{\omega \neq 0, D}$ to obtain

$$\begin{aligned}\langle \frac{d}{dt} q_h(t), \hat{q}_h \rangle &= -a_\star \langle \nabla_h^{\mathrm{c}} \cdot [f_h(u_h), f_h(v_h)], \hat{r}_h \rangle_{\mathcal{P}_r} - a_\star \langle \partial_{x,h}[f_h(r_h)], \hat{u}_h \rangle_{\mathcal{D}_u} - a_\star \langle \partial_{y,h}[f_h(r_h)], \hat{v}_h \rangle_{\mathcal{D}_v} \\ &\quad + \omega \langle f_h(v_h), \hat{u}_h \rangle_{\mathcal{D}_u} - \omega \langle f_h(u_h), \hat{v}_h \rangle_{\mathcal{D}_v} + \nu_u \langle \partial_{x,h}[\nabla_h^{\mathrm{v}} \cdot (\mathbf{u}_h)], \hat{u}_h \rangle_{\mathcal{D}_u} + \nu_u \langle \partial_{y,h}[\nabla_h^{\mathrm{v}} \cdot (\mathbf{u}_h)], \hat{v}_h \rangle_{\mathcal{D}_v} \\ &\quad + \nu_r \langle \nabla_h^{\mathrm{c}} \cdot (\partial_{x,h} r_h - \frac{\omega}{a_\star} v_h, \partial_{y,h} r_h + \frac{\omega}{a_\star} u_h), \hat{r}_h \rangle_{\mathcal{P}_r}.\end{aligned}$$

By using the discrete integration by part formula and properties of the discrete kernel, we also have

$$\begin{aligned}-a_\star \langle \nabla_h^{\mathrm{c}} \cdot [f_h(u_h), f_h(v_h)], \hat{r}_h \rangle_{\mathcal{P}_r} &= \langle \partial_{x,h}[f_h(\hat{r}_h)], u_h \rangle_{\mathcal{D}_u} + \langle \partial_{y,h}[f_h(\hat{r}_h)], v_h \rangle_{\mathcal{D}_v} \\ &= \langle \omega f_h(\hat{v}_h), u_h \rangle_{\mathcal{D}_u} - \langle \omega f_h(\hat{u}_h), v_h \rangle_{\mathcal{D}_v} \\ &= \omega \langle f_h(u_h), \hat{v}_h \rangle_{\mathcal{D}_v} - \omega \langle f_h(v_h), \hat{u}_h \rangle_{\mathcal{D}_u},\end{aligned}$$

$$\langle \partial_{x,h}[f_h(r_h)], \hat{u}_h \rangle_{\mathcal{D}_u} + \langle \partial_{y,h}[f_h(r_h)], \hat{v}_h \rangle_{\mathcal{D}_v} = -\langle \nabla_h^{\mathrm{v}} \cdot (\hat{\mathbf{u}}_h), f_h(r_h) \rangle_{\mathcal{D}_\phi} = 0$$

and

$$\langle \partial_{x,h}[\nabla_h^{\mathrm{v}} \cdot (\mathbf{u}_h)], \hat{u}_h \rangle_{\mathcal{D}_u} + \langle \partial_{y,h}[\nabla_h^{\mathrm{v}} \cdot (\mathbf{u}_h)], \hat{v}_h \rangle_{\mathcal{D}_v} = -\langle \nabla_h^{\mathrm{v}} \cdot (\mathbf{u}_h), \nabla_h^{\mathrm{v}} \cdot (\hat{\mathbf{u}}_h) \rangle_{\mathcal{D}_\phi} = 0.$$

Therefore, the staggered scheme on the D grid (5.19) is orthogonality preserving if

$$\langle q_h, \hat{q}_h \rangle = \nu_r \langle \nabla_h^{\mathrm{c}} \cdot (\partial_{x,h} r_h - \frac{\omega}{a_\star} v_h, \partial_{y,h} r_h + \frac{\omega}{a_\star} u_h), \hat{r}_h \rangle_{\mathcal{P}_r} = 0 \quad \forall \hat{q}_h \in \mathcal{E}^{\square}_{\omega \neq 0, D},$$

which is the case if $\nu_r = 0$. This leads to Point *(ii)*. □

5.2.3 Behavior of the solutions of the staggered schemes

The discrete Hodge decompositions on B or D grids allow us to define the discrete orthogonal projection

$$\mathbb{P}_h : \left\{ \begin{array}{ccc} \mathbb{R}^{3N} & \longrightarrow & \mathcal{E}^{\square}_{\omega \neq 0} \\ q_h & \longmapsto & \hat{q}_h \end{array} \right. \tag{5.27}$$

Lemma 5.7. *Let $q_{\nu,h}(t)$ be the solution of the semi-discrete scheme (5.8) on B grids or (5.19) on D grids. Then, with $\nu_r = 0$, we obtain*

$$\forall C_1 \in \mathbb{R}^+, \quad \text{if} \quad \|q_h^0 - \mathbb{P}_h(q_h^0)\| = C_1 M \quad \text{then} \quad \|q_{\nu,h}(t) - \mathbb{P}_h(q_h^0)\| \leq C_1 M, \quad \forall t \geq 0,$$

which means that the LF-DP scheme is accurate at low Froude number at any time.

Proof. By linearity, the solution of semi-discrete staggered scheme $q_{\nu,h}(t)$ can be written as

$$q_{\nu,h}(t) = q^a_{\nu,h}(t) + q^b_{\nu,h}(t)$$

where $q^a_{\nu,h}(t)$ and $q^b_{\nu,h}(t)$ are the solutions of (5.8) or (5.19) with the initial condition respectively given by

$$q^a_{\nu,h}(0) = \mathbb{P}_h(q_h^0) \quad \text{and} \quad q^b_{\nu,h}(0) = q_h^0 - \mathbb{P}_h(q_h^0).$$

Then, we have

$$\|q_{\nu,h}(t) - \mathbb{P}_h(q_h^0)\| = \|q^a_{\nu,h}(t) + q^b_{\nu,h}(t) - \mathbb{P}_h(q_h^0)\| \leq \|q^a_{\nu,h}(t) - \mathbb{P}_h(q_h^0)\| + \|q^b_{\nu,h}(t)\|$$

Moreover, when $\nu_r = 0$, the dissipation of the semi-discrete staggered schemes leads to the conclusion that $\|q^b_{\nu,h}(t)\| \leq \|q^b_{\nu,h}(0)\|$. For this reason, the accuracy of the scheme is linked to the behavior of $q^a_{\kappa,h}(t)$. Since the semi-discrete schemes (5.8) or (5.19) are well-balanced schemes, we obviously have $q^a_{\nu,h}(t) = \mathbb{P}_h(q_h^0)$. Therefore, we obtain

$$\forall t \geq 0, \quad \|q_{\nu,h}(t) - \mathbb{P}_h(q_h^0)\| \leq C_1 M.$$

□

Remark 5.5. *Since it is difficult to prove the dissipation of the energy for the semi-discrete scheme (5.8) or (5.19) when $\nu_r \neq 0$, we do not have enough evidence to conclude that the well-balanced schemes based on the Apparent Topography method, like the AT-DP or AT-LF schemes are accurate at low Froude number at any time. However, we can prove the conditional stability of those schemes at the fully discrete level and, from the numerical point of view, the well-balanced schemes with the Apparent Topography method for the diffusion on the pressure equation are still accurate at low Froude number.*

5.2.4 Fourier analysis for the semi-discrete staggered schemes

In this subsection, we perform the Fourier analysis which is a useful tool to analyze the influence of the discrete scheme on some important quantities such as dispersion law, damping error, phase and group velocities. We note that this method was already used to study the behavior of numerical schemes applied to linear wave equation with Coriolis force that use finite differences (e.g. [41, 54]), finite elements ([11]) and finite volumes ([55]).

We now look for the solution of this semi-discrete scheme under the form of discrete Fourier modes

$$r_{i,j}(t) = \varphi_r(t)e^{\boldsymbol{i}(k_x x_i + k_y y_j)}, \quad u_{i,j}(t) = \varphi_u(t)e^{\boldsymbol{i}(k_x x_i + k_y y_j)} \quad \text{and} \quad v_{i,j}(t) = \varphi_v(t)e^{\boldsymbol{i}(k_x x_i + k_y y_j)}. \tag{5.28}$$

Substituting discrete Fourier modes (5.28) into the semi-discrete scheme (5.8) or (5.19), we obtain the following linear system of differential equation

$$\begin{pmatrix} \varphi_r'(t) \\ \varphi_u'(t) \\ \varphi_v'(t) \end{pmatrix} + \begin{pmatrix} \nu_r\left(\frac{\alpha^2}{\Delta x^2}+\frac{\beta^2}{\Delta y^2}\right) & ia_\star\eta\frac{\alpha}{\Delta x} - i\nu_r\frac{\omega}{a_\star}\frac{\beta}{\Delta y} & ia_\star\eta\frac{\beta}{\Delta y} + i\nu_r\frac{\omega}{a_\star}\frac{\alpha}{\Delta x} \\ ia_\star\eta\frac{\alpha}{\Delta x} & \nu_u\frac{\alpha^2}{\Delta x^2} & \nu_u\frac{\alpha}{\Delta x}\frac{\beta}{\Delta y} - \omega\eta \\ ia_\star\eta\frac{\beta}{\Delta y} & \nu_u\frac{\alpha}{\Delta x}\frac{\beta}{\Delta y} + \omega\eta & \nu_u\frac{\beta^2}{\Delta y^2} \end{pmatrix} \begin{pmatrix} \varphi_r(t) \\ \varphi_u(t) \\ \varphi_v(t) \end{pmatrix} = 0 \tag{5.29}$$

where parameters α, β an η are specified in Table 5.1 depending on the scheme under study.

Staggered type scheme	α	β	η
B grid	$2\sin(\frac{k_x\Delta x}{2})\cos(\frac{k_y\Delta y}{2})$	$2\sin(\frac{k_y\Delta y}{2})\cos(\frac{k_x\Delta x}{2})$	1
D grid	$2\sin(\frac{k_x\Delta x}{2})$	$2\sin(\frac{k_y\Delta y}{2})$	$\cos(\frac{k_x\Delta x}{2})\cos(\frac{k_y\Delta y}{2})$

Table 5.1: *Parameters α, β an η in the Fourier analysis of the semi-discrete staggered schemes.*

We shall denote the amplification matrix of (5.29) by $\mathcal{A}(\nu, \Delta x, \Delta y)$. One eigenvalue of this amplification matrix is $\lambda_1 = 0$ corresponding to the stationary state and the other eigenvalues corresponding to the inertia-gravity modes are given by

$$\lambda = \frac{\nu_r+\nu_u}{2}\left(\frac{\alpha^2}{\Delta x^2}+\frac{\beta^2}{\Delta y^2}\right) \pm i\sqrt{\omega^2\eta^2 + a_\star^2\eta^2\left(\frac{\alpha^2}{\Delta x^2}+\frac{\beta^2}{\Delta y^2}\right) - \left(\frac{\nu_r-\nu_u}{2}\right)^2\left(\frac{\alpha^2}{\Delta x^2}+\frac{\beta^2}{\Delta y^2}\right)^2}.$$

The real part $\Re(\lambda)$ of the eigenvalues indicates the damping rate of the Fourier modes and the imaginary part $\Im(\lambda)$ represents the propagation. Moreover, the quantity $\frac{\Im(\lambda)}{\omega}$ is the numerical dispersion law of the scheme. In case $\nu_r = \nu_u$, the dispersion law of the staggered scheme reduces to

$$\frac{\Im(\lambda)}{\omega} = \sqrt{R_d^2\eta^2\left(\frac{\alpha^2}{\Delta x^2}+\frac{\beta^2}{\Delta y^2}\right)+\eta^2},$$

where $Rd = \frac{a_\star}{\omega}$ stands for the Rossby deformation radius. Let us note that in this case, the numerical dispersion law only depends on parameters $k_x\Delta x$, $k_y\Delta y$, $\frac{R_d}{\Delta x}$, and $\frac{R_d}{\Delta y}$.
Let us denote h be the grid size assumed to the same in x and y direction; various choices of the ratio $\frac{Rd}{h}$ are discussed in this section.

Figure 5.3a indicates that the dispersion relation of the exact model is a monotone function and we do not recover this property at the discrete level for the staggered schemes, as shown in Figures 5.3b, 5.3c and 5.3d. However, we can observe that the dispersion relation of the

well-balanced staggered schemes is monotonic in the region of interest, which is the low frequency region. Since the energy transfers a lot in this region, i.e. for waves with long wavelengths, it is a strong requirement for the numerical scheme to possess as many good properties as possible in this important region. In the other regions, the dispersion law of the numerical scheme is usually not the same as that of the continuous model, and we therefore need a strong damping effect by the numerical viscosity to ensure the that the waves move in correct direction. Moreover, on B grids, the dispersion law of the AT-DP scheme is better than that of the LF-DP scheme. We can also notice that the dispersion law of the B grid scheme is more accurate than that of the scheme on D grid in the region of low frequency in x and high frequency in y direction or vice versa. One explanation for this result is that on the B grid scheme, it is easy to evaluate the Coriolis force without averaging since the velocities u and v are located at the same points while we need the average for this source term on D grids.

In consideration of the damping error, the damping rate of the AT-DP is twice larger than that of the LF-DP scheme on both B and D grids. Figure 5.4 shows that unlike the AT-DP scheme on B grids, the AT-DP scheme on D grids has a strong damping rate in the region of high frequencies, i.e for short wavelengths. This is one evidence that we may expect fewer oscillations for the waves with short wavelengths with the staggered type schemes on D grids.

Figure 5.5 and 5.6 show the dispersion law of AT-DP scheme on B and D grids with different vales of the ratio between the Rossby deformation radius Rd and the space step (h). These figures indicate that the dispersion law of the AT-DP scheme on B grids has the same shape for both resolved ($\frac{Rd}{h} = 2$) and under-resolved cases ($\frac{Rd}{h} = \frac{1}{2}$), while on D grids, the dispersion law is better in the resolved case.

5.3 Analysis of fully discrete staggered schemes

We now introduce two new parameters θ_1 and θ_2 involved in the time discretization of the Coriolis source term. For the sake of simplicity, we denote

$$\boldsymbol{u}^{\perp,\theta} = \begin{pmatrix} -\theta_1 v^n - (1-\theta_1) v^{n+1} \\ \theta_2 u^n + (1-\theta_2) u^{n+1} \end{pmatrix}.$$

We now propose the following time discretizations for the staggered scheme on B grids

$$\begin{cases} r^{n+1}_{i+1/2,j+1/2} = r^n_{i+1/2,j+1/2} - a_\star \Delta t \nabla_h . (\boldsymbol{u}^n_h)_{i+1/2,j+1/2} + \nu_r \Delta t \nabla_h . \left[\nabla_h (r^n_h) + \frac{\omega}{a_\star} \boldsymbol{u}^{\perp,n}_h \right]_{i+1/2,j+1/2} \\ \boldsymbol{u}^{n+1}_{i,j} = \boldsymbol{u}^n_{i,j} - a_\star \Delta t \nabla_h (r^n_h)_{i,j} + \nu_u \Delta t \nabla_h [\nabla_h . (\boldsymbol{u}^n_h)]_{i,j} - \omega \Delta t \boldsymbol{u}^{\perp,\theta}_{i,j}. \end{cases} \tag{5.30}$$

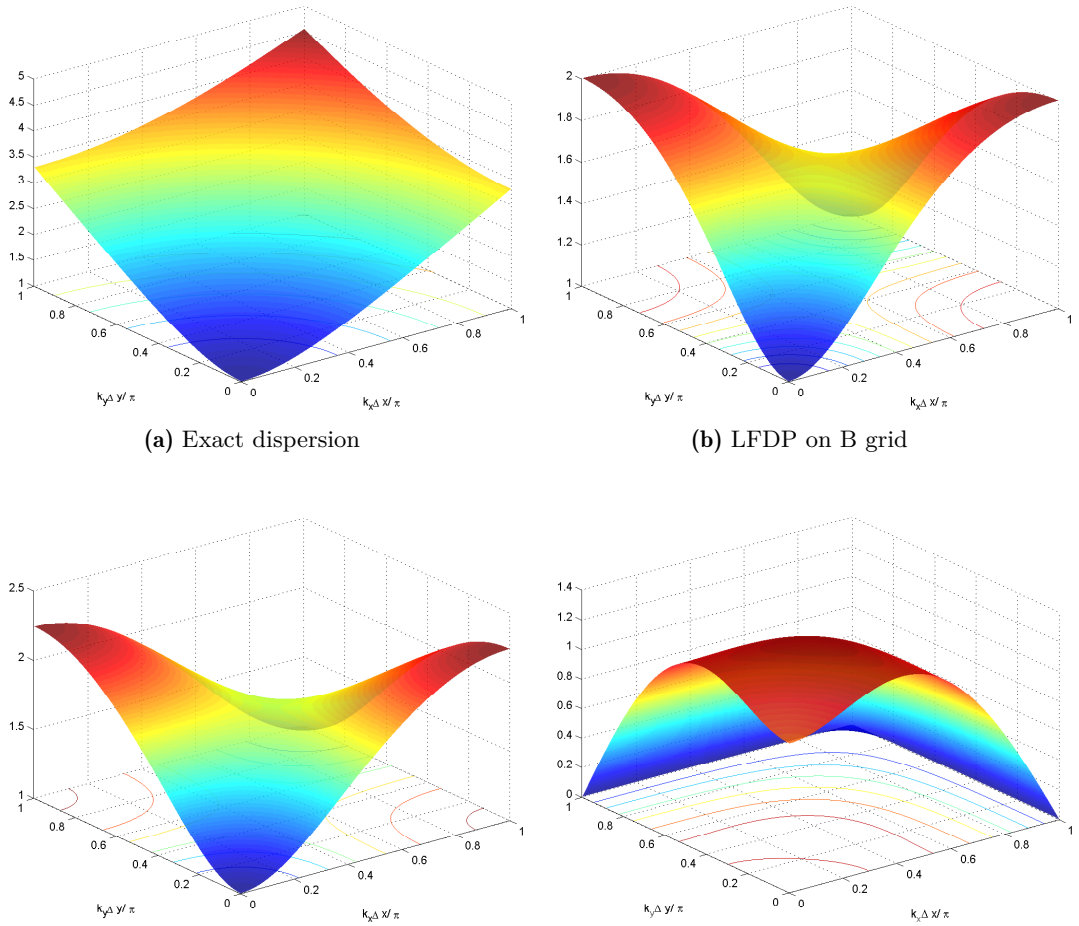
and on D grids

$$\begin{cases} r^{n+1}_{i,j} = r^n_{i,j} - a_\star \Delta t \nabla^{\mathrm{c}}_h \cdot [f_h(u^n_h), f_h(v^n_h)]_{i,j} + \nu_r \Delta t \nabla^{\mathrm{c}}_h \cdot (\partial_{x,h} r^n_h - \frac{\omega}{a_\star} v^n_h, \partial_{y,h} r^n_h + \frac{\omega}{a_\star} u^n_h)_{i,j} \\ u^{n+1}_{i,j+1/2} = u^n_{i,j+1/2} - a_\star \Delta t \partial_{x,h} [f_h(r^n_h)]_{i,j+1/2} + \nu_u \Delta t \partial_{x,h} [\nabla^{\mathrm{v}}_h \cdot (\mathbf{u}^n_h)]_{i,j+1/2} + \omega f_h(v^\theta_h)_{i,j+1/2} \\ v^{n+1}_{i+1/2,j} = v^n_{i+1/2,j} - a_\star \Delta t \partial_{y,h} [f_h(r^n_h)]_{i+1/2,j} + \nu_u \Delta t \partial_{y,h} [\nabla^{\mathrm{v}}_h \cdot (\mathbf{u}^n_h)]_{i+1/2,j} - \omega f_h(u^n_h)_{i+1/2,j} \end{cases} \tag{5.31}$$

5.3.1 Stability condition of the fully discrete scheme

In this subsection, for the sake of simplicity, we only consider the case

$$\kappa_r = \kappa^x_r = \kappa^y_r = \eta^x_r = \eta^y_r \quad \text{and} \quad \kappa = \kappa_u = \kappa_v = \eta_u = \eta_v.$$



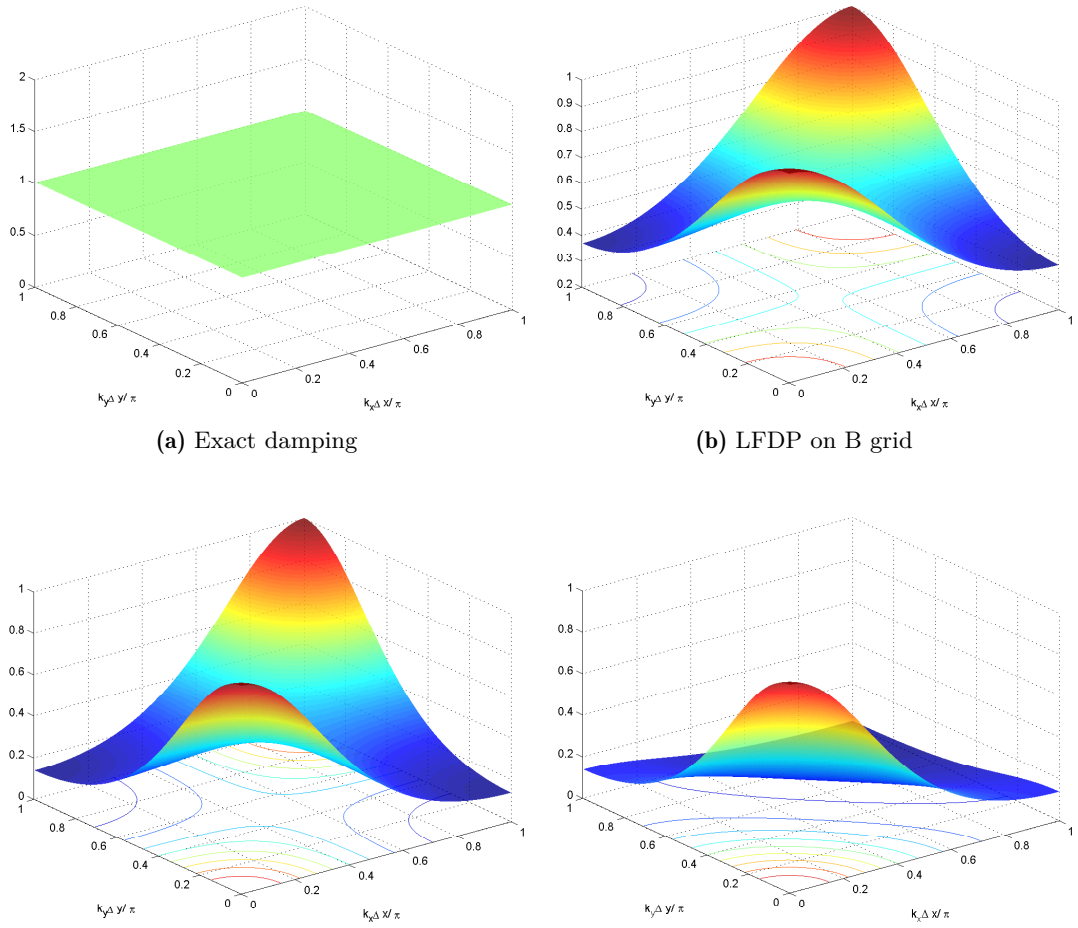
(a) Exact dispersion

(b) LFDP on B grid

(c) ATDP on B grid

(d) ATDP on D grid

Figure 5.3: *Dispersion relation* $\frac{\Im(\lambda)}{\omega}$ *of the staggered type schemes, depicted as a function of* $\frac{k_x \Delta x}{\pi}$ *and* $\frac{k_y \Delta y}{\pi}$ *with* $\frac{Rd}{\Delta x} = \frac{Rd}{\Delta y} = 1$.



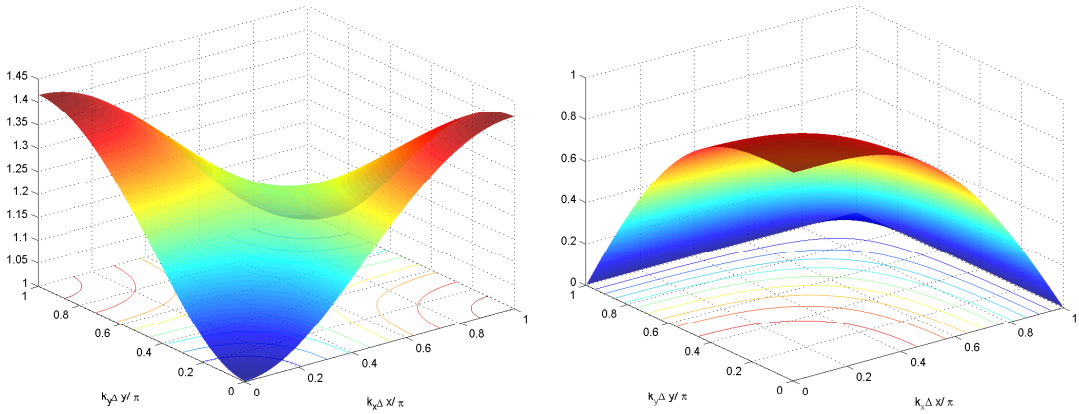
(a) Exact damping

(b) LFDP on B grid

(c) ATDP on B grid

(d) ATDP on D grid

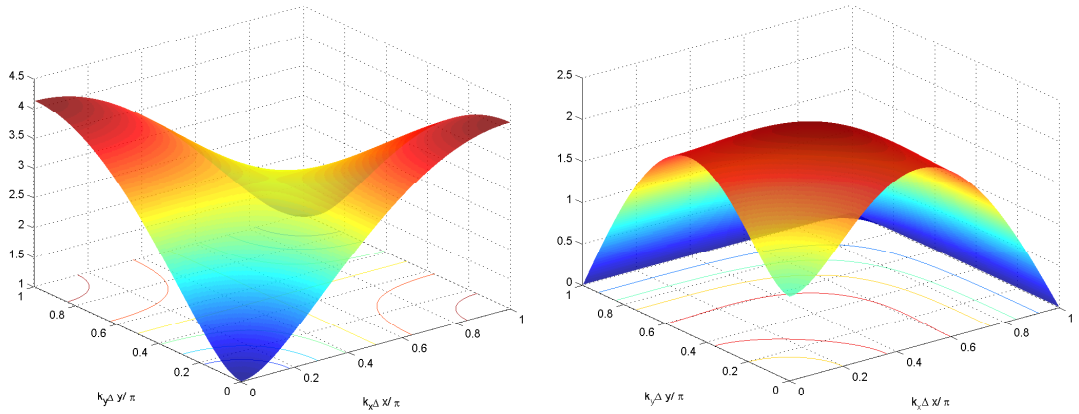
Figure 5.4: *Damping error* $e^{-\Re(\lambda)}$ *of the staggered type schemes, depicted as a function of* $\frac{k_x \Delta x}{\pi}$ *and* $\frac{k_y \Delta y}{\pi}$ *with* $\frac{Rd}{\Delta x} = \frac{Rd}{\Delta y} = 1$.



(a) ATDP on B grid

(b) ATDP on D grid

Figure 5.5: *Dispersion relation of the staggered type schemes, depicted as a function of* $\frac{k_x \Delta x}{\pi}$ *and* $\frac{k_y \Delta y}{\pi}$ *with* $\frac{Rd}{\Delta x} = \frac{Rd}{\Delta y} = \frac{1}{2}$.



(a) ATDP on B grid **(b)** ATDP on D grid

Figure 5.6: *Dispersion relation of the staggered type schemes, depicted as a function of* $\frac{k_x\Delta x}{\pi}$ *and* $\frac{k_y\Delta y}{\pi}$ *with* $\frac{Rd}{\Delta x}=\frac{Rd}{\Delta y}=2$.

It is important to use a discretization of the Coriolis force which is implicit enough in order to ensure the stability of the numerical scheme. Therefore, we only consider the parameters θ_1 and θ_2 belonging to the domain $\theta_1+\theta_2\leq\frac{1}{2}$. We mention [37] for more details of this stability region.

Lemma 5.8. *For a uniform mesh* $\Delta x=\Delta y=h$, *the LF-DP schemes (*i.e.(5.30) *with* $\nu_r=0$*) are stable under the following conditions*

$$\Delta t\leq\min\{\Delta t_a,\Delta t_b\}\quad\text{where}\quad\Delta t_a:=\frac{\kappa_u h}{2a_*}\ \text{ and }\ \Delta t_b:=\frac{2}{\omega|\theta_2-\theta_1|}.$$

Let us set $\varphi(x)=\frac{\sqrt{1+x^2}-1}{x^2}$. *The AT-LF and AT-DP schemes are stable provided that the time step is smaller than*

Scheme	$(\theta_1=0,\theta_2=0)$	$(\theta_1=1,\theta_2=0)$ *or* $(\theta_1=0,\theta_2=1)$	$(\theta_1=1/2,\theta_2=1/2)$
AT-LF	$\frac{\kappa_r}{2}\frac{h}{a_*}$	$\min\left\{\frac{2}{\omega},\frac{\kappa_r h}{4a_*}\varphi\left(\frac{\kappa_r\omega h}{4a_*}\right),\frac{4h}{\kappa_r a_*}\varphi\left(\frac{2\omega h}{\kappa_r a_*}\right)\right\}$	$\frac{\kappa_r h}{a_*}\varphi\left(\frac{\kappa_r\omega h}{2a_*}\right)$
AT-DP	$\frac{2\kappa}{2+\kappa^2}\frac{h}{a_*}$	$\min\left\{\frac{\kappa}{2+\kappa^2}\frac{h}{a_*},\frac{1}{\omega}\right\}$	$\min\left\{\frac{\kappa}{2+\kappa^2}\frac{h}{a_*},\frac{2}{\omega}\right\}$

Proof. The characteristic polynomial of the fully discrete staggered schemes can be obtained from that of the collocated vertex-based scheme in [53] by using $\eta=1$ instead of $\eta=\cos(\frac{k_x\Delta x}{2})\cos(\frac{k_y\Delta y}{2})$. Therefore, the proof is similar to the one in [53]. □

Remark 5.6. *Let us note that with the staggered scheme on D grids, because of the structure of the discrete kernel at the interface, it is essential to use the average for the Coriolis force. As a consequence, it is necessary to use either* $\theta_1=1,\theta_2=0$ *or* $\theta_1=0,\theta_2=1$ *to ensure that the proposed scheme is totally explicit. Hence, the stability condition of that scheme really depends on the Coriolis parameter* ω. *However, with the staggered scheme on B grids, we can overcome this drawback by the fact that the velocity* u *and* v *are defined at the same place. As a result, we can use a larger domain for the parameters* θ_1 *and* θ_2. *For instance, one can choose* $\theta_1=\theta_2\leq\frac{1}{2}$ *to ensure that the stability condition of the LF-DP scheme does not depend on the Coriolis source term.*

5.3.2 Orthogonality preserving scheme

Let us introduce two new parameters τ_1 and τ_2 involved in the time discretization of the velocity divergence in the pressure equation. We shall denote

$$\boldsymbol{u}^{\tau} = \begin{pmatrix} \tau_1 u^n + (1-\tau_1)u^{n+1} \\ \tau_2 v^n + (1-\tau_2)v^{n+1} \end{pmatrix}.$$

Then, the LF-DP-τ scheme on B grids can be written as

$$\begin{cases} r^{n+1}_{i+1/2,j+1/2} = r^n_{i+1/2,j+1/2} - a_\star \Delta t \nabla_h \cdot (\boldsymbol{u}^{\tau}_h)_{i+1/2,j+1/2} \\ \boldsymbol{u}^{n+1}_{i,j} = \boldsymbol{u}^n_{i,j} - a_\star \Delta t \nabla_h (r^n_h)_{i,j} + \nu_u \Delta t \nabla_h [\nabla_h \cdot (\boldsymbol{u}^n_h)]_{i,j} - \omega \Delta t \boldsymbol{u}^{\perp,\theta}_{i,j}. \end{cases} \tag{5.32}$$

The LF-DP-τ scheme on D grids is given by

$$\begin{cases} r^{n+1}_{i,j} = r^n_{i,j} - a_\star \Delta t \nabla^{\mathrm{c}}_h \cdot [f_h(u^{\tau}_h), f_h(v^{\tau}_h)]_{i,j} \\ u^{n+1}_{i,j+1/2} = u^n_{i,j+1/2} - a_\star \Delta t \partial_{x,h} [f_h(r^n_h)]_{i,j+1/2} + \nu_u \Delta t \partial_{x,h} [\nabla^{\mathrm{v}}_h \cdot (\mathbf{u}^n_h)]_{i,j+1/2} + \omega f_h(v^{\theta}_h)_{i,j+1/2} \\ v^{n+1}_{i+1/2,j} = v^n_{i+1/2,j} - a_\star \Delta t \partial_{y,h} [f_h(r^n_h)]_{i+1/2,j} + \nu_u \Delta t \partial_{y,h} [\nabla^{\mathrm{v}}_h \cdot (\mathbf{u}^n_h)]_{i+1/2,j} - \omega f_h(u^n_h)_{i+1/2,j} \end{cases} \tag{5.33}$$

Remark 5.7. *The LF-DP-τ scheme (5.32) on B grids or (5.33) on D grids is still explicit although the velocity field $\boldsymbol{u}^{n+1}$ appears in the pressure equation. In fact, we can compute the velocity field first and then use it to compute $\boldsymbol{u}^{\tau}$ in the pressure equation without having to solve any linear system.*

Lemma 5.9. *The Low Froude - Divergence penalization-τ scheme (LF-DP-τ) is an orthogonality preserving scheme if*

$$\tau_1 = \theta_2 \quad \text{and} \quad \tau_2 = \theta_1,$$

which means that the velocity field in the Coriolis source term and in the pressure equation must be computed using the same time strategy.

Proof. We note that the proof of this property for both B and D grid schemes are very similar, so we only present the explanation for the B grid scheme. By taking the product of the fully discrete scheme (5.32) with the steady state $\hat{q}_h \in \mathcal{E}^{\square}_{\omega\neq 0,B}$, using periodic boundary conditions, the discrete integration by part formula and the properties of elements of the discrete kernel, we will obtain

$$\begin{aligned} \left\langle q^{n+1}_h, \hat{q}_h \right\rangle &= -a_\star \Delta t \langle \nabla_h \cdot (\boldsymbol{u}^{\tau}_h), \hat{r}_h \rangle_{\mathcal{D}} - \omega \Delta t \langle \boldsymbol{u}^{\perp,\theta}_h, \hat{\boldsymbol{u}}_h \rangle_{\mathcal{P}} \\ &= \langle a_\star \Delta t \nabla_h \hat{r}_h, \boldsymbol{u}^{\tau}_h \rangle_{\mathcal{P}} + \left\langle \omega \Delta t \hat{\boldsymbol{u}}^{\perp}_h, \boldsymbol{u}^{\theta}_h \right\rangle_{\mathcal{P}}. \end{aligned}$$

Therefore, using that when $\tau_1 = \theta_2$ and $\tau_2 = \theta_1$, we get $\boldsymbol{u}^{\tau}_h = \boldsymbol{u}^{\theta}_h$, it follows that

$$\left\langle q^{n+1}_h, \hat{q}_h \right\rangle = 0, \forall \hat{q}_h \in \mathcal{E}^{\square}_{\omega\neq 0,B}.$$

□

5.4 Numerical test case

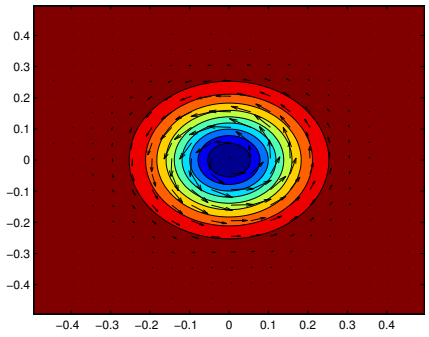
5.4.1 Well-balanced test case

In this test case, we investigate the behavior of the Godunov type schemes with a geostrophic equilibrium as initial condition. Particularly, we consider the stationary vortex in the square domain $\mathbb{T}^2 = [-0.5, 0.5] \times [-0.5, 0.5]$ with the initial pressure r^0 given by

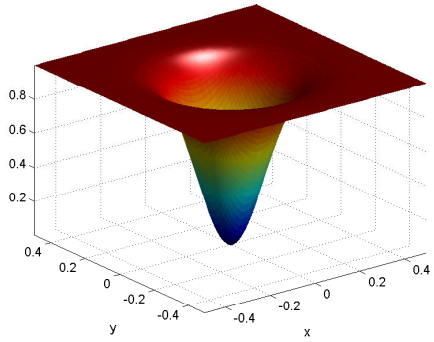
$$r(x, y, t = 0) = 1 - \exp\left[-\left(\frac{3x}{0.5}\right)^2 - \left(\frac{3y}{0.5}\right)^2\right].$$

Then we construct the initial velocity field $\boldsymbol{u}^0$ by using the definition of the discrete kernel (5.15) so that we can obtain a nontrivial stationary state which is shown in Figure 5.7.
In what follows, various schemes are tested; they are denominated as X-Y schemes, the first letter referring to the diffusion strategy in the pressure equation, the second to the diffusion strategy in the velocity equation. The "Classical" strategy indicates that the standard formulations with diffusion $-\nu_r \Delta r$ in the pressure equation or $-\nu_u(\partial_{xx}u, \partial_{yy}v)^T$ in the velocity equation are used.

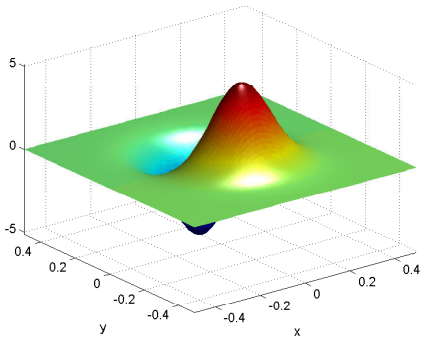
Figure 5.8 indicates that the Classical–Classical (C-C), Apparent Topography–Classical (AT-C) and Classical–Divergence Penalization (C-DP) schemes are unable to capture the discrete steady state. However, we can see that the correction related to the velocity diffusion in this test case is more important than the correction related to the pressure diffusion because the C-DP scheme is more accurate than the AT-C scheme. On the contrary, the AT-LF, LF-DP and AT-DP schemes are well-balanced since their discrete kernel includes the discrete geostrophic equilibrium.
Figure 5.9 shows that the structure of the vortex of the non well-balanced schemes (C-C and AT-C) is different from that of the initial condition while they are exactly the same with all well-balanced schemes (AT-LF, LF-DP and AT-DP). Moreover, this figure also indicates that unlike the other non well-balanced schemes, the C-DP scheme can preserve the structure of the solution. This is another evidence to say that in this test case, the correction related to the velocity diffusion is really important.



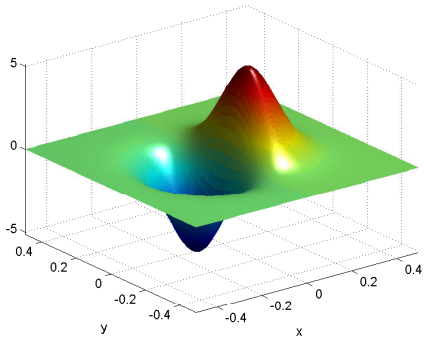
(a) Contours of $r(x,y,t)$ and vector field at $t=0$



(b) $r(x,y,t)$ at $t=0$

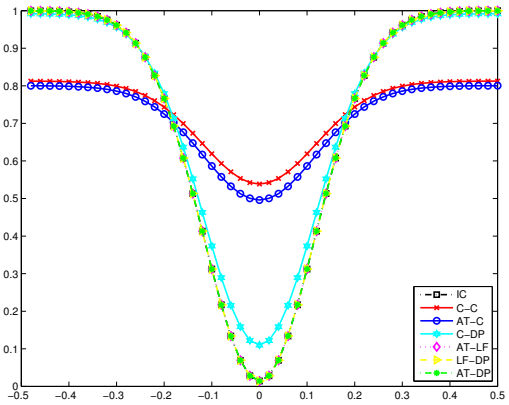


(c) $u(x,y,t)$ at $t=0$

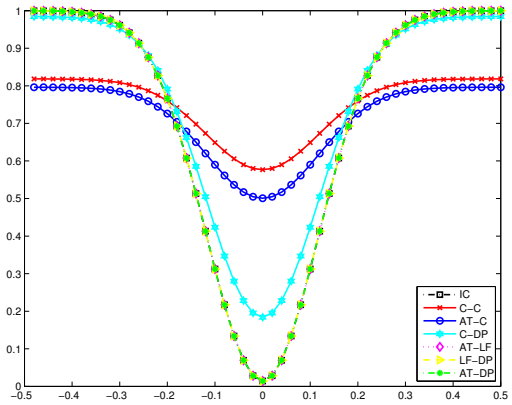


(d) $v(x,y,t)$ at $t=0$

Figure 5.7: *A stationary vortex as initial condition with* 100×100 *grid cells.*

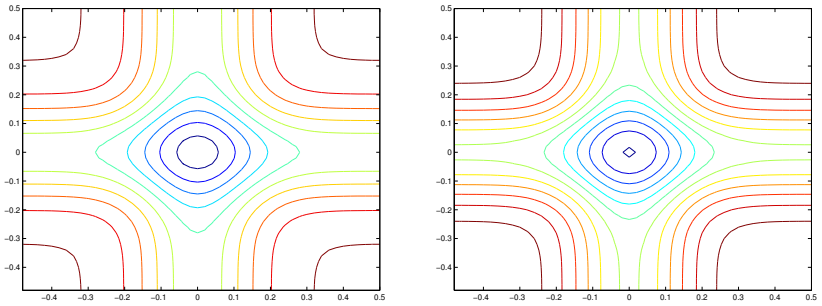


(a) $r(x,y,t=10)$ with 50×50 grid cells



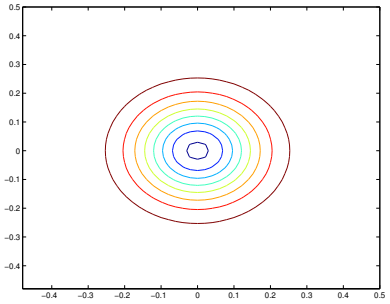
(b) $r(x,y,t=20)$ with 50×50 grid cells

Figure 5.8: *1D-cut of stationary vortex for staggered B type schemes.*

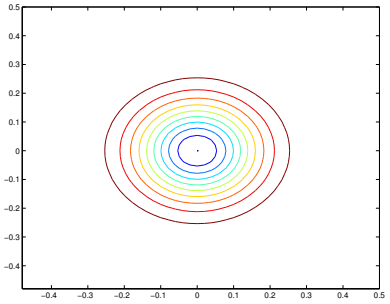


(a) $C - C$ scheme

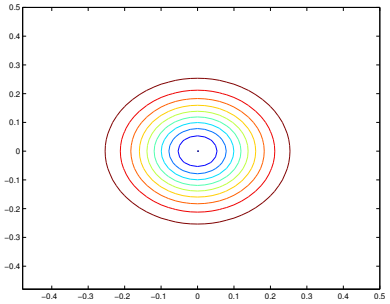
(b) $AT - C$ scheme



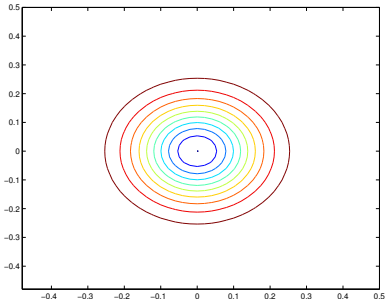
(c) $C - DP$ scheme



(d) $AT - LF$ scheme



(e) $LF - DP$ scheme



(f) $AT - DP$ scheme

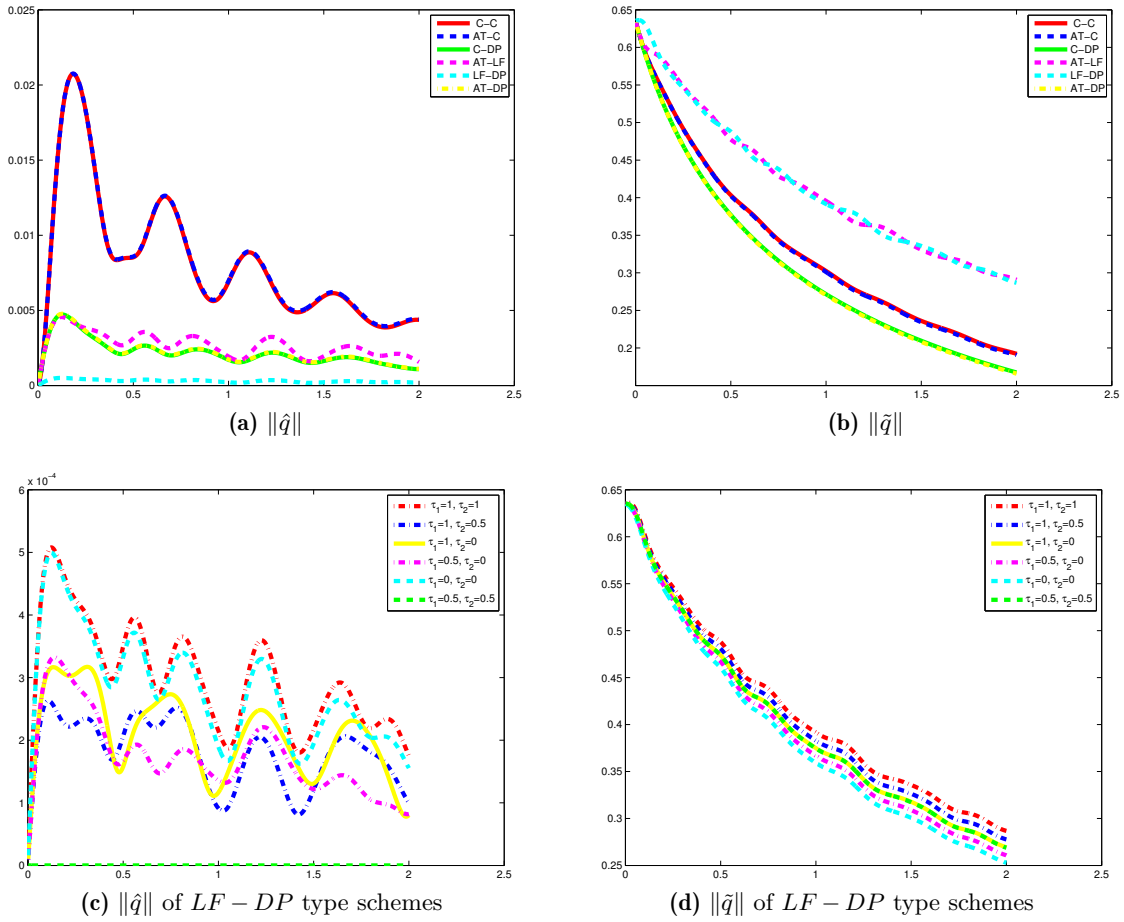
Figure 5.9: *Contours of* r *at* $t = 20$ *for Godunov type schemes with* 50×50 *grid cells.*

5.4.2 Orthogonality preserving test case

In this test case, we consider periodic boundary conditions and the initial vector field is given by

$$\begin{cases} u(x,y,t=0) = \frac{1}{2}\exp\left[-\left(\frac{4x}{0.4}\right)^2 - \left(\frac{4y}{0.8}\right)^2\right] \\ v(x,y,t=0) = \frac{1}{2}\exp\left[-\left(\frac{4x}{0.8}\right)^2 - \left(\frac{4y}{0.4}\right)^2\right]. \end{cases} \tag{5.34}$$

in the domain $\mathbb{T}^2 = [-0.5, 0.5] \times [-0.5, 0.5]$. Then the initial pressure $r(x,y,t=0)$ is constructed by using the definition of the discrete orthogonal subspace.
Figure (5.10a) indicates that all presented schemes are not orthogonality preserving schemes because they create some waves in the discrete kernel. The LF-DP scheme is better than the other schemes since it creates very small errors. Figure (5.10b) shows that the orthogonal parts of the solutions obtained by the AT-LF and LF-DP schemes damp slower than those obtained by the other schemes. On the other hand, (5.10c) points out that only the LF-DP scheme with $\tau_1 = \tau_2 = \frac{1}{2}$ (the parameters τ_1 and τ_2 correspond to the time discretization for the divergence of the velocity field in the pressure equation) preserves the orthogonality property. Moreover, the damping rate of the orthogonal part which depends on the parameters τ_1 and τ_2 is presented in Figure (5.10d). However, the effect of those parameters is small.



(a) $\|\hat{q}\|$ **(b)** $\|\tilde{q}\|$

(c) $\|\hat{q}\|$ of $LF-DP$ type schemes **(d)** $\|\tilde{q}\|$ of $LF-DP$ type schemes

Figure 5.10: *Orthogonality preserving test case: the evolution of the kernel and orthogonal parts with* 50×50 *grid cells and* $\theta_1 = \theta_2 = \frac{1}{2}$.

5.4.3 Accuracy at low Froude number test case

We now consider an initial condition close to the discrete kernel up to a perturbation of size M. This initial condition is simply given by

$$q_h^0 = \hat{q}_h^0 + M\frac{\tilde{q}_h^0}{\|\tilde{q}_h^0\|}$$

where $\hat{q}_h^0$ stands for the kernel part given in § 5.4.1 and $\tilde{q}_h^0$ is the orthogonal part considered in § 5.4.2.

Figure 5.11 confirms the analysis in section § 5.2.3 . While the C-C, AT-C and C-DP schemes are not accurate at low Froude number since the norm of $\|q - \mathbb{P}q^0\|$ is not of size $\mathcal{O}(M)$ (figure (5.11b)), all the well-balanced schemes (AT-LF, LF-DP and AT-DP schemes) are accurate at low Froude number because the norms of the total deviation in Figures 5.11c and 5.11d remain of order $\mathcal{O}(M)$. Moreover, in consideration of the C-C and AT-C schemes, we can see in Figure 5.11b, that the norm of the total deviation is an increasing function of time. Since the initial projection $\mathbb{P}q^0$ is exactly a non-trivial steady state, the numerical solutions obtained by the C-C and AT-C schemes are far from the correct kernel. On the other hand, Figure 5.11a shows that the norm of the spurious waves of the C-C and AT-C schemes tend to decrease with time. This phenomena implies that the numerical solution of those schemes tend to the trivial steady state.

Figure 5.12 shows that when the Froude number decrease, the $\max_t \|q - \mathbb{P}q^0\|(t)$ of all presented non well-balanced schemes seem to be constants. The well-balanced schemes show a good behavior since the $\max_t \|q - \mathbb{P}q^0\|(t)$ is a function decreasing linearly with the Froude number.

5.4.4 Water column test case

In this test case, we consider a discontinuous initial condition which is given by

$$\begin{cases} r(t=0,x,y) = \begin{cases} 2, & \text{if} \quad x^2+y^2 \leq 1 \\ 1, & \text{if} \quad x^2+y^2 > 1, \end{cases} \\ u(t=0,x,y) = 0, \\ v(t=0,x,y) = 0. \end{cases}$$

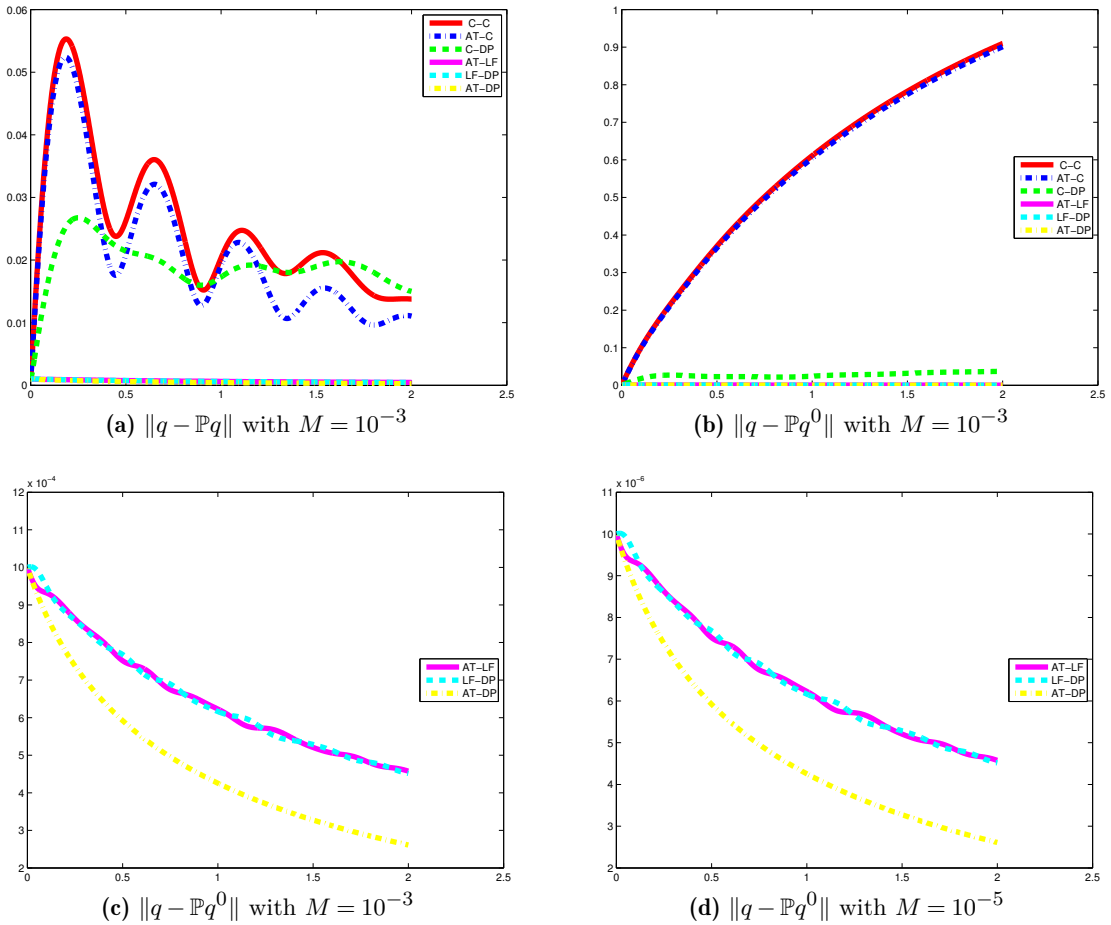
with periodic boundary conditions on the domain $[-5,5]\times[-5,5]$. This initial condition corresponds to a circular dam break and is very far from the geostrophic equilibrium (5.15) and (5.25).

Figure 5.13 shows the final state of all numerical staggered type schemes on B and D grids. The three well balanced schemes, namely the AT-DP, LF-DP and AT-LF schemes tend to the geostrophic equilibrium while the solution obtained from the C-C and C-DP schemes seem to tend to a constant state. Although the solution obtained with the AT-C scheme is better than that obtained with the classical scheme, this strategy does not converge the non-trivial steady state. This is in agreement with the results of these strategies on A grids (collocated schemes), as presented in [53].

Figure 5.14 presents the contours of the solutions obtained by the AT-DP scheme on B and D grids. As can be seen, there is small oscillation in the numerical solution on B grids while we do not have this problem on D grids. One possible explanation is that the scheme on B grids has damping rate for the waves with shortest wavelengths that is lower than that on D grids. We can also observe this small oscillation in the final state shown in Figure 5.15.

5.5 Conclusion

This study deals with the ability of staggered type schemes on Cartesian meshes to capture discrete non-trivial steady states characterized by the balance between the pressure gradient and



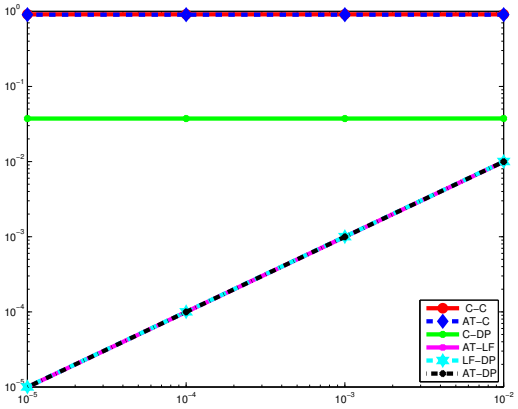
(a) $\|q - \mathbb{P}q\|$ with $M = 10^{-3}$ **(b)** $\|q - \mathbb{P}q^0\|$ with $M = 10^{-3}$

(c) $\|q - \mathbb{P}q^0\|$ with $M = 10^{-3}$ **(d)** $\|q - \mathbb{P}q^0\|$ with $M = 10^{-5}$

Figure 5.11: *Evolution of the spurious wave and total deviation with* 50×50 *grid cells, with an initial condition close to the discrete kernel.*

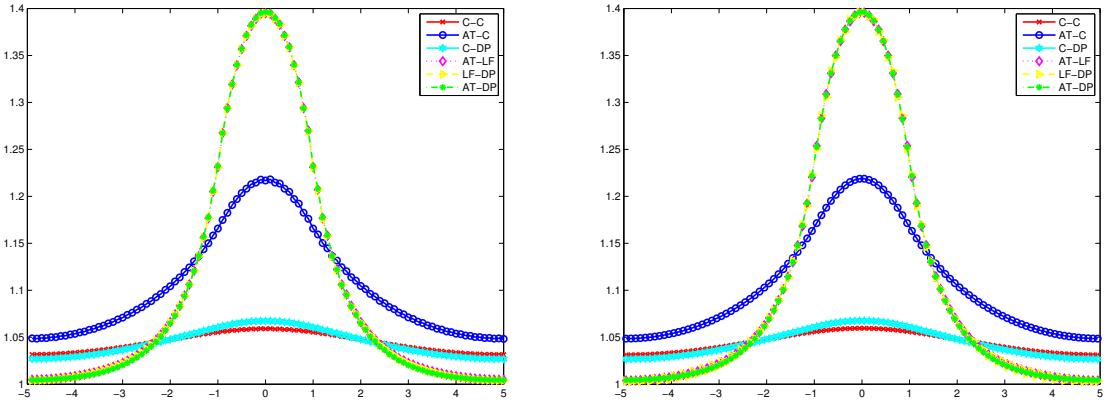
the Coriolis force. The AT-DP and LF-DP scheme on both B and D grid- are well-balanced schemes, but only the LF-DP strategy with the same time discretization for the velocity field in the Coriolis source term and in the pressure equation can preserve the orthogonal subspace of the kernel.

Unlike on B grids, the horizontal and vertical velocity components on D grids are not defined on the same cell, so we need some approximation for the Coriolis force. As a consequence, the dispersion laws on D grids are less accurate than on B grids. However, since the D grid schemes have a larger damping rate on the waves with shortest wavelengths, the obtained schemes on D grids present fewer oscillations.



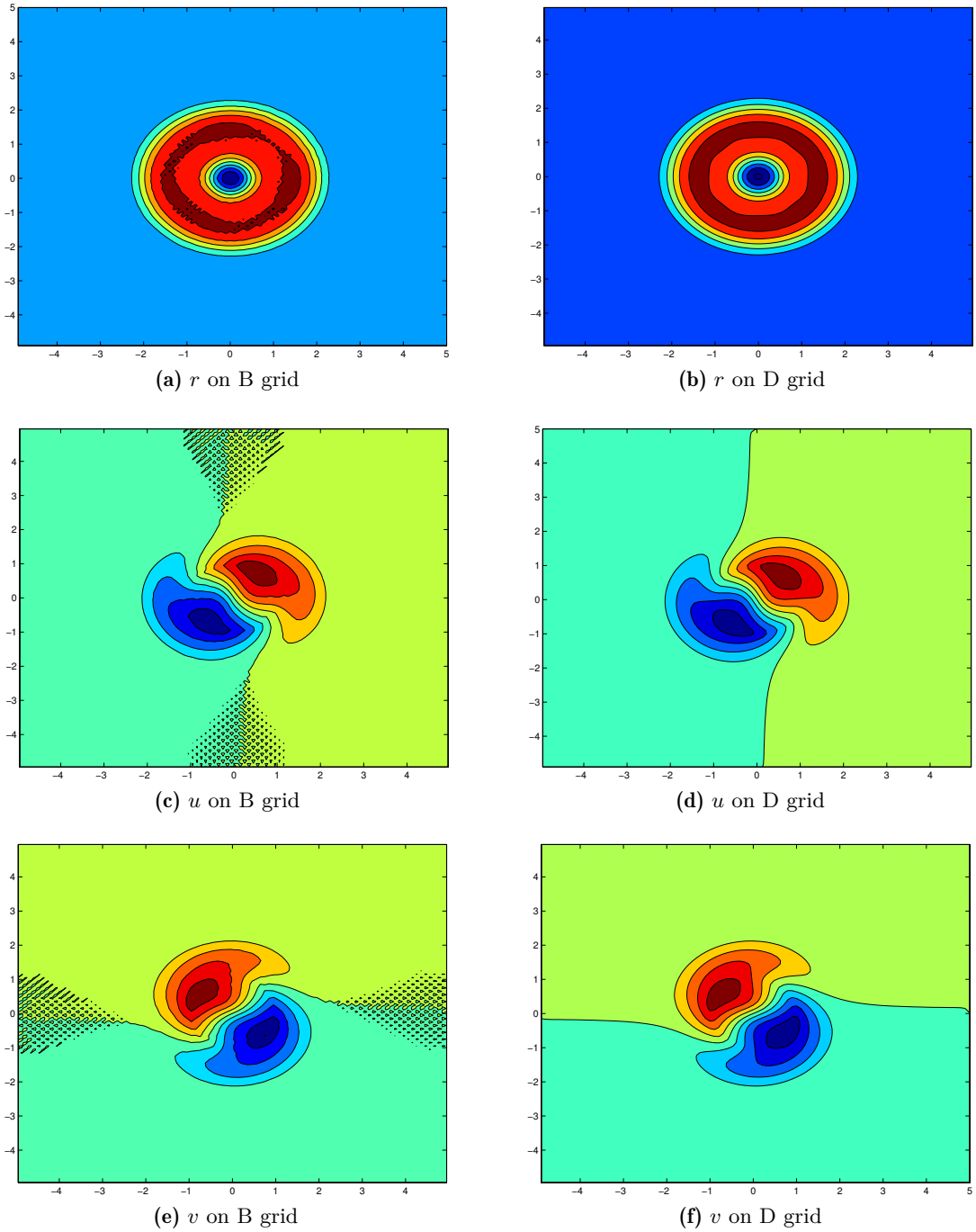
(a)

Figure 5.12: *The log-log graph of* $\max_t \|q - \mathbb{P}q^0\|(t)$ *for* $t = 2$ *and Froude number*$= 10^{-2}$, 10^{-3}, 10^{-4} *and* 10^{-5}.



(a) r on B grid **(b)** r on D grid

Figure 5.13: *Cross section of the pressure* r *at* $y = 0$ *at time* $t = 100$.



(a) r on B grid

(b) r on D grid

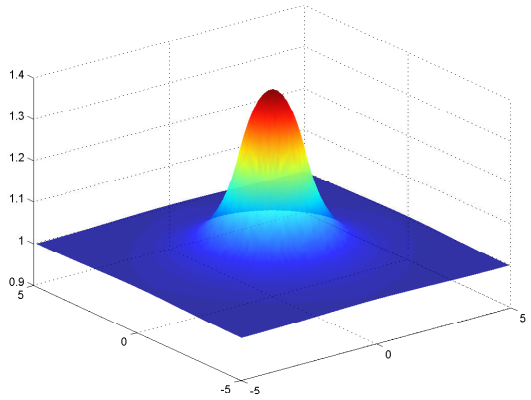
(c) u on B grid

(d) u on D grid

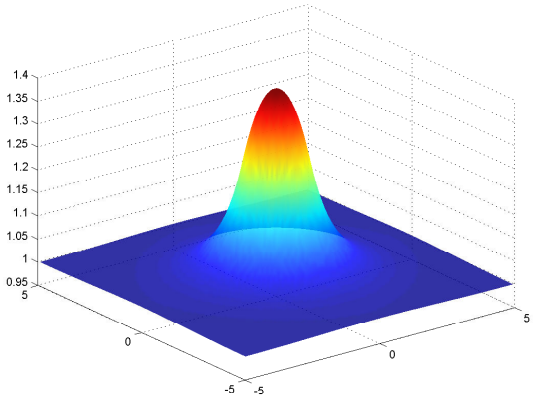
(e) v on B grid

(f) v on D grid

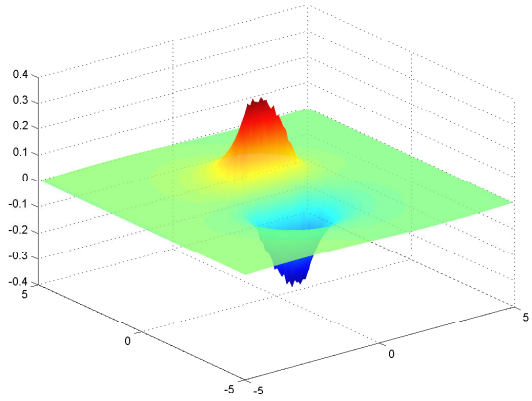
Figure 5.14: *Contours of AT-DP solutions on B and D grids at time* $t = 1$.



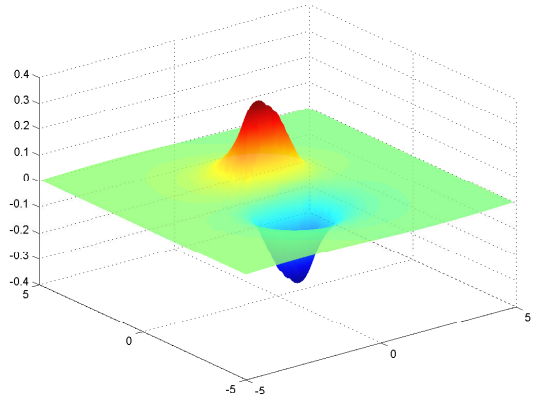
(a) r on B grid



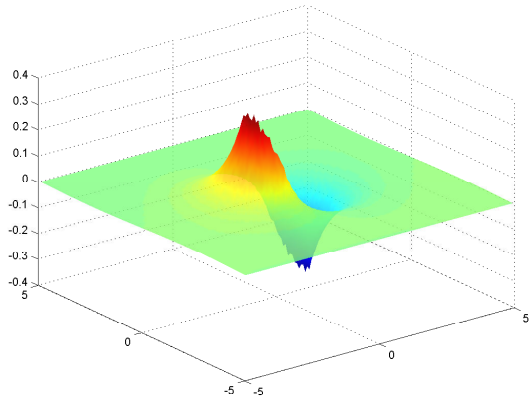
(b) r on D grid



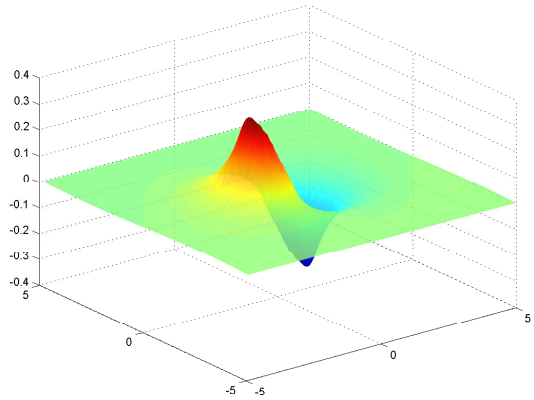
(c) u on B grid



(d) u on D grid



(e) v on B grid



(f) v on D grid

Figure 5.15: *AT-DP solutions on B and D grids at time* $t = 100$.

CHAPTER 6

Analysis of staggered type schemes applied to the linear wave equation with Coriolis source term. Part 2: on triangular meshes

In order to succeed,
we must first believe that we can.

Nikos Kazantzakis.

Abstract

By analyzing the discrete kernel of the linear acoustic operator of the Godunov type schemes applied to the linear wave equation with Coriolis source term, we clearly show that there are some drawbacks in the collocated Godunov scheme on triangular meshes. To overcome this difficulty, we propose a staggered strategy which computes the pressure field at the vertices of the triangles (centers of dual cells) and the velocity field at the (primary) cell centers. Our analysis shows that, with periodic boundary conditions, and unlike the Cartesian mesh case, the numerical diffusion on the velocity equations is no more the reason of the inaccuracy problem. Therefore, in order to capture the discrete geostrophic equilibrium, we only have to correct the diffusion on the pressure equation by simply deleting this diffusion term (see [37]) or applying the Apparent Topography method introduced in [13].

Chapter content

6.1 Introduction

The Navier-Stokes equations can be reduced to the shallow water equation in large scale oceanic phenomena. This is because in this model, the horizontal scale is much larger than the vertical one. The mathematical details of this reduction procedure can be found in [1]. More importantly, in large scale atmospheric circulations, the effect of the Earth's rotation is clearly noticeable and the balance between the pressure gradient and Coriolis force leads to the so-called geostrophic equilibrium. As a matter of fact, circulations normally take place around this geostrophic equilibrium, so it is very important to have appropriate numerical schemes which can correctly describe the waves under these circumstances. Obviously, it is highly expected that the numerical schemes can capture a discrete version of the geostrophic equilibrium or at least produce numerical solutions that remain close to this state with acceptable small errors. For the purpose of deriving numerical schemes with such kind of properties, we begin with the investigation of the dimensionless shallow water system of equations on the rotating frame which is given by

$$\begin{cases} \partial_t h + \nabla \cdot (h\bar{\mathbf{u}}) = 0, & \text{(6.1a)} \\ St\partial_t(h\bar{\mathbf{u}}) + \nabla \cdot (h\bar{\mathbf{u}} \otimes \bar{\mathbf{u}}) + \dfrac{1}{Fr^2}\nabla\left(\dfrac{h^2}{2}\right) = -\dfrac{1}{Fr^2}h\nabla b - \dfrac{1}{Ro}h\bar{\mathbf{u}}^{\perp}. & \text{(6.1b)} \end{cases}$$

In System (6.1) unknowns h and $\bar{\mathbf{u}}$ respectively denote the water depth and the velocity of the water column and function $b(x)$ denotes the topography of the considered oceanic basin and is a given function. Dimensionless numbers St, Fr and Ro respectively stand for the Strouhal, the Froude and the Rossby numbers. As mentioned above, we are interested in the large scale oceanographic flows, so we will focus on the case in which the Froude and Rossby numbers are small. In particularly, we shall assume that

$$Ro = \mathcal{O}(M) \quad \text{and} \quad Fr = \mathcal{O}(M)$$

with M a small parameter (typical values lead to $M \sim 10^{-2}$). For a Strouhal number of order $\mathcal{O}(\frac{1}{M})$, *i.e.*for small time scale, the solution of system (6.1) satisfies at the leading order the linear wave equation with Coriolis source term

$$\begin{cases} \partial_t r + a_\star \nabla . \boldsymbol{u} = 0 \\ \partial_t \boldsymbol{u} + a_\star \nabla r = -\omega \boldsymbol{u}^{\perp} \end{cases} \tag{6.2}$$

where $\boldsymbol{u} = (u, v)^T$, and $\boldsymbol{u}^{\perp} = (-v, u)^T$. The parameters $a_\star$ and ω are constants of order one, respectively related to the wave velocity and to the rotating velocity. The stationary state corresponding to Equation (6.2) is the geostrophic equilibrium which is given by

$$a_\star \nabla r = -\omega \boldsymbol{u}^{\perp}. \tag{6.3}$$

It is well known that the classical Godunov scheme applied to (6.2) will fail to capture the discrete steady state (6.3). Unlike in the 1D case, for which the reason for the accuracy problem is only linked to the numerical diffusion in the pressure equation [37], the numerical diffusion in the velocity equations in the 2D case is also a reason for the inaccurate behavior of the Godunov type schemes. This comes from the fact that the steady state (6.3) also implies the divergence free condition $\nabla \cdot \boldsymbol{u} = 0$, while the numerical diffusion in the velocity equations of the Godunov scheme, on 2D cartesian meshes, does not vanish on velocities satisfying a discrete equivalent of this condition. Therefore, modified Godunov schemes on Cartesian meshes are introduced in [53] to correctly capture the discrete steady state; these schemes use a combination of the Apparent Topography strategy in [25] and of the idea named Divergence Penalisation in [20].

The present work is strongly motivated by the fact that [21] pointed out that there is no more problem related to the divergence free constraint on the velocity solution of the Godunov type schemes applied to the homogeneous linear wave equation on *triangular* meshes. Reference [21] provides a detailed explanation based on the discrete kernel of the Godunov scheme on triangular meshes. Therefore, one may think that we can apply the Apparent Topography strategy on (6.2) on triangular meshes to be able to correctly capture discrete steady states. However, the main purpose of this work is to point out that the collocated scheme applied to the rotating linear wave equation (6.2) still suffers from some problems, and we propose a staggered strategy to avoid these problems.

This work is organized as follows. In Section 6.2, we clearly present the problem we encounter with the collocated Godunov scheme on triangular meshes by showing the discrete kernel of the collocated scheme. Then we propose a staggered scheme in Section 6.3 to overcome this difficulty. Particularly, we define discrete differential operators which satisfy mimetic properties and perform the analysis of the staggered scheme, in which we focus on some essential properties such as the preservation of the kernel and of its orthogonal subspace. We then take into account in Section 6.4 the time discretization to introduce the appropriate one step as well as splitting schemes, such that they still possess the same good properties as the semi-discrete scheme. Some numerical test cases are shown in Section 6.5 to confirm the analysis led in the theoretical part and we discuss some perspectives. Concluding remarks complete the study in Section 6.6.

6.2 Explanation of the wrong behavior of collocated schemes

In this section, we perform the analysis of the Godunov collocated scheme applied to the linear wave equation (6.2) to clearly point out that it is necessary to use a staggered scheme to capture well the geostrophic equilibrium (6.3).

We shall denote by T_i a generic triangular cell used for the discretization of the computational domain. Let A_{ij} be the common edge of the neighboring cells T_i and T_j, and $\boldsymbol{n}_{ij}$ the unit normal vector to A_{ij} pointing from T_i to T_j. Moreover, we also denote the area of the cell T_i by $|T_i|$ and the length of the edge A_{ij} by $|A_{ij}|$. Then, the semi-discrete collocated scheme for the linear wave equation (6.2) can be written as

$$\begin{cases} \frac{d}{dt} r_i + \frac{a_\star}{2|T_i|} \displaystyle\sum_{A_{ij}\subset\partial T_i} |A_{ij}| \left[(\boldsymbol{u}_i + \boldsymbol{u}_j)\cdot \boldsymbol{n}_{ij} + \kappa_r (r_i - r_j) \right] = 0 \\ \frac{d}{dt} \boldsymbol{u}_i + \frac{a_\star}{2|T_i|} \displaystyle\sum_{A_{ij}\subset\partial T_i} |A_{ij}| \left[(r_i + r_j) + \kappa_{\boldsymbol{u}} (\boldsymbol{u}_i - \boldsymbol{u}_j)\cdot \boldsymbol{n}_{ij} \right] \boldsymbol{n}_{ij} = -\omega \boldsymbol{u}_i^\perp \end{cases} \tag{6.4}$$

where κ_r and $\kappa_{\boldsymbol{u}}$ represent the parameters of the diffusion terms. The classical Godunov scheme corresponds to the case $\kappa_r = \kappa_{\boldsymbol{u}} = 1$. We mention [21] for the construction of the scheme applied to the homogeneous wave equation. Here, we take into account the effect of the Coriolis force in the right-hand side of the semi-discrete scheme.

Let us also note that the semi-discrete collocated Godunov scheme (6.4) can be written in the following compact form

$$\frac{d}{dt} q_h + L_{\kappa,h} q_h = 0$$

where

$$q_h := \begin{pmatrix} r_h \\ \boldsymbol{u}_h \end{pmatrix} \in \mathbb{R}^{3N} \quad \text{and} \quad L^i_{\kappa,h} q_h = \begin{pmatrix} \frac{a_\star}{2|T_i|} \displaystyle\sum_{A_{ij}\subset\partial T_i} |A_{ij}| [(\boldsymbol{u}_i + \boldsymbol{u}_j)\cdot \boldsymbol{n}_{ij} + \kappa_r (r_i - r_j)] \\ \frac{a_\star}{2|T_i|} \displaystyle\sum_{A_{ij}\subset\partial T_i} |A_{ij}| [(r_i + r_j) + \kappa_{\boldsymbol{u}} (\boldsymbol{u}_i - \boldsymbol{u}_j)\cdot \boldsymbol{n}_{ij}] \boldsymbol{n}_{ij} + \omega \boldsymbol{u}_i^\perp \end{pmatrix}.$$

Let $r_h = (r_i)$ be in $\mathbb{R}^N$ and $\boldsymbol{u}_h = (\boldsymbol{u}_i)$ in $\mathbb{R}^{2N}$ where N is the number of triangles. To be convenient, we now define the discrete gradient of r_h and divergence of velocity $\boldsymbol{u}_h$ respectively by the following formulas

$$(\nabla_h r_h)_i = \frac{1}{|T_i|} \sum_{A_{ij} \subset \partial T_i} |A_{ij}| \frac{(r_i + r_j)}{2} \boldsymbol{n}_{ij} \tag{6.5}$$

and

$$(\nabla_h \cdot \boldsymbol{u}_h)_i = \frac{1}{|T_i|} \sum_{A_{ij} \subset \partial T_i} |A_{ij}| \frac{(\boldsymbol{u}_i + \boldsymbol{u}_j)}{2} \cdot \boldsymbol{n}_{ij}. \tag{6.6}$$

Moreover, considering two cell-centered vectors $q_h^1 = (r_h^1, \boldsymbol{u}_h^1)$ and $q_h^2 = (r_h^2, \boldsymbol{u}_h^2)$, we also define discrete scalar products by

$$\langle r_h^1, r_h^2 \rangle_c := \sum_{i=1}^{N} |T_i| r_i^1 r_i^2 \ , \ \langle \boldsymbol{u}_h^1, \boldsymbol{u}_h^2 \rangle_c := \sum_{i=1}^{N} \sum_{i=1}^{N} |T_i| \boldsymbol{u}_i^1 \cdot \boldsymbol{u}_i^2 \ , \ \langle q_h^1, q_h^2 \rangle_c := \langle r_h^1, r_h^2 \rangle_c + \langle \boldsymbol{u}_h^1, \boldsymbol{u}_h^2 \rangle_c. \tag{6.7}$$

We use the same notations for these three scalar products, but the context in which they are used avoids any ambiguity.

Proposition 6.1. *With the discrete operators defined by (6.5), (6.6) and the discrete scalar products in (6.7), and using periodic boundary conditions, we obtain the following discrete integration by part formula on a collocated mesh:*

$$\langle \nabla_h \cdot \boldsymbol{u}_h, r_h \rangle_c = -\langle \nabla_h r_h, \boldsymbol{u}_h \rangle_c. \tag{6.8}$$

Proof. First, by using the fact that $\sum_{A_{ij} \subset \partial T_i} |A_{ij}| \boldsymbol{n}_{ij} = 0$, we get

$$\sum_{A_{ij} \subset \partial T_i} |A_{ij}| \boldsymbol{u}_i \cdot \boldsymbol{n}_{ij} = 0 \quad \text{and} \quad \sum_{A_{ij} \subset \partial T_i} |A_{ij}| r_i \boldsymbol{n}_{ij} = 0.$$

Therefore, using that $\boldsymbol{n}_{ji} = -\boldsymbol{n}_{ij}$, we obtain

$$\begin{aligned}
\langle \nabla_h \cdot \boldsymbol{u}_h, r_h \rangle_c &= \frac{1}{2} \sum_{i}^{N} \sum_{A_{ij} \subset \partial T_i} |A_{ij}| [(\boldsymbol{u}_i + \boldsymbol{u}_j) \cdot \boldsymbol{n}_{ij}] r_i = \frac{1}{2} \sum_{i}^{N} \sum_{A_{ij} \subset \partial T_i} |A_{ij}| \boldsymbol{u}_j \cdot \boldsymbol{n}_{ij} r_i \\
&= \frac{1}{2} \sum_{A_{ij}} |A_{ij}| \boldsymbol{u}_j \cdot \boldsymbol{n}_{ij} r_i + \frac{1}{2} \sum_{A_{ij}} |A_{ij}| \boldsymbol{u}_i \cdot \boldsymbol{n}_{ji} r_j = -\frac{1}{2} \sum_{A_{ij}} |A_{ij}| r_i \boldsymbol{u}_j \cdot \boldsymbol{n}_{ji} - \frac{1}{2} \sum_{A_{ij}} |A_{ij}| r_j \boldsymbol{u}_i \cdot \boldsymbol{n}_{ij} \\
&= -\frac{1}{2} \sum_{i}^{N} \sum_{A_{ij}} |A_{ij}| r_j \boldsymbol{n}_{ij} \cdot \boldsymbol{u}_i = -\frac{1}{2} \sum_{i}^{N} \sum_{A_{ij}} |A_{ij}| (r_i + r_j) \boldsymbol{n}_{ij} \cdot \boldsymbol{u}_i \\
&= -\langle \nabla_h r_h, \boldsymbol{u}_h \rangle_c.
\end{aligned}$$

□

Lemma 6.1. *On a triangular mesh, with the collocated Godunov scheme, we have:*

$$Ker L_{\kappa_r \neq 0, h} = \left\{ q_h := \begin{pmatrix} r_h \\ \boldsymbol{u}_h \end{pmatrix} \in \mathbb{R}^{3N} \text{ such that } \exists a \in \mathbb{R} : r_i = a \text{ and } \boldsymbol{u} = 0 \right\}. \tag{6.9}$$

Moreover, we also have

$$Ker L_{\kappa_r = 0, h} = C_h^1 \cap C_h^2 \tag{6.10}$$

with the following definitions for C_h^1 and C_h^2:

$$C_h^1 := \left\{ q_h := \begin{pmatrix} r_h \\ \boldsymbol{u}_h \end{pmatrix} \in \mathbb{R}^{3N} \text{such that } \frac{a_\star}{|T_i|} \sum_{A_{ij} \subset \partial T_i} |A_{ij}| \frac{r_i + r_j}{2} \boldsymbol{n}_{ij} = -\omega \boldsymbol{u}_i^\perp \right\}.$$

$$C_h^2 := \left\{ q_h := \begin{pmatrix} r_h \\ \boldsymbol{u}_h \end{pmatrix} \in \mathbb{R}^{3N} \text{such that } \exists \phi_h^{\mathrm{L}} \in \mathcal{C}_\sharp^0(\overline{\Omega}), (\phi_h^{\mathrm{L}})_{|T_i} \in P^1(T_i), \boldsymbol{u}_i = (\nabla \times \phi_h^{\mathrm{L}})_{|T_i}, \forall i \in [1, N] \right\},$$

where $\mathcal{C}_\sharp^0(\overline{\Omega})$ denotes the space of periodic continuous functions over $\overline{\Omega}$.

Proof. We have

$$\sum_i^N \sum_{A_{ij} \subset \partial T_i} |A_{ij}| [\kappa_r (r_i - r_j)] r_i = \sum_{A_{ij}} |A_{ij}| \kappa_r (r_i - r_j) r_i + \sum_{A_{ij}} |A_{ij}| \kappa_r (r_j - r_i) r_j = \kappa_r \sum_{A_{ij}} |A_{ij}| |r_i - r_j|^2. \tag{6.11}$$

and

$$\begin{aligned} &\sum_i^N \sum_{A_{ij} \subset \partial T_i} |A_{ij}| [\kappa_{\boldsymbol{u}} (\boldsymbol{u}_i - \boldsymbol{u}_j) \cdot \boldsymbol{n}_{ij}] \boldsymbol{n}_{ij} \cdot \boldsymbol{u}_i = \\ &\sum_{A_{ij}} |A_{ij}| [\kappa_{\boldsymbol{u}} (\boldsymbol{u}_i - \boldsymbol{u}_j) \cdot \boldsymbol{n}_{ij}] \boldsymbol{n}_{ij} \cdot \boldsymbol{u}_i + \sum_{A_{ij}} |A_{ij}| [\kappa_{\boldsymbol{u}} (\boldsymbol{u}_j - \boldsymbol{u}_i) \cdot \boldsymbol{n}_{ji}] \boldsymbol{n}_{ji} \cdot \boldsymbol{u}_j = \\ &\kappa_{\boldsymbol{u}} \sum_{A_{ij}} |A_{ij}| |(\boldsymbol{u}_i - \boldsymbol{u}_j) \cdot \boldsymbol{n}_{ij}|^2. \end{aligned} \tag{6.12}$$

Moreover, we also have the energy conservation for the Coriolis force

$$\langle \boldsymbol{u}_h^\perp, \boldsymbol{u}_h \rangle = 0. \tag{6.13}$$

Let us denote $q_h := \begin{pmatrix} r_h \\ \boldsymbol{u}_h \end{pmatrix}$, then by using (6.8),(6.11), (6.12) and (6.13), we obtain

$$\langle L_{\kappa,h} q_h, q_h \rangle = \frac{a_\star \kappa_r}{2} \sum_{A_{ij}} |A_{ij}| |r_i - r_j|^2 + \frac{a_\star \kappa_{\boldsymbol{u}}}{2} \sum_{A_{ij}} |A_{ij}| |(\boldsymbol{u}_i - \boldsymbol{u}_j).\boldsymbol{n}_{ij}|^2. \tag{6.14}$$

We now suppose that $\kappa_{\boldsymbol{u}} \neq 0$ and we will consider the influence of κ_r on the structure of the kernel $KerL_{\kappa,h}$. From (6.14), a necessary condition to be in the kernel is

$$\forall (i,j) : (\boldsymbol{u}_i - \boldsymbol{u}_j) \cdot \boldsymbol{n}_{ij} = 0. \tag{6.15}$$

This obviously implies that

$$\sum_{A_{ij} \subset \partial T_i} |A_{ij}| (\boldsymbol{u}_i - \boldsymbol{u}_j) \cdot \boldsymbol{n}_{ij} = 0.$$

But since $\sum_{A_{ij} \subset \partial T_i} |A_{ij}| \boldsymbol{n}_{ij} = 0$, this also implies that

$$\sum_{A_{ij} \subset \partial T_i} |A_{ij}| (\boldsymbol{u}_i + \boldsymbol{u}_j) \cdot \boldsymbol{n}_{ij} = 0.$$

In the case $\kappa_r \neq 0$, (6.14) also implies that

$$\forall (i,j), r_i = r_j$$

and it follows that r must be a constant: $\exists a \in \mathbb{R}$ such that $r_i = a$ for all $i \in [1,N]$. Then, the equation $L_{\kappa,h} q_h = 0$ reduces to

$$\frac{a_\star}{2|T_i|} \sum_{A_{ij} \subset \partial T_i} |A_{ij}|(r_i + r_j)\boldsymbol{n}_{ij} = -\omega \boldsymbol{u}_i^\perp.$$

But since $r_i = r_j = a$ we obtain

$$2a\frac{a_\star}{2|T_i|} \sum_{A_{ij} \subset \partial T_i} |A_{ij}|\boldsymbol{n}_{ij} = -\omega \boldsymbol{u}_i^\perp \quad \text{and so} \quad \boldsymbol{u}_i^\perp = 0. \tag{6.16}$$

This leads to the conclusion that

$$Ker L_{\kappa_r \neq 0,h} = \left\{ q_h := \begin{pmatrix} r_h \\ \boldsymbol{u}_h \end{pmatrix} \in \mathbb{R}^{3N} \text{such that } \exists a \in \mathbb{R} : r_i = a \text{ and } \boldsymbol{u} = 0 \right\}.$$

In the other case , when $\kappa_r = 0$, we recall from [21, Lemma 5.1] that (6.15) is equivalent to the fact that there exists a periodic Lagrange piecewise P^1 conforming function, denoted by ϕ_h^{L}, defined by its values at the vertices of the mesh, and constants $(b,c) \in \mathbb{R}^2$ such that on each triangle T_i there holds

$$\boldsymbol{u}_i = (b,c)^T + (\nabla \times \phi_h^{\mathrm{L}})_{|T_i}. \tag{6.17}$$

Moreover, from $L_{\kappa,h} q_h = 0$ and (6.15), we get that

$$\frac{a_\star}{|T_i|} \sum_{A_{ij} \subset \partial T_i} |A_{ij}| \frac{(r_i + r_j)}{2} \boldsymbol{n}_{ij} = -\omega \boldsymbol{u}_i^\perp, \tag{6.18}$$

which means that $q_h \in C_h^1$. Moreover, by multiplying (6.18) by $|T_i|$ and summing over all $i \in [1,N]$, we obtain by periodicity that $\sum_{i=1,N} |T_i| u_i = 0$, which implies that $(b,c)^T$ in (6.17) vanishes. Therefore, (6.17) with $b = c = 0$ implies that $q_h \in C_h^2$. In conclusion, we obtain

$$Ker L_{\kappa_r = 0,h} \subset C_h^1 \cap C_h^2$$

and the converse inclusion is also true. □

Remark 6.1. *With the discrete energy defined by*

$$E_h = \sum_i |T_i|(r_i^2 + |\boldsymbol{u}_i|^2),$$

relation (6.14) implies the following L^2-stability result for the collocated Godunov type schemes:

$$\frac{1}{2}\frac{d}{dt} E_h = -\frac{a_\star \kappa_r}{2} \sum_{A_{ij}} |A_{ij}||r_i - r_j|^2 - \frac{a_\star \kappa_{\boldsymbol{u}}}{2} \sum_{A_{ij}} |A_{ij}||(\boldsymbol{u}_i - \boldsymbol{u}_j).\boldsymbol{n}_{ij}|^2 \leq 0.$$

Remark 6.2. *The kernel given by (6.9) clearly shows that the standard Godunov scheme ($\kappa_r \neq 0$) is not able to preserve general geostrophic equilibria. On the other hand, the situation is less clear when the diffusion in the pressure equation vanishes. Indeed, let us consider (6.10). An important property is that*

$$\frac{1}{|T_i|} \sum_{A_{ij} \subset \partial T_i} |A_{ij}| \frac{r_i + r_j}{2} \boldsymbol{n}_{ij}$$

is exactly the gradient on the cell T_i of the non-conforming P^1 *finite element function $\tilde{r}_h^{\mathrm{CR}}$ which is defined by the values $\frac{r_i+r_j}{2}$ at the midpoints of the mesh edges A_{ij}. So C_h^1 tells us that u_h is, up to a multiplicative constant, the curl of the non-conforming P^1 function $\tilde{r}_h^{\mathrm{CR}}$, while C_h^2 tells us that u_h is the curl of a conforming P^1 function ϕ_h^{L}. The equality*

$$(\mathbf{curl}\,\phi_h^{\mathrm{L}})_{|T_i} = \frac{a_\star}{\omega}(\mathbf{curl}\,\tilde{r}_h^{\mathrm{CR}})_{|T_i} \tag{6.19}$$

for all $i \in [1,N]$ provides $2N$ constraints while r has N degrees of freedom but ϕ_h^{CR} only has a number of degrees of freedom equal to the number of vertices of the mesh. In a periodic triangular mesh with no holes, the number of vertices is half of the number of cells, so that (6.19) is actually probably too constraining to admit non trivial solutions.

With the previous remark, it becomes obvious that one possibility is to define the pressure field at the vertices and to replace the gradient of $\tilde{r}_h^{\mathrm{CR}}$ by the gradient of the conforming Lagrange P^1 function defined by these values at the vertices. Of course then, we have to change the pressure evolution equation accordingly. This new staggered scheme is what we explore in the sequel of this chapter.

6.3 Analysis of the semi-discrete staggered schemes

6.3.1 Definition of the discrete operators and the semi-discrete staggered scheme

We first define the discrete version of the gradient and divergence operators. As mentioned above, the discrete gradient will be defined over the cells from values at the vertices. The discrete divergence operator will then be defined such that a discrete integration by parts holds; in this way we shall still be able to prove stability of the scheme by energy estimates.

Let $(r_k)_{k\in[1,N_r]}$ be a discrete scalar field defined by its values at the vertices of the mesh, where N_r is the number of vertices. Let us denote by r_h the globally continuous function that is piecewise P^1 on each cell and defined by the values $(r_k)_{k\in[1,N_r]}$ at the vertices of the mesh. We can define the discrete gradient on the (primal) cells by using the following formula

$$(\nabla_h^T r_h)_i = \frac{1}{|T_i|}\sum_{A_{ij}\subset\partial T_i} |A_{ij}|\frac{r_h(S_{ij})+r_h(N_{ij})}{2}\boldsymbol{n}_{ij}, \tag{6.20}$$

where for any edge A_{ij}, we have denoted by N_{ij} and S_{ij} its extremities (see Figure 6.1(b)). It is easily checked that (6.20) is the gradient of the P^1 Lagrange function r_h.

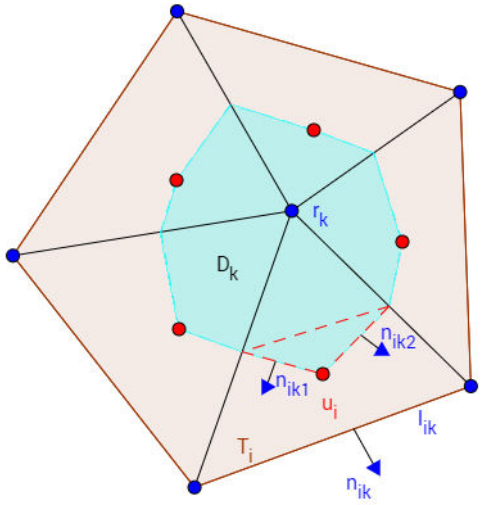
Let $(\boldsymbol{u}_i)_{i\in[1,N]}$ be a discrete vector field defined by its values on the triangular cells. We shall define its divergence on a dual mesh constructed as follows. To each vertex is associated a dual cell obtained by joining the barycenters of the cells which share the vertex to the midpoints of the edges (see Figure 6.1(a)). Let us denote the area of the dual cell D_k by $|D_k|$. Then, we can define the discrete divergence $(\nabla_h \cdot \boldsymbol{u}_h)$ on the dual cell by the following formula

$$(\nabla_h^D \cdot \boldsymbol{u}_h)_k = \frac{1}{|D_k|}\sum_{T_i|T_i\cap D_k\neq\emptyset} \boldsymbol{u}_i \cdot \frac{1}{2} l_{ik}\boldsymbol{n}_{ik}, \tag{6.21}$$

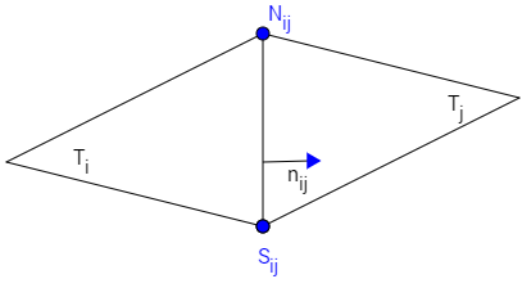
where for a triangle T_i and a vertex k, l_{ik} is the length of the edge of T_i that is opposite to vertex k and $\boldsymbol{n}_{ik}$ the unit normal vector pointing outside T_i on this edge. This edge will be named $\mathcal{A}_{ik}$ below.

Moreover, we can define the discrete curl of the vector field by

$$(\nabla_h^D \times \boldsymbol{u}_h)_k = -(\nabla_h^D \cdot \boldsymbol{u}_h^\perp)_k = -\frac{1}{|D_k|}\sum_{T_i|T_i\cap D_k\neq\emptyset} \boldsymbol{u}_i^\perp \cdot \frac{1}{2} l_{ik}\boldsymbol{n}_{ik}. \tag{6.22}$$



(a) Dual cell



(b) Two neighbor triangles

Figure 6.1: *Staggered scheme.*

On the other hand, we can define the discrete scalar product between $q_h^1 = (r_h^1, \boldsymbol{u}_h^1)$ and $q_h^2 = (r_h^2, \boldsymbol{u}_h^2)$ by

$$\langle q_h^1, q_h^2 \rangle := \langle r_h^1, r_h^2 \rangle_{\mathcal{D}} + \langle \boldsymbol{u}_h^1, \boldsymbol{u}_h^2 \rangle_{\mathcal{P}} := \sum_{k=1}^{N_r} |D_k| r_k^1 r_k^2 + \sum_{i=1}^{N} |T_i| \boldsymbol{u}_i^1 \cdot \boldsymbol{u}_i^2. \tag{6.23}$$

We propose the following semi-discrete staggered scheme applied to the linear wave equation with Coriolis source term.

$$\begin{cases} \frac{d}{dt} r_k(t) + a_\star (\nabla_h^D \cdot \boldsymbol{u}_h)_k - \nu_r \left[\nabla_h^D \cdot \left(\nabla_h^T r_h + \frac{\omega}{a_\star} \boldsymbol{u}_h^\perp \right) \right]_k = 0 \\ \frac{d}{dt} \boldsymbol{u}_i(t) + a_\star (\nabla_h^T r_h)_i - \nu_u (\nabla^{\mathrm{CR}} [\![\boldsymbol{u}_h \cdot \boldsymbol{n}]\!])_i = -\omega \boldsymbol{u}_i^\perp, \end{cases} \tag{6.24}$$

where $\nu_r = \frac{\kappa_r a_\star h}{2}$, with h a typical mesh size and $\nu_u = \frac{\kappa_u a_\star}{2}$ represent the parameters that control the diffusion terms and $(\nabla^{\mathrm{CR}} [\![\boldsymbol{u}_h \cdot \boldsymbol{n}]\!])_i$ is given by the following expression

$$(\nabla^{\mathrm{CR}} [\![\boldsymbol{u}_h \cdot \boldsymbol{n}]\!])_i := \frac{1}{|T_i|} \sum_{A_{ij} \subset \partial T_i} |A_{ij}| [(\boldsymbol{u}_j - \boldsymbol{u}_i) \cdot \boldsymbol{n}_{ij}] \boldsymbol{n}_{ij}. \tag{6.25}$$

We also note that the Low Froude (LF) staggered scheme corresponds to $\nu_r = 0$ and the Apparent Topography (AT) scheme corresponds to $\nu_r > 0$. Moreover, the term defined by (6.25)

is the standard diffusive upwinding term in the velocity equation. The notation ∇^{CR} is here to recall that the right-hand side in (6.25) is the gradient of the non-conforming P^1 function with value $[(\boldsymbol{u}_j - \boldsymbol{u}_i)\cdot \boldsymbol{n}_{ij}]$ at the midpoint of edge A_{ij}.

6.3.2 Properties of the discrete operators

Proposition 6.2. *With the discrete divergence, curl, gradient and scalar product defined respectively by (6.21), (6.22), (6.20) and (6.23), we have the following properties for the semi-discrete staggered scheme (6.24):*

i. Discrete integration by part (energy conservation for the pressure gradient force)

$$\langle \nabla_h^D \cdot \boldsymbol{u}_h, r_h\rangle_{\mathcal{D}} = -\langle \nabla_h^T r_h, \boldsymbol{u}_h\rangle_{\mathcal{P}} \tag{6.26}$$

ii. Energy conservation for the Coriolis force

$$\langle \boldsymbol{u}_h^{\perp}, \boldsymbol{u}_h\rangle_{\mathcal{P}} = 0. \tag{6.27}$$

iii. No vorticity production for the pressure gradient force

$$\nabla_h^D \times \left(\nabla_h^T r_h\right) = 0. \tag{6.28}$$

Proof. Using periodic boundary conditions, we obtain

$$\begin{aligned}
\langle \nabla_h^T r_h, \boldsymbol{u}_h\rangle_{\mathcal{P}} &= \sum_{i=1}^{N} |T_i| (\nabla_h^T r_h)_i \cdot \boldsymbol{u}_i = \sum_{i=1}^{N} \sum_{A_{ij}\subset \partial T_i} \frac{|A_{ij}|}{2} [r_h(S_{ij}) + r_h(N_{ij})] \boldsymbol{n}_{ij}\cdot \boldsymbol{u}_i \\
&= \sum_{k=1}^{N_r} \sum_{T_i | T_i\cap D_k \neq \emptyset} \frac{1}{2} r_k \boldsymbol{u}_i \cdot \sum_{\substack{j | A_{ij}\subset \partial T_i \\ A_{ij}\neq \mathcal{A}_{ik}}} |A_{ij}| \boldsymbol{n}_{ij} = \sum_{k=1}^{N_r} \sum_{T_i | T_i\cap D_k \neq \emptyset} \frac{1}{2} r_k \boldsymbol{u}_i \cdot (-l_{ik}\boldsymbol{n}_{ik}) \\
&= -\sum_{k=1}^{N_r} \left(\sum_{T_i | T_i\cap D_k \neq \emptyset} \boldsymbol{u}_i \cdot \frac{1}{2} l_{ik} \boldsymbol{n}_{ik} \right) r_k = -\langle \nabla_h^D \cdot \boldsymbol{u}_h, r_h\rangle_{\mathcal{D}}.
\end{aligned}$$

which proves point *(i)*.
Point *(ii)* is obvious and now we turn to point *(iii)*; we have

$$\begin{aligned}
\left[\nabla_h^D \times \left(\nabla_h^T r_h\right)\right]_k &= -\left[\nabla_h^D \cdot \left(\nabla_h^T r_h\right)^{\perp}\right]_k = -\frac{1}{|D_k|} \sum_{T_i | T_i\cap D_k \neq \emptyset} \left(\nabla_h^T r_h\right)_i^{\perp} \cdot \frac{1}{2} l_{ik} \boldsymbol{n}_{ik} \\
&= \frac{1}{|D_k|} \sum_{T_i | T_i\cap D_k \neq \emptyset} \left(\nabla_h^T r_h\right)_i \cdot \frac{1}{2} l_{ik} \boldsymbol{n}_{ik}^{\perp} \\
&= \frac{1}{2|D_k|} \sum_{T_i | T_i\cap D_k \neq \emptyset} [r_h(N_{ik}) - r_h(S_{ik})] = 0,
\end{aligned}$$

□

where N_{ik} and S_{ik} are the vertices of edge $\mathcal{A}_{ik}$ oriented such that $\overrightarrow{N_{ik}S_{ik}} = l_{ik}\boldsymbol{n}_{ik}^{\perp}$ and where we have used that, since r_h is a P^1 function

$$\left(\nabla_h^T r_h\right)_i \cdot l_{ik}\boldsymbol{n}_{ik}^{\perp} = \left(\nabla_h^T r_h\right)_i \cdot \overrightarrow{N_{ik}S_{ik}} = \nabla r_h \cdot \overrightarrow{N_{ik}S_{ik}} = r_h(N_{ik}) - r_h(S_{ik}).$$

6.3.3 Evolution of the discrete energy

Lemma 6.2. *With $\nu_r = 0$ and the discrete energy defined with the following expression*

$$E_h(t) = \langle q_h, q_h \rangle = \sum_{k=1}^{N_r} |D_k| r_k^2 + \sum_{i=1}^{N} |T_i| \, |\boldsymbol{u}_i|^2, \tag{6.29}$$

we obtain

$$\frac{d}{dt} E_h(t) \leq 0$$

which means that the Low Froude scheme dissipates the discrete energy.

Proof. We take the scalar product of the staggered scheme (6.24) with $q_h = (r_h, \boldsymbol{u}_h)$ to obtain

$$\frac{1}{2}\frac{d}{dt} E_h(t) + \left\langle \nabla_h^D \cdot \boldsymbol{u}_h, r_h \right\rangle_{\mathcal{D}} + \left\langle \nabla_h^T r_h, \boldsymbol{u}_h \right\rangle_{\mathcal{P}} + \left\langle \boldsymbol{u}_h^{\perp}, \boldsymbol{u}_h \right\rangle_{\mathcal{P}} + \nu_u \left\langle \nabla^{\mathrm{CR}} [\![\boldsymbol{u}_h \cdot \boldsymbol{n}]\!], \boldsymbol{u}_h \right\rangle_{\mathcal{P}}$$

Moreover, with periodic boundary conditions, we also have (see (6.12))

$$-\left\langle \nabla^{\mathrm{CR}} [\![\boldsymbol{u}_h \cdot \boldsymbol{n}]\!], \boldsymbol{u}_h \right\rangle_{\mathcal{P}} = \sum_{A_{ij}} |A_{ij}| |(\boldsymbol{u}_i - \boldsymbol{u}_j).\boldsymbol{n}_{ij}|^2. \tag{6.30}$$

Therefore, using (6.26), (6.27) and (6.30), we get

$$\frac{d}{dt} E_h(t) = -2\nu_u \sum_{A_{ij}} |A_{ij}| |(\boldsymbol{u}_i - \boldsymbol{u}_j).\boldsymbol{n}_{ij}|^2.$$

which means that the energy of the LF scheme is decreasing in time. □

6.3.4 Analysis of the discretized steady-states and their orthogonal subspace

We now define a set of discretized steady-states with staggered variables on triangular meshes by the following expression

$$\mathcal{E}_{\omega \neq 0}^{\triangle} = \left\{ \hat{q}_h = (\hat{r}_h, \hat{\boldsymbol{u}}_h) \in \mathbb{R}^{N_r} \times \mathbb{R}^{2N} : a_\star (\nabla_h^T \hat{r}_h)_i = -\omega \hat{\boldsymbol{u}}_i^{\perp} \right\} \tag{6.31}$$

which is a consistent discretization of the geostrophic equilibrium (6.3). Then we have the following discrete Hodge decomposition:

Lemma 6.3. *The orthogonal space of $\mathcal{E}_{\omega \neq 0}^{\triangle}$ is given by*

$$\mathcal{E}_{\omega \neq 0}^{\triangle, \perp} = \left\{ \tilde{q}_h = (\tilde{r}_h, \tilde{\boldsymbol{u}}_h) \in \mathbb{R}^{N_r} \times \mathbb{R}^{2N} : \quad a_\star (\nabla_h^D \times \tilde{\boldsymbol{u}}_h)_k = \omega \tilde{r}_k \right\}. \tag{6.32}$$

Proof. First of all, we define the set $\mathcal{A}_h$ by

$$\mathcal{A}_h = \left\{ q_h = (r_h, \boldsymbol{u}_h) \in \mathbb{R}^{N_r} \times \mathbb{R}^{2N} : \quad a_\star (\nabla_h^D \times \boldsymbol{u}_h)_k = \omega r_k \right\}.$$

For each $\hat{q}_h = (\hat{r}_h, \hat{\boldsymbol{u}}_h) \in \mathcal{E}_{\omega \neq 0}^{\triangle}$ and arbitrary $\tilde{q}_h = (\tilde{r}_h, \tilde{\boldsymbol{u}}_h) \in \mathbb{R}^{N_r} \times \mathbb{R}^{2N}$, we use the discrete integration by part formula (6.26) to obtain

$$\begin{aligned}
\langle \hat{q}_h, \tilde{q}_h \rangle &= \langle \hat{r}_h, \tilde{r}_h \rangle_{\mathcal{D}} + \langle \hat{\boldsymbol{u}}_h, \tilde{\boldsymbol{u}}_h \rangle_{\mathcal{P}} = \langle \hat{r}_h, \tilde{r}_h \rangle_{\mathcal{D}} + \langle \hat{\boldsymbol{u}}_h^{\perp}, \tilde{\boldsymbol{u}}_h^{\perp} \rangle_{\mathcal{P}} \\
&= \langle \hat{r}_h, \tilde{r}_h \rangle_{\mathcal{D}} - \frac{a_\star}{\omega} \left\langle \nabla_h^T \hat{r}_h, \tilde{\boldsymbol{u}}_h^{\perp} \right\rangle_{\mathcal{P}} = \langle \hat{r}_h, \tilde{r}_h \rangle_{\mathcal{D}} + \frac{a_\star}{\omega} \left\langle \hat{r}_h, \nabla_h^D \cdot \tilde{\boldsymbol{u}}_h^{\perp} \right\rangle_{\mathcal{D}} \\
&= \left\langle \hat{r}_h, \tilde{r}_h - \frac{a_\star}{\omega} \nabla_h^D \times \tilde{\boldsymbol{u}}_h \right\rangle_{\mathcal{D}}.
\end{aligned}$$

Hence, if $\tilde{q}_h \in \mathcal{A}_h$, we obviously have $\langle \hat{q}_h, \tilde{q}_h \rangle = 0$ which leads to $\mathcal{A}_h \subset \mathcal{E}^{\triangle,\perp}_{\omega\neq 0}$. On the other hand, since $\hat{r}_h$ can be arbitrary in $\mathbb{R}^{N_r}$ when $\hat{q}_h \in \mathcal{E}^{\triangle}_{\omega\neq 0}$, then the equality $\left\langle \hat{r}_h, \tilde{r}_h - \frac{a_\star}{\omega}\nabla_h^D \times \tilde{\boldsymbol{u}}_h \right\rangle_{\mathcal{D}} = 0$ for all $\hat{q}_h \in \mathcal{E}^{\triangle}_{\omega\neq 0}$ implies that $\tilde{r}_h - \frac{a_\star}{\omega}\nabla_h^D \times \tilde{\boldsymbol{u}}_h = 0$ and thus $\tilde{q}_h \in \mathcal{A}_h$. It follows that $\mathcal{E}^{\triangle,\perp}_{\omega\neq 0} \subset \mathcal{A}_h$. □

Remark 6.3. *The discrete Hodge decomposition allows us to define the discrete orthogonal projection*

$$\mathbb{P}_h : \left\{ \begin{array}{ccc} \mathbb{R}^{N_r+2N} & \longrightarrow & \mathcal{E}^{\triangle}_{\omega\neq 0} \\ q_h & \longmapsto & \hat{q}_h \end{array} \right.$$

and we can construct $\hat{q}_h$ by what follows.

Let $q_h = (r_h, \boldsymbol{u}_h)$ be given in $\mathbb{R}^{N_r+2N}$. For all $(\hat{p}_h, \hat{\boldsymbol{v}}_h) \in \mathcal{E}^{\triangle}_{\omega\neq 0}$, using orthogonality, we have

$$\langle \hat{r}_h, \hat{p}_h \rangle_{\mathcal{D}} + \langle \hat{\boldsymbol{u}}_h, \hat{\boldsymbol{v}}_h \rangle_{\mathcal{P}} = \langle r_h, \hat{p}_h \rangle_{\mathcal{D}} + \langle \boldsymbol{u}_h, \hat{\boldsymbol{v}}_h \rangle_{\mathcal{P}}.$$

We then use the definition of the discrete steady-states and the discrete integration by part formula to get

$$\langle \hat{r}_h, \hat{p}_h \rangle_{\mathcal{D}} - \left(\frac{a_\star}{\omega}\right)^2 \left\langle \nabla_h^D \cdot \left(\nabla_h^T \hat{r}_h\right), \hat{p}_h \right\rangle_{\mathcal{D}} = \langle r_h, \hat{p}_h \rangle_{\mathcal{D}} - \frac{a_\star}{\omega} \langle \nabla_h^D \times \boldsymbol{u}_h, \hat{p}_h \rangle_{\mathcal{D}}.$$

As a result, since $\hat{p}_h$ can be arbitrary in $\mathbb{R}^{N_r}$, it is possible to find $\hat{r}_h$ by solving the following linear system

$$\hat{r}_k - \left(\frac{a_\star}{\omega}\right)^2 \left[\nabla_h^D \cdot \left(\nabla_h^T \hat{r}_h\right)\right]_k = r_k - \frac{a_\star}{\omega} (\nabla_h^D \times \boldsymbol{u}_h)_k. \tag{6.33}$$

Then, by the definition of the discrete steady-states, the part of the velocity field in $\mathcal{E}^{\triangle}_{\omega\neq 0}$ is given by

$$\hat{\boldsymbol{u}}_i = \frac{a_\star}{\omega} (\nabla_h^T \hat{r}_h)_i^{\perp}.$$

Finally, the orthogonal component is simply given by $\tilde{q}_h = q_h - \hat{q}_h$. Moreover, the linear system (6.33) defines a unique solution since $-\nabla_h^D \cdot$ and ∇_h^T are adjoint operators as shown by the discrete integration by part formula (6.26).

6.3.5 Well-balanced and orthogonality preserving properties

Definition 6.1. *A semi-discrete scheme is said to be well-balanced if*

$$q_h^0 \in \mathcal{E}^{\triangle}_{\omega\neq 0} \quad \Rightarrow \quad \forall t \geq 0, \quad q_h(t) = q_h^0 \in \mathcal{E}^{\triangle}_{\omega\neq 0}.$$

Definition 6.2. *A semi-discrete scheme is said to be orthogonality preserving if*

$$q_h^0 \in \mathcal{E}^{\triangle,\perp}_{\omega\neq 0} \quad \Rightarrow \quad \forall t \geq 0, \quad q_h(t) \in \mathcal{E}^{\triangle,\perp}_{\omega\neq 0}.$$

Lemma 6.4. *We have:*

i. The semi-discrete staggered type scheme (6.24) is a well-balanced scheme in the sense that it can capture the discrete steady state (6.31).

ii. The Low Froude semi-discrete staggered scheme ($\nu_r = 0$) is an orthogonality preserving scheme.

Proof. With the discrete steady state (6.31), the velocity field can be written as

$$\hat{\boldsymbol{u}}_i = \frac{a_\star}{\omega}(\nabla_h^T \hat{r}_h)_i^\perp.$$

Since we have no vorticity production of the gradient term (see (6.28)), we get

$$(\nabla_h^D \cdot \hat{\boldsymbol{u}}_h)_k = \frac{a_\star}{\omega}\left[\nabla_h^D \cdot (\nabla_h^T \hat{r}_h)^\perp\right]_k = -\frac{a_\star}{\omega}\left[\nabla_h \times (\nabla_h^T \hat{r}_h)\right]_k = 0. \tag{6.34}$$

On the other hand, since $\boldsymbol{u}$ is the curl of a discrete piecewise P^1, globally continuous function, then its normal jumps through edges vanishes, as recalled in the lines before (6.17):

$$(\hat{\boldsymbol{u}}_i - \hat{\boldsymbol{u}}_j) \cdot \boldsymbol{n}_{ij} = 0. \tag{6.35}$$

Therefore, the definition of the discrete kernel (6.31), the divergence free property (6.34) and the vanishing of the jumps of the velocity field through the edges A_{ij} (6.35) imply the well-balanced property of the semi-discrete staggered scheme (6.24). This proves Point *(i)*.
In consideration of the orthogonality preserving property, by taking the discrete scalar product of the semi-discrete staggered scheme with the stationary state $\hat{q}_h \in \mathcal{E}^{\triangle}_{\omega\neq 0}$, we obtain

$$\begin{aligned}\left\langle \frac{d}{dt} q_h(t), \hat{q}_h \right\rangle =& - a_\star \langle \nabla_h^D \cdot \boldsymbol{u}_h, \hat{r}_h \rangle_{\mathcal{D}} + \nu_r \left\langle \nabla_h^D \cdot \left(\nabla_h^T r_h + \frac{\omega}{a_\star} \boldsymbol{u}_h^\perp \right), \hat{r}_h \right\rangle_{\mathcal{D}} \\ & - a_\star \langle \nabla_h^T r_h, \hat{\boldsymbol{u}}_h \rangle_{\mathcal{P}} + \nu_u \langle \nabla^{\mathrm{CR}} [\![\boldsymbol{u}_h \cdot \boldsymbol{n}]\!], \hat{\boldsymbol{u}}_h \rangle_{\mathcal{P}} - \omega \langle \boldsymbol{u}_h^\perp, \hat{\boldsymbol{u}}_h \rangle_{\mathcal{P}}.\end{aligned}$$

By using the discrete integration by part formula and (6.34) , we have

$$\langle \nabla_h^T r_h, \hat{\boldsymbol{u}}_h \rangle_{\mathcal{P}} = -\left\langle r_h, \nabla_h^D \cdot \hat{\boldsymbol{u}}_h \right\rangle_{\mathcal{D}} = 0.$$

Moreover, by calculations similar to (6.12), we get, thanks to (6.35):

$$\begin{aligned}-\left\langle \nabla^{\mathrm{CR}} [\![\boldsymbol{u}_h \cdot \boldsymbol{n}]\!], \hat{\boldsymbol{u}}_h \right\rangle_{\mathcal{P}} &= \sum_{A_{ij}} |A_{ij}| [(\boldsymbol{u}_i - \boldsymbol{u}_j) \cdot \boldsymbol{n}_{ij}] \boldsymbol{n}_{ij} \cdot \hat{\boldsymbol{u}}_i + \sum_{A_{ij}} |A_{ij}| [(\boldsymbol{u}_j - \boldsymbol{u}_i) \cdot \boldsymbol{n}_{ji}] \boldsymbol{n}_{ji} \cdot \hat{\boldsymbol{u}}_j \\ &= \sum_{A_{ij}} |A_{ij}| [(\boldsymbol{u}_i - \boldsymbol{u}_j).\boldsymbol{n}_{ij}] [(\hat{\boldsymbol{u}}_i - \hat{\boldsymbol{u}}_j) \cdot \boldsymbol{n}_{ij}] = 0.\end{aligned}$$

Using a final discrete integration by part formula and the fact that $\hat{q}_h \in \mathcal{E}^{\triangle}_{\omega\neq 0}$, we get

$$-a_\star \langle \nabla_h^D \cdot \boldsymbol{u}_h, \hat{r}_h \rangle_{\mathcal{D}} - \omega \langle \boldsymbol{u}_h^\perp, \hat{\boldsymbol{u}}_h \rangle_{\mathcal{P}} = \left\langle a_\star \nabla_h^T \hat{r}_h + \omega \hat{\boldsymbol{u}}_h^\perp, \boldsymbol{u}_h \right\rangle_{\mathcal{P}} = 0.$$

As a result, the condition to ensure the orthogonality preserving property of the semi-discrete staggered scheme is given by

$$\forall \hat{q}_h \in \mathcal{E}^{\triangle}_{\omega\neq 0}, \quad \nu_r \left\langle \nabla_h^D \cdot \left(\nabla_h^T r_h + \frac{\omega}{a_\star} \boldsymbol{u}_h^\perp \right), \hat{r}_h \right\rangle_{\mathcal{D}} = 0.$$

Therefore, the semi-discrete staggered scheme is orthogonal preserving when we have no diffusion on the pressure equation $\nu_r = 0$. □

Remark 6.4. *Although both Low Froude and Apparent Topography staggered schemes can capture the discrete geostrophic equilibrium (6.31), the behavior of each strategy is very different. Let us decompose the numerical solution into two parts*

$$q_h(t) = \hat{q}_h(t) + \tilde{q}_h(t).$$

Since the Low Froude scheme is orthogonality preserving, the orthogonal part $\tilde{q}_h(t)$ does not move into the kernel. As a result, we can ensure that

$$\hat{q}_h(t) = \mathbb{P}_h q_h^0 \ , \ \forall t \geq 0.$$

This means that the kernel part of the numerical solution is always equal to the initial projection $\mathbb{P}_h q_h^0$ which is also the final state of the numerical solution when the orthogonal part is damped to zero by the numerical diffusion of the scheme.
However, when the numerical scheme has some diffusion on the pressure equation like the Apparent Topography scheme, it is impossible to ensure the orthogonality preserving property. As a consequence, the orthogonal part not only damps out, but also partly moves into the kernel. Therefore, the kernel part of this scheme may be changed at each time step until there is no energy left in the orthogonal of the kernel.

6.3.6 Behavior of the solution of the staggered scheme

Lemma 6.5. *Let $q_{\nu,h}(t)$ be the solution of the semi-discrete scheme (6.24). Then, with $\nu_r = 0$, we obtain*

$$\forall C_1 \in \mathbb{R}^+, \ \textit{if } \|q_h^0 - \mathbb{P}_h(q_h^0)\| = C_1 M, \ \textit{then } \|q_{\nu,h}(t) - \mathbb{P}_h(q_h^0)\| \leq C_1 M,$$

which means that the LF scheme is accurate at low Froude number at anytime.

Proof. By linearity, the solution of semi-discrete staggered scheme $q_{\nu,h}(t)$ can be written as

$$q_{\nu,h}(t) = q^a_{\nu,h}(t) + q^b_{\nu,h}(t)$$

where $q^a_{\nu,h}(t)$ and $q^b_{\nu,h}(t)$ are the solution of (6.24) with initial conditions respectively given by

$$q^a_{\nu,h}(0) = \mathbb{P}_h(q_h^0) \quad \text{and} \quad q^b_{\nu,h}(0) = q_h^0 - \mathbb{P}_h(q_h^0).$$

Then, we have

$$\|q_{\nu,h}(t) - \mathbb{P}_h(q_h^0)\| = \|q^a_{\nu,h}(t) + q^b_{\nu,h}(t) - \mathbb{P}_h(q_h^0)\| \leq \|q^a_{\nu,h}(t) - \mathbb{P}_h(q_h^0)\| + \|q^b_{\nu,h}(t)\|.$$

Moreover, when $\nu_r = 0$, the dissipation of the semi-discrete staggered scheme proved in Lemma 6.2 leads to the conclusion that $\|q^b_{\nu,h}(t)\| \leq \|q^b_{\nu,h}(0)\|$. For this reason, the accuracy of the scheme is linked to the behavior of $q^a_{\kappa,h}(t)$. Since the semi-discrete scheme (6.24) is a well-balanced scheme, we obviously have $q^a_{\nu,h}(t) = \mathbb{P}_h(q_h^0)$. Therefore, we obtain

$$\forall t \geq 0, \quad \|q_{\nu,h}(t) - \mathbb{P}_h(q_h^0)\| \leq C_1 M.$$

□

Remark 6.5. *Since it is difficult to prove the dissipation of the energy for the semi-discrete scheme (6.24) when $\nu_r \neq 0$, we do not have enough evidence to conclude that the well-balanced scheme based on the Apparent Topography method is accurate at low Froude number at anytime.*

6.4 Analysis of fully discrete staggered schemes

We present two types of time discretization for the scheme (6.24). The first is a one-step scheme, while the second, inspired by the so-called "Boris push" [56] from plasma physics, splits the velocity rotation due to the Coriolis force from the other effects.

6.4.1 The fully discrete one step scheme

We now introduce two new parameters θ_1 and θ_2 corresponding to the time discretization of the Coriolis source term. To be convenient, we denote

$$\boldsymbol{u}^{\perp,\theta} = \begin{pmatrix} -\theta_1 v^n - (1-\theta_1) v^{n+1} \\ \theta_2 u^n + (1-\theta_2) u^{n+1} \end{pmatrix}.$$

The fully discrete one step staggered scheme can be written as

$$\begin{cases} r_k^{n+1} = r_k^n - a_\star \Delta t (\nabla_h^D \cdot \boldsymbol{u}_h^n)_k + \nu_r \Delta t \left[\nabla_h^D \cdot \left(\nabla_h^T r_h^n + \frac{\omega}{a_\star} (\boldsymbol{u}_h^n)^\perp \right) \right]_k \\ \boldsymbol{u}_i^{n+1} = \boldsymbol{u}_i^n - a_\star \Delta t (\nabla_h^T r_h^n)_i + \nu_u \Delta t \left(\nabla^{\mathrm{CR}} [\![\boldsymbol{u}_h^n \cdot \boldsymbol{n}]\!] \right)_i - \omega \Delta t \boldsymbol{u}_i^{\perp,\theta}. \end{cases} \tag{6.36}$$

Well-balanced scheme

Lemma 6.6. *The fully discrete one step scheme (6.36) is a well-balanced scheme.*

Proof. We now assume that at time $t^n = n\Delta t$, the numerical solution is in the discrete kernel which reads

$$a_\star (\nabla_h^T r_h^n)_i = -\omega (\boldsymbol{u}_i^n)^\perp. \tag{6.37}$$

Then, we will show that the numerical does not change in the next time step:

$$r^{n+1} = r^n, \quad u^{n+1} = u^n \quad \text{and} \quad v^{n+1} = v^n.$$

We note that the state (6.37) also implies that

$$(\nabla_h^D \cdot \boldsymbol{u}_h^n)_k = 0 \quad \text{and} \quad \nabla^{\mathrm{CR}} [\![\boldsymbol{u}_h^n \cdot \boldsymbol{n}]\!]_i = 0. \tag{6.38}$$

Therefore, the pressure equation of the fully discrete one step scheme (6.36) reduces to $r_k^{n+1} = r_k^n$. Moreover, the velocity equation leads to

$$\begin{pmatrix} 1 & -\omega\Delta t(1-\theta_1) \\ \omega\Delta t(1-\theta_2) & 1 \end{pmatrix} \begin{pmatrix} u_i^{n+1} \\ v_i^{n+1} \end{pmatrix} = \begin{pmatrix} 1 & -\omega\Delta t(1-\theta_1) \\ \omega\Delta t(1-\theta_2) & 1 \end{pmatrix} \begin{pmatrix} u_i^n \\ v_i^n \end{pmatrix}$$

which implies that $u_i^{n+1} = u_i^n$ and $v_i^{n+1} = v_i^n$. □

Orthogonality preserving scheme

Let us introduce two new parameters τ_1 and τ_2 used for the time discretization of the divergence field in the pressure equation. We shall denote

$$\boldsymbol{u}^\tau = \begin{pmatrix} \tau_1 u^n + (1-\tau_1) u^{n+1} \\ \tau_2 v^n + (1-\tau_2) v^{n+1} \end{pmatrix}.$$

Then, we specialize to the case $\nu_r = 0$ and introduce the LF-τ scheme which we define by

$$\begin{cases} r_k^{n+1} = r_k^n - a_\star \Delta t (\nabla_h^D \cdot \boldsymbol{u}_h^\tau)_k \\ \boldsymbol{u}_i^{n+1} = \boldsymbol{u}_i^n - a_\star \Delta t (\nabla_h^T r_h^n)_i + \nu_u \Delta t \left(\nabla^{\mathrm{CR}} [\![\boldsymbol{u}_h^n \cdot \boldsymbol{n}]\!] \right)_i - \omega \Delta t \boldsymbol{u}_i^{\perp,\theta}. \end{cases} \tag{6.39}$$

Remark 6.6. *The LF-τ scheme (6.39) is still explicit although the velocity field $\boldsymbol{u}^{n+1}$ appears in the pressure equation. In fact, we can compute the velocity field first and then it is used to compute $\boldsymbol{u}^\tau$ in the pressure equation without having to solve any linear system.*

Lemma 6.7. *The Low Froude-τ scheme (LF-τ) is an orthogonality preserving scheme if*

$$\tau_1 = \theta_2 \quad \textit{and} \quad \tau_2 = \theta_1.$$

which means that the velocity field used in the Coriolis source term in the velocity equation and that used to compute the divergence term in the pressure equation should be the same.

Proof. By taking the product of the fully discrete scheme (6.39) with the steady state $\hat{q}_h \in \mathcal{E}^{\triangle}_{\omega\neq 0}$, using periodic boundary conditions, the discrete integration by part formula and the properties of elements in the discrete kernel, we will obtain

$$\begin{aligned}\left\langle q_h^{n+1}, \hat{q}_h \right\rangle &= -a_\star \Delta t \langle \nabla_h^D \cdot \boldsymbol{u}_h^\tau, \hat{r}_h \rangle_{\mathcal{D}} - \omega \Delta t \langle \boldsymbol{u}_h^{\perp,\theta}, \hat{\boldsymbol{u}}_h \rangle_{\mathcal{P}} \\ &= \left\langle a_\star \Delta t \nabla_h^T \hat{r}_h, \boldsymbol{u}_h^\tau \right\rangle_{\mathcal{P}} + \left\langle \omega \Delta t \hat{\boldsymbol{u}}_h^{\perp}, \boldsymbol{u}_h^\theta \right\rangle_{\mathcal{P}}.\end{aligned}$$

We realize that when $\tau_1 = \theta_2 \quad \text{and} \quad \tau_2 = \theta_1$, we get $\boldsymbol{u}_h^\tau = \boldsymbol{u}_h^\theta$, from which it follows that

$$\left\langle q_h^{n+1}, \hat{q}_h \right\rangle = 0, \forall \hat{q}_h \in \mathcal{E}^{\triangle}_{\omega\neq 0}.$$

Therefore, we conclude that $q_h^{n+1} \in \mathcal{E}^{\triangle,\perp}_{\omega\neq 0}$. □

6.4.2 The fully discrete splitting scheme

Let us define a four-step staggered scheme applied to (6.2) by using a strategy that splits the effect of rotation induced by the Coriolis force from the rest of the effects:

- 1^{st} step ($1-\theta$ velocity advance without rotation):

$$\boldsymbol{u}_i^* = \boldsymbol{u}_i^n - (1-\theta) a_\star \Delta t (\nabla_h^T r_h^n)_i + (1-\theta)\nu_u \Delta t \left(\nabla^{\mathrm{CR}} [\![\boldsymbol{u}_h^n \cdot \boldsymbol{n}]\!]\right)_i \tag{6.40}$$

- 2^{nd} step (velocity full rotation):

$$\boldsymbol{u}^{**} - \boldsymbol{u}^* = -\omega \Delta t \, [\theta \boldsymbol{u}^* + (1-\theta)\boldsymbol{u}^{**}]^\perp \tag{6.41}$$

- 3^{rd} step (θ velocity advance without rotation) :

$$\boldsymbol{u}_i^{n+1} = \boldsymbol{u}_i^{**} - \theta a_\star \Delta t (\nabla_h^T r_h^n)_i + \theta \nu_u \Delta t \left(\nabla^{\mathrm{CR}} [\![\boldsymbol{u}_h^n \cdot \boldsymbol{n}]\!]\right)_i \tag{6.42}$$

- 4^{th} step (pressure update):

$$r_k^{n+1} = r_k^n - a_\star \Delta t \left(\nabla_h^D \cdot [\theta \boldsymbol{u}_h^* + (1-\theta)\boldsymbol{u}_h^{**}]\right)_k + \nu_r \Delta t \left[\nabla_h^D \cdot \left(\nabla_h^T r_h^n + \frac{\omega}{a_\star}(\boldsymbol{u}_h^n)^\perp\right)\right]_k. \tag{6.43}$$

Lemma 6.8. *We have:*

i. The four-step splitting staggered scheme is a well-balanced scheme.

ii. When $\nu_r = 0$, the four-step splitting staggered scheme is an orthogonality preserving scheme.

Proof. We first assume that at time $t^n = n\Delta t$, the numerical solution is in the discrete kernel (6.31). Therefore, we have (6.37) and (6.38). Hence, the first step reduces to

$$\boldsymbol{u}_i^* = \boldsymbol{u}_i^n - (1-\theta)a_\star \Delta t(\nabla_h^T r_h^n)_i \tag{6.44}$$

and the third to

$$\boldsymbol{u}_i^{n+1} = \boldsymbol{u}_i^{**} - \theta a_\star \Delta t(\nabla_h^T r_h^n)_i. \tag{6.45}$$

Using (6.37), the rotation step can be written as

$$\begin{aligned}
\boldsymbol{u}_i^{**} + (1-\theta)\omega\Delta t(\boldsymbol{u}_i^{**})^\perp &= \boldsymbol{u}_i^n - (1-\theta)a_\star\Delta t(\nabla_h^T r_h^n)_i - \theta\omega\Delta t\left[\boldsymbol{u}_i^n - (1-\theta)a_\star\Delta t(\nabla_h^T r_h^n)_i\right]^\perp \\
&= \boldsymbol{u}_i^n + \theta a_\star\Delta t(\nabla_h^T r_h^n)_i + \omega\Delta t\boldsymbol{u}_i^{n,\perp} - \theta\omega\Delta t\left[\boldsymbol{u}_i^n - (1-\theta)a_\star\Delta t(\nabla_h^T r_h^n)_i\right]^\perp \\
&= \left[\boldsymbol{u}_i^n + \theta a_\star\Delta t(\nabla_h^T r_h^n)_i\right] + (1-\theta)\omega\Delta t\left[\boldsymbol{u}_i^n + \theta a_\star\Delta t(\nabla_h^T r_h^n)_i\right]^\perp.
\end{aligned}$$

By a straightforward uniqueness argument, this leads to

$$\boldsymbol{u}_i^{**} = \boldsymbol{u}_i^n + \theta a_\star\Delta t(\nabla_h^T r_h^n)_i. \tag{6.46}$$

Thanks to (6.45), this implies that $\boldsymbol{u}_i^{n+1} = \boldsymbol{u}_i^n$. On the other hand, from (6.44) and (6.46), we easily obtain

$$\left(\nabla_h^D \cdot [\theta\boldsymbol{u}_h^* + (1-\theta)\boldsymbol{u}_h^{**}]\right)_k = \left(\nabla_h^D \cdot \boldsymbol{u}_h^n\right)_k = 0,$$

from which it follows that $r_k^{n+1} = r_k^n$. This proves Point *(i)*.
In consideration of the orthogonality preserving property of the four-step staggered scheme, we assume that at time t^n, the numerical solution verifies $q_h^n \in \mathcal{E}_{\omega\neq 0}^{\triangle,\perp}$ and we will show that at next time step, the numerical solution q_h^{n+1} is still in this subspace. For any state $\hat{q} = (\hat{r}_h, \hat{\boldsymbol{u}}_h) \in \mathcal{E}_{\omega\neq 0}^{\triangle}$, by the discrete integration by part formula, we have

$$\langle \nabla_h^T r_h, \hat{\boldsymbol{u}}_h\rangle_{\mathcal{P}} = -\left\langle r_h, \nabla_h^D \cdot \hat{\boldsymbol{u}}_h\right\rangle_{\mathcal{D}} = 0,$$

and

$$\left\langle \nabla^{\mathrm{CR}}[\![\boldsymbol{u}_h \cdot \boldsymbol{n}]\!], \hat{\boldsymbol{u}}_h\right\rangle_{\mathcal{P}} = -\sum_{A_{ij}} |A_{ij}|[(\boldsymbol{u}_i - \boldsymbol{u}_j).\boldsymbol{n}_{ij}][(\hat{\boldsymbol{u}}_i - \hat{\boldsymbol{u}}_j).\boldsymbol{n}_{ij}] = 0.$$

Hence, from (6.40) and (6.42), we clearly get

$$\langle \boldsymbol{u}_h^*, \hat{\boldsymbol{u}}_h\rangle = \langle \boldsymbol{u}_h^n, \hat{\boldsymbol{u}}_h\rangle \quad \text{and} \quad \langle \boldsymbol{u}_h^{n+1}, \hat{\boldsymbol{u}}_h\rangle = \langle \boldsymbol{u}_h^{**}, \hat{\boldsymbol{u}}_h\rangle$$

Then, the rotation step (6.41) leads to

$$\begin{aligned}
\langle \boldsymbol{u}_h^{n+1}, \hat{\boldsymbol{u}}_h\rangle = \langle \boldsymbol{u}_h^{**}, \hat{\boldsymbol{u}}_h\rangle &= \langle \boldsymbol{u}_h^*, \hat{\boldsymbol{u}}_h\rangle - \omega\Delta t\left\langle [\theta\boldsymbol{u}_h^* + (1-\theta)\boldsymbol{u}_h^{**}]^\perp, \hat{\boldsymbol{u}}_h\right\rangle \\
&= \langle \boldsymbol{u}_h^n, \hat{\boldsymbol{u}}_h\rangle + \omega\Delta t\left\langle [\theta\boldsymbol{u}_h^* + (1-\theta)\boldsymbol{u}_h^{**}], \hat{\boldsymbol{u}}_h^\perp\right\rangle.
\end{aligned} \tag{6.47}$$

On the other hand, when $\nu_r = 0$, the 4^{th} step (6.43) gives us

$$\begin{aligned}
\langle r_h^{n+1}, \hat{r}_h\rangle &= \langle r_h^n, \hat{r}_h\rangle - a_\star\Delta t\left\langle \nabla_h^D \cdot [\theta\boldsymbol{u}_h^* + (1-\theta)\boldsymbol{u}_h^{**}], \hat{r}_h\right\rangle \\
&= \langle r_h^n, \hat{r}_h\rangle + a_\star\Delta t\left\langle [\theta\boldsymbol{u}_h^* + (1-\theta)\boldsymbol{u}_h^{**}], \nabla_h^T\hat{r}_h\right\rangle.
\end{aligned} \tag{6.48}$$

Therefore, gathering (6.47) and (6.48), using the fact that $q_h^n \in \mathcal{E}_{\omega\neq 0}^{\triangle,\perp}$ and the fact that $\hat{q}_h \in \mathcal{E}_{\omega\neq 0}^{\triangle}$, we obtain

$$\langle q_h^{n+1}, \hat{q}_h\rangle = \langle r_h^n, \hat{r}_h\rangle + \langle \boldsymbol{u}_h^n, \hat{\boldsymbol{u}}_h\rangle + \Delta t \left\langle [\theta \boldsymbol{u}_h^* + (1-\theta)\boldsymbol{u}_h^{**}], a_\star \nabla_h^T \hat{r}_h + \omega \hat{\boldsymbol{u}}_h^\perp \right\rangle = 0.$$

In conclusion, we have

$$\langle q_h^{n+1}, \hat{q}_h\rangle = 0\ ,\ \forall \hat{q}_h \in \mathcal{E}_{\omega\neq 0}^{\triangle} \quad \text{which implies} \quad q_h^{n+1} \in \mathcal{E}_{\omega\neq 0}^{\triangle,\perp}.$$

□

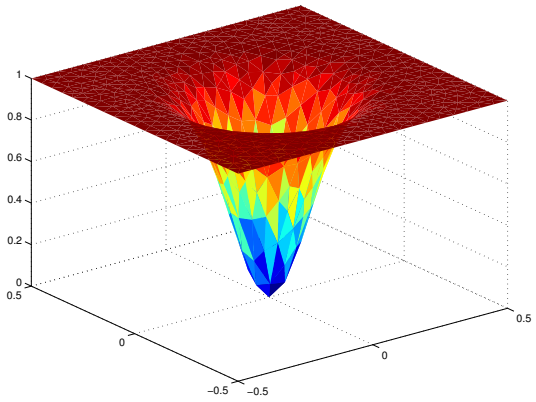
6.5 Numerical test cases

6.5.1 Well-balanced test case

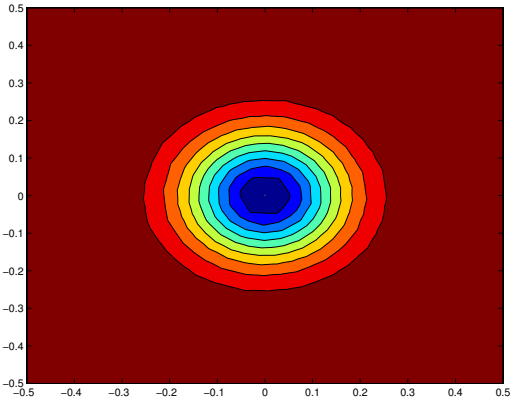
In this test case, we investigate the behavior of the Godunov type schemes with a geostrophic equilibrium as in initial condition. Particularly, we consider the stationary vortex in the square domain $\mathbb{T}^2 = [-0.5, 0.5] \times [-0.5, 0.5]$ with initial pressure r^0 given by

$$r(x, y, t = 0) = 1 - \exp\left[-\left(\frac{3x}{0.5}\right)^2 - \left(\frac{3y}{0.5}\right)^2\right],$$

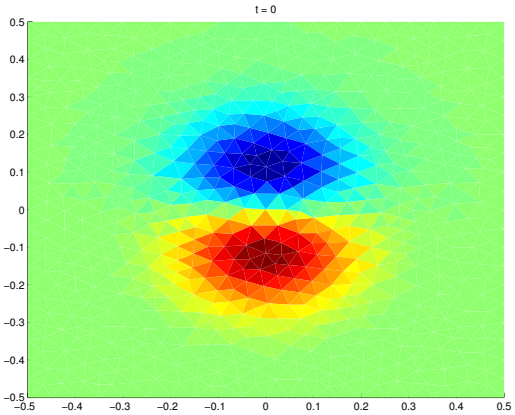
and we construct the discrete initial pressure by interpolating this pressure field at the mesh vertices. Then, we construct the initial velocity field $\boldsymbol{u}^0$ by using the definition of the discrete kernel (6.31) so that we can obtain a discrete stationary state (see Fig. 6.2).
Figure 6.3 indicates that the Classical (C) staggered scheme is not well-balanced since it produces some spurious wave in the orthogonal subspace and also damps the kernel part. On the contrary, the LF and AT staggered schemes are well-balanced schemes. This is because the orthogonal part of those scheme is always equal to zero (Figure 6.3b) during the computation and the kernel part of these scheme is a constant function in time (Figure 6.3a). On the other hand, Figure 6.4 clearly shows that the contour of the pressure of the classical scheme is different from that of the other schemes. This is another evidence to show that the classical scheme is unable to capture the steady state. However, as can be seen and unlike the classical scheme on Cartesian meshes (see [53]), this scheme still preserves the structure of the vortex on triangular mesh. One possible explanation for this property is the influence of the cell geometry. Indeed, the discrete divergence free velocity space is a much better approximation of the continuous one on triangular meshes than on cartesian ones, so that fewer problems are expected. Then, from the numerical point of view, the classical staggered scheme on triangular meshes is similar to the collocated Classical - Divergence Penalization (C-DP) scheme on Cartesian meshes, as introduced in [53] (that scheme also has a satisfactory discrete divergence free velocity space).



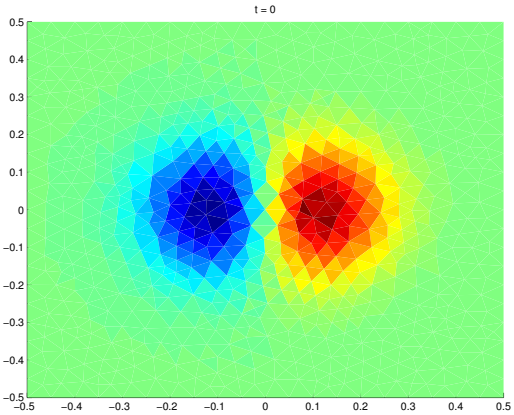
(a) Pressure $r(x,y,t)$ at $t=0$



(b) Contours of $r(x,y,t)$ at $t=0$

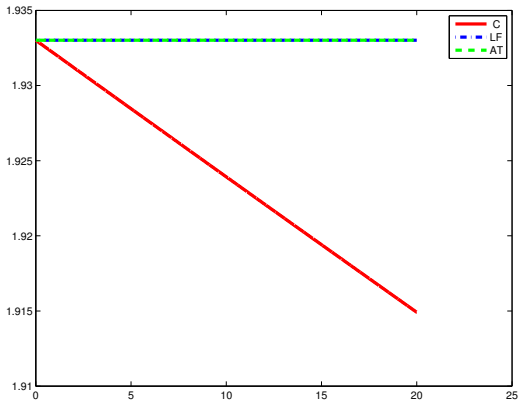


(c) Contours of $u(x,y,t)$ at $t=0$

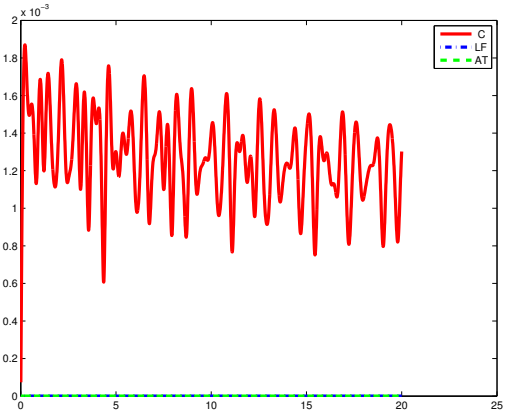


(d) Contours of $v(x,y,t)$ at $t=0$

Figure 6.2: *Stationary vortex as initial condition.*

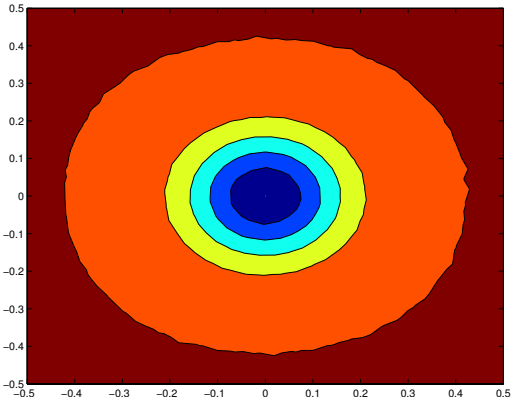


(a) $\|\hat{q}\|$

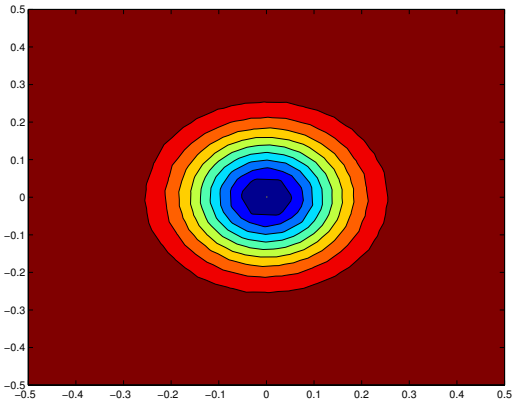


(b) $\|\tilde{q}\|$

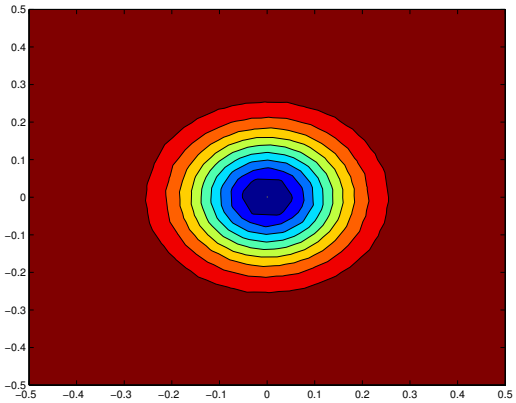
Figure 6.3: *Vortex test case: evolution of the kernel and orthogonal parts.*



(a) Classical scheme



(b) LF scheme



(c) AT scheme

Figure 6.4: *Pressure contours* $r(x,y,t)$ *at time* $t = 20$ *obtained from staggered type schemes.*

6.5.2 Orthogonality preserving test case

In this test case, we consider periodic boundary conditions and an initial vector field given by

$$u(x,y,t=0)=\frac{1}{2}\exp\left[-\left(\frac{4x}{0.4}\right)^2-\left(\frac{4y}{0.8}\right)^2\right] \text{ and } v(x,y,t=0)=\frac{1}{2}\exp\left[-\left(\frac{4x}{0.8}\right)^2-\left(\frac{4y}{0.4}\right)^2\right]$$

in the domain $\mathbb{T}^2=[-0.5,0.5]\times[-0.5,0.5]$ and we construct the discrete initial velocity by interpolating this velocity field at the cell centers. Then the initial pressure $r(x,y,t=0)$ is constructed at the mesh vertices by using the definition of the discrete orthogonal subspace (6.32). In all cases, we choose $\theta_1=\theta_2=\frac{1}{2}$ for the time discretization of the Coriolis force i (6.36). Figure 6.5a and 6.5c show that the classical, Apparent Topography and even Low Froude schemes are not orthogonality preserving since the kernel components of these schemes are not equal to zero. Moreover, since the kernel part is updated at each time step, it is not a constant in time. Only the LF-τ scheme with $\tau_1=\tau_2=\frac{1}{2}$ is able to preserve the orthogonal kernel since this strategy does not create waves in the incompressible subspace during the computational process, as proved by Lemma 6.7. Figure 6.5b shows that the damping rate of the orthogonal parts of the Classical and Apparent Topography schemes is larger than that of the Low Froude scheme. On the other hand, the damping rate of the orthogonal part of the LF-τ scheme shown in Figure 6.5d indicates that when the scheme gets more implicit for the divergence velocity field on the pressure equation, then the damping of the numerical schemes increases. Although the classical and Apparent Topography schemes are not orthogonality preserving, these strategies on triangular grid create waves with much smaller amplitudes in the kernel than that on the Cartesian grid (see [53]).

6.5.3 Accuracy at low Froude number test case

We now consider an initial condition close to the discrete kernel, up to a perturbation of size M. This initial condition is simply given by

$$q_h^0=\hat{q}_h^0+M\frac{\tilde{q}_h^0}{\|\tilde{q}_h^0\|},$$

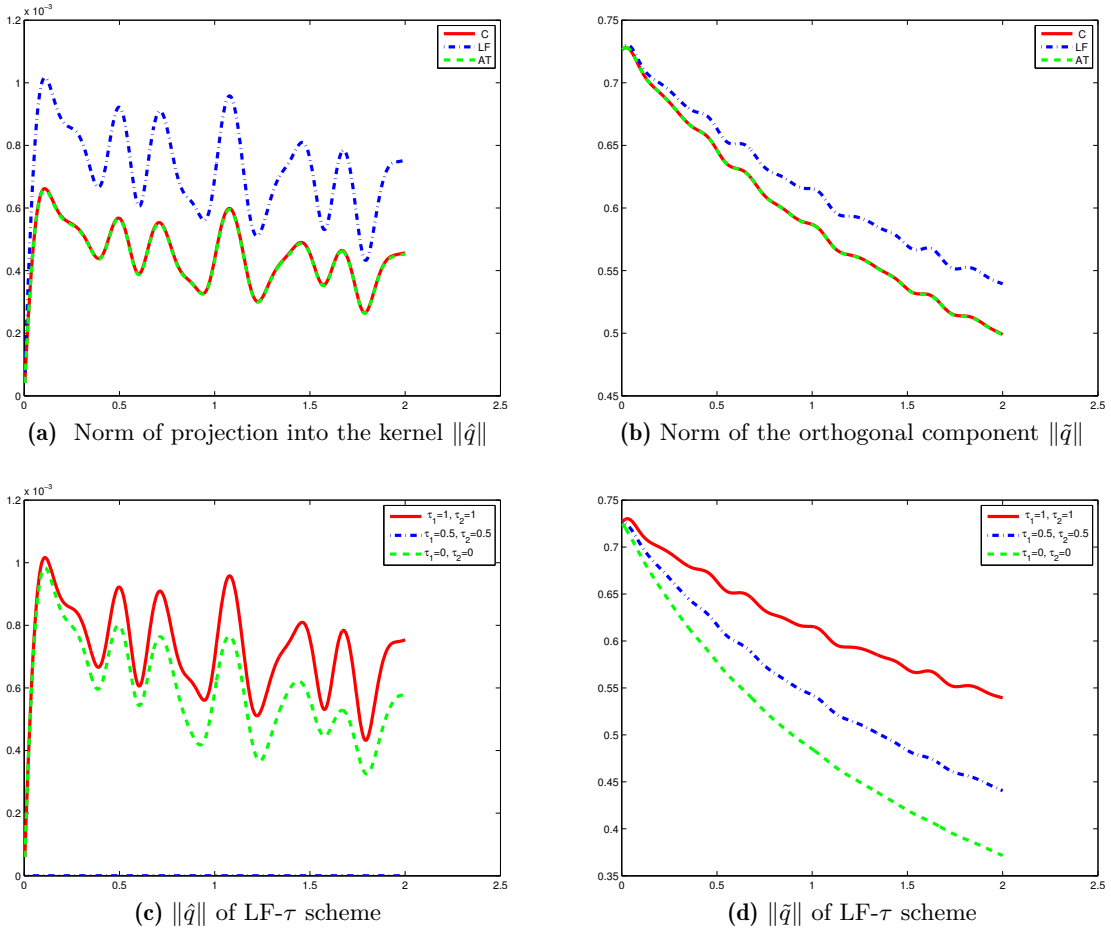
where $\hat{q}_h^0$ stands for the kernel part given in Section 6.5.1 and $\tilde{q}_h^0$ is the orthogonal part considered in Section 6.5.2.

Figure 6.6b and 6.6c indicate that the classical scheme is not accurate at low Froude number because the norm of $\|q(t)-\mathbb{P}q^0\|$ does not depend on the parameter M. By contrast, the proposed LF and AT schemes are accurate at low Froude number because the norm of the wave in the orthogonal space remains of order $\mathcal{O}(M)$ (see Figure 6.6d and 6.6e). On the other hand, we can observe in Figure 6.6a that the norm of the orthogonal component of the solution of the classical scheme globally decreases in time, while the total deviation is an increasing function. This implies that the solution of the classical scheme will tend in long time to a stationary solution in the wrong discrete kernel, which will be different from the projection of the initial condition.

6.5.4 Circular dam-break test case

In this test case, we consider the initial condition which is given by

$$\begin{cases} r(x,y,t=0)=\begin{cases}2, & \text{if} \quad x^2+y^2\leq 1\\ 1, & \text{if} \quad x^2+y^2>1.\end{cases}\\ u(x,,y,t=0)=0,\\ v(x,,y,t=0)=0.\end{cases}$$



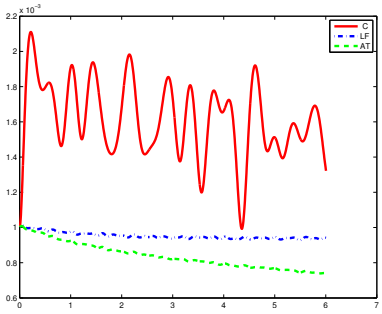
(a) Norm of projection into the kernel $\|\hat{q}\|$

(b) Norm of the orthogonal component $\|\tilde{q}\|$

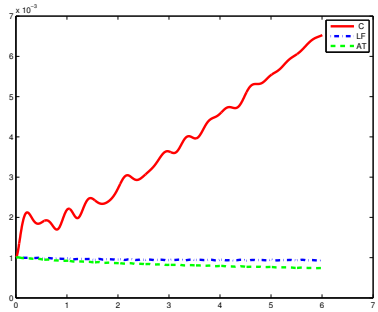
(c) $\|\hat{q}\|$ of LF-τ scheme

(d) $\|\tilde{q}\|$ of LF-τ scheme

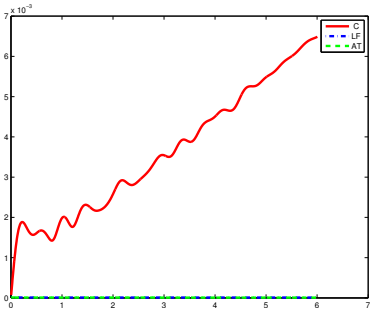
Figure 6.5: *Orthogonality preserving test case: evolution of the kernel and orthogonal parts with* $\theta_1 = \theta_2 = \frac{1}{2}$ *for the time discretization of the Coriolis force.*



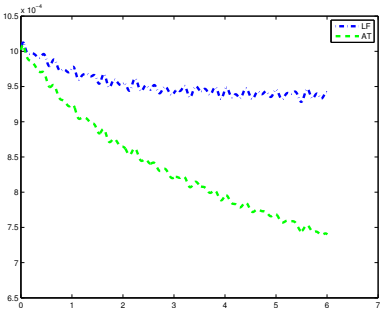
(a) $\|q(t) - \mathbb{P}q\|$ with $M = 10^{-3}$



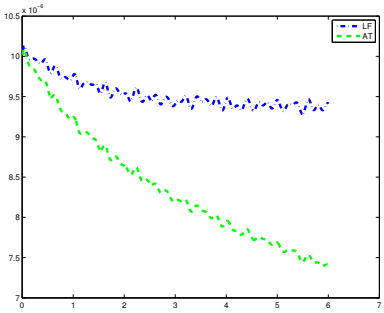
(b) $\|q(t) - \mathbb{P}q^0\|$ with $M = 10^{-3}$



(c) $\|q(t) - \mathbb{P}q^0\|$ with $M = 10^{-5}$



(d) $\|q(t) - \mathbb{P}q^0\|$ with $M = 10^{-3}$



(e) $\|q(t) - \mathbb{P}q^0\|$ with $M = 10^{-5}$

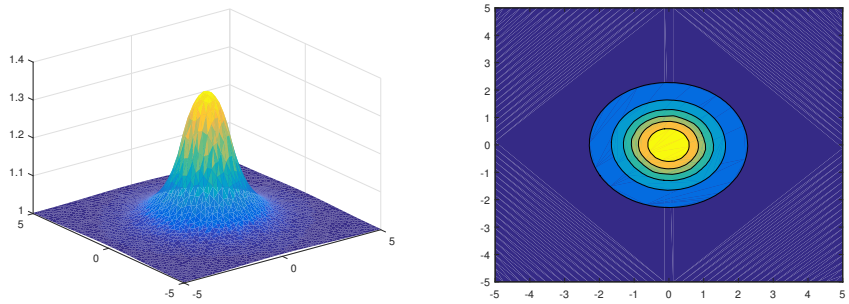
Figure 6.6: *Evolution of the orthogonal component and deviation from the initial condition, when the initial condition is close to the discrete kernel.*

with periodic boundary condition, $a_\star = 1$ and the domain $[-5,5] \times [-5,5]$. Figure 6.7 presents the projection into the kernel of the initial pressure field.

Figure 6.8 presents the evolution of the pressure when using the classical scheme with one initial condition which is very far from the geostrophic equilibrium. This figure clearly shows that in long time, the solution of the classical scheme tends to a trivial steady sate which consists of only a constant pressure field. This is an evidence to show that, although the classical scheme on triangular grids behaves much better than that on Cartesian grids in the previous test cases, this scheme eventually tends to the wrong kernel. On the contrary, the Low Froude scheme pressure solution presented in Figure 6.9 tends to the geostrophic equilibrium, since the final state is similar to projection $\mathbb{P}q^0$ of the initial condition, as shown in Figure 6.7.

On the other hand, Figure 6.10 clearly shows that the kernel part of the classical scheme is damped during the simulation while it is nearly a constant with the LF or AT schemes. Moreover, the distance between the numerical solution of the classical scheme and the initial projection (total deviation) increases in time, while it tends to zero with the other schemes. This is another evidence to conclude that the classical scheme will tend to the wrong kernel, while the other

schemes have solutions that, as expected, tend to the correct geostrophic equilibrium.



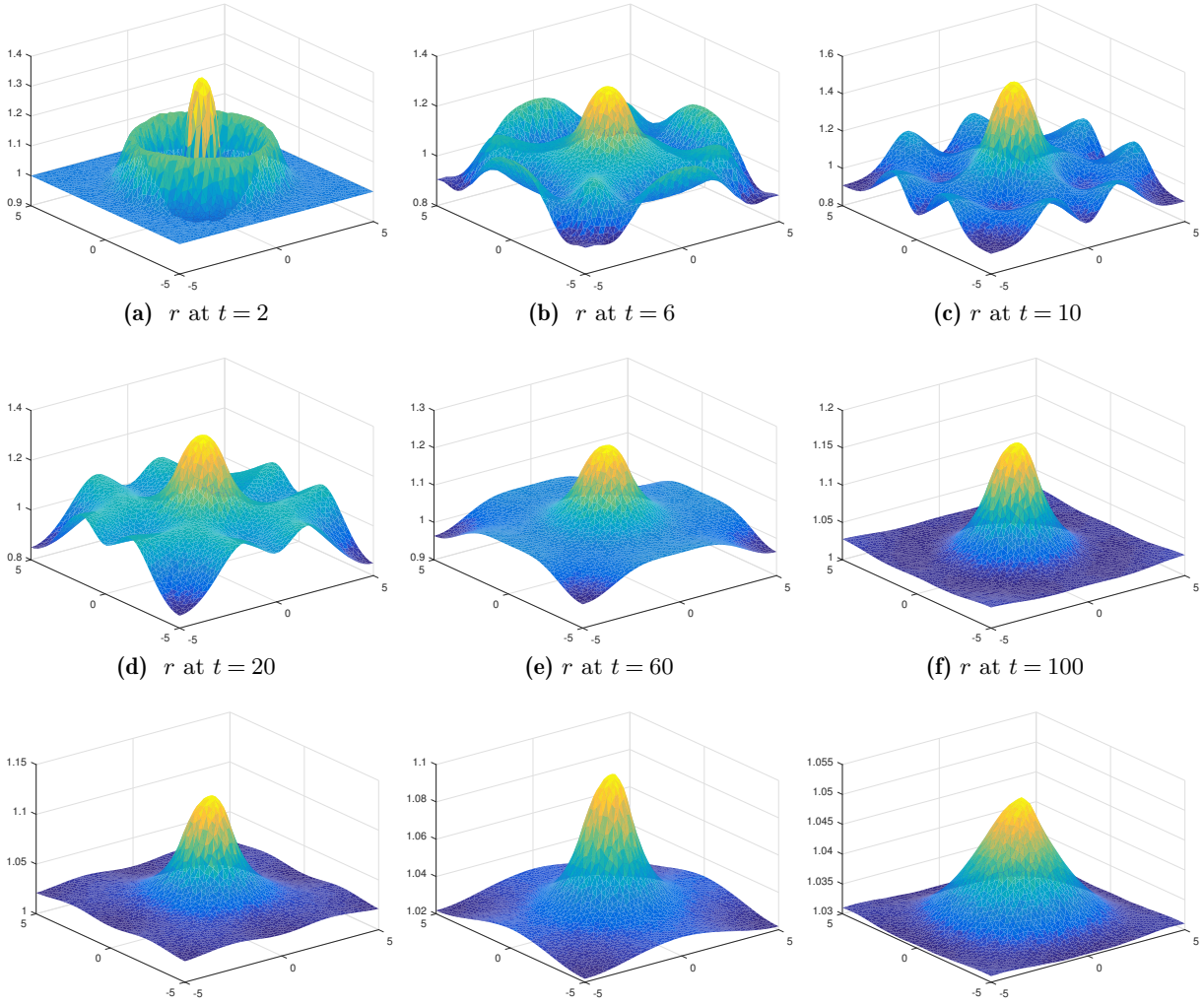
(a) $\hat{r}(x,y)$ at $t=0$ **(b)** Contour of $\hat{r}(x,y)$ at $t=0$

Figure 6.7: *The initial projection of* $r(x,y)$.

6.6 Conclusion

In this work, we explain the disadvantage of the collocated Godunov scheme applied to the linear wave equation with Coriolis force on triangular meshes. Then, we propose new staggered type schemes that are accurate for the simulation of flows near the geostrophic equilibrium. The construction of these schemes is based on the adaptation of the Low Froude and Apparent Topography strategies on triangular grids. The time discretization leads to two strategies: the one step and the four-step splitting schemes. Theoretical analysis and numerical results show that the new schemes are well-balanced. On the other hand, unlike the Apparent Topography method, the Low Froude one step or spitting scheme is orthogonality preserving under an appropriate discretization in time.

Future works will be dedicated to the optimal time step choice of these schemes and the extensions to the nonlinear shallow water equations.

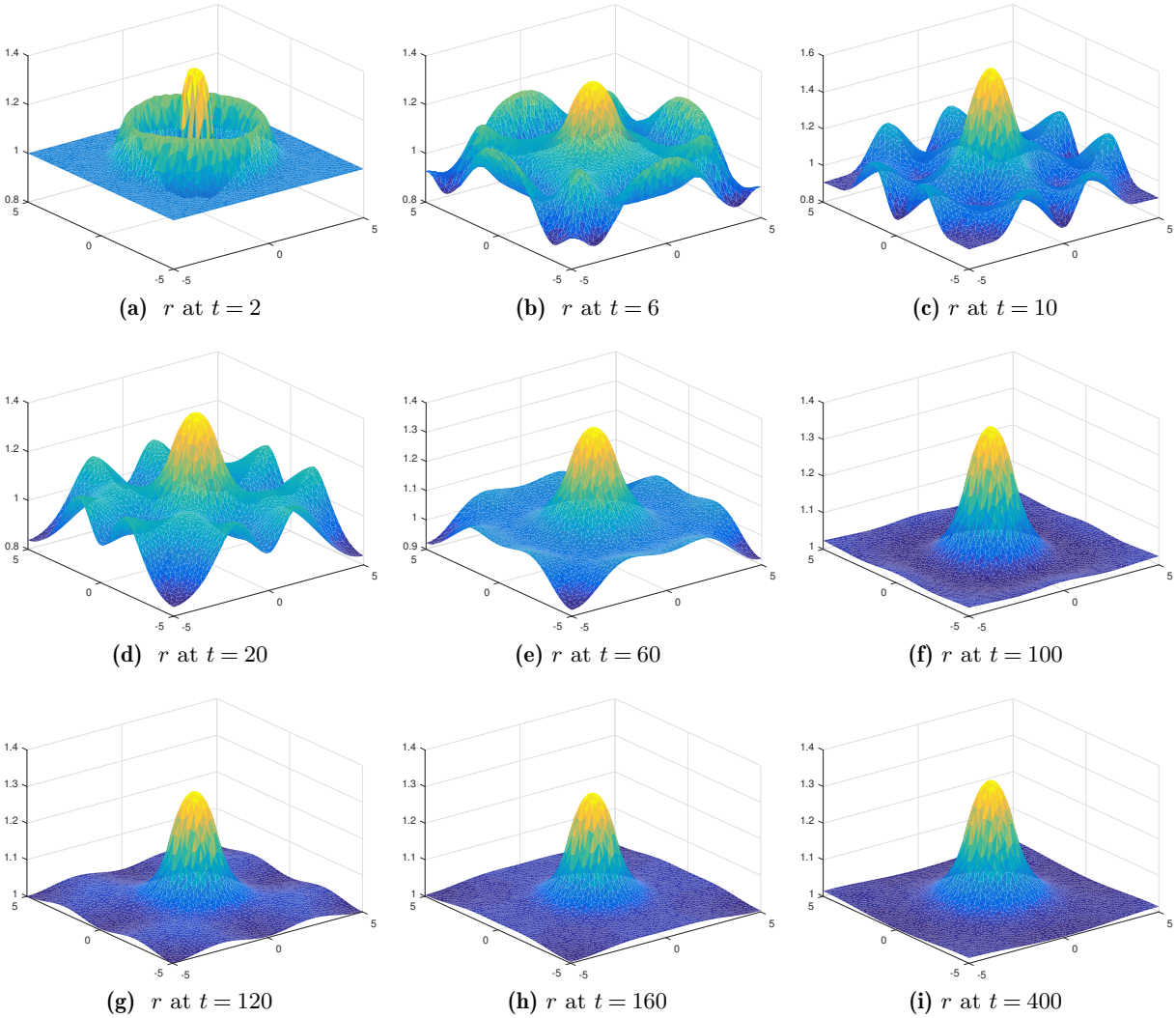


(a) r at $t = 2$ **(b)** r at $t = 6$ **(c)** r at $t = 10$

(d) r at $t = 20$ **(e)** r at $t = 60$ **(f)** r at $t = 100$

(g) r at $t = 120$ **(h)** r at $t = 160$ **(i)** r at $t = 400$

Figure 6.8: *The pressure solution $r(x,y,t)$ of the Classical staggered scheme.*

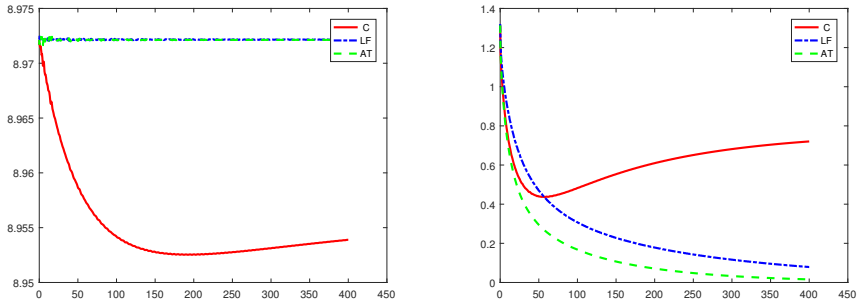


(a) r at $t = 2$ **(b)** r at $t = 6$ **(c)** r at $t = 10$

(d) r at $t = 20$ **(e)** r at $t = 60$ **(f)** r at $t = 100$

(g) r at $t = 120$ **(h)** r at $t = 160$ **(i)** r at $t = 400$

Figure 6.9: *The pressure solution $r(x, y, t)$ of the Low Froude staggered scheme.*



(a) $\|\mathbb{P}q(t)\|$ **(b)** $\|q(t) - \mathbb{P}q^0\|$

Figure 6.10: *The evolution of the kernel part and total deviation from the initial condition.*

Part III

Analysis of numerical schemes for non-linear shallow water equations with Coriolis force

CHAPTER 7

Godunov type schemes for nonlinear rotating shallow water equation

You cannot have a positive life
and a negative mind.

Joyce Meyer.

Abstract

The shallow water equations, a simplified model from Euler or Navier-Stokes equations, can be used to model many phenomena in geophysical fluid mechanics. At large scale phenomena, the Coriolis force, due to the Earth's rotation, plays an important role and the atmospheric or oceanic circulations are frequently observed to take place near the geostrophic equilibrium, which is the balance between the pressure gradient and the Coriolis force. The analysis and the numerical preservation of trivial equilibria with zero velocity (i.e. lake at rest equilibrium) has been studied in the last fifteenth years and the study of non trivial equilibria with non-zero velocity receives now a great attention but is still an open question. We are interested in designing a modified Godunov scheme that captures the discrete version of this geostrophic equilibrium or that can be accurate around this state with acceptable small errors, in order to improve the accuracy of the classical Godunov scheme significantly.

Chapter content

7.1 Introduction

To model complex systems such as the atmosphere or an ocean, it is very common to use the shallow water equations which is a simplified system obtained from the fully compressible Euler equations in the case where the horizontal length scale is much larger than the vertical one, but still retains some main characteristics of the original system. For instance, they share very similar conservation laws. In the atmospheric and ocean modeling, the typical length scales are hundreds of kilometers and it is essential to consider the geometry of the earth and its rotation as source terms for the shallow water equations since they become important at that scale. Moreover, the observed phenomena usually take place around the so called geostrophic equilibrium which is the balance between the pressure gradient and the Coriolis force. In order to study the rotating shallow water model, we introduce the two-dimensional system of partial differential equations

$$\begin{cases} \partial_t h + \nabla \cdot (h\mathbf{u}) = 0, & \text{(7.1a)} \\ \partial_t (h\mathbf{u}) + \nabla \cdot (h\mathbf{u} \otimes \mathbf{u}) + \nabla \left(g \dfrac{h^2}{2} \right) = -gh\nabla b - h\Omega \mathbf{u}^{\perp}, & \text{(7.1b)} \end{cases}$$

where h and $\mathbf{u} = (u, v)$ are functions of time $t > 0$ and space $(x, y) \in \mathbb{R}^2$. These variables denote respectively the vertical height of the water and the horizontal velocity. In this model, the Coriolis parameter Ω stands for the angular velocity and $\mathbf{u}^{\perp} = (-v, u)$ is the orthogonal velocity. The first numerical method which can be applied to this model is the finite difference method. We mention [36, 57] and references therein for this approach. One common advantage of this method is its simplicity of implementation and the fact that it produces results with good accuracy for regions with smooth solutions. However, since the rotating shallow water system admits shock waves [58, 59], in these regions with non-smooth solutions, the finite difference scheme usually introduces unphysical oscillations.

It is crucially important for numerical schemes to capture exactly or at least very accurately some particular solutions of system (7.1) at the discrete level. Without Coriolis source term ($\Omega = 0$), the above system (7.1) reduces to the model with topography, and the observations are small perturbations around the lake at rest equilibrium which is given by

$$\mathbf{u} = 0 \quad \text{and} \quad h + b = \text{cst}. \tag{7.2}$$

Among numerical strategies that are able to preserve the particular solution (7.2), we mention the so called well-balanced schemes, see, e.g [5, 6], and we refer to [9, 10] for schemes that can handle more general one dimensional moving equilibriums ($hu = \text{cst}$ and $\frac{|u|^2}{2} + g(h + b) = \text{cst}$). In the case with Coriolis source term, it is a challenge to design schemes able to capture the nontrivial geostrophic equilibrium given by

$$\nabla \cdot \mathbf{u} = 0 \quad \text{and} \quad g\nabla h = -\Omega \mathbf{u}^{\perp}. \tag{7.3}$$

In [13], Bouchut et al. introduced a technique named Apparent Topography method to obtain a well-balanced scheme with Coriolis force in the one-dimensional case. The idea is to modify the numerical diffusion introduced by standard schemes in the pressure equation so that it applies only on states that do not verify the geostrophic equilibrium. This method is then extended to the two dimensional case on Cartesian grids in [14]. However, the 2D geostrophic equilibrium is more complex in 2D than in 1D, because it implies that the velocity field is divergence free, while standard schemes also fail to maintain divergence free velocities. As a result, it was shown [53] by studying the linear wave equation that it is necessary to modify not only the diffusion on the pressure equation, but also on the velocity equations to design schemes that are able to capture discrete equivalents of the nontrivial steady state (7.3). This gives rise to strategies combining

the Apparent Topography method with the Divergence Penalisation or Low Froude strategies introduced in [20].

In this work, we extend the results obtained for the linear wave equation in [53] to the nonlinear rotating shallow water system. The outline of this work is as follows: in Section 7.2, we investigate the properties of All Froude type schemes for linear wave equations with Coriolis source term. In particular, we focus on the accuracy around geostrophic equilibriums at low Froude number. We then explain in Section 7.3 how to use the combination of the Apparent Topography method with other corrections such as the All Froude and Divergence Penalisation techniques in the nonlinear case. Moreover, the time discretization of the Coriolis source term is also investigated in this section. Many numerical results are then shown in Section 7.4 to illustrate the fact that the proposed scheme is more accurate than the classical ones.

7.2 Behavior of All Froude type schemes applied to the linear wave equation with Coriolis source term

By analyzing the discrete kernel of the modified equation associated to the discretization of the linear acoustic operator by the classical Godunov scheme, the work in [53] shows that it is essential to modify the numerical diffusion terms in both pressure and velocity equations since they are responsible for the inaccuracy problems encountered by standard Godunov type schemes. That work also proposes some numerical strategies that combine the Apparent Topography method presented in [13], with the Divergence Penalization and Low Froude methods inspired by [20, 37]. However, for stability reasons, the Low Froude strategy may not be a good option in the nonlinear case since it is proved in [22] to be unstable when used on the standard Euler system with no Coriolis source term. Therefore, in this section we will replace the Low Froude strategy by the All Froude one, as advocated in [22]. To investigate the effect of the All Froude strategy, let us start with the compact form of the modified equation of the various schemes which is given by

$$\begin{cases} \partial_t q + \mathcal{L} q = 0, & \text{(7.4a)} \\ q(t=0,\boldsymbol{x}) = q^0(\boldsymbol{x}). & \text{(7.4b)} \end{cases}$$

The spatial operator is defined by $\mathcal{L} = L_\omega - \mathcal{B}_{\kappa,\eta}$, with

$$L_\omega q = \begin{pmatrix} a_\star \nabla \cdot \boldsymbol{u} \\ a_\star \nabla r + \omega \boldsymbol{u}^\perp \end{pmatrix}$$

and

$$\mathcal{B}_{\kappa,\eta} q = \begin{pmatrix} \frac{\kappa_r^x a_\star \Delta x}{2} \frac{\partial^2 r}{\partial x^2} + \frac{\kappa_r^y a_\star \Delta y}{2} \frac{\partial^2 r}{\partial y^2} \\ \frac{\kappa_u a_\star \Delta x}{2} \frac{\partial^2 u}{\partial x^2} \\ \frac{\kappa_v a_\star \Delta y}{2} \frac{\partial^2 v}{\partial y^2} \end{pmatrix} + \begin{pmatrix} -\frac{\eta_r^x a_\star \Delta x}{2} \frac{\omega}{a_\star} \frac{\partial v}{\partial x} + \frac{\eta_r^y a_\star \Delta y}{2} \frac{\omega}{a_\star} \frac{\partial u}{\partial y} \\ \frac{\eta_u a_\star \Delta x}{2} \frac{\partial^2 v}{\partial x \partial y} \\ \frac{\eta_v a_\star \Delta y}{2} \frac{\partial^2 u}{\partial y \partial x} \end{pmatrix}. \tag{7.5}$$

The Classical (C) Godunov scheme uses $\kappa_x^r = \kappa_y^r = \kappa_u = \kappa_v = 1$ and $\eta_r^x = \eta_r^y = \eta_u = \eta_v = 0$. The Apparent Topography (AT) strategy uses $\eta_r^x = \kappa_r^x$ and $\eta_r^y = \kappa_r^y$ to typically transform the diffusion operator in the pressure equation from $-\frac{1}{2} a_\star h \Delta r$ into $-\frac{1}{2} h \nabla \cdot (a_\star \nabla r + \omega \boldsymbol{u}^\perp)$, where h is the mesh step. The Divergence Penalization (DP) strategy uses $\eta_u = \kappa_u$ and $\eta_v = \kappa_v$ to transform the $-\frac{1}{2} a_\star h (\partial_{xx} u, \partial_{xx} v)^T$ diffusion term in the velocity equation into $-\frac{1}{2} a_\star h \nabla (\nabla \cdot \boldsymbol{u})$. The All Froude (AF) strategy can be applied to the pressure *or* to the velocity equation (but not to both equations at the same time for stability reasons) and amounts to replace κ_r^x and κ_r^y *or* κ_u and κ_v by $\mathcal{O}(M)$ functions, instead of 0 in the Low Froude strategy.

AF schemes	κ_r	κ_u	κ_v	η_r	η_u	η_v
AT-AF	$\mathcal{O}(1)$	$\mathcal{O}(M)$	$\mathcal{O}(M)$	κ_r	0	0
C-AF	$\mathcal{O}(1)$	$\mathcal{O}(M)$	$\mathcal{O}(M)$	0	0	0
AF-DP	$\mathcal{O}(M)$	$\mathcal{O}(1)$	$\mathcal{O}(1)$	0	κ_u	κ_v
AF-C	$\mathcal{O}(M)$	$\mathcal{O}(1)$	$\mathcal{O}(1)$	0	0	0

Table 7.1: *Parameters of All Froude type schemes.*

The choices of parameters in (7.4) corresponding to the All Froude strategies are summarised in Table 7.1.

Let us note that the behavior of the All Froude schemes are similar to the corresponding Low Froude schemes in the linear case when $M \ll 1$. Although the AT-AF and AF-DP do not have a correct kernel, their behavior around the geostrophic equilibrium remains satisfactory. Particularly, we obtain

Lemma 7.1. *The solution $q_{\kappa,\eta}$ of the modified equation for the AF-DP parameters is accurate at low Froude number locally in time.*

Proof. Let $\mathbb{P}$ be the orthogonal projector on the sets of states that verify $L_\omega q = 0$, and let $q^a_{\kappa,\eta}(t)$ and $q^b_{\kappa,\eta}(t)$ be the solutions of (7.4) with initial conditions respectively given by $\mathbb{P}q^0(x)$ and $q^0(x) - \mathbb{P}q^0(x)$. Then, by linearity, the solution of (7.4) is simply given by

$$q_{\kappa,\eta}(t) = q^a_{\kappa,\eta}(t) + q^b_{\kappa,\eta}(t).$$

Moreover, the dissipation of energy by the AF-DP scheme leads to

$$\|q^b_{\kappa,\eta}(t)\| \leq \|q^b_{\kappa,\eta}(0)\| = \|q^0 - \mathbb{P}q^0\|. \tag{7.6}$$

We also notice that the triangular inequality leads us to

$$\|q_{\kappa,\eta}(t) - \mathbb{P}q^0\| = \|q^a_{\kappa,\eta}(t) + q^b_{\kappa,\eta}(t) - \mathbb{P}q^0\| \leq \|q^a_{\kappa,\eta}(t) - \mathbb{P}q^0\| + \|q^b_{\kappa,\eta}(t)\|. \tag{7.7}$$

We now try to find a bound for the quantity $\|q^a_{\kappa,\eta}(t) - \mathbb{P}q^0\|$. By using the fact that $\mathbb{P}q^0 \in KerL_\omega$, we have $L_\omega(\mathbb{P}q^0) = 0$. This implies that

$$\partial_t(q^a_{\kappa,\eta}(t) - \mathbb{P}q^0) + L_\omega(q^a_{\kappa,\eta}(t) - \mathbb{P}q^0) = B_{\kappa,\eta}(q^a_{\kappa,\eta}(t) - \mathbb{P}q^0) + B_{\kappa,\eta}\mathbb{P}q^0.$$

By multiplying the above equation with $q^a_{\kappa,\eta}(t) - \mathbb{P}q^0$ and integrating over $\mathbb{T}_2$, we will get

$$\langle \partial_t(q^a_{\kappa,\eta}(t) - \mathbb{P}q^0), q^a_{\kappa,\eta}(t) - \mathbb{P}q^0\rangle = \langle B_{\kappa,\eta}(q^a_{\kappa,\eta}(t) - \mathbb{P}q^0), q^a_{\kappa,\eta}(t) - \mathbb{P}q^0\rangle + \langle B_{\kappa,\eta}\mathbb{P}q^0, q^a_{\kappa,\eta}(t) - \mathbb{P}q^0\rangle.$$

Moreover, we also have $\langle B_{\kappa,\eta}(q^a_{\kappa,\eta}(t) - \mathbb{P}q^0), q^a_{\kappa,\eta}(t) - \mathbb{P}q^0\rangle \leq 0$. Thus, we obtain

$$\frac{1}{2}\frac{d}{dt}\|q^a_{\kappa,\eta}(t) - \mathbb{P}q^0\|^2_{L^2} \leq \left|\langle B_{\kappa,\eta}\mathbb{P}q^0, q^a_{\kappa,\eta}(t) - \mathbb{P}q^0\rangle\right| \leq \|B_{\kappa,\eta}\mathbb{P}q^0\|_{L^2}\|q^a_{\kappa,\eta}(t) - \mathbb{P}q^0\|_{L^2}$$

which allows us to write that

$$\frac{d}{dt}\|q^a_{\kappa,\eta}(t) - \mathbb{P}q^0\|_{L^2} \leq \|B_{\kappa,\eta}\mathbb{P}q^0\|_{L^2}.$$

It follows that

$$\|q^a_{\kappa,\eta}(t) - \mathbb{P}q^0\|_{L^2} \leq t\|B_{\kappa,\eta}\mathbb{P}q^0\|_{L^2}. \tag{7.8}$$

Therefore, from (7.7), (7.6) and (7.8), we obtain

$$\|q_{\kappa,\eta}(t) - \mathbb{P}q^0\| \leq \|q^0 - \mathbb{P}q^0\| + t\|B_{\kappa,\eta}\mathbb{P}q^0\|_{L^2}. \tag{7.9}$$

For the sake of convenience, we denote $\mathbb{P}q^0 = \hat{q}^0 = (\hat{r}^0, \hat{u}^0, \hat{v}^0)$. From the numerical diffusion term (7.5), we clearly get for the AF-DP scheme

$$\|B_{\kappa,\eta}\hat{q}^0\|_{L^2} \leq \nu_r \|\Delta \hat{r}^0\|_{L^2} \tag{7.10}$$

where $\nu_r = \frac{\kappa_r^x a_\star \Delta x}{2} = \frac{\kappa_r^y a_\star \Delta y}{2} = \mathcal{O}(M\Delta x)$. Therefore, when the initial condition is close to the kernel, i.e. when $\|q^0 - \mathbb{P}q^0\| = \mathcal{O}(M)$, (7.9) and (7.10) show that the numerical solution of the AF-DP scheme is still close to the kernel locally in time $t = \mathcal{O}(1)$. □

As a numerical illustration, we now consider an initial condition which is close to the discrete kernel up to a perturbation of size $\mathcal{O}(M)$. In particular, this initial condition is given by

$$q_h^0 = \hat{q}_h^0 + M\frac{\tilde{q}_h^0}{\|\tilde{q}_h^0\|},$$

where $\hat{q}_h^0$ belongs to the kernel and $\tilde{q}_h^0$ is in the orthogonal subspace. They are respectively given by

$$\hat{q}^0 = \begin{cases} \hat{r}(t=0,x,y) = 1 - \exp\left[-\left(\frac{3x}{0.5}\right)^2 - \left(\frac{3y}{0.5}\right)^2\right] \\ \hat{u}(t=0,x,y) = -\frac{6y}{0.5}\exp\left[-\left(\frac{3x}{0.5}\right)^2 - \left(\frac{3y}{0.5}\right)^2\right] \\ \hat{v}(t=0,x,y) = \frac{6x}{0.5}\exp\left[-\left(\frac{3x}{0.5}\right)^2 - \left(\frac{3y}{0.5}\right)^2\right] \end{cases}$$

and

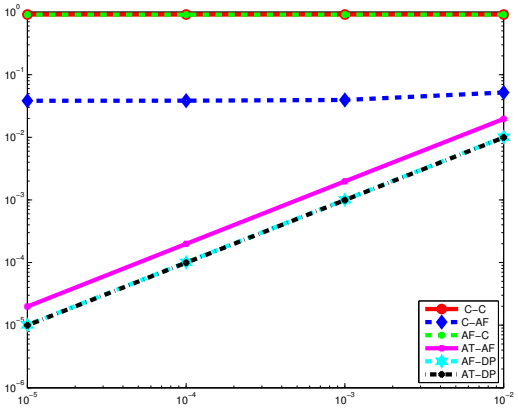
$$\tilde{q}^0 = \begin{cases} \tilde{r}(t=0,x,y) = -\frac{4x}{0.8}\exp\left[-\left(\frac{4x}{0.8}\right)^2 - \left(\frac{4y}{0.4}\right)^2\right] + \frac{4y}{0.8}\exp\left[-\left(\frac{4x}{0.4}\right)^2 - \left(\frac{4y}{0.8}\right)^2\right] \\ \tilde{u}(t=0,x,y) = \frac{1}{2}\exp\left[-\left(\frac{4x}{0.4}\right)^2 - \left(\frac{4y}{0.8}\right)^2\right] \\ \tilde{v}(t=0,x,y) = \frac{1}{2}\exp\left[-\left(\frac{4x}{0.8}\right)^2 - \left(\frac{4y}{0.4}\right)^2\right] \end{cases}$$

in the periodic domain $\mathbb{T}^2 = [-0.5, 0.5] \times [-0.5, 0.5]$.
In Fig. 7.1, we present the maximum value of the deviation from the initial projection $\hat{q}_h^0$ over the time interval with different values of M. This figure shows that like the well-balanced AT-DP scheme, the total deviation of both AT-AF and AF-DP strategies is proportional to M whereas it remains constant for the other strategies and we also notice that the constant is smaller for the C-AF scheme than for the C-C and AF-C schemes.

7.3 Modified Godunov type schemes applied to the nonlinear shallow water equation

In this section, we present some methods to modify the classical Godunov type schemes in order to derive new numerical strategies which are more accurate around the geostrophic equilibrium and in the adjustment process. To discretize the shallow water equations, we first exhibit the general form of the finite volume method which has received a great interest in the context of hyperbolic conservation laws since this method is conservative, among other properties.
The shallow water equation (7.1) can be expressed in terms of conservative variables as

$$\frac{\partial U}{\partial t} + \nabla \cdot F(U) = S(U)$$



(a)

Figure 7.1: $\max_{t\in[0,2]} \|q - \mathbb{P}q^0\|(t)$ *as a function of the Froude number (log-log scale)*

where $F(U) = (F_x(U), F_y(U)^T$ and

$$U = \begin{pmatrix} h \\ hu \\ hv \end{pmatrix}, \quad F_x(U) = \begin{pmatrix} hu \\ hu^2 + \frac{1}{2}gh^2 \\ huv \end{pmatrix}, \quad F_y(U) = \begin{pmatrix} hv \\ huv \\ hv^2 + \frac{1}{2}gh^2 \end{pmatrix}, \quad S(U) = \begin{pmatrix} 0 \\ \Omega hv \\ -\Omega hu \end{pmatrix}.$$

Let us now suppose that the domain $\mathbb{T}^2$ is discretized into N cells T_i. Let A_{ij} be the common edge of two neighboring cells T_i and T_j. The notation $\mathbf{n}_{ij}$ stands for the unit normal vector to A_{ij} pointing from T_i to T_j. Considering the homogeneous equations and integrating over the space domain T_i, we obtain the following relation

$$\frac{d}{dt}U_i(t) + \frac{1}{|T_i|}\sum_{A_{ij}\subset\partial T_i}\int_{A_{ij}} F(U)\cdot\mathbf{n}_{ij}ds = 0,$$

where $U_i(t)$ is an average of the unknowns on cell T_i given by $U_i(t) = \frac{1}{|T_i|}\int_{T_i} U(\mathbf{x},t)\ \mathbf{dx}$. For the sake of simplicity, let us denote the numerical flux by

$$\Phi_{ij} = \Phi(U_i, U_j, \mathbf{n}_{ij}) \approx \frac{1}{|A_{ij}|}\int_{A_{ij}} F(U)\cdot\boldsymbol{n}_{ij}ds.$$

Then the general form of finite volume scheme can be written as

$$\frac{d}{dt}U_i(t) + \frac{1}{|T_i|}\sum_{A_{ij}\subset\partial T_i} |A_{ij}|\Phi_{ij} = 0. \tag{7.11}$$

The purpose of the finite volume method is to update the cell average of the unknown at every time step by computing fluxes across cell interfaces. There are several kinds of flux Φ_{ij} which can be used in (7.11). In this work, we are interested in the modification of the so called Roe approximate Riemann soler [60] for the interface fluxes. To do so, let us note that the Roe flux

can be written as

$$\begin{aligned}\Phi_{ij}^{\text{Roe}} &= \frac{1}{2}\begin{pmatrix}(h_i\mathbf{u}_i + h_j\mathbf{u_j})\cdot\mathbf{n}_{ij}\\ h_i(\mathbf{u}_i\cdot\mathbf{n}_{ij})\mathbf{u}_i + h_j(\mathbf{u}_j\cdot\mathbf{n}_{ij})\mathbf{u}_j\end{pmatrix} + \frac{1}{2}\begin{pmatrix}0\\ \frac{g}{2}(h_i^2+h_j^2)\mathbf{n}_{ij}\end{pmatrix}\\ &- \frac{1}{4}|\mathbf{u}_{ij}\cdot\mathbf{n}_{ij} - c_{ij}|\left(\Delta h - \frac{h_{ij}}{c_{ij}}\Delta(\mathbf{u}\cdot\mathbf{n}_{ij})\right)\begin{pmatrix}1\\ \mathbf{u}_{ij}-c_{ij}\mathbf{n}_{ij}\end{pmatrix}\\ &- \frac{1}{2}|\mathbf{u}_{ij}\cdot\mathbf{n}_{ij}|h_{ij}\Delta(\mathbf{u}\cdot\mathbf{n}_{ij}^{\perp})\begin{pmatrix}0\\ \mathbf{n}_{ij}^{\perp}\end{pmatrix} - \frac{1}{4}|\mathbf{u}_{ij}\cdot\mathbf{n}_{ij} + c_{ij}|\left(\Delta h + \frac{h_{ij}}{c_{ij}}\Delta(\mathbf{u}\cdot\mathbf{n}_{ij})\right)\begin{pmatrix}1\\ \mathbf{u}_{ij}+c_{ij}\mathbf{n}_{ij}\end{pmatrix},\end{aligned} \tag{7.12}$$

(see Appendix 7.B for more details).
In the subsonic case $|\mathbf{u}_{ij}\cdot\mathbf{n}_{ij}| \le c_{ij}$, the Roe flux becomes

$$\begin{aligned}\Phi_{ij}^{\text{Roe}} &= \frac{1}{2}\begin{pmatrix}(h_i\mathbf{u}_i + h_j\mathbf{u_j})\cdot\mathbf{n}_{ij}\\ h_i(\mathbf{u}_i\cdot\mathbf{n}_{ij})\mathbf{u}_i + h_j(\mathbf{u}_j\cdot\mathbf{n}_{ij})\mathbf{u}_j\end{pmatrix} + \frac{1}{2}\begin{pmatrix}0\\ \frac{g}{2}(h_i^2+h_j^2)\mathbf{n}_{ij}\end{pmatrix}\\ &- \frac{1}{2}\begin{pmatrix}\frac{h_{ij}}{c_{ij}}\mathbf{u}_{ij}\cdot\mathbf{n}_{ij}\Delta(\mathbf{u}\cdot\mathbf{n}_{ij})\\ c_{ij}\Delta h\,[\mathbf{u}_{ij} + (\mathbf{u}_{ij}\cdot\mathbf{n}_{ij})\mathbf{n}_{ij}] + \frac{h_{ij}}{c_{ij}}\mathbf{u}_{ij}\cdot\mathbf{n}_{ij}\Delta(\mathbf{u}\cdot\mathbf{n}_{ij})\mathbf{u}_{ij} + h_{ij}|\mathbf{u}_{ij}\cdot\mathbf{n}_{ij}|\Delta(\mathbf{u}\cdot\mathbf{n}_{ij}^{\perp})\mathbf{n}_{ij}^{\perp}\end{pmatrix}\\ &- \frac{1}{2}\begin{pmatrix}c_{ij}\Delta h\\ c_{ij}h_{ij}\Delta(\mathbf{u}\cdot\mathbf{n}_{ij})\mathbf{n}_{ij}\end{pmatrix}.\end{aligned} \tag{7.13}$$

7.3.1 The correction for the mass equation

Apparent Topography strategy

The hydrostatic reconstruction has been introduced in [5] for the shallow water equation with non-flat topography based on a local reconstruction at the interface. By using this technique, we can obtain a well-balanced scheme from any solver of the homogeneous problem. In the presence of Coriolis source term, the work in [13] (see also [25]) adapts this strategy to derive the scheme named *Apparent Topography scheme*. Before going into the detail of this method, let us review the hydrostatic reconstruction with topography $b(\mathbf{x})$ on a Cartesian mesh by considering the semi-discrete scheme applied to the shallow water equations

$$\frac{\partial}{\partial t}U_{i,j}(t) + \frac{1}{\Delta x}\left(F_{i+1/2,j} - F_{i-1/2,j}\right) + \frac{1}{\Delta y}\left(G_{i,j+1/2} - G_{i,j-1/2}\right) = S_{i,j}$$

where the numerical fluxes are given by

$$F_{i+1/2,j} = \mathcal{F}\left(U_{i+1/2_-,j}, U_{i+1/2_+,j}\right), \quad \text{and} \quad G_{i,j+1/2} = \mathcal{G}\left(U_{i,j+1/2_-}, U_{i,j+1/2_+}\right).$$

In the above formula, the conservative variables are defined by

$$U_{i+1/2_-,j} = \begin{pmatrix}h_{i+1/2_-,j}\\ h_{i+1/2_-,j}\mathbf{u}_{i,j}\end{pmatrix}, \quad U_{i+1/2_+,j} = \begin{pmatrix}h_{i+1/2_+,j}\\ h_{i+1/2_+,j}\mathbf{u}_{i+1,j}\end{pmatrix}$$

and

$$U_{i,j+1/2_-} = \begin{pmatrix}h_{i,j+1/2_-}\\ h_{i,j+1/2_-}\mathbf{u}_{i,j}\end{pmatrix}, \quad U_{i,j+1/2_+} = \begin{pmatrix}h_{i,j+1/2_+}\\ h_{i,j+1/2_+}\mathbf{u}_{i,j+1}\end{pmatrix}$$

where the hydrostatic reconstruction is defined by

$$h_{i+1/2_-,j} = (h_{i,j} + b_{i,j} - \max\{b_{i,j}, b_{i+1,j}\})_+, \quad h_{i+1/2_+,j} = (h_{i+1,j} + b_{i+1,j} - \max\{b_{i,j}, b_{i+1,j}\})_+$$

$$h_{i,j+1/2_-} = (h_{i,j} + b_{i,j} - \max\{b_{i,j}, b_{i,j+1}\})_+, \quad h_{i,j+1/2_+} = (h_{i,j+1} + b_{i,j+1} - \max\{b_{i,j}, b_{i,j+1}\})_+.$$

Then, in order to ensure the well balanced property, the topography source term is approximated by

$$S_{i,j} = \begin{pmatrix} 0 \\ \frac{g}{2} h^2_{i+1/2_-,j} - \frac{g}{2} h^2_{i-1/2_+,j} \\ \frac{g}{2} h^2_{i,j+1/2_-} - \frac{g}{2} h^2_{i,j-1/2_+} \end{pmatrix}.$$

In order to apply this technique to the Coriolis source term, the apparent topography scheme has been introduced in [25] by considering the Coriolis as a new topography

$$g\Delta b^1_{i+1/2,j} = -\omega \frac{v_{i,j} + v_{i+1,j}}{2} \Delta x, \quad \text{and} \quad g\Delta b^1_{i,j+1/2} = \omega \frac{u_{i,j} + u_{i,j+1}}{2} \Delta y$$

Remark 7.1. *It is important to note that with* $\Delta b_{i+1/2,j} = b_{i+1,j} - b_{i,j}$ *and* $\Delta b_{i,j+1/2} = b_{i,j+1} - b_{i,j}$, *the hydrostatic reconstruction can be written as*

$$h_{i+1/2_-,j} = \Big(h_{i,j} - (\Delta b_{i+1/2,j})_+\Big)_+, \quad h_{i+1/2_+,j} = \Big(h_{i+1,j} - (-\Delta b_{i+1/2,j})_+\Big)_+$$

and

$$h_{i,j+1/2_-} = \Big(h_{i,j} - (\Delta b_{i,j+1/2})_+\Big)_+, \quad h_{i,j+1/2_+} = \Big(h_{i,j+1} - (-\Delta b_{i,j+1/2})_+\Big)_+,$$

which implies that we do not have to solve for the explicit new topography when we apply this technique to the Coriolis source term.

The All Froude strategy

We now propose two ways to adapt the All Froude strategy to the mass equation. The first is simply to keep all diffusion terms on this equation "small enough". In particular, we multiply these diffusion terms by the local Froude number and the Roe flux becomes

$$\begin{aligned} \Phi_{ij}^{\text{AF1, Roe}} = & \frac{1}{2} \begin{pmatrix} (h_i \mathbf{u}_i + h_j \mathbf{u_j}) \cdot \mathbf{n}_{ij} \\ h_i (\mathbf{u}_i \cdot \mathbf{n}_{ij}) \mathbf{u}_i + h_j (\mathbf{u}_j \cdot \mathbf{n}_{ij}) \mathbf{u}_j \end{pmatrix} + \frac{1}{2} \begin{pmatrix} 0 \\ \frac{g}{2}(h_i^2 + h_j^2) \mathbf{n}_{ij} \end{pmatrix} \\ & - \frac{1}{4} |\mathbf{u}_{ij} \cdot \mathbf{n}_{ij} - c_{ij}| \left(\Delta h - \frac{h_{ij}}{c_{ij}} \Delta(\mathbf{u} \cdot \mathbf{n}_{ij}) \right) \begin{pmatrix} \theta_{ij} \\ \mathbf{u}_{ij} - c_{ij} \mathbf{n}_{ij} \end{pmatrix} \\ & - \frac{1}{2} |\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}| h_{ij} \Delta(\mathbf{u} \cdot \mathbf{n}^\perp_{ij}) \begin{pmatrix} 0 \\ \mathbf{n}^\perp_{ij} \end{pmatrix} - \frac{1}{4} |\mathbf{u}_{ij} \cdot \mathbf{n}_{ij} + c_{ij}| \left(\Delta h + \frac{h_{ij}}{c_{ij}} \Delta(\mathbf{u} \cdot \mathbf{n}_{ij}) \right) \begin{pmatrix} \theta_{ij} \\ \mathbf{u}_{ij} + c_{ij} \mathbf{n}_{ij} \end{pmatrix}, \end{aligned} \tag{7.14}$$

where $\theta_{ij} = \theta(M_{ij})$ with $\theta(M) = \min\{M, 1\}$ and M_{ij} stands for the local Froude number.

The other way is to introduce anti-diffusion related to the numerical viscosity in the pressure equation. To derive an All Mach (Froude) scheme in this way, the classical flux of any X-scheme is modified as follows

$$\Phi_{ij}^{AF,X} = \Phi_{ij}^X + (\theta_{ij} - 1) \frac{c_{ij}}{2} \begin{pmatrix} h_i - h_j \\ 0 \end{pmatrix}.$$

In other words, only the final term in (7.13) is modified by multiplication with the local Froude

number. Then, in the subsonic case, this modified Roe flux can be written as

$$\Phi_{ij}^{\mathrm{AF2,Roe}} = \frac{1}{2}\begin{pmatrix} (h_i\mathbf{u}_i + h_j\mathbf{u_j})\cdot\mathbf{n}_{ij} \\ h_i(\mathbf{u}_i\cdot\mathbf{n}_{ij})\mathbf{u}_i + h_j(\mathbf{u}_j\cdot\mathbf{n}_{ij})\mathbf{u}_j \end{pmatrix} + \frac{1}{2}\begin{pmatrix} 0 \\ \frac{g}{2}(h_i^2+h_j^2)\mathbf{n}_{ij} \end{pmatrix}$$
$$- \frac{1}{2}\begin{pmatrix} \frac{h_{ij}}{c_{ij}}\mathbf{u}_{ij}\cdot\mathbf{n}_{ij}\Delta(\mathbf{u}\cdot\mathbf{n}_{ij}) \\ c_{ij}\Delta h\left[\mathbf{u}_{ij} + (\mathbf{u}_{ij}\cdot\mathbf{n}_{ij})\mathbf{n}_{ij}\right] + \frac{h_{ij}}{c_{ij}}\mathbf{u}_{ij}\cdot\mathbf{n}_{ij}\Delta(\mathbf{u}\cdot\mathbf{n}_{ij})\mathbf{u}_{ij} + h_{ij}|\mathbf{u}_{ij}\cdot\mathbf{n}_{ij}|\Delta(\mathbf{u}\cdot\mathbf{n}_{ij}^{\perp})\mathbf{n}_{ij}^{\perp} \end{pmatrix}$$
$$- \frac{1}{2}\begin{pmatrix} \theta_{ij}c_{ij}\Delta h \\ c_{ij}h_{ij}\Delta(\mathbf{u}\cdot\mathbf{n}_{ij})\mathbf{n}_{ij} \end{pmatrix}. \tag{7.15}$$

7.3.2 The correction for the velocity equation

All Froude strategy

In [20], the inaccuracy of Godunov type schemes applied to the compressible Euler system is investigated. In particular, the author points out that the numerical viscosity of the classical Godunov type schemes in the velocity equation is responsible for the inaccuracy problem at low Mach number. As a consequence, the numerical solution of the classical Godunov scheme may be far form the incompressible solution at low Mach number.

The All Mach (Froude) Godunov scheme proposed in [22] is as follows: The classical flux of any X-scheme is modified by setting

$$\Phi_{ij}^{AF,X} = \Phi_{ij}^{X} + (\theta_{ij}-1)\frac{h_{ij}c_{ij}}{2}\begin{pmatrix} 0 \\ \left[(\mathbf{u}_i-\mathbf{u}_j)\cdot\mathbf{n}_{ij}\right]\mathbf{n}_{ij} \end{pmatrix}, \tag{7.16}$$

where $\theta_{ij} = \theta(M_{ij})$ with $\theta(M) = \min\{M,1\}$ and M_{ij}, h_{ij}, c_{ij} respectively stand for the local Froude number, water depth and sound velocity. Therefore, in the subsonic case, the modified Roe flux can be written as

$$\Phi_{ij}^{\mathrm{C,AF}} = \frac{1}{2}\begin{pmatrix} (h_i\mathbf{u}_i + h_j\mathbf{u_j})\cdot\mathbf{n}_{ij} \\ h_i(\mathbf{u}_i\cdot\mathbf{n}_{ij})\mathbf{u}_i + h_j(\mathbf{u}_j\cdot\mathbf{n}_{ij})\mathbf{u}_j \end{pmatrix} + \frac{1}{2}\begin{pmatrix} 0 \\ \frac{g}{2}(h_i^2+h_j^2)\mathbf{n}_{ij} \end{pmatrix}$$
$$- \frac{1}{2}\begin{pmatrix} \frac{h_{ij}}{c_{ij}}\mathbf{u}_{ij}\cdot\mathbf{n}_{ij}\Delta(\mathbf{u}\cdot\mathbf{n}_{ij}) \\ c_{ij}\Delta h\left[\mathbf{u}_{ij} + (\mathbf{u}_{ij}\cdot\mathbf{n}_{ij})\mathbf{n}_{ij}\right] + \frac{h_{ij}}{c_{ij}}\mathbf{u}_{ij}\cdot\mathbf{n}_{ij}\Delta(\mathbf{u}\cdot\mathbf{n}_{ij})\mathbf{u}_{ij} + h_{ij}|\mathbf{u}_{ij}\cdot\mathbf{n}_{ij}|\Delta(\mathbf{u}\cdot\mathbf{n}_{ij}^{\perp})\mathbf{n}_{ij}^{\perp} \end{pmatrix}$$
$$- \frac{1}{2}\begin{pmatrix} c_{ij}\Delta h \\ \theta_{ij}c_{ij}h_{ij}\Delta(\mathbf{u}\cdot\mathbf{n}_{ij})\mathbf{n}_{ij} \end{pmatrix}. \tag{7.17}$$

Remark 7.2. *The correction (7.16) is different from a low Mach number fix introduced in [19] by multiplying the jump in the normal velocity component of the Riemann problem with the local Mach number. In fact, this LMRoe flux can be written as*

$$\Phi_{ij}^{LMRoe} = \frac{1}{2}\begin{pmatrix} (h_i\mathbf{u}_i + h_j\mathbf{u_j})\cdot\mathbf{n}_{ij} \\ h_i(\mathbf{u}_i\cdot\mathbf{n}_{ij})\mathbf{u}_i + h_j(\mathbf{u}_j\cdot\mathbf{n}_{ij})\mathbf{u}_j \end{pmatrix} + \frac{1}{2}\begin{pmatrix} 0 \\ \frac{g}{2}(h_i^2+h_j^2)\mathbf{n}_{ij} \end{pmatrix}$$
$$- \frac{1}{4}|\mathbf{u}_{ij}.\mathbf{n}_{ij} - c_{ij}|\left(\Delta h - \theta_{ij}\frac{h_{ij}}{c_{ij}}\Delta(\mathbf{u}\cdot\mathbf{n}_{ij})\right)\begin{pmatrix} 1 \\ \mathbf{u}_{ij} - c_{ij}\mathbf{n}_{ij} \end{pmatrix}$$
$$- \frac{1}{2}|\mathbf{u}_{ij}.\mathbf{n}_{ij}|h_{ij}\Delta(\mathbf{u}\cdot\mathbf{n}_{ij}^{\perp})\begin{pmatrix} 0 \\ \mathbf{n}_{ij}^{\perp} \end{pmatrix} - \frac{1}{4}|\mathbf{u}_{ij}.\mathbf{n}_{ij} + c_{ij}|\left(\Delta h + \theta_{ij}\frac{h_{ij}}{c_{ij}}\Delta(\mathbf{u}\cdot\mathbf{n}_{ij})\right)\begin{pmatrix} 1 \\ \mathbf{u}_{ij} + c_{ij}\mathbf{n}_{ij} \end{pmatrix}.$$

Divergence Penalisation

Instead of using the classical numerical diffusion in the velocity equation, this diffusion is replaced by $-\nabla(h\nabla\cdot\boldsymbol{u})$. It is essential to use values of $\nabla\cdot\boldsymbol{u}$ computed at the corners of the cells to define the numerical diffusion at the cell centers. In order to design this diffusion term, let us first define the discrete operators at the vertices of each cell (i,j)

$$[\nabla^{\mathrm{v}}\cdot\boldsymbol{u}]_{i+1/2,j+1/2}=\frac{(u_{i+1,j+1}+u_{i+1,j})-(u_{i,j+1}+u_{i,j})}{2\Delta x}+\frac{(v_{i+1,j+1}+v_{i,j+1})-(v_{i+1,j}+v_{i,j})}{2\Delta y}$$

$$[f^{\mathrm{v}}(h)]_{i+1/2,j+1/2}=\frac{h_{i+1,j+1}+h_{i,j+1}+h_{i+1,j}+h_{i,j}}{4}.$$

We shall also need dual operators that enable to switch from the grid vertices to the cell centers. For any φ defined at the vertices, we define the following discrete gradient

$$[\nabla^{\mathrm{c}}\varphi]_{i,j}=\frac{1}{2}\begin{pmatrix}\frac{\varphi_{i+1/2,j+1/2}-\varphi_{i-1/2,j+1/2}}{\Delta x}\\ \frac{\varphi_{i+1/2,j+1/2}-\varphi_{i+1/2,j-1/2}}{\Delta y}\end{pmatrix}+\frac{1}{2}\begin{pmatrix}\frac{\varphi_{i+1/2,j-1/2}-\varphi_{i-1/2,j-1/2}}{\Delta x}\\ \frac{\varphi_{i-1/2,j+1/2}-\varphi_{i-1/2,j-1/2}}{\Delta y}\end{pmatrix}.$$

Therefore, the discrete version of the term $\nabla(h\nabla.\boldsymbol{u})$ can be obtained by applying the operator ∇^{c} on φ defined by

$$\varphi_{i+1/2,j+1/2}=[f^{\mathrm{v}}(h)\nabla^{\mathrm{v}}\cdot\boldsymbol{u}]_{i+1/2,j+1/2}\,.$$

7.3.3 Time discretization method

For the purpose of our study, we now only consider Cartesian grids. Then, the fully discrete finite volume scheme for the nonlinear shallow water equation is given by

$$U_{i,j}^{n+1}=U_{i,j}^{n}-\frac{\Delta t}{\Delta x}\left(F_{i+1/2,j}^{n}-F_{i-1/2,j}^{n}\right)-\frac{\Delta t}{\Delta y}\left(G_{i,j+1/2}^{n}-G_{i,j-1/2}^{n}\right)+\Delta t S_{i,j}^{\theta} \tag{7.18}$$

where $U_{i,j}^{n}$ is the cell average at time level t^n, $F_{i+1/2,j}^{n}$ and $G_{i,j+1/2}^{n}$ are respectively the numerical flux in x and y directions defined at the interface between two neighboring cells. The term $S_{i,j}^{\theta}$ is an approximation of the Coriolis source term which is defined by

$$S_{i,j}^{\theta}=\begin{pmatrix}0\\ \theta_1 S_u^n+(1-\theta_1)S_u^{n+1}\\ \theta_2 S_v^n+(1-\theta_2)S_v^{n+1}\end{pmatrix},$$

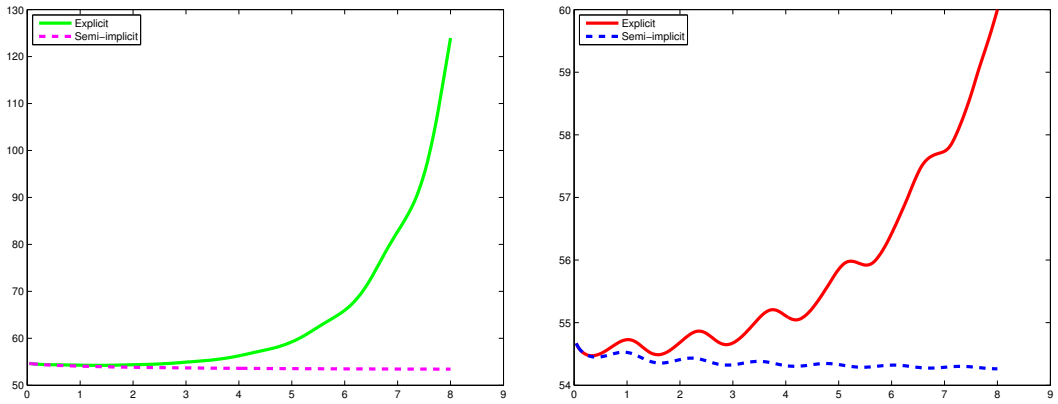
where the parameters θ_1 and θ_2 are both in $[0,1]$. The terms S_u and S_v represent the approximations of the Coriolis source term in horizontal and vertical directions, respectively. Particularly, for the classical approximation, we have

$$S_u=h_{i,j}v_{i,j}\quad\text{and}\quad S_v=-h_{i,j}u_{i,j}.$$

However, for the Apparent Topography strategy, these terms are defined by

$$S_u=\frac{g}{2}h_{i+1/2_-,j}^2-\frac{g}{2}h_{i-1/2_+,j}^2\quad\text{and}\quad S_v=\frac{g}{2}h_{i,j+1/2_-}^2-\frac{g}{2}h_{i,j-1/2_+}^2.$$

Let us emphasize that the totally explicit scheme ($\theta_1=\theta_2=1$) is unstable (see Figure 7.2). Therefore, for stability reasons, we have to consider a discretization in time for the Coriolis force which will be implicit enough. In fact, the parameters θ_1 and θ_2 must be in the stability region $\theta_1+\theta_2\leq 1$ (see [37] for more details in the context of the linear wave equation). On the other hand, let us note that due to the approximation of the Coriolis force in the Apparent Topography method, we restrict our study to the two cases $\theta_1=1,\theta_2=0$ and $\theta_1=0,\theta_2=1$ to ensure that the proposed schemes are still explicit without having to solve any algebraic system of equations at each time step.



(a) Classical Roe scheme **(b)** Apparent Topography scheme

Figure 7.2: *The total energy of the explicit scheme ($\theta_1 = \theta_2 = 1$) and semi-implicit scheme ($\theta_1 = 1, \theta_2 = 0$) with initial velocity fluid at rest and initial height given by $h^0(\mathbf{x}) = 1 + \chi_{[-1,1]}(\mathbf{x})$.*

7.4 Numerical results

7.4.1 Stationary vortex test case

We now begin with the stationary vortex test case proposed in [61] (see also in [62]) to investigate the behaviour of the finite volume scheme applied to the rotating shallow water equations. By using this test case, we can compare all proposed schemes in terms of error between the numerical solution and the exact solution. We consider the stationary vortex in the square domain $\Omega = [-0.5, 0.5] \times [-0.5, 0.5]$ with periodic boundary conditions and the initial velocity field given by

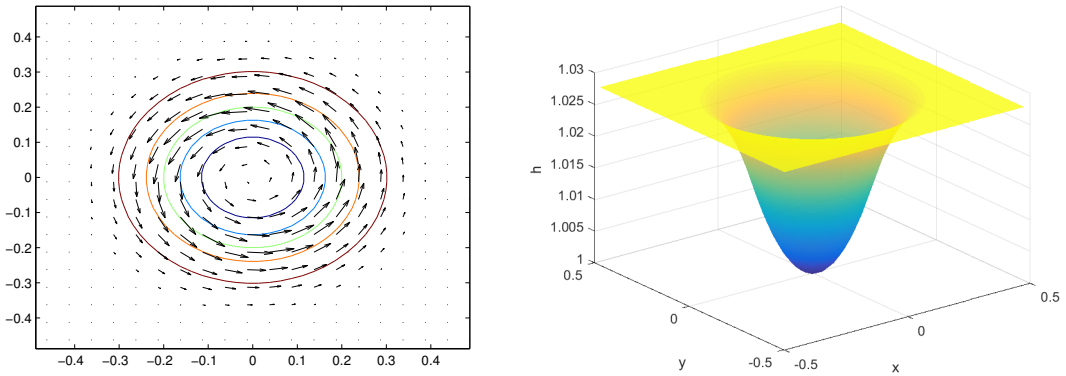
$$\boldsymbol{u}^0(r,\theta) = \nu_\theta(r)\bar{e}_\theta, \quad \nu_\theta(r) = \varepsilon \left[5r\chi\left(r < \frac{1}{5}\right) + (2 - 5r)\chi\left(\frac{1}{5} \leq r < \frac{2}{5}\right) \right] \tag{7.19}$$

where r stands for the distance to the center of the domain and χ is the characteristic function. Let us note that the vortex is a stationary solution of the shallow water equation if the water depth is the solution of the ODE

$$\partial_r h^0 = \frac{1}{g}\left(\omega\nu_\theta + \frac{\nu_\theta^2}{r}\right) \tag{7.20}$$

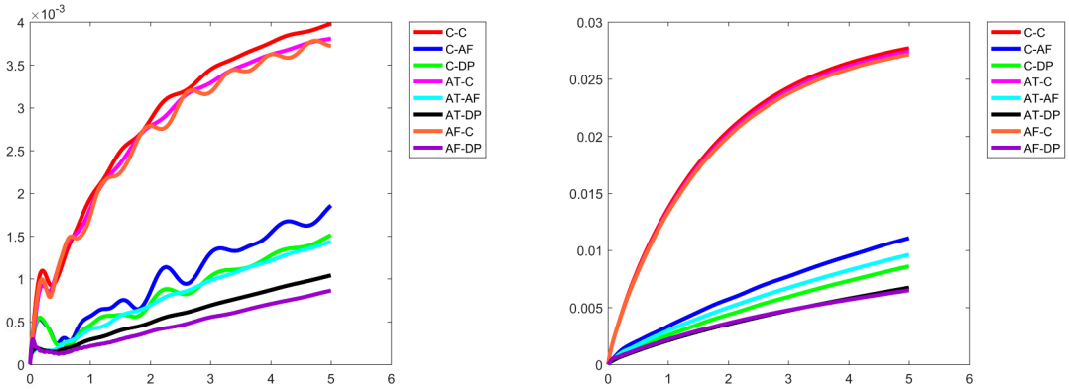
Figure 7.3 shows such kind of initial condition with parameter $\varepsilon = 0.1$. Let us emphasize that when the water depth and Coriolis parameter ω are of order $\mathcal{O}(1)$, then the parameter ε has a strong impact on the order of the Froude number and Rossby number. In particular, we obtain $\mathsf{Fr} = \mathsf{Ro} \approx \mathcal{O}(\varepsilon)$. We will investigate the influence of ε on the modified Godunov type schemes proposed in the previous sections.

Figures 7.5 and 7.6 show a cut of the water fluid depth along the x-axis at $y = 0$ for some values of the parameter ε at time $t = 5$ and $t = 10$. These figures together with Figure 7.4 clearly show that the AT-DP, AT-AF and AF-DP schemes which are obtained by using a combination of the corrections for both mass and velocity equations have smaller errors than the other schemes. Since the behavior of the AT-C and AF-C are similar to the C-C scheme, we can say that the correction in the mass equation, which works well in the one dimensional case, does not improve too much the accuracy of numerical schemes in the 2D case. Moreover, as can be seen on these figures, the correction on the velocity equation seems to be much more important than the correction on the mass equation. This is a conclusion which is similar to the study of the



(a) Contour of h and vector field at $t = 0$ **(b)** h at $t = 0$

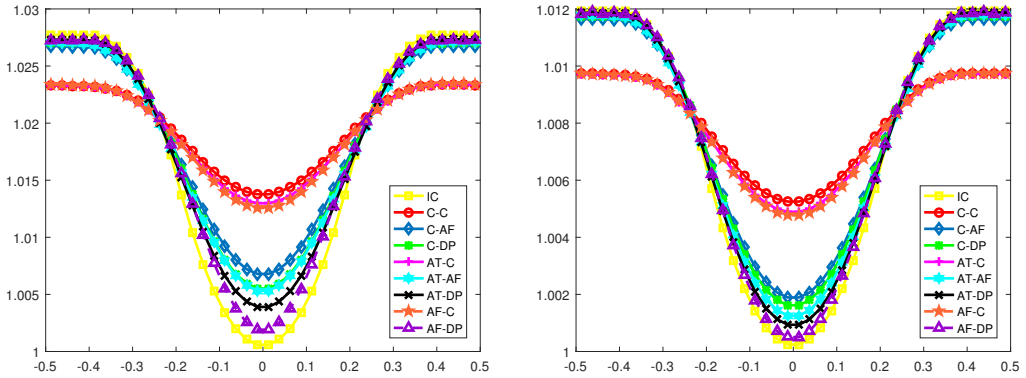
Figure 7.3: *Initial condition with* 40×40 *grid cells and* $\varepsilon = 0.1$.



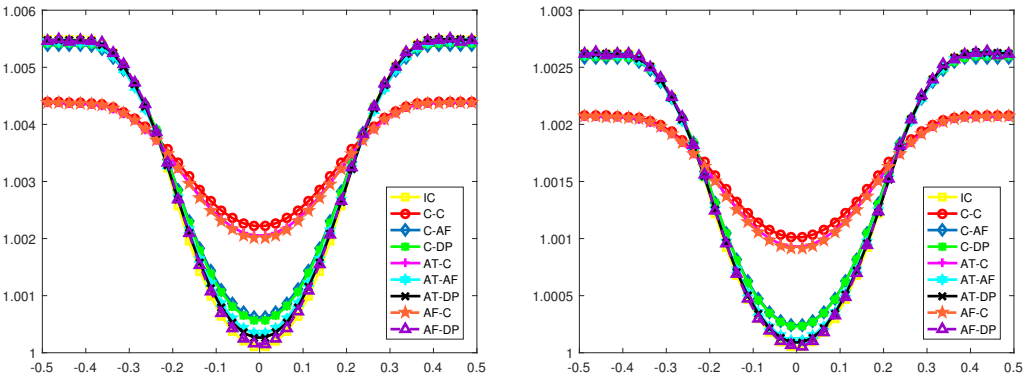
(a) h **(b)** $h\boldsymbol{u}$

Figure 7.4: *Time evolution of the water depth and discharge errors up to time* $t = 5$ *with parameter* $\varepsilon = 0.1$

approximation of the linear wave equation in 2D by these schemes. We mention [53] for more details of the analysis of these strategies in the linear case. Figure 7.4 also shows that the error is smaller for the water depth (10^{-3}) than the discharge $h\mathbf{u}$ (10^{-2}).
On the other hand, we also observe that the smaller the value of ε, the better the approximation of the exact solution by the solutions of the AT-DP, AT-AF and AF-DP schemes. The reason for this is that when ε tends to zero, the nonlinear convection term gets significantly small as compared to the other terms, and the behavior of the rotating shallow water equations is really dominated by the acoustic linear wave equation. Therefore, the numerical schemes which work well with the linear wave equation provide some convincing results.
Moreover, another evidence to illustrate that those schemes are better than the other schemes is that the errors of those schemes behave like ε^2, as can be seen from the log-log graphs displayed in Figure 7.7, while the errors of the classical scheme or even of the schemes with the correction on the pressure equation behave roughly like ε.

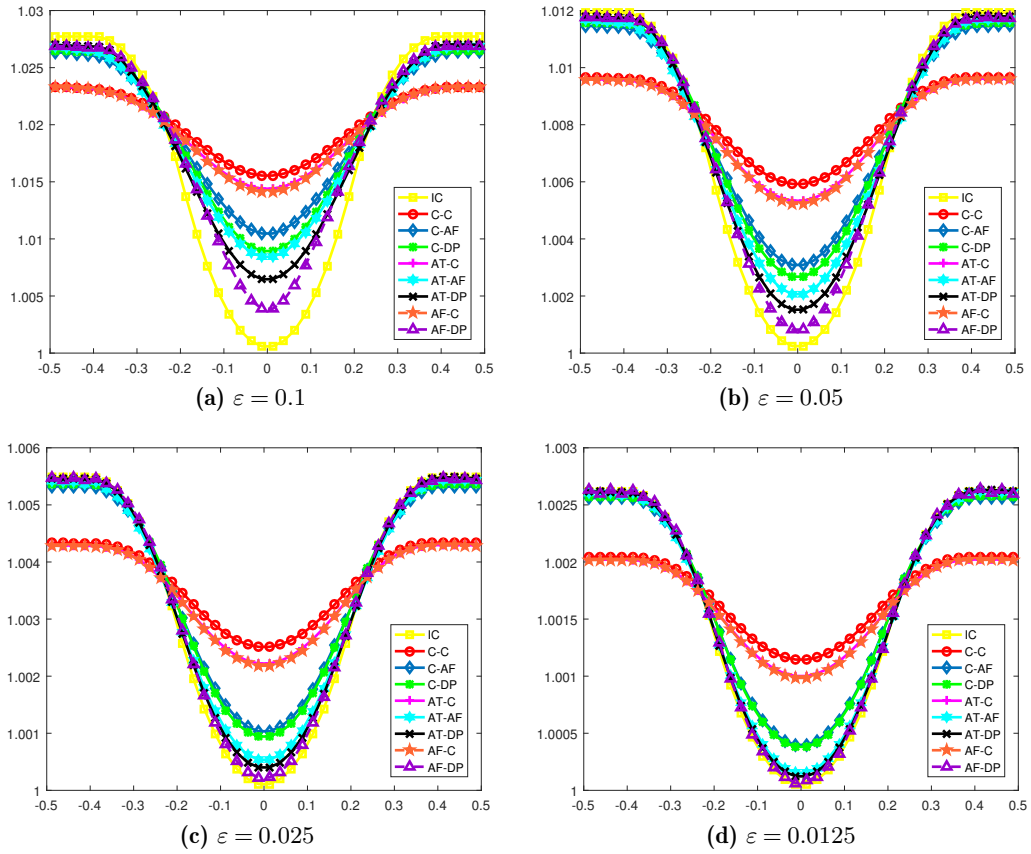


(a) $\varepsilon = 0.1$ **(b)** $\varepsilon = 0.05$



(c) $\varepsilon = 0.025$ **(d)** $\varepsilon = 0.0125$

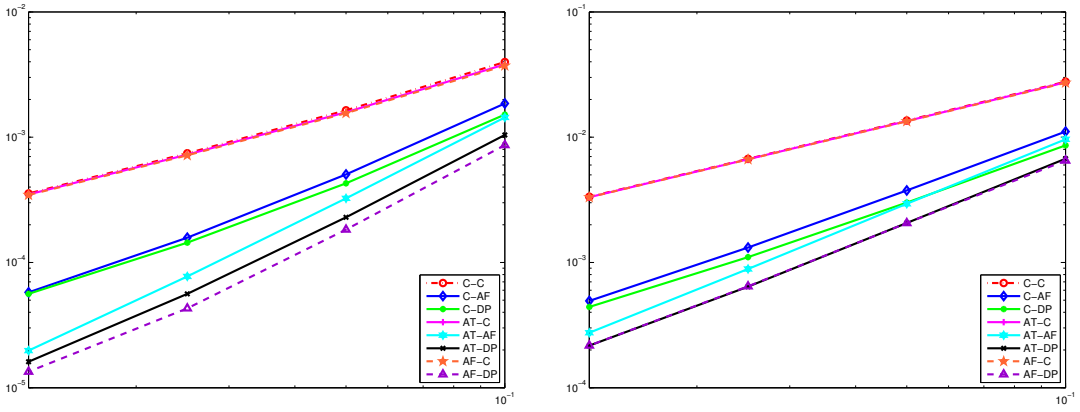
Figure 7.5: *Horizontal cut of water depths computed by Godunov type schemes at time* $t = 5$



(a) $\varepsilon = 0.1$ **(b)** $\varepsilon = 0.05$

(c) $\varepsilon = 0.025$ **(d)** $\varepsilon = 0.0125$

Figure 7.6: *Horizontal cut of water depths computed by Godunov type schemes at time* $t = 10$



(a) h **(b)** hu

Figure 7.7: *Log-log graph of the error at time* $t = 5$ *as a function of* ε

7.4.2 Nonlinear geostrophic adjustment simulation

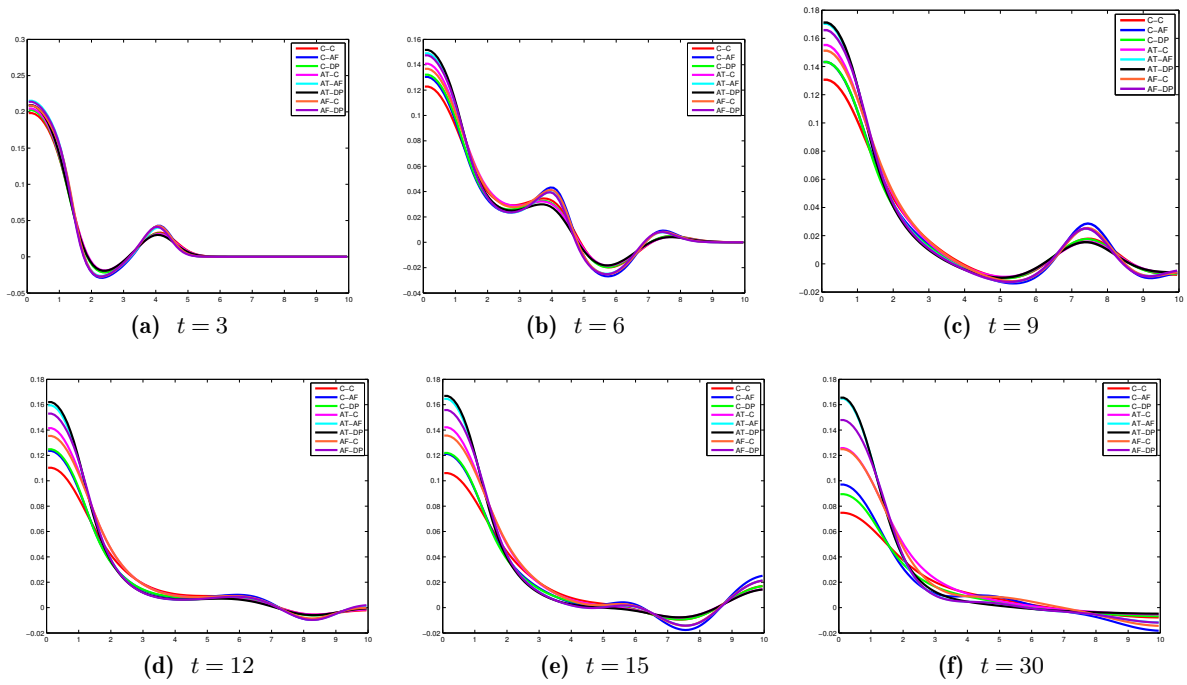
We now consider the test case proposed in [59] (see also in [27]) with the initial condition given by

$$h(x,y,0) = 1 + \frac{A_0}{2}\left[1 - \tanh\left(\frac{\sqrt{(\sqrt{\lambda}x)^2 + (\frac{y}{\sqrt{\lambda}})^2} - R_i}{R_E}\right)\right], \quad u(x,y,0) = 0, \quad v(x,y,0) = 0$$

where the parameters A_0, λ, R_E and R_i respectively stand for the amplitude of the initial unbalanced height field, the aspect ratio, the edge width and the initial radius of the mass imbalance. Let us also note that positive and negative values of parameter A_0 respectively correspond to elevations and depressions of the height perturbation. The parameter R_E is strongly related to the smoothness of the initial imbalance and we will fix $R_E = 0.1$ for the purpose of this test case.
In this test case, we consider a uniform mesh with 200×200 grid cells for the domain $[-10,10] \times [-10,10]$ with periodic boundary conditions. The gravity and the Coriolis force have been fixed to $g = \omega = 1$.
Since the rotating shallow water equation admit shock solutions, the work in [59] uses a very high order scheme with a fine mesh (500×500 grid cells) on the purpose to accurately capture the behavior of the time adjustment process. In particular, they adapt the weighted essentially non-oscillatory method (WENO) from [63] to obtain a fully discrete scheme with fifth-order accuracy in space in smooth regions, third-order accurate near discontinuities and use a fourth-order Runge-Kutta integration scheme in time. The main goal in this test case is to show that the first-order modified Godunov type schemes with appropriate corrections for numerical diffusions can obtain promising results in comparison to the ones obtained in [59].

Figures 7.8 and 7.9 show the axisymmetric adjustment ($\lambda = 1$) of the elevation and depression. As can be seen, all presented solutions look similar during the transient state, but totally different at the final state. The final adjustment of the solution obtained by the classical scheme is far from the other solutions and the top height of this scheme is much lower than the reference solution in [59]. Moreover, the correction on the velocity equation now becomes less important than in the previous test case, since the result obtained here with this strategy is just a little higher than that of the classical scheme, while we can gain more with the correction on the mass equation. Since the top height of the final adjustment computed with the AT-DP, AT-AF and AF-DP schemes are close to the reference one in [59], we have an evidence that they behave much better than the other schemes. On the other hand, Figure 7.8 shows that the AT-AF and AT-DP schemes seem to be better than the AF-DP scheme for the elevation while the converse holds for the depression, as shown in Figure 7.9.

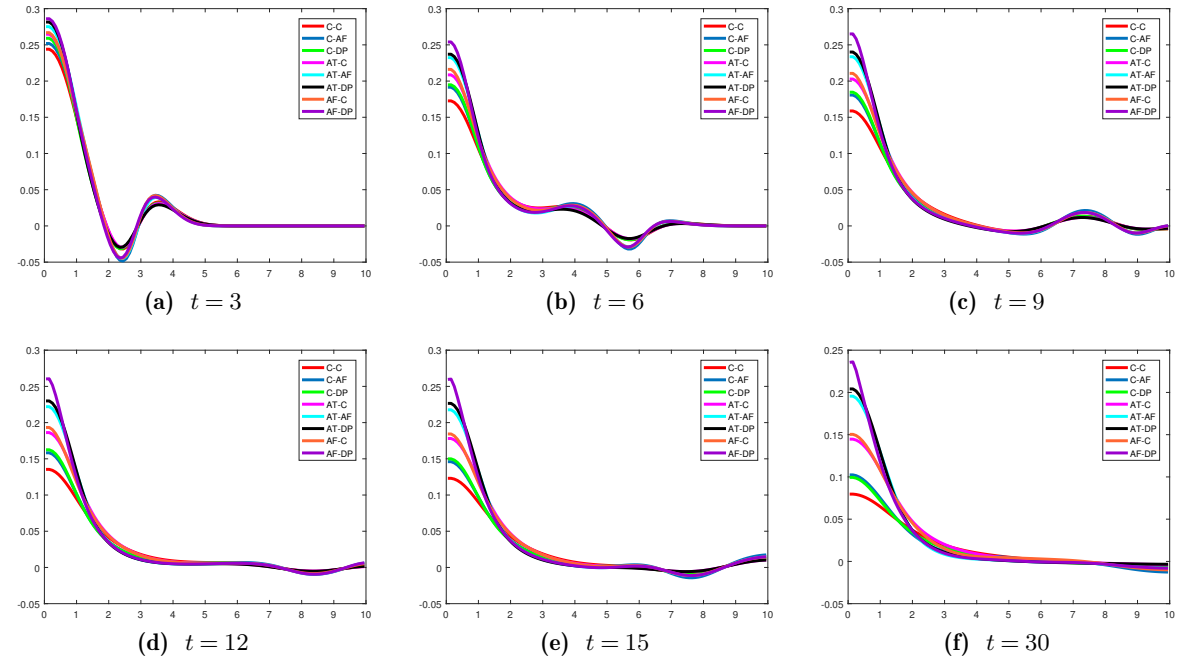
We now turn to non-axisymmetric mass adjustment with an elliptical initial condition obtained by increasing the aspect ratio to $\lambda = 2.5$. The initial velocity is also set up with zero. The time dependent mass adjustment computed by the AT-DP scheme is plotted in a sequence of times in Figures 7.10 (elevation) and 7.11 (depression). Similar results for the AT-AF scheme are shown in Figures 7.12 and 7.13. Those figures show that the elevation leads to two shock waves propagating outwards while the depression leads to a rarefaction wave followed by a shock. On the other hand, as can be seen in Figures 7.14 and 7.15, when time evolves, we observe a clockwise rotation in the test case with $A_0 = 0.5$, while a counter-clockwise rotation occurs with $A_0 = -0.5$. This is an agreement with the results in [59].



(a) $t = 3$ **(b)** $t = 6$ **(c)** $t = 9$

(d) $t = 12$ **(e)** $t = 15$ **(f)** $t = 30$

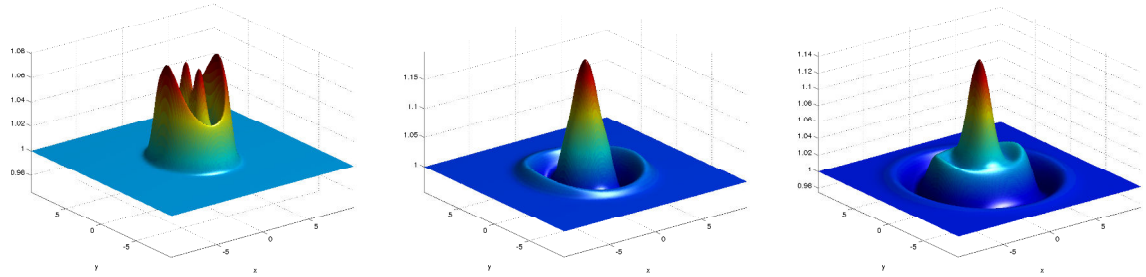
Figure 7.8: *Time-dependent mass adjustment with initial elevation: evolution of the perturbation height h with parameters $A_0 = 0.5$, $\lambda = 1$, $R_E = 0.1$ and $R_i = 1$.*



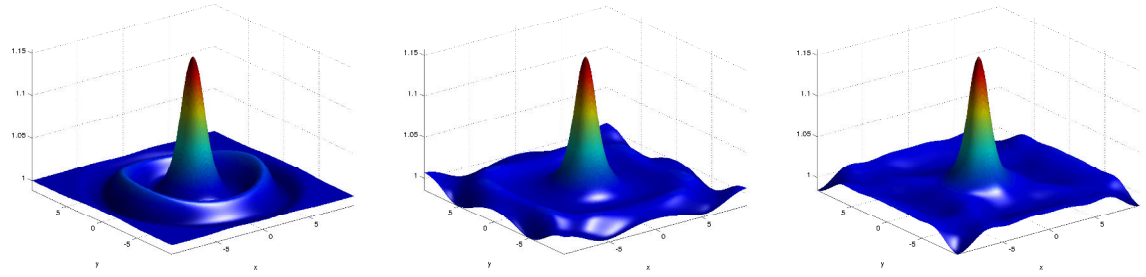
(a) $t = 3$ **(b)** $t = 6$ **(c)** $t = 9$

(d) $t = 12$ **(e)** $t = 15$ **(f)** $t = 30$

Figure 7.9: *Time-dependent mass adjustment with initial depression: evolution of the perturbation height h with parameters $A_0 = -0.5$, $\lambda = 1$, $R_E = 0.1$ and $R_i = 1$.*

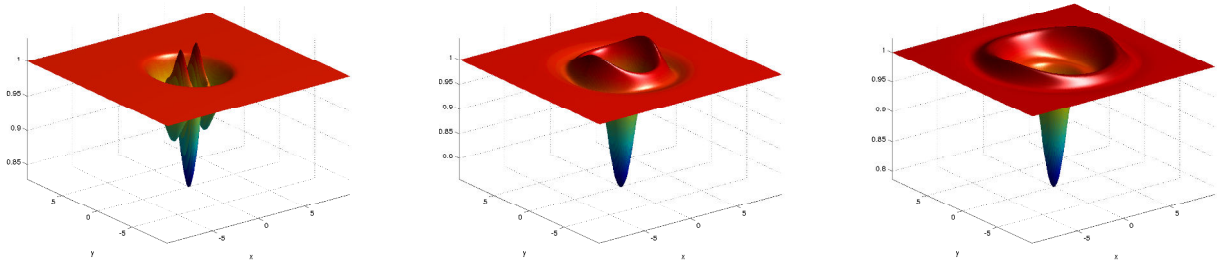


(a) $t = 2$ **(b)** $t = 4$ **(c)** $t = 6$

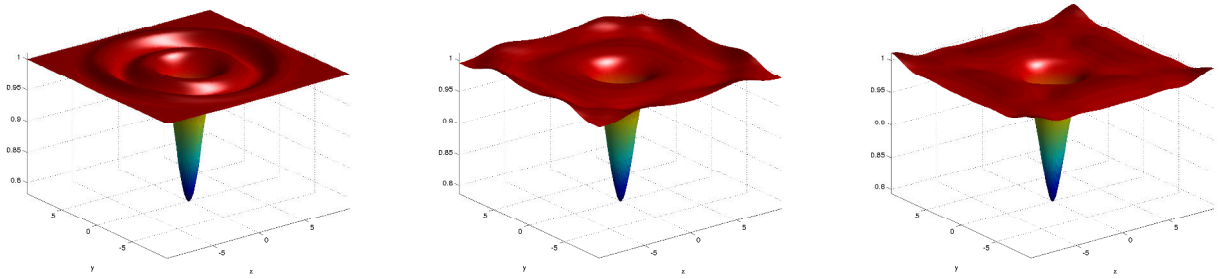


(d) $t = 8$ **(e)** $t = 14$ **(f)** $t = 20$

Figure 7.10: *Time-dependent mass adjustment with initial elevation (AT-DP scheme): evolution of the perturbation height h with $\lambda = 2.5$*

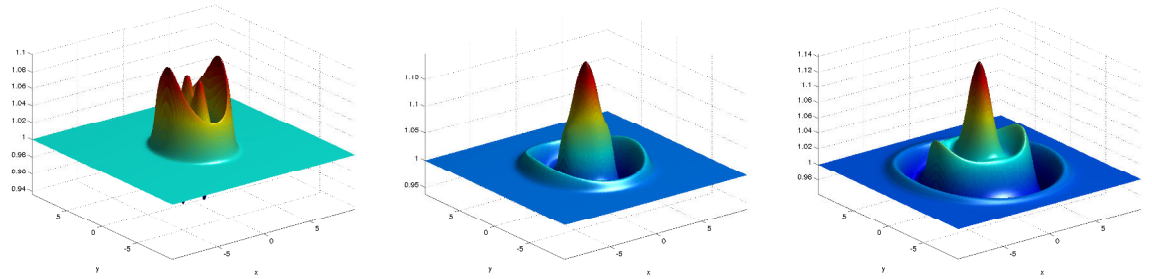


(a) $t = 2$ **(b)** $t = 4$ **(c)** $t = 6$

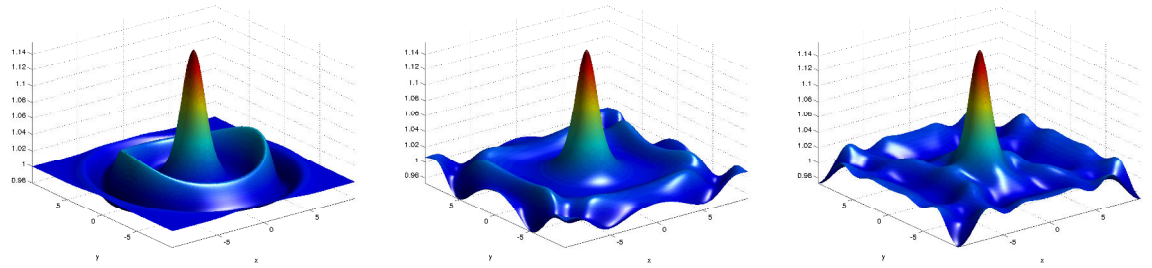


(d) $t = 8$ **(e)** $t = 14$ **(f)** $t = 20$

Figure 7.11: *Time-dependent mass adjustment with initial depression (AT-DP scheme): evolution of the perturbation height h with $\lambda = 2.5$*

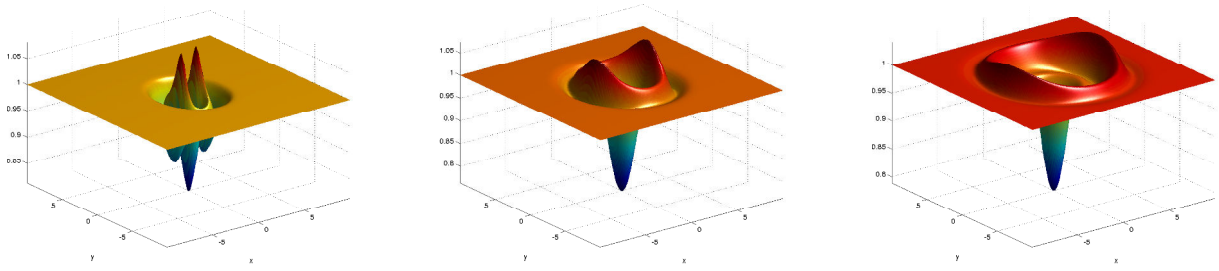


(a) $t = 2$ **(b)** $t = 4$ **(c)** $t = 6$

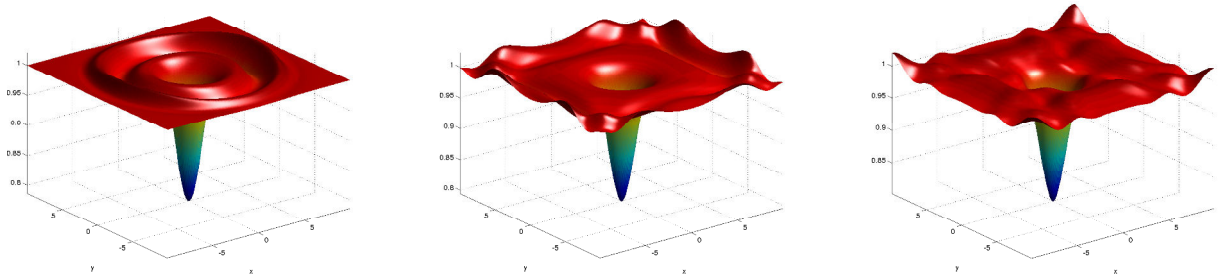


(d) $t = 8$ **(e)** $t = 14$ **(f)** $t = 20$

Figure 7.12: *Time-dependent mass adjustment with initial elevation (AT-AF scheme): evolution of the perturbation height h with $\lambda = 2.5$*

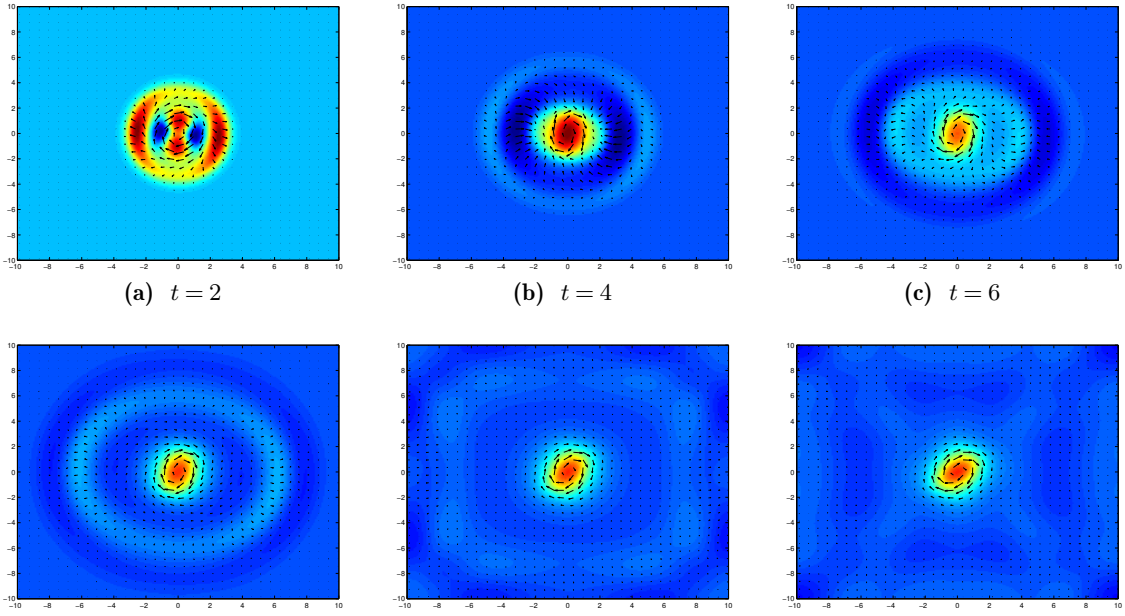


(a) $t = 2$ **(b)** $t = 4$ **(c)** $t = 6$



(d) $t = 8$ **(e)** $t = 14$ **(f)** $t = 20$

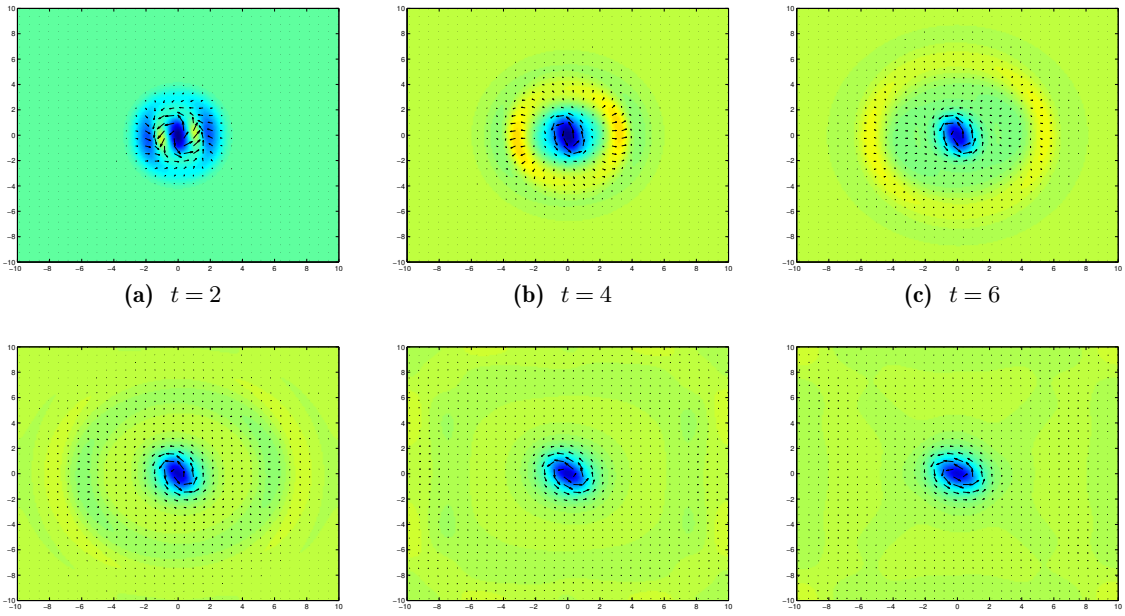
Figure 7.13: *Time-dependent mass adjustment with initial depression (AT-AF scheme): evolution of the perturbation height h with $\lambda = 2.5$*



(a) $t = 2$ **(b)** $t = 4$ **(c)** $t = 6$

(d) $t = 8$ **(e)** $t = 14$ **(f)** $t = 20$

Figure 7.14: *Time-dependent mass adjustment with initial elevation (AT-DP scheme): evolution of the perturbation height (flat view) and velocity field with* $\lambda = 2.5$



(a) $t = 2$ **(b)** $t = 4$ **(c)** $t = 6$

(d) $t = 8$ **(e)** $t = 14$ **(f)** $t = 20$

Figure 7.15: *Time-dependent mass adjustment with initial depression (AT-DP scheme): evolution of the perturbation height (flat view) and velocity field with* $\lambda = 2.5$

7.4.3 Water column test case with discontinuous initial condition (circular dam-break test case)

In this test case, we consider the initial condition given by

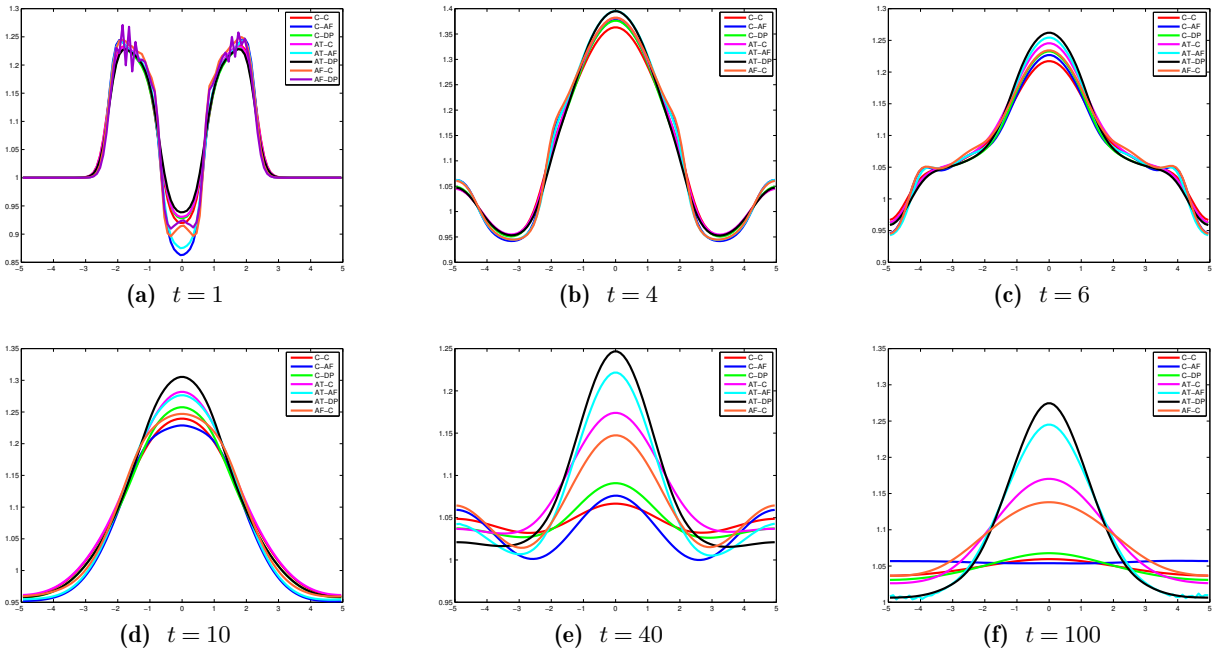
$$\begin{cases} h(x,y,t=0) = \begin{cases} 1+A_0, & \text{if} \quad x^2+y^2 \leq R_0 \\ 1, & \text{if} \quad x^2+y^2 > R_0. \end{cases} \\ u(x,,y,t=0) = 0, \\ v(x,,y,t=0) = 0. \end{cases} \tag{7.21}$$

with the domain $\Omega = [-5,5] \times [-5,5]$. The gravity and Coriolis force are fixed to $g = \Omega = 1$. Parameters A_0 and R_0 respectively stand for the amplitude of the initial mass imbalance and for its radius. Let us note that this initial condition is very similar to the one in the previous test case with the smoothness parameter $R_E = 0$. The purpose of this test case is to check the stability of all proposed schemes with discontinuous initial condition and the long time behavior of the numerical solution. Since we use periodic boundary conditions in this test case, whenever a wave goes out, another wave goes in the domain. However, the waves are damped by numerical diffusion and after long time the remaining wave is nearly the geostrophic equilibrium.

In Figure 7.16, we present the evolution of the water depth for different schemes. There is a drawback with the AF-DP scheme since this strategy introduces some oscillations at time $t = 1$. Therefore, for the rest of the figures, we do not show the result of this scheme. We can observe that the transient states of the proposed schemes are quite similar for short times, but totally different for the longtime behavior. In particular, the AT-DP scheme which captures well the discrete geostrophic equilibrium in the linear case has the top of the water height higher than the other schemes. Moreover, since the C-DP and C-AF solutions are very close to that of the classical scheme in general, the correction on the mass equation seems to be more important than that on the velocity equations. In consideration of the modified schemes with correction on the mass equation, the water height of the Apparent Topography scheme with All Froude correction for velocity equations (AT-AF) is higher than with the Apparent Topography - Classical scheme (AT-C). On the other hand, it is more useful to modify the scheme by the Apparent Topography or Divergence Penalisation methods than by the All Froude technique, since those strategies have diffusion terms that exactly cancel on geostrophic equilibriums in the linear wave equation (see [53] for more detail), while the All Froude strategy retains a small but non-zero diffusion which therefore has a noticeable effect in the long run.
In Figure 7.17, we plot the final state of three different types of schemes: C-C, AT-C (correction only for the mass equation) and AT-DP (correction for both mass and velocity equations). As can be seen, the water height of the classical scheme is very close to a constant state and its velocity filed is nearly equal to zero. The horizontal velocity u and the vertical velocity v of the AT-C scheme are close to a constant in the x and y direction respectively. The solution of the AT-DP scheme is very close to the geostrophic equilibrium. These results for the nonlinear rotating shallow water equations are similar to the results of those schemes applied to the linear wave equation with Coriolis force [53].

7.5 Conclusion

In this work, we point out that the preservation of the geostrophic equilibrium including the divergence constraint is a difficult issue. It is necessary to combine the Apparent Topography strategy with some correction to the Low Froude number problem, such as the All Froude strategy and Divergence Penalisation technique in the aim to derive new schemes which are more accurate than the classical ones. Some numerical test cases are investigated to show that the AT-DP and AT-AF have good behavior around geostrophic equilibriums while the classical schemes totally



(a) $t = 1$ (b) $t = 4$ (c) $t = 6$

(d) $t = 10$ (e) $t = 40$ (f) $t = 100$

Figure 7.16: *Water column test case: evolution of the water height h with $A_0 = R_0 = 1$ and 100×100 grid cells*

fail to capture this important phenomena.
In a future work, we aim to study a theoretical proof for the stability of the proposed schemes. Another interesting investigation is the influence of the cell geometry on the Godunov type scheme applied to the nonlinear rotating shallow water equations. It is motivated by [21] with the fact that the divergence constraint is no more a problem on triangular meshes.

7.A Conservation properties of rotating shallow water equation.

Let us define $\nabla^\perp = (-\partial_y, \partial_x)$ and if φ is a scalar field, we can define the orthogonal gradient as

$$\nabla^\perp \varphi = \left(-\frac{\partial \varphi}{\partial y}, \frac{\partial \varphi}{\partial x} \right).$$

For a velocity vector field $\boldsymbol{u} = (u, v)$, the relative vorticity is defined as

$$\zeta = curl(\boldsymbol{u}) = \nabla^\perp \cdot \boldsymbol{u} = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}.$$

We now note that the convection term of the momentum equation can be written as $\boldsymbol{u} \cdot \nabla \boldsymbol{u} = \nabla \left(\frac{|\boldsymbol{u}|^2}{2} \right) + \zeta \boldsymbol{u}^\perp$ which leads to

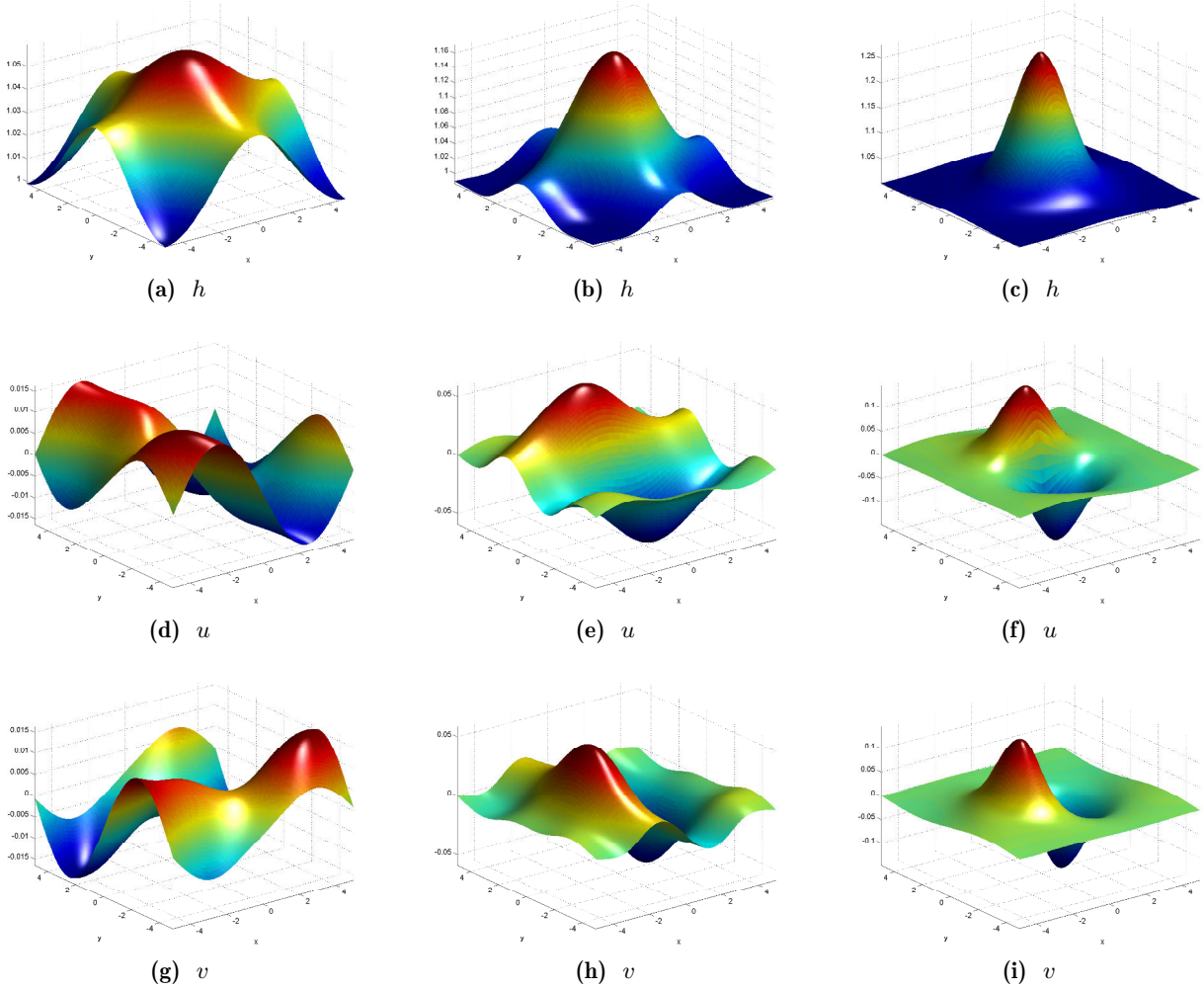
$$curl(\boldsymbol{u} \cdot \nabla \boldsymbol{u}) = \nabla^\perp \cdot (\zeta \boldsymbol{u}^\perp) = \nabla \cdot (\zeta \boldsymbol{u}).$$

This implies that the momentum equation can be written as the evolution of vorticity

$$\frac{\partial \zeta}{\partial t} + \nabla \cdot [(\zeta + \Omega)\, \boldsymbol{u}] = 0.$$

If we assume that $\partial_t \Omega = 0$, we have the conservation form of the absolute vorticity

$$\frac{\partial}{\partial t}(\zeta + \Omega) + \nabla \cdot [(\zeta + \Omega)\, \boldsymbol{u}] = 0. \tag{7.A.1}$$



(a) h **(b)** h **(c)** h

(d) u **(e)** u **(f)** u

(g) v **(h)** v **(i)** v

Figure 7.17: *Water column test case: final states of C-C (left column), AT-C (center) and AT-DP (right column) schemes with* $A_0 = R_0 = 1$ *and* 100×100 *grid cells*

We now denote the material derivative as $\frac{D}{Dt}() = \frac{\partial}{\partial t}() + \boldsymbol{u} \cdot \nabla()$; then the equation for the absolute vorticity can be written as

$$\frac{D}{Dt}(\zeta + \Omega) + (\zeta + \Omega)\nabla \cdot \boldsymbol{u} = 0. \tag{7.A.2}$$

By multiplying the mass equation with $\frac{\zeta+\Omega}{h}$ and subtracting to the equation (7.A.2), we obtain

$$\frac{D}{Dt}(\zeta + \Omega) - \left(\frac{\zeta + \Omega}{h}\right)\frac{D}{Dt}h = 0.$$

We then divide this equation by h to get

$$\frac{D}{Dt}\left(\frac{\zeta + \Omega}{h}\right) = 0. \tag{7.A.3}$$

Therefore, the potential vorticity $q = \frac{\zeta+\Omega}{h}$ is conservative. If we now assume that the Coriolis parameter Ω is a constant, the conservation of the potential vorticity provides us with a relation between the water depth h and the vorticity ζ. For example, when h increases, the vorticity must increase to ensure the conservation.

Another important property of the rotating shallow water equation is the conservation of energy. We now denote the kinetic energy and potential energy respectively by

$$\mathcal{K}_{\mathcal{E}} = \frac{h}{2}|\boldsymbol{u}|^2 \quad \text{and} \quad \mathcal{P}_{\mathcal{E}} = \frac{gh\,(h+2b)}{2}.$$

By using the mass equation, we have

$$\begin{aligned}
\frac{\partial}{\partial t}\left(\frac{h}{2}|\boldsymbol{u}|^2\right) + \nabla\cdot\left[\left(\frac{h}{2}|\boldsymbol{u}|^2\right)\boldsymbol{u}\right] &= \frac{|\boldsymbol{u}|^2}{2}\left(\frac{\partial}{\partial t}h + \nabla\cdot(h\boldsymbol{u})\right) + h\left[\frac{\partial}{\partial t}\left(\frac{|\boldsymbol{u}|^2}{2}\right) + \boldsymbol{u}\cdot\nabla\left(\frac{|\boldsymbol{u}|^2}{2}\right)\right] \\
&= h\boldsymbol{u}\cdot\left(\frac{\partial}{\partial t}\boldsymbol{u} + \boldsymbol{u}\cdot\nabla\boldsymbol{u}\right) = h\boldsymbol{u}\cdot\left(-g\nabla(h+b) - \Omega\boldsymbol{u}^{\perp}\right) \\
&= -g\boldsymbol{u}\cdot\nabla\left(\frac{h^2}{2}\right) - gh\boldsymbol{u}\cdot\nabla b. \qquad (7.A.4)
\end{aligned}$$

Moreover, we also have

$$\begin{aligned}
\frac{\partial}{\partial t}\left(\frac{1}{2}gh(h+2b)\right) + \nabla\cdot\left(\frac{1}{2}gh(h+2b)\boldsymbol{u}\right) &= h\left[\frac{\partial}{\partial t}\left(\frac{g(h+2b)}{2}\right) + \boldsymbol{u}\cdot\nabla\left(\frac{g(h+2b)}{2}\right)\right] \\
&= -\frac{1}{2}gh^2\nabla\cdot\boldsymbol{u} + gh\boldsymbol{u}\cdot\nabla b. \qquad (7.A.5)
\end{aligned}$$

We now define the total energy as the sum of the kinetic energy and the potential energy $E = \mathcal{K}_{\mathcal{E}} + \mathcal{P}_{\mathcal{E}}$; then, adding (7.A.4) and (7.A.5), we obtain

$$\frac{\partial}{\partial t}E + \nabla\cdot(E\boldsymbol{u}) + \nabla\cdot\left(\frac{1}{2}gh^2\boldsymbol{u}\right) = 0. \qquad (7.A.6)$$

If we consider the rotating shallow water equation on the domain $\mathbb{T}$ with periodic boundary condition, we get

$$\frac{d}{dt}\int_{\mathbb{T}} E(\boldsymbol{x},t)\ \mathrm{d}\boldsymbol{x} = 0. \qquad (7.A.7)$$

For any function $\mathcal{G}(q)$ of the potential vorticity, we have

$$\frac{\partial}{\partial t}\left(h\mathcal{G}(q)\right) = \mathcal{G}(q)\frac{\partial}{\partial t}h + h\frac{\partial}{\partial t}\mathcal{G}(q) = -\mathcal{G}(q)\nabla\cdot(h\boldsymbol{u}) - h\boldsymbol{u}\cdot\nabla\mathcal{G}(q)$$

which leads to the conservation of the quantity $h\mathcal{G}(q)$

$$\frac{\partial}{\partial t}\left(h\mathcal{G}(q)\right) + \nabla\cdot\left(h\mathcal{G}(q)\boldsymbol{u}\right) = 0. \qquad (7.A.8)$$

As a result, we obtain

$$\frac{\partial}{\partial t}\int_{\mathbb{T}} h\mathcal{G}(q)\ \mathrm{d}\mathbf{x} = 0. \qquad (7.A.9)$$

We can apply this property to $\mathcal{G}(q) = q^2$ to obtain the conservation of the total enstrophy

$$\frac{\partial}{\partial t}\int_{\mathbb{T}} hq^2\ \mathrm{d}\mathbf{x} = 0.$$

7.B The Roe solver applied to the shallow water equation

To derive the formula of the numerical flux, we need to evaluate the flux $F_x(U)n_x + F_y(U)n_y$ over the borders A_{ij} separating the cell T_i ans its neighbors the cell T_j. On the purpose to do that we rewrite the homogeneous shallow water equation as quasi-linear form

$$\frac{\partial U}{\partial t} + \frac{\partial F_x(U)}{\partial U}\frac{\partial U}{\partial x} + \frac{\partial F_y(U)}{\partial U}\frac{\partial U}{\partial y} = 0$$

where $\frac{\partial F_x(U)}{\partial U}$ and $\frac{\partial F_y(U)}{\partial U}$ are the Jacobians of flux functions respectively given by

$$\frac{\partial F_x(U)}{\partial U} = \begin{pmatrix} 0 & 1 & 0 \\ -u^2+gh & 2u & 0 \\ -uv & v & u \end{pmatrix} \quad \text{and} \quad \frac{\partial F_y(U)}{\partial U} = \begin{pmatrix} 0 & 0 & 1 \\ -uv & v & u \\ -v^2+gh & 0 & 2v \end{pmatrix}.$$

Let us denote $c = \sqrt{gh}$, then we have

$$A = \frac{\partial (F(U).\boldsymbol{n})}{\partial U} = \begin{pmatrix} 0 & n_x & n_y \\ (c^2-u^2)n_x - uvn_y & 2un_x + vn_y & un_y \\ (c^2-v^2)n_y - uvn_x & vn_x & un_x + 2vn_y \end{pmatrix}.$$

The eigenvalues of the Jacobian matrix A are verified by

$$\lambda_1 = un_x + vn_y - c, \qquad \lambda_2 = un_x + vn_y, \qquad \text{and } \lambda_3 = un_x + vn_y + c.$$

Let denote R be the matrix of right eigenvectors

$$R = \begin{pmatrix} 1 & 0 & 1 \\ u - cn_x & -n_y & u + cn_x \\ v - cn_x & n_x & v + cn_y \end{pmatrix}.$$

Then, the inverse of this matrix is given by

$$R^{-1} = \frac{1}{2c}\begin{pmatrix} c + un_x + vn_y & -n_x & -n_y \\ 2c(un_y - vn_x) & -2cn_y & 2cn_x \\ c - un_x - vn_y & n_x & n_y \end{pmatrix} = \begin{pmatrix} \frac{1}{2} + \frac{\mathbf{u}.\mathbf{n}}{2c} & \frac{-n_x}{2c} & \frac{-n_y}{2c} \\ -\mathbf{u}.\mathbf{n}^{\perp} & -n_y & n_x \\ \frac{1}{2} - \frac{\mathbf{u}.\mathbf{n}}{2c} & \frac{n_x}{2c} & \frac{n_y}{2c} \end{pmatrix}$$

One possible way to compute the numerical flux at the interface $\partial T_i \cap \partial T_j$ is using the solution of the local one-dimensional Riemann problem

$$\begin{cases} \partial_t U + \frac{\partial (F(U).\boldsymbol{n})}{\partial U}\partial_\xi(U) = 0 \\ U(t=0,\xi) = \begin{cases} U_L & \text{if } \xi < 0 \\ U_R & \text{if } \xi > 0 \end{cases} \end{cases} \tag{7.B.10}$$

We now denote the approximation of the above Jacobian matrix by $A_{\mathbf{n_{ij}}}(U_i, U_j)$ and this approximation must fulfill some classical requirements. In particular:

i. It must depend on right and left value U_j and U_i.

ii. $(F(U_j) - F(U_i))\,.\mathbf{n}_{ij} = A_{\mathbf{n_{ij}}}(U_i, U_j)\,(U_j - U_i)$.

iii. It must have distinct real eigenvalues and a complete set of eigenvectors.

iv. It must become exact flux Jacobian when $U_j = U_i$

The Roe solver in [60] satisfies those requirements and this upwind type scheme can be written as

$$\Phi_{ij}^{\text{Roe}} = \frac{F(U_i)+F(U_j)}{2}.\boldsymbol{n}_{ij} - \frac{|A_{\mathbf{n_{ij}}}(U_i,U_j)|}{2}(U_j - U_i)$$

where the Roe matrix $A_{\mathbf{n_{ij}}}(U_i,U_j) = A_{\mathbf{n_{ij}}}(U_{ij})$ is defined with the help of the Roe averages

$$h_{ij} = \sqrt{h_i h_j}, \quad u_{ij} = \frac{\sqrt{h_i}u_i + \sqrt{h_j}u_j}{\sqrt{h_i}+\sqrt{h_j}}, \quad v_{ij} = \frac{\sqrt{h_i}v_i + \sqrt{h_j}v_j}{\sqrt{h_i}+\sqrt{h_j}}, \quad c_{ij} = \sqrt{g h_{ij}}.$$

Let us note that the approximate Jacobian $A_{\mathbf{n_{ij}}}(U_{ij})$ has the shape of A, but it is evaluated at Roe average states

$$A_{\mathbf{n_{ij}}}(U_{ij}) = \begin{pmatrix} 0 & n_x & n_y \\ (c_{ij}^2 - u_{ij}^2)n_x - u_{ij}v_{ij}n_y & 2u_{ij}n_x + v_{ij}n_y & u_{ij}n_y \\ (c_{ij}^2 - v_{ij}^2)n_y - u_{ij}v_{ij}n_x & v_{ij}n_x & u_{ij}n_x + 2v_{ij}n_y \end{pmatrix}$$

On the other hand, we also notice that

$$|A_{\mathbf{n_{ij}}}(U_{ij})| = R|\Lambda|R^{-1}$$

where the diagonal matrix Λ is simply given by

$$\Lambda_{\mathbf{n_{ij}}}(U_{ij}) = \begin{pmatrix} |\mathbf{u}_{ij}.\mathbf{n}_{ij} - c_{ij}| & 0 & 0 \\ 0 & |\mathbf{u}_{ij}.\mathbf{n}_{ij}| & 0 \\ 0 & 0 & |\mathbf{u}_{ij}.\mathbf{n}_{ij} + c_{ij}| \end{pmatrix}$$

As a result, with notation $\Delta(\cdot) = (\cdot)_j - (\cdot)_i$ and $\mathbf{n}_{ij}^{\perp} = (-n_y, n_x)^T$, the Roe flux can be rewritten as

$$\begin{aligned}
\Phi_{ij}^{\text{Roe}} &= \frac{F(U_i)+F(U_j)}{2}.\boldsymbol{n}_{ij} - \frac{R(U_{ij},\mathbf{n}_{ij})|\Lambda_{\mathbf{n}_{ij}}(U_{ij})|R^{-1}(U_{ij},\mathbf{n}_{ij})}{2}(U_j - U_i) \\
&= \frac{F(U_i)+F(U_j)}{2}.\boldsymbol{n}_{ij} - \frac{1}{4}|\mathbf{u}_{ij}.\mathbf{n}_{ij} - c_{ij}|\begin{pmatrix} \Delta h - \dfrac{h_{ij}}{c_{ij}}\Delta(\mathbf{u}\cdot\mathbf{n}_{ij}) \end{pmatrix}\begin{pmatrix} 1 \\ \mathbf{u}_{ij} - c_{ij}\mathbf{n}_{ij} \end{pmatrix} \\
&\quad - \frac{1}{2}|\mathbf{u}_{ij}.\mathbf{n}_{ij}|h_{ij}\Delta(\mathbf{u}\cdot\mathbf{n}_{ij}^{\perp})\begin{pmatrix} 0 \\ \mathbf{n}_{ij}^{\perp} \end{pmatrix} - \frac{1}{4}|\mathbf{u}_{ij}.\mathbf{n}_{ij} + c_{ij}|\begin{pmatrix} \Delta h + \dfrac{h_{ij}}{c_{ij}}\Delta(\mathbf{u}\cdot\mathbf{n}_{ij}) \end{pmatrix}\begin{pmatrix} 1 \\ \mathbf{u}_{ij} + c_{ij}\mathbf{n}_{ij} \end{pmatrix}.
\end{aligned}$$

The well known disadvantage of the Roe scheme is that it is does not satisfy entropy condition, so we need entropy fix to overcome this problem. The entropy fix according to Harten and Hyman in [64] is applied to the acoustic waves ($k = 1$ and $k = 3$)

$$Q^H(\lambda_k) = \begin{cases} \frac{1}{2}\left(\frac{\lambda_k^2}{\delta_k} + \delta_k\right) & \text{if} \quad |\lambda_k| \le \delta_k \\ |\lambda_k| & \text{if} \quad |\lambda_k| > \delta_k \end{cases}$$

where

$$\delta_k = \max\{0, \lambda_k - \lambda_k(U_L), \lambda_k(U_R) - \lambda_k\}.$$

The HLL solver proposed by Harten , Lax and van Leer in [65] is also used to calculate the flux at the interface. It can be expressed as

$$\Phi_{ij}^{\text{HLL}} = \begin{cases} F(U_i).\boldsymbol{n}_{ij} & \text{if } S_L \geq 0 \\ \frac{S_R F(U_i).\boldsymbol{n}_{ij} - S_L F(U_j).\boldsymbol{n}_{ij} + S_R S_L (U_j - U_i)}{S_R - S_L} & \text{if } S_L \leq 0 \leq S_R \\ F(U_j).\boldsymbol{n}_{ij} & \text{if } S_R \leq 0. \end{cases} \tag{7.B.11}$$

where S_L and S_R are verified by

$$S_L = \min\{U_i.\boldsymbol{n}_{ij} - \sqrt{gh_i}, U_j.\boldsymbol{n}_{ij} - \sqrt{gh_j}\}$$

and

$$S_R = \max\{U_i.\boldsymbol{n}_{ij} + \sqrt{gh_i}, U_j.\boldsymbol{n}_{ij} + \sqrt{gh_j}\}.$$

Part IV

Outlooks and conclusion

Conclusion

> *Any intelligent fool can make things bigger, more complex, and more violent. It takes a touch of genius and a lot of courage to move in the opposite direction.*
>
> E. F. Schumacher

The ocean circulation is strongly influenced by the rotation of the Earth through the fictitious force known as the Coriolis force. The aim of this dissertation is to develop finite volume schemes which are able to preserve the well known geostrophic equilibrium which is a balance between the horizontal Coriolis force and pressure gradient. The main part of this thesis is devoted to the study of the linear wave equation with Coriolis source term which is obtained from the rotating shallow water equations by using an asymptotic expansion. We follow the framework proposed in [20] to analyse the behaviour of the Godunov type scheme applied to this linear wave equation. The main contributions of this work can be summarised as:

In the one dimensional case, unlike the homogeneous equations, the inaccuracy problem of the Godunov scheme appears in the presence of Coriolis source term. By analysing the kernel of the modified equation associated to the classical scheme, this work pointed out that the numerical diffusion on the pressure equation is responsible for this drawback. One simple correction consists in making this diffusion term vanish which provides a "Low Froude" scheme. Another correction is to adapt the "Apparent Topography" technique in [13]. Both schemes are proven to well capture the discrete 1D geostrophic equilibrium. However, the kernel of the Low Froude scheme is defined at cell centers while it is located at the cell interfaces for the Apparent Topography scheme.

Fourier analysis is also performed to compare these strategies on collocated meshes in terms of dispersion relation and damping error. We also provide the stability condition of those schemes by using the Von Neumann method. These stability conditions are less restrictive than the classical ones using an appropriate discretisation in time of the Coriolis source term. Due to the structure of the discrete kernel and to ensure the obtained schemes are totally explicit, the time step of the Apparent Topography scheme strongly depends on the Coriolis parameter. On the other hand, since the dispersion law of the collocated schemes is not a monotonic curve like in the continuous model, we propose a staggered strategy based on the aforementioned corrections in order to obtain the staggered type schemes. They turn out to have a better dispersion law than that on collocated grids. More importantly, we can ensure that the dispersion of staggered schemes is a monotonic function which helps us avoid the oscillation of the shortest wave. Some numerical results lead to the conclusion that the Low Froude staggered scheme seems to be the best candidate since its dispersion law is robust with time step as well as the relation between Rossby deformation radius and mesh sizes. This scheme also preserves the subspace which is orthogonal to the kernel.

In the two dimensional case, it is more challenging for numerical schemes to preserve the non-trivial steady state (2D geostrophic equilibrium). In this dissertation, we point out that unlike the 1D case, the problem for the inaccuracy is also linked to the numerical viscosity on the velocity equation. This is because the 2D geostrophic equilibrium also implies the free divergence constraint and the classical scheme are unable to handle it. To improve the accuracy of numerical schemes, this thesis proposes the extension of the Apparent Topography scheme by a coupling with Low Mach and Divergence Penalisation strategies mentioned in [20].

For collocated meshes, we propose two strategies: cell-centered and vertex-based schemes. Moreover, we also introduce the staggered type schemes on both B and D grids. Fourier analysis indicates that the dispersion law of B grid scheme is more accurate than that of D grid. This is because the B grid allows to evaluate the Coriolis force easily, without averaging while the average is required for the Coriolis source term on D grid. However, the D grid scheme has a good damping rate in the short wave region. As a result, the staggered schemes on D grid induces fewer oscillations than the one on B grid. We also provide some CFL conditions of both collocated and staggered schemes. Unlike the B grid scheme, the stability condition of staggered scheme on D grid actually depends on the magnitude of the Coriolis parameter. The proposed staggered type schemes on both B and D grid are proven to well capture the corresponding discrete kernel and only the Low Froude–Divergence Penalisation scheme with suitable time discretisation for the velocity field on the Coriolis force and pressure equation is an orthogonality preserving scheme.

In this dissertation, we investigated **the influence of cell geometry** on the Godunov type scheme applied to the linear wave equation with Coriolis source term. This work clearly points out the disadvantage of the collocated scheme on triangular grids since the kernel of this scheme requires that the gradient of a P1 conforming function is equal to that of a non-conforming function. To overcome this obstacle, we propose some new staggered strategies where the velocity field is computed at the primary cell centers and the pressure at the vertices. By the fact that all jumps of normal velocity components are equal to zero, the divergence constraint is no more an issue in this case and we can use all the developed strategies in the one dimensional case to cure the inaccuracy. The resulting scheme is proven to be a well-balanced scheme. Moreover, we also show that unlike the Apparent Topography scheme, the Low Froude scheme is orthogonality preserving.

Finally, we extended satisfying strategies in the linear case to **the non-linear rotating shallow water equations**. Various numerical test cases such as stationary vortex and geostrophic adjustment are investigated on the purpose of showing that the AT-DP (Apparent Topography–Divergence Penalisation) and AT-AF (Apparent Topography–All Froude) are more accurate for the simulations around the geostrophic equilibrium than the classical or even AT-C schemes.

There are some natural extensions of this thesis. The first thing is to prove that an energy estimate is satisfied by the Apparent Topography scheme. This would help ensure that the obtained scheme is accurate at low Froude number at any (or locally in) time which is only investigated with numerical test cases. The second point is the optimal time step for the staggered scheme on triangular grids. Also, the other kind of staggered mesh like C grid must be investigated since this grid with finite difference schemes provides good dispersion relations. An important task is to develop the theoretical analysis for the collocated scheme applied to the rotating shallow water equations as well as to extend the staggered scheme on triangular meshes to the non-linear case.

APPENDIX A

Inertial Oscillation

Many of life's failures are people who did not realize how close they were to success when they gave up.

Thomas A. Edison

Abstract

The inertial oscillation is one of the simplest model to express the time dependent motion under the Coriolis force due to the Earth's rotation. In this work, we review some basic properties of the inertial oscillation and perform the analysis for the θ-scheme applied to the Coriolis source term.

Chapter content

A.1 Preliminary result

Lemma A.1. *Let us consider the second-order polynomial $P = X^2 + BX + C$. The roots $X_\pm$ of P lie in the unit circle (*i.e.$|X_\pm| \leq 1$*) iff*

$$|C| \leq 1 \quad \textit{and} \quad |B| \leq 1 + C. \tag{A.1}$$

A.2 Basic properties of the inertial oscillation

The inertial oscillation is prescribed by the following system

$$\begin{cases} \dfrac{\partial u}{\partial t} - \omega v = 0, \\ \dfrac{\partial v}{\partial t} + \omega u = 0. \end{cases} \tag{A.2}$$

which shows the relation between the horizontal velocity u and the vertical velocity v under the Coriolis force characterised by the parameter ω. In particular, System (A.2) indicates that because of the positivity of the angular velocity ($\omega > 0$), a change in y-direction velocity v causes a change in x-direction velocity u and vice-versa. The general solution of System (A.2) is

$$u(t) = V\sin(\omega t + \phi) \quad \text{and} \quad v(t) = V\cos(\omega t + \phi)$$

where ϕ is an arbitrary constant and the speed V is also a constant determined by using the conservation of the energy

$$V = \sqrt{u(t)^2 + v(t)^2} = \sqrt{u(0)^2 + v(0)^2}. \tag{A.3}$$

The trajectory of a particle governed by this velocity field is defined by

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = u(t) \quad \text{and} \quad \frac{\mathrm{d}}{\mathrm{d}t}y(t) = v(t),$$

which implies that

$$(x - a_0)^2 + (y - b_0)^2 = \frac{V^2}{\omega^2}$$

where $(a_0, b_0) = \left(x(0) + \frac{V}{\omega}\cos\phi, y(0) - \frac{V}{\omega}\sin\phi\right)$. It means that the trajectory of a particle is a circle whose center is (a_0, b_0) and radius $r = \frac{V}{\omega}$.

A.3 Analysis of θ-scheme applied to the inertial oscillation

The θ-scheme based on the weighted average between explicit and implicit scheme apply to the system (A.2)

$$\begin{cases} \dfrac{u^{n+1} - u^n}{\Delta t} = \omega\left[\theta_1 v^n + (1-\theta_1)v^{n+1}\right], \\ \dfrac{v^{n+1} - v^n}{\Delta t} = -\omega\left[\theta_2 u^n + (1-\theta_2)u^{n+1}\right] \end{cases} \tag{A.4}$$

where $0 \leq \theta_1, \theta_2 \leq 1$ and $t^n = n\Delta t$. When $\theta_1 = \theta_2 = 1$, the scheme (A.4) is totally explicit (forward Euler method). On the contrary, the case $\theta_1 = \theta_2 = 0$ corresponds to the implicit scheme (backward Euler method). Moreover, the case $\theta_1 = \theta_2 = \frac{1}{2}$ is known as the Crank-Nicolson scheme (semi-implicit). One of the main advantages of the explicit scheme is its simplicity since the next values are calculated by using the current values. However, this algorithm has an unacceptable behaviour since the kinetic energy increases. In other words, the explicit scheme is *unstable*. To go further, we prove the following result:

Lemma A.2.

i. When $\theta_1+\theta_2>1$, the θ-scheme (A.4) is unstable. Therefore, a necessary condition for stability is $\theta_1+\theta_2\leq 1$.

ii. When $0\leq\theta_1,\theta_2\leq\frac{1}{2}$, the θ-scheme (A.4) is stable.

iii. When $(1-2\theta_1)(1-2\theta_2)<0$, the θ-scheme (A.4) is stable provided that

$$\Delta t\leq\frac{2}{\omega\sqrt{|(1-2\theta_1)(1-2\theta_2)|}}. \tag{A.5}$$

Proof. From (A.6), we have

$$\begin{pmatrix}u^{n+1}\\v^{n+1}\end{pmatrix}=\frac{1}{1+(\omega\Delta t)^2(1-\theta_1)(1-\theta_2)}\begin{pmatrix}1-(\omega\Delta t)^2\theta_2(1-\theta_1) & \omega\Delta t\\ -\omega\Delta t & 1-(\omega\Delta t)^2\theta_1(1-\theta_2)\end{pmatrix}\begin{pmatrix}u^n\\v^n\end{pmatrix}. \tag{A.6}$$

Therefore, the characteristic equation of the amplification matrix in (A.6) reads

$$\lambda^2+\xi\lambda+\zeta=0 \tag{A.7}$$

where

$$\xi=-\frac{2-(\omega\Delta t)^2(\theta_1+\theta_2-2\theta_1\theta_2)}{1+(\omega\Delta t)^2(1-\theta_1)(1-\theta_2)}\quad\text{and}\quad\zeta=\frac{1+(\omega\Delta t)^2\theta_1\theta_2}{1+(\omega\Delta t)^2(1-\theta_1)(1-\theta_2)}.$$

In order to ensure that the roots of (A.7) are in the unit circle ($|\lambda_\pm|\leq 1$), we apply Lemma A.1:

- The first condition $|\zeta|\leq 1$ leads to $(\omega\Delta t)^2\left[1-(\theta_1+\theta_2)\right]\geq 0$. This proves Point 1.
- The next condition $-\xi\leq 1+\zeta$ reduces to $(\omega\Delta t)^2\geq 0$ which does not imply any additional constraint upon Δt.
- The final condition $\xi\leq 1+\zeta$ reads

  $$-(\omega\Delta t)^2(1-2\theta_1)(1-2\theta_2)\leq 4.$$

  This implies Points 2 and 3.

□

Remark A.1. *When $(1-2\theta_1)(1-2\theta_2)\geq 0$, the CFL condition of the θ-scheme does not depend on the magnitude of the Coriolis parameter ω. Otherwise, the time step must satisfy the condition (A.5).*

Remark A.2. *When $\theta_1=\theta_2=\theta$, the choice of the parameter θ has a strong impact on the kinetic energy. In particular, from Equation (A.6), we obtain*

$$\left(u^{n+1}\right)^2+\left(v^{n+1}\right)^2=\frac{1+(\omega\Delta t)^2\theta}{1+(\omega\Delta t)^2(1-\theta)}\left[(u^n)^2+(v^n)^2\right].$$

Therefore, the kinetic energy of the inertial oscillation increases, remains constant or decreases over the time if the parameter θ is greater, equal to or less than $\frac{1}{2}$.

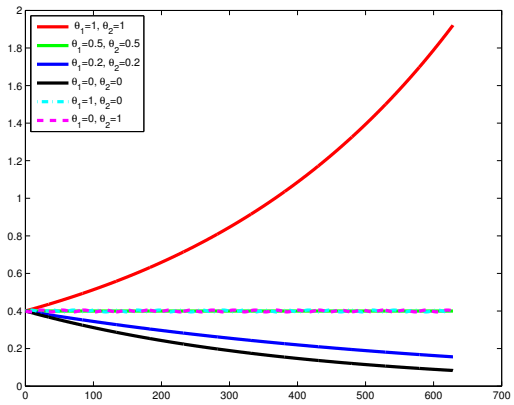
A.4 Numerical test

We now investigate the behaviour of the θ-scheme by considering the domain $[-2,2]\times[-2,2]$ with 201×201 grid cells and the Coriolis parameter $\omega = 0.1$.

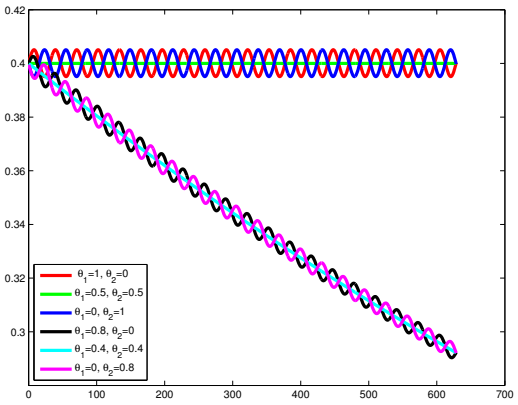
Figure A.1 confirms the result from Lemma A.2: the explicit scheme is unstable since it makes the kinetic energy increase. We have the conservation of the energy with the case $\theta_1 = \theta_2 = \frac{1}{2}$ and strictly decreasing energy for the case $\theta_1 = \theta_2 < \frac{1}{2}$ and the more implicit, the more damped the energy. For the case $\theta_1 \neq \theta_2$ and $\theta_1 + \theta_2 \leq 1$, Figure (A.1b) indicates that the energy is not strictly decreasing, but the scheme is still stable under the CFL condition (A.5). Moreover, from the numerical point of view, this figure also shows that the behaviour of the numerical energy depends on the quantity $\theta_1 + \theta_2$.

In Figure A.2 we fix the initial velocity field $u^0 = 0.1$, $v^0 = 0$ and show the influence of the parameters θ_1 and θ_2 on the trajectory of a particle. Trajectories associated to the Crank-Nicolson scheme ($\theta_1 = \theta_2 = 0.5$) is a circle at a center $(a_0, b_0) = (0, -1)$ with the radius $r = 1$. The radius of the particle changes in the other cases. Particularly, it increases in time for the explicit case since this scheme is unstable and it decreases in time for the case $\theta_1 = \theta_2 \leq \frac{1}{2}$ by the fact that this scheme makes the kinetic energy decrease.

Figure A.3 indicates that the initial condition has a strong impact on the trajectory of the particle since the circle center and the radius is prescribed by the initial condition.

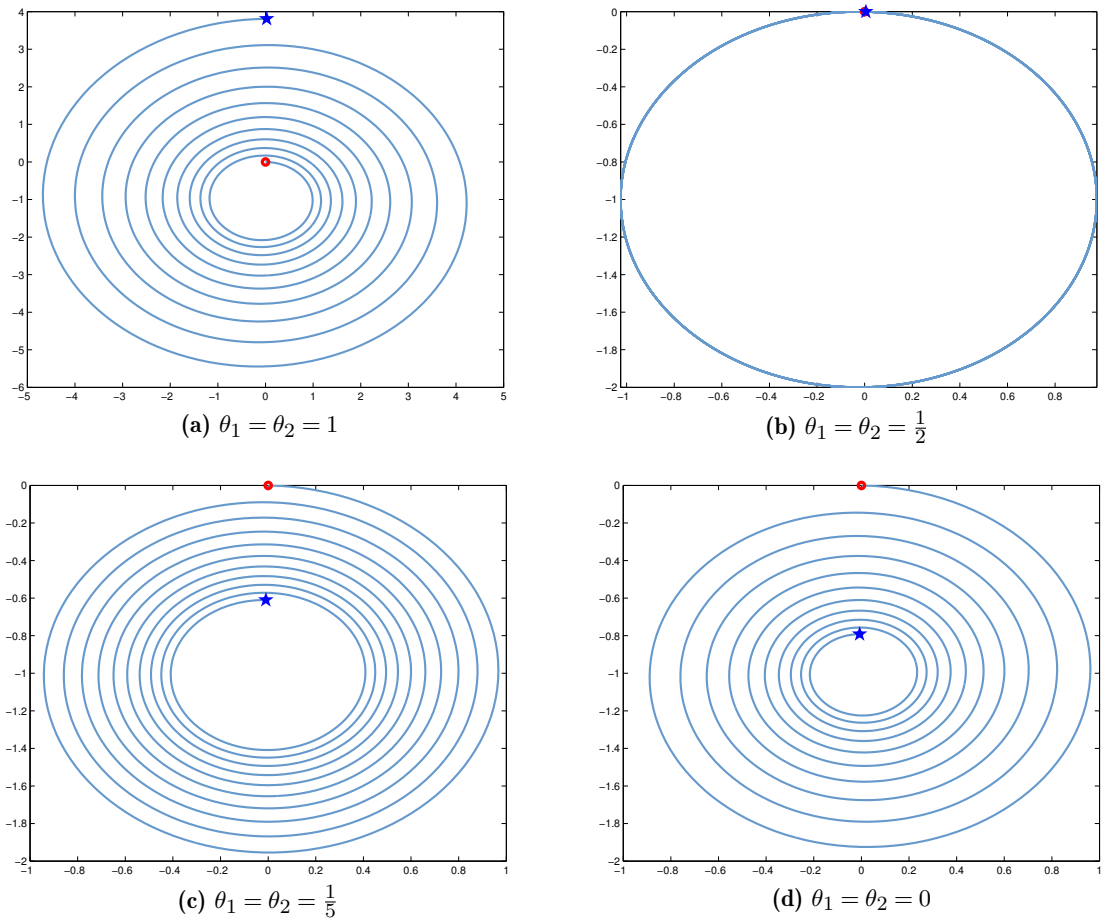


(a)



(b)

Figure A.1: *The kinetic energy of the inertial oscillation with various values of θ_1 and θ_2 for the initial condition $u^0 = 0.1, v^0 = 0$ and the time step $\Delta t = 0.5$.*



(a) $\theta_1 = \theta_2 = 1$ (b) $\theta_1 = \theta_2 = \frac{1}{2}$

(c) $\theta_1 = \theta_2 = \frac{1}{5}$ (d) $\theta_1 = \theta_2 = 0$

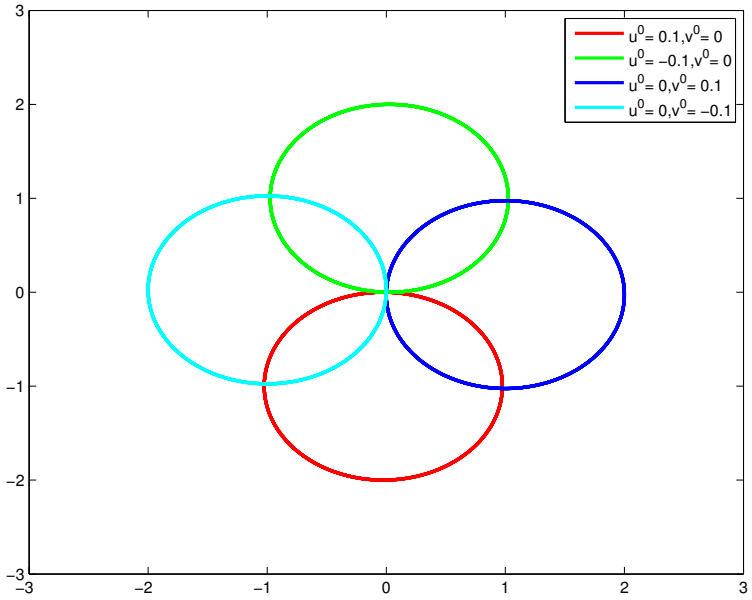
Figure A.2: *Trajectory of the particle with various values of θ_1 and θ_2 with starting point (red circle), final point (blue star) and the initial condition $u^0 = 0.1, v^0 = 0$.*

A.5 Conclusion

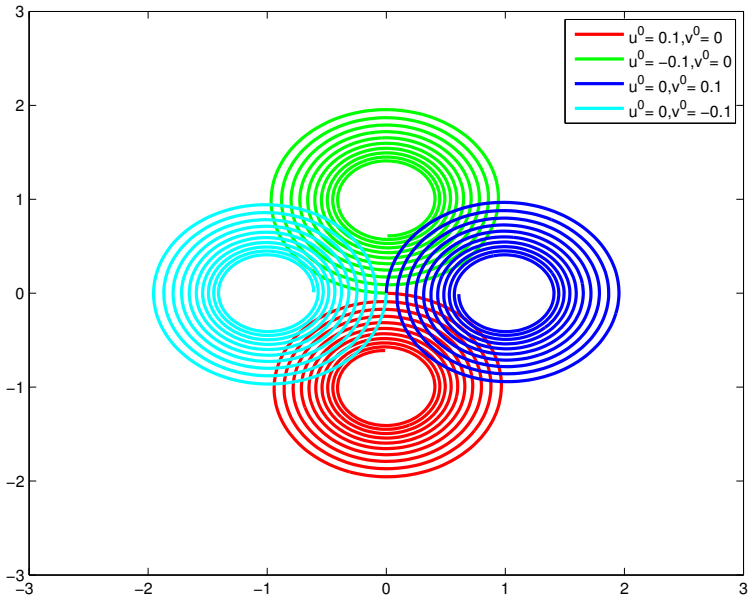
The explicit scheme applied to the inertial oscillation is unstable since it makes the kinetic energy increase. Consequently so does the radius of the trajectory of a moving particle. Therefore, it is essential to make the Coriolis source term implicit enough when we apply the θ-scheme to the inertial oscillation. In particular, the parameters θ_1 and θ_2 must lie in the stability region $\theta_1 + \theta_2 \leq 1$.

Moreover, the parameters θ_1 and θ_2 has a strong influence on the kinetic energy. The more implicit, the more damped the energy. Especially it is a constant like the continuous model with the choice $\theta_1 = \theta_2 = \frac{1}{2}$.

There is no CFL condition for the θ-scheme when the parameters involved in the discretisation of the Coriolis source term satisfy $0 \leq \theta_1, \theta_2 \leq \frac{1}{2}$. In any other case, we need a restriction for the time step which depends on the magnitude of the Coriolis.

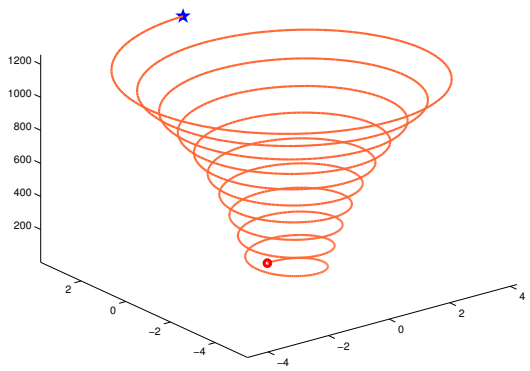


(a) $\theta_1 = \theta_2 = 0.5$

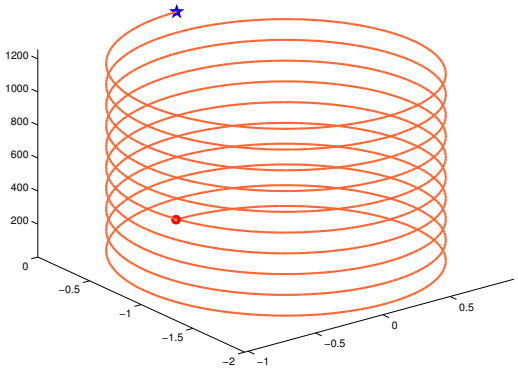


(b) $\theta_1 = \theta_2 = \frac{1}{2}$

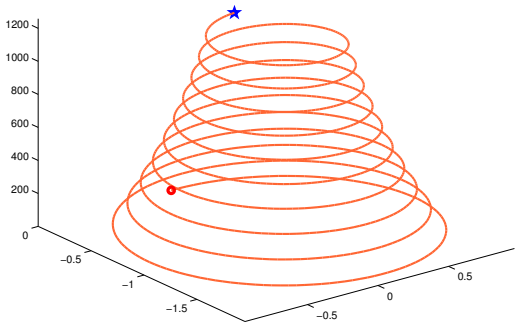
Figure A.3: *Trajectory of the particle with various values of the initial condition.*



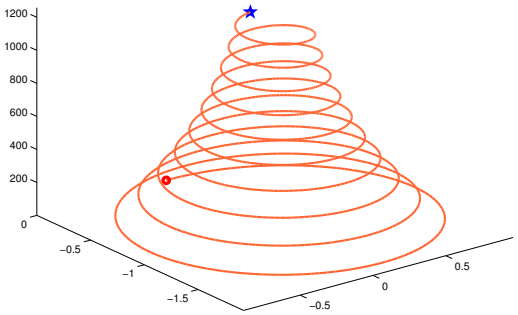
(a) $\theta_1 = \theta_2 = 1$



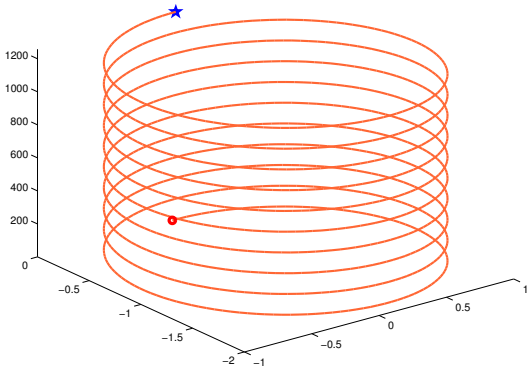
(b) $\theta_1 = \theta_2 = \frac{1}{2}$



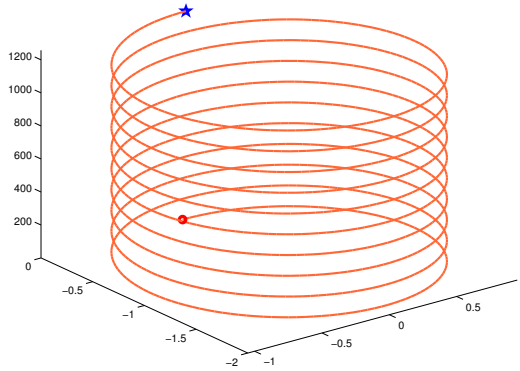
(c) $\theta_1 = \theta_2 = \frac{1}{5}$



(d) $\theta_1 = \theta_2 = 0$



(e) $\theta_1 = 1, \theta_2 = 0$



(f) $\theta_1 = 0, \theta_2 = 1$

Figure A.4: *Trajectory of the particle in for the view in 3D with various values of* θ_1 *and* θ_2 *with starting point (red circle), final point (blue star) and the initial condition* $u^0 = 0.1, v^0 = 0$.

Bibliography

[1] Joseph Pedlosky. *Geophysical fluid dynamics*. Springer Science & Business Media, 2013 *Cited on pages 1, 157.*

[2] Benoit Cushman-Roisin and Jean-Marie Beckers. *Introduction to geophysical fluid dynamics: physical and numerical aspects*. Vol. 101. Academic Press, 2011 *Cited on pages 1, 127.*

[3] Eleuterio F Toro. *Riemann solvers and numerical methods for fluid dynamics: a practical introduction*. Springer Science & Business Media, 2013 *Cited on page 3.*

[4] Edwige Godlewski and Pierre-Arnaud Raviart. *Numerical approximation of hyperbolic systems of conservation laws*. Vol. 118. Springer Science & Business Media, 2013 *Cited on page 3.*

[5] E. Audusse et al. "A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows". In: *SIAM J. Sci. Comput.* 25.6 (2004), pp. 2050–2065 *Cited on pages 3, 14, 44, 55, 91, 185, 190.*

[6] Sebastian Noelle et al. "Well-balanced finite volume schemes of arbitrary order of accuracy for shallow water flows". In: *Journal of Computational Physics* 213.2 (2006), pp. 474–499 *Cited on pages 3, 185.*

[7] T Morales de Luna, MJ Castro Diaz, and Carlos Parés. "Reliability of first order numerical schemes for solving shallow water system over abrupt topography". In: *Applied Mathematics and Computation* 219.17 (2013), pp. 9012–9032 *Cited on page 3.*

[8] Guoxian Chen and Sebastian Noelle. "A new hydrostatic reconstruction scheme based on subcell reconstructions". In: *SIAM Journal on Numerical Analysis* 55.2 (2017), pp. 758–784 *Cited on page 3.*

[9] Ulrik S Fjordholm, Siddhartha Mishra, and Eitan Tadmor. "Well-balanced and energy stable schemes for the shallow water equations with discontinuous topography". In: *Journal of Computational Physics* 230.14 (2011), pp. 5587–5609 *Cited on pages 3, 185.*

[10] Christophe Berthon and Christophe Chalons. "A fully well-balanced, positive and entropy-satisfying Godunov-type method for the shallow-water equations". In: *Mathematics of Computation* 85.299 (2016), pp. 1281–1307 *Cited on pages 3, 185.*

[11] Daniel Y Le Roux, Virgile Rostand, and Benoit Pouliot. "Analysis of numerically induced oscillations in 2d finite-element shallow-water models part I: inertia-gravity waves". In: *SIAM Journal on Scientific Computing* 29.1 (2007), pp. 331–360 *Cited on pages 3, 139.*

[12] D.Y. Le Roux. "Spurious inertial oscillations in shallow-water models". In: *J. Comput. Phys.* 231.24 (2012), pp. 7959–7987 *Cited on pages 3, 13, 91.*

[13] F. Bouchut, J. Le Sommer, and V. Zeitlin. "Frontal geostrophic adjustment and nonlinear wave phenomena in one-dimensional rotating shallow water. II. High-resolution numerical simulations". In: *J. Fluid Mech.* 514 (2004), pp. 35–63 *Cited on pages 3, 5, 14, 43, 44, 52, 53, 55, 56, 82, 91, 99, 106, 120, 125, 155, 185, 186, 190, 213.*

[14] V. Zeitlin. *Nonlinear dynamics of rotating shallow water: Methods and advances*. Vol. 2. Elsevier Science, 2007 *Cited on pages 3, 91, 185.*

[15] H. Guillard and C. Viozat. "On the behaviour of upwind schemes in the low Mach number limit". In: *Comput. Fluids* 28.1 (1999), pp. 63–86 *Cited on pages 3, 14, 91.*

[16] Eli Turkel. "Preconditioned methods for solving the incompressible and low speed compressible equations". In: *Journal of computational physics* 72.2 (1987), pp. 277–298 *Cited on page 3.*

[17] E. Turkel, A. Fiterman, and Van Leer. B. "Preconditioning and the limit of the compressible to the incompressible flow equations for Finite Difference schemes". In: *Frontiers of computational fluid dynamics*. Wiley, Chichester, 1995, pp. 215–234 *Cited on page 3.*

[18] Xue-song Li and Chun-wei Gu. "An all-speed Roe-type scheme and its asymptotic analysis of low Mach number behaviour". In: *Journal of Computational Physics* 227.10 (2008), pp. 5144–5159 *Cited on page 3.*

[19] F. Rieper. "A low-Mach number fix for Roe's approximate Riemann solver". In: *J. Comput. Phys.* 230.13 (2011), pp. 5263–5287 *Cited on pages 4, 14, 192.*

[20] S. Dellacherie. "Analysis of Godunov type schemes applied to the compressible Euler system at low Mach number". In: *J. Comput. Phys.* 229.4 (2010), pp. 978–1016 *Cited on pages 4, 6, 14, 18, 20, 26, 34, 37, 43, 62, 91, 92, 95, 99, 120, 125, 157, 186, 192, 213, 214.*

[21] S. Dellacherie, P. Omnes, and F. Rieper. "The influence of cell geometry on the Godunov scheme applied to the linear wave equation". In: *J. Comput. Phys.* 229.14 (2010), pp. 5315–5338 *Cited on pages 4, 6, 14, 18, 62, 91, 95, 158, 161, 204.*

[22] S. Dellacherie et al. "Construction of modified Godunov type schemes accurate at any Mach number for the compressible Euler system". In: *Math. Models Methods Appl. Sci.* 26.13 (2016), pp. 2525–2615 *Cited on pages 4, 7, 62, 91, 186, 192.*

[23] F. Rieper and G. Bader. "The influence of cell geometry on the accuracy of upwind schemes in the low Mach number regime". In: *J. Comput. Phys.* 228.8 (2009), pp. 2918–2933 *Cited on pages 4, 14.*

[24] H. Guillard. "On the behavior of upwind schemes in the low Mach number limit. IV: P0 approximation on triangular and tetrahedral cells". In: *Comput. Fluids* 38.10 (2009), pp. 1969–1972 *Cited on pages 4, 14.*

[25] F. Bouchut. *Nonlinear stability of finite volume methods for hyperbolic conservation laws and well-balanced schemes for sources*. Frontiers in Mathematics. Birkhäuser Verlag, Basel, 2004 *Cited on pages 5, 13, 14, 24, 36, 53, 91, 157, 190, 191.*

[26] George W Platzman. "Some response characteristics of finite-element tidal models". In: *Journal of Computational Physics* 40.1 (1981), pp. 36–63 *Cited on pages 5, 55, 64.*

[27] M.J. Castro, J.A. López, and C. Parés. "Finite volume simulation of the geostrophic adjustment in a rotating shallow-water system". In: *SIAM J. Sci. Comput.* 31.1 (2008), pp. 444–477 *Cited on pages 14, 36, 43, 48, 55, 66, 91, 99, 109, 198.*

[28] M. Lukacova-Medvidova, S. Noelle, and M. Kraft. "Well-balanced finite volume evolution Galerkin methods for the shallow water equations". In: *J. Comput. Phys.* 221.1 (2007), pp. 122–147 *Cited on pages 14, 36, 91.*

[29] Y. Moguen et al. "Pressure–velocity coupling allowing acoustic calculation in low Mach number flow". In: *J. Comput. Phys.* 231.16 (2012), pp. 5522–5541 *Cited on page 14.*

[30] M. Parisot and J.-P. Vila. “Numerical scheme for multilayer shallow-water model in the low-Froude number regime”. In: *C. R. Acad. Sci. Ser. I Math.* 352.11 (2014), pp. 953–957 *Cited on page 14.*

[31] F. Couderc, A. Duran, and J.-P. Vila. “An explicit asymptotic preserving low Froude scheme for the multilayer shallow water model with density stratification”. In: *J. Comput. Phys.* 343 (2017), pp. 235–270 *Cited on page 14.*

[32] H. Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. Springer, New York, 2011 *Cited on page 15.*

[33] S. Alinhac. *Hyperbolic partial differential equations*. Springer, Dordrecht, 2009 *Cited on page 17.*

[34] M. Marden. *Geometry of polynomials*. 1966 *Cited on pages 23, 30.*

[35] R.J. LeVeque. *Finite volume methods for hyperbolic problems*. Vol. 31. Cambridge Univ. Press, 2002 *Cited on page 24.*

[36] R.D. Richtmyer and K.W. Morton. *Difference methods for initial-value problems*. second. Robert E. Krieger Publishing Co., Inc., Malabar, FL, 1994 *Cited on pages 37, 185.*

[37] Emmanuel Audusse et al. “Godunov type scheme for the linear wave equation with Coriolis source term”. In: *ESAIM: Proceedings and Surveys* 58 (2017), pp. 1–26 *Cited on pages 43, 44, 46, 48, 49, 52, 53, 55, 59, 82, 91, 94, 95, 143, 155, 157, 186, 193.*

[38] E. Audusse et al. “Analysis of Apparent Topography Scheme for the Linear Wave Equation with Coriolis Force”. In: *FVCA VIII. Hyperbolic, Elliptic and Parabolic Problems*. Vol. 200. 2017, pp. 209–217 *Cited on pages 55, 57, 59, 65, 91, 94, 99.*

[39] Dellacherie, Stéphane, Jung, Jonathan, and Omnes, Pascal. “Preliminary results for the study of the godunov scheme applied to the linear wave equation with porosity at low mach number”. In: *ESAIM: ProcS.* 52 (2015), pp. 105–126 *Cited on page 62.*

[40] Akio Arakawa and Vivian R Lamb. “Computational design of the basic dynamical processes of the UCLA general circulation model”. In: *Methods in computational physics* 17 (1977), pp. 173–265 *Cited on pages 82, 127.*

[41] John K Dukowicz. “Mesh effects for Rossby waves”. In: *Journal of Computational Physics* 119.1 (1995), pp. 188–194 *Cited on pages 82, 127, 139.*

[42] D. Olbers, J. Willebrand, and C. Eden. *Ocean dynamics*. Springer Science & Business Media, 2012 *Cited on page 91.*

[43] P. Azérad and F. Guillén. “Mathematical justification of the hydrostatic approximation in the primitive equations of geophysical fluid dynamics”. In: *SIAM J. Math. Anal.* 33.4 (2001), pp. 847–859 *Cited on page 91.*

[44] C. Hu, R. Temam, and M. Ziane. “The primitive equations on the large scale ocean under the small depth hypothesis”. In: *Discrete Contin. Dyn. Syst.* 9.1 (2003), pp. 97–131 *Cited on page 91.*

[45] E. Audusse, C. Chalons, and P. Ung. “A simple well-balanced and positive numerical scheme for the shallow-water system”. In: *Commun. Math. Sci.* 13.5 (2015), pp. 1317–1332 *Cited on page 91.*

[46] H. Zakerzadeh. “The RS-IMEX scheme for the rotating shallow water equations with the Coriolis force”. In: *FVCA VIII. Hyperbolic, Elliptic and Parabolic Problems*. 2017, pp. 199–207 *Cited on page 91.*

[47] J. Thuburn and C.J. Cotter. “A framework for mimetic discretization of the rotating shallow-water equations on arbitrary polygonal grids”. In: *SIAM J. Sci. Comput.* 34.3 (2012), B203–B225 *Cited on pages 91, 92, 104.*

[48] R. Klein. “Semi-implicit extension of a Godunov-type scheme based on low Mach number asymptotics I: One-dimensional flow”. In: *J. Comput. Phys.* 121.2 (1995), pp. 213–237 *Cited on page 91.*

[49] H. Guillard and A. Murrone. “On the behavior of upwind schemes in the low Mach number limit: II. Godunov type schemes”. In: *Comput. Fluids* 33.4 (2004), pp. 655–675 *Cited on page 91.*

[50] S. Vater and R. Klein. “Stability of a cartesian grid projection method for zero Froude number shallow water flows”. In: *Numer. Math.* 113.1 (2009), pp. 123–161 *Cited on page 91.*

[51] R.F. Warming and B.J. Hyett. “The modified equation approach to the stability and accuracy analysis of finite-difference methods”. In: *J. Comput. Phys.* 14.2 (1974), pp. 159–179 *Cited on page 100.*

[52] Christopher Eldred. “Linear and nonlinear properties of numerical methods for rotating shallow water equations”. PhD thesis. Colorado State University, 2015 *Cited on page 104.*

[53] E. Audusse et al. “Analysis of modified Godunov schemes for the linear wave equation with Coriolis source terrm on cartesian meshes”. preprint. *Cited on pages 125, 127, 129, 143, 149, 157, 172, 175, 185, 186, 195, 203.*

[54] David A Randall. “Geostrophic adjustment and the finite-difference shallow-water equations”. In: *Monthly Weather Review* 122.6 (1994), pp. 1371–1377 *Cited on pages 127, 139.*

[55] AM Mohammadian and DY Le Roux. “Fourier analysis of a class of upwind schemes in shallow water systems for gravity and Rossby waves”. In: *International journal for numerical methods in fluids* 57.4 (2008), pp. 389–416 *Cited on page 139.*

[56] J.P. Boris. “Relativistic plasma simulation – optimization of a hybrid code”. In: *Proceedings of the 4th conference on numerical simulation of Plasmas*. Naval Res. Lab., Wash., D. C., 1970, pp. 3–67 *Cited on page 168.*

[57] Bjørn Gjevik, Halvard Moe, and Atle Ommundsen. “Idealized model simulations of barotropic flow on the Catalan shelf”. In: *Continental Shelf Research* 22.2 (2002), pp. 173–198 *Cited on page 185.*

[58] Allen C Kuo and Lorenzo M Polvani. “Time-dependent fully nonlinear geostrophic adjustment”. In: *Journal of Physical Oceanography* 27.8 (1997), pp. 1614–1634 *Cited on page 185.*

[59] A.C. Kuo and L.M. Polvani. “Nonlinear geostrophic adjustment, cyclone ,anticyclone asymmetry, and potential vorticity rearrangement.” In: *Phys. Fluids* 12.5 (2000), pp. 1087–1100 *Cited on pages 185, 198.*

[60] Philip L Roe. “Approximate Riemann solvers, parameter vectors, and difference schemes”. In: *Journal of computational physics* 43.2 (1981), pp. 357–372 *Cited on pages 189, 208.*

[61] E. Audusse, R. Klein, and A. Owinoh. “Conservative discretization of Coriolis force in a finite volume framework”. In: *J. Comput. Phys.* 228.8 (2009), pp. 2934–2950 *Cited on page 194.*

[62] E. Audusse et al. “Preservation of the discrete geostrophic equilibrium in shallow-water flows”. In: *FVCA VI. Problems & perspectives*. Vol. 4. Springer Proc. Math. Springer, Heidelberg, 2011, pp. 59–67 *Cited on page 194.*

[63] Guang-Shan Jiang and Chi-Wang Shu. “Efficient implementation of weighted ENO schemes”. In: *Journal of computational physics* 126.1 (1996), pp. 202–228 *Cited on page 198.*

[64] A. Harten and J.M. Hyman. “Self-adjusting grid methods for one-dimensional hyperbolic conservation laws”. In: *J. Comput. Phys.* 50.2 (1983), pp. 235–269 *Cited on page 208.*

[65] A. Harten, P. Lax, and B. van Leer. “On upstream differencing and Godunov-type schemes for hyperbolic conservation laws”. In: *SIAM Rev.* 25.1 (1983), pp. 35–61 *Cited on page 209.*

Titre : **Analyse mathématique de schémas volume finis pour la simulation des écoulements quasi-géostrophiques à bas nombre de Froude.**

Mots clefs : équilibre géostrophique, bas nombre de Froude, système hyperbolique, méthode de volumes finis, schéma de Godunov, diffusion numérique, schéma équilibre, force de Coriolis.

Résumé : Le système de Saint-Venant joue un rôle important dans la simulation de modèles océaniques, d'écoulements côtiers et de ruptures de barrages. Plusieurs sortes de termes sources peuvent être pris en compte dans ce modèle, comme la topographie, les effets de friction de Manning et la force de Coriolis. Celle-ci joue un rôle central dans les phénomènes à grande échelle spatiale car les circulations atmosphériques ou océaniques sont souvent observées autour de l'équilibre géostrophique qui correspond à l'équilibre du gradient de pression et de cette force. La capacité des schémas numériques à bien reproduire le lac au repos a été largement étudiée; en revanche, la question de l'équilibre géostrophique (incluant la contrainte de vitesse à divergence nulle) est beaucoup plus complexe et peu de travaux lui ont été consacrés.
Dans cette thèse, nous concevons des schémas volumes finis qui préservent les équilibres géostrophiques discrets dans le but d'améliorer significativement la précision des simulations numériques de perturbations autour de ces équilibres. Nous développons tout d'abord des schémas colocalisés et décalés sur des maillages rectangulaires ou triangulaires pour une linéarisation du modèle d'origine. Le point commun décisif de ces méthodes est d'adapter et de combiner les stratégies dites "topographie apparente", "bas Mach" et "pénalisation de divergence" pour contrôler l'effet de la diffusion numérique contenue dans les schémas, de telle sorte qu'elle ne détruise pas les équilibres géostrophiques. Enfin, nous étendons ces stratégies au cas non-linéaire et montrons des résultats prometteurs.

Title : **Analysis of finite volume schemes for the quasi-geostrophic flows at low Froude number.**

Keywords : Geostrophic equilibrium, low Froude number, hyperbolic system, finite volume method, Godunov scheme, numerical diffusion, well-balanced scheme, Coriolis force.

Abstract : The shallow water system plays an important role in the numerical simulation of oceanic models, coastal flows and dam-break floods. Several kinds of source terms can be taken into account in this model, such as the influence of bottom topography, Manning friction effects and Coriolis force. For large scale oceanic phenomena, the Coriolis force due to the Earth's rotation plays a central role since the atmospheric or oceanic circulations are frequently observed around the so-called geostrophic equilibrium which corresponds to the balance between the pressure gradient and the Coriolis source term. The ability of numerical schemes to well capture the lake at rest, has been widely studied. However, the geostrophic equilibrium issue, including the divergence free constraint on the velocity, is much more complex and only few works have been devoted to its preservation.
In this manuscript, we design finite volume schemes that preserve the discrete geostrophic equilibrium in order to improve significantly the accuracy of numerical simulations of perturbations around this equilibrium. We first develop collocated and staggered schemes on rectangular and triangular meshes for a linearized model of the original shallow water system. The crucial common point of the various methods is to adapt and combine several strategies known as the Apparent Topography, the Low Mach and the Divergence Penalisation methods, in order to handle correctly the numerical diffusions involved in the schemes on different cell geometries, so that they do not destroy geostrophic equilibria. Finally, we extend these strategies to the non-linear case and show convincing numerical results.

Université Paris-13
Laboratoire Analyse, Géométrie et Application
UMR CNRS 7539, Villetaneusse, France