Analysis and numerical approximation of some mathematical models of free-surface flows

Chourouk El Hassanieh

Sorbonne Université
Université Libanaise
INRIA

Doctoral School **Sciences Mathématiques de Paris Centre**

University Department **Laboratoire Jacques-Louis Lions**

Thesis defended by **Chourouk El Hassanieh**

Defended on **December 7, 2023**

In order to become Doctor from Sorbonne Université and from Université Libanaise

Academic Field **Applied Mathematics**

Analysis and numerical approximation of some mathematical models of free-surface flows

Thesis supervised by	Jacques Sainte-Marie	Supervisor
	Samer Israwi	Supervisor
	Bernard Di Martino	Co-supervisor
	Julien Guillod	Co-advisor

Committee members

Referees	Marguerite Gisclon	Associate professor at Université Savoie Mont Blanc	
	Sebastian Noelle	Professor at Aachen University	
Examiners	Wael Youssef	Associate professor at Université Libanaise	
	Carlos Parés	Professor at Universidad de Málaga	
	Hélène Mathis	Professor at Université de Montpellier	Committee President
Guests	Bernard Di Martino	Associate professor at Université de Corse	
	Edwige Godlewski	Professor emeritus at Sorbonne Université	
Supervisors	Jacques Sainte-Marie	Research Director at Sorbonne Université, INRIA	
	Samer Israwi	Professor at Université Libanaise	
	Julien Guillod	Associate professor at Sorbonne Université	

Colophon

Doctoral dissertation entitled “Analysis and numerical approximation of some mathematical models of free-surface flows”, written by Chourouk El Hassanieh, completed on December 12, 2023, typeset with the document preparation system LaTeX and the yathesis class dedicated to theses prepared in France.

Sorbonne Université
Université Libanaise
INRIA

École doctorale **Sciences Mathématiques de Paris Centre**

Unité de recherche **Laboratoire Jacques-Louis Lions**

Thèse présentée par **Chourouk El Hassanieh**

Soutenue le **7 décembre 2023**

En vue de l'obtention du grade de docteur de Sorbonne Université et de l'Université Libanaise

Discipline **Mathématiques Appliquées**

Analyse et approximation numérique de quelques modèles mathématiques d'écoulements à surface libre

Thèse dirigée par	Jacques Sainte-Marie	Directeur
	Samer Israwi	Directeur
	Bernard Di Martino	Co-directeur
	Julien Guillod	Co-encadrant

Composition du jury

Rapporteurs	Marguerite Gisclon	Maître de conférence à l'Université Savoie Mont Blanc	
	Sebastian Noelle	Professeur à Aachen University	
Examinateurs	Wael Youssef	Maître de conférence à l'Université Libanaise	
	Carlos Parés	Professeur à l'Universidad de Málaga	
	Hélène Mathis	Professeure à l'Université de Montpellier	Présidente du Jury
Invités	Bernard Di Martino	Maître de conférence à l'Université de Corse	
	Edwige Godlewski	Professeure émertie à Sorbonne Université	
Directeurs de thèse	Jacques Sainte-Marie	Directeur de Recherche à Sorbonne Université, INRIA	
	Samer Israwi	Professeur à l'Université Libanaise	
	Julien Guillod	Maître de conférence à Sorbonne Université	

Acknowledgements

This work would not have been possible without the support and encouragement of all those who believed in me. Thank you for being an integral part of this step. I could write pages about all of the people who inspired me and stood by my side throughout those few challenging years and it is difficult to summarize all the gratitude and joy.

It was a pleasure to have worked under the supervision of *Bernard Di Martino*, *Edwige Godlewski*, *Samer Israwi*, *Julien Guillod*, and *Jacques Sainte-Marie*. *Edwige*, thank you for always taking the time to explain, for your patience to re-explain, and for making your office a welcoming place for discussions. I admire and respect you and I am so happy to have had the opportunity to learn from you. *Jacques*, I cannot thank you enough for making this thesis happen. Thank you for your kindness, patience, and sincerity. I know that you have always had my best interest in mind and your contributions extend beyond what may be immediately evident and that have made a significant impact. *Bernard*, I have learned a lot from you starting from my internship at INRIA, thank you for always being available and patient, and for all the advice and motivation. Your kindness and positivity have not gone unnoticed. Thank you *Julien* for supervising me throughout this tough part of the thesis. I have learned a lot from you in so little time. You have unknowingly helped me overcome a very difficult period during my thesis where I needed that kind of motivation. Thank you *Samer* for making this collaboration possible. The year I have spent in Lebanon was one of the most complicated times, despite that I was given the opportunity to learn and advance in my thesis.

I would like to thank *Marguerite Gisclon*, *Hélène Mathis*, *Sebastian Noelle*, *Carlos Parés*, and *Wael Youssef* for kindly accepting to take part in my thesis jury.

I could not be more thankful for the members of my thesis committee, *Anne-Laure Dalibard* and *Nicolas Seguin*. Thank you for taking the time to monitor my progress during the thesis and making sure that I was in good working conditions. I would like to take this opportunity in order to say thank you *Anne-Laure Dalibard* for all the advice, motivation, and action taken with my well-being and best interest in mind. It has been a pleasure.

Of course I could not forget all the doctors, professors, and permanent members of *LJLL* and *INRIA*. Thank you all for the several fruitful discussions and advice. A special thank you to *Julien Guieu*, *Corentin Lacombe*, and *Jean-François Venuti* for their help in the several time-consuming administrative processes. Thank you for always welcoming me with a smile despite my time consuming requests and several questions. Mille mercis à *Frédérique Concord* pour son aide et pour avoir pris le temps d'écouter et de proposer des solutions. Merci *Axelle*, *Isabelle*, and *Rim*.

The following part is less formal, dedicated to all my friends and colleagues in no specific order, I will just go with the flow and thank you one by one. It was difficult to summarize and I hope that I have not missed a name. Thank you for making my experience at work such a beautiful one, I cannot thank you enough for all the nice memories.

Maria, thankyou for the beautiful friendship, wise advice, and all the laughs. We did it madame! *Apolline*, thankyou for teaching me French phrases and slang that has not been used since the last century and for the very long chatting breaks. Thankyou both for being there through ups and downs and for always turning bad situations into laughs. *Elena*, thankyou for all the love and motivation, your sweet words made my day everyday! *Claudia*, thankyou for the kindness and care, and the competitive spirit during game nights! *Suney*, thankyou for the lovely invitations and delicious food. *Alexiane* and *Pauline*, thankyou for the constant encouragement and thankyou *Kani-Sira* for bringing such a joy around. *Antoine*, thankyou for the friendship and beautiful times (and for the famous "La Churukita" title, it remains my favorite). I already miss "Le Salon" and "Le Diner" and the style you brought to our office . *Marwa*, *Maya*, and *Lama*, thankyou for the very long phone calls and honest feedback. *Giorgia, Noemi, Emma, Chiara*, thankyou for bringing a lovely spirit to the working place and for the great parties. *Rui*, thankyou for constantly sharing candy, tea, and complaining about the visa struggles. *Jesús*, thankyou for your contribution in the research around mosquitoes, it helps me sleep at night. *Ramón*, thankyou for sharing my taste in music (es una lata el trabajar!), the 13h30 crous group, and the lovely walks. Thankyou *Nicolás* for all the useful and non useful information. Thankyou *Mathieu* for sharing my taste in classical music, for the friendship, and the lovely time we spent working on our common project. *Pierre* I have went through the first salt bag already! Thank you for contributing to my journey towards high blood pressure. Thank you for always taking the time to answer my questions, although some of this credit goes to the constant RER B delays.

Thank you *Liangying*, *Matthieu*, *Yvonne*, and *Nicolaï*, for the lovely days and the 15-25 discussions. Thank you *Matthias*, *Siwar*, *Haibo*, *Jg*, *Van-Thanh*, *Edouard*, and *Suraj* for the sweet lunch-time at the cantine and *Juliette* for the discussions, laughs, and for being there at almost every conference ;) It is always comforting to find a familiar face. Thank you *Allen* and it Liu Di for the beautiful company and thoughtful texts. *Anna* and *Emilio* thank you for the friendship and for the chance to explore Lyon. *Anatole*, *Thomas*, *Robin*, *Charles*, *Cristóbal* ,*Gong*, *Ludovic*, *Lucas*^{2}, *Mingyue*, *Nga*, and *Roxanne* thank you for all the kindness and sweet conversations. *Jules*, thankyou for all the opinions that were asked for and not asked for. *Nathalie*, thankyou for the humor and sweet 15-25 conversations. *Nílo* and *Alheli*, thankyou for the hospitality and all the love and for the long lengthy mathematical discussions on transport equations. *Violetta* and *Annina* for the friendship and good times. *Marcel*, *Lucia*, and *Zhengping* thank you for sharing my last few weeks at 15-25-324. Gracias *Facundo* por las fotos con la torre Eiffel y la amistad. Gracias *Adriana*, *Hector*, *Cami*, *Lucas*, y *Mercedes* por todo el amor y la motivación.

Quería tomarme un momento para agradecerte desde lo más profundo de mi corazón. No hay palabras suficientes para expresar lo agradecida que estoy por tenerte a mi lado, tienes el secreto de mi sonrisa. *Hanane* and *Maen* thank you for the love, this thesis is dedicated to you.

Analysis and numerical approximation of some mathematical models of free-surface flows

Abstract

This thesis is dedicated to the study of some partial differential equations describing free-surface flows in fluid mechanics and it consists of three interrelated projects. The first project investigates the implementation of numerical schemes for the Saint-Venant system using a kinetic approach, with a primary focus on the one-dimensional case. By adopting an implicit-in-time kinetic approach, this work offers a computational advantage over traditional implicit schemes, since it presents an explicit expression for the inverse of the matrix. The implicit kinetic scheme preserves the positivity of the water height and satisfies an entropy inequality. The second contribution delves into the stability analysis of the hydrostatic Euler equations. A transformation is introduced to rewrite these equations as a generalized quasi-linear system with an integral operator, establishing equivalence under specific conditions. This transformation allows for deeper insights into the spectrum of the matrix operator. Furthermore, we propose an exact multi-layer $\mathbb{P}_0$-discretization, which could be used to solve numerically the transformed system and we analyse its spectrum. The third and final contribution is a work in progress aiming to provide a mathematical justification for the mechanical balance laws of the two-dimensional Boussinesq system. This system is widely used in nearshore zone applications and it is useful to assess its accuracy in terms of fundamental principles such as mass, momentum, and energy conservation. We give estimates to quantify the errors introduced by these approximations, offering valuable insights into the accuracy of the Boussinesq system.

Keywords: Free-surface flows, Euler equations, Numerical approximation, Implicit schemes, Boussinesq equations

---

Analyse et approximation numérique de quelques modèles mathématiques d'écoulements à surface libre

Résumé

Cette thèse est dédiée à l'étude de certaines équations aux dérivées partielles décrivant les écoulements à surface libre en mécanique des fluides et se compose de trois projets interconnectés. Le premier projet étudie l'implémentation de schémas numériques pour le système de Saint-Venant en utilisant une approche cinétique, en se concentrant principalement sur le cas unidimensionnel. En adoptant une approche cinétique implicite en temps, ce travail offre un avantage de calcul par rapport aux schémas implicites traditionnels, puisqu'il présente une expression explicite pour l'inverse de la matrice. Le schéma cinétique implicite préserve la positivité de la hauteur d'eau et satisfait une inégalité d'entropie. La deuxième contribution traite de l'analyse de la stabilité des équations d'Euler hydrostatiques. Une transformation est introduite pour réécrire ces équations comme un système quasi-linéaire généralisé avec un opérateur intégral, établissant l'équivalence dans des conditions spécifiques. Cette transformation permet de mieux comprendre le spectre de l'opérateur matriciel. En outre, nous proposons une discrétisation exacte multicouche de $\mathbb{P}_0$, qui pourrait être utilisée pour résoudre numériquement le système transformé et nous analysons son spectre. La troisième et dernière contribution est un travail en cours visant à fournir une justification mathématique des lois d'équilibre mécanique du système bidimensionnel de Boussinesq. Ce système est largement utilisé dans les applications des zones littorales et il est utile d'évaluer sa précision en termes de principes fondamentaux tels que la conservation de la masse, de la quantité de mouvement et de l'énergie. Nous donnons des estimations pour quantifier les erreurs introduites par ces approximations, offrant ainsi des indications précieuses sur la précision du système de Boussinesq.

Mots clés : Équations à surface libre, Équations d'Euler, Approximation numérique, Schémas implicites, Équations de Boussinesq

---

Laboratoire Jacques-Louis Lions – Sorbonne Université – Campus Pierre et Marie Curie – 4 place Jussieu – 75005 Paris – France
INRIA – 2 rue Simone Iff – 75012 Paris – France
Laboratoire de Mathématiques – Université Libanaise – Campus Rafic Hariri – Hadath – Al Chouf – Liban

This thesis has been prepared at the following research units.

Laboratoire Jacques-Louis Lions

Sorbonne Université
Campus Pierre et Marie Curie
4 place Jussieu
75005 Paris
France

Web Site `https://ljll.math.upmc.fr/`

INRIA

2 rue Simone Iff
75012 Paris
France

Web Site `https://www.inria.fr/`

Laboratoire de Mathématiques

Université Libanaise
Campus Rafic Hariri
Hadath
Al Chouf
Liban

Summary

This thesis revolves around partial differential equations describing free-surface flows in fluid mechanics. The contributions of this work are summarized into three separate yet interconnected projects.

Using kinetic schemes is a practical way to implement numerical schemes that allow to satisfy strong stability properties among which are preserving the positivity of the water height and satisfying an entropy inequality which are two properties of the continuous system. The first contribution of this thesis is exploring an implicit-in-time kinetic approach to the Saint-Venant system in the one-dimensional and two-dimensional frameworks. We mainly focus on the one-dimensional case for flat and variable topographies. It is important to note that when using an implicit scheme, such as in the context of the Saint-Venant system, there is typically a requirement to invert an operator, often represented by a matrix, at each time step. However, the use of a kinetic solver provides a better scenario where we have an explicit expression for the inverse of the operator. Hence, one can hardly imagine an implicit scheme with a better computational cost than that of a kinetic solver. The implicit kinetic scheme satisfies a discrete entropy inequality without any restriction on the time step that is required for explicit schemes [5, 76, 79]. However, the CFL constraint required by explicit schemes is replaced by computational costs which lead us to evaluate the practical interest of the implicit scheme with respect to its explicit counter-part.

The hydrostatic Euler equations also form an integral part of this thesis, representing the second contribution of this research. Understanding the stability of these equations is crucial for predicting the behavior of various natural phenomena that occur in the ocean and the atmosphere and it remains a challenging and ongoing research area due to the complex nature of these equations which do not fall under any classical type. However, following [91, 95] we introduce a transformation that allows us to reduce this complexity by rewriting the hydrostatic Euler equations as a generalized quasi-linear system with an integral operator and we show that the two systems are equivalent under certain conditions. Although these conditions are not optimal, they allow us to justify the claims in [29, 30, 86, 90, 91, 95] by proving the equivalence between the two systems for a certain class of solutions. We then focus on the generalized quasi-linear system and we give a full decomposition of the spectrum of the matrix operator. In this work, we additionally demonstrate the existence of the two real eigenvalues mentioned in [30, 32] given certain limiting conditions of Hölder regularity beyond which this claim is no longer valid. This work is complemented by a multi-layer discretization as described in [4, 7], allowing us to study the hyperbolicity properties of the discretized system.

The third contribution of this thesis constitutes in giving a mathematical justification of the mechanical balance laws of the two-dimensional Boussinesq system following [57]. A class of Boussinesq-type systems has been derived [14, 16] and supported by theoretical and numerical justifications [39, 43, 63, 67]. Since these models find practical applications in the nearshore zone, it becomes crucial to estimate their accuracy in terms of the fundamental principles of mass, momentum, and energy balance laws. The balance laws provide a quantitative understanding of how different phys-

ical quantities, such as mass, momentum, and energy, interact and evolve within the system. By studying these balance laws, we can identify key conservation properties and constraints governing the dynamics of the system. However, the Boussinesq equations do not satisfy exact mechanical balance laws such as exact mass conservation, exact momentum conservation, or exact energy conservation as a result of the approximations used to derive these set of equations. The introduced approximations entail a degree of error, which can be measured by employing precise error estimates on the norm-difference between velocity fields, potentials, and pressure contributions of the approximated quantities. In [57] this methodology was employed to justify approximate mechanical balance laws for the Korteweg-de Vries equation.

Résumé

Cette thèse porte sur les équations aux dérivées partielles décrivant les écoulements à surface libre en mécanique des fluides. Les contributions de ce travail sont résumées en trois projets distincts interconnectés.

L'utilisation de schémas cinétiques est un moyen pratique de mettre en œuvre des schémas numériques qui permettent de satisfaire des propriétés de stabilité fortes, parmi lesquelles la préservation de la positivité de la hauteur d'eau et une inégalité d'entropie, qui sont deux propriétés du système continu. La première contribution de cette thèse est l'exploration d'une approche cinétique implicite en temps du système de Saint-Venant dans les cadres unidimensionnel et bidimensionnel. Nous nous concentrons principalement sur le cas unidimensionnel pour des topographies plates et variables. Il est important de noter que lors de l'utilisation d'un schéma implicite, comme dans le contexte du système de Saint-Venant, il est généralement nécessaire d'inverser un opérateur, souvent représenté par une matrice, à chaque pas de temps. Cependant, l'utilisation d'un solveur cinétique offre un meilleur scénario dans la mesure où nous disposons d'une expression explicite pour l'inverse de l'opérateur. On peut donc difficilement imaginer un schéma implicite avec un meilleur coût de calcul que celui d'un solveur cinétique. Le schéma cinétique implicite satisfait une inégalité d'entropie discrète sans aucune restriction sur le pas de temps, chose requise pour les schémas explicites [5,76,79]. Cependant, la contrainte CFL requise par les schémas explicites est remplacée par des coûts de calcul qui nous amènent à évaluer l'intérêt pratique du schéma implicite par rapport à son homologue explicite.

Les équations d'Euler hydrostatiques sont aussi une partie intégrante de cette thèse, représentant la deuxième contribution de cette recherche. Comprendre la stabilité de ces équations est crucial pour prédire le comportement de divers phénomènes naturels qui se produisent dans l'océan et l'atmosphère et cela reste un domaine de recherche difficile et actuel en raison de la nature complexe de ces équations qui ne relèvent d'aucun type classique. Cependant, en suivant [91,95], nous introduisons une transformation qui nous permet de réduire cette complexité en réécrivant les équations d'Euler hydrostatiques comme un système quasi-linéaire généralisé avec un opérateur intégral et nous montrons que les deux systèmes sont équivalents sous certaines conditions. Bien que ces conditions ne soient pas optimales, elles nous permettent de justifier les affirmations de [29,30,86,90,91,95] en prouvant l'équivalence entre les deux systèmes pour une certaine classe de solutions. Nous nous concentrons ensuite sur le système quasi-linéaire généralisé et nous donnons une décomposition complète du spectre de l'opérateur matriciel. Dans ce travail, nous démontrons en outre l'existence des deux valeurs propres réelles mentionnées dans [30,32] sous certaines conditions limites de régularité de Hölder au-delà desquelles cette affirmation n'est plus valide. Ce travail est complété par une discrétisation multicouche telle que décrite dans [4,7] qui nous permet d'étudier les propriétés d'hyperbolicité du système discrétisé.

La troisième contribution de cette thèse consiste à donner une justification mathématique des lois de conservation mécanique du système bidimensionnel de Boussinesq introduit dans [57]. Une classe

de systèmes de type Boussinesq a été dérivée dans [14,16] et soutenue par des justifications théoriques et numériques [39, 43, 63, 67]. Comme ces modèles trouvent des applications pratiques dans la zone littorale, il devient crucial d'estimer leur précision en termes des lois de conservation fondamentales de la masse, de la quantité de mouvement et de l'énergie. Les lois de conservation fournissent une compréhension quantitative de la manière dont les différentes quantités physiques, telles que la masse, la quantité de mouvement et l'énergie, interagissent et évoluent au sein du système. L'étude de ces lois de conservation permet d'identifier les principales propriétés de conservation et les contraintes qui régissent la dynamique du système. Cependant, les équations de Boussinesq ne satisfont pas les lois de conservation mécanique exactes telles que la conservation exacte de la masse, de la quantité de mouvement ou de l'énergie en raison des approximations utilisées pour dériver cet ensemble d'équations. Les approximations introduites impliquent un certain degré d'erreur, qui peut être mesuré en utilisant des estimations d'erreur précises sur la différence entre les champs de vitesse, les potentiels et les contributions de pression des quantités approchées. Dans [57], cette méthodologie a été employée pour justifier des lois de conservation mécanique approximatives pour l'équation de Korteweg-de Vries.

Contents

Chapter 1

General Introduction

Outline of the current chapter

This thesis is devoted to the mathematical study of free-surface fluid flows in different contexts modeled by the Saint-Venant equations, the hydrostatic Euler equations, and the Boussinesq equations. Despite their similarities, each set of equations addresses specific phenomena under certain physical assumptions.

Let us begin by a brief motivation for the previously mentioned topics. Before delving into the physical motivation of this thesis, consider the following three questions: Why free-surface flows? What are the difficulties encountered in such models? How do we circumvent these difficulties? While answering these questions, you will put together the complete picture. The study

of free-surface flows is the general motivation of the thesis. This includes the study of ocean hydrodynamics and is closely linked with forecasting natural disasters as well as energy production, water management and agriculture. All of which, provide better managing of water resources while preserving the ecosystem and allow us to predict future risks while mitigating the consequences. At that position, the free-surface Navier-Stokes system [49] (19^{th} century) is yet the most precise mathematical model that describes the evolution problem of such physical flows in a time-dependent domain. These equations have a high level of complexity and the existence and uniqueness of these equations is an open problem yet to be resolved. Numerically, they prove to be one other challenge as the complexity involved makes it difficult to derive good numerical approximations that yield reasonable computational costs. Circumventing this complexity requires certain simplifications relative to specific physical scenarios. We are mainly concerned with flows in the absence of viscosity which leads to the simplification of the Navier-Stokes equations into the Euler equations; otherwise known as the water waves problem. These equations inherit the complexity of the former system and their well-posedness remains an open problem which is why further simplification is required.

This thesis is a walk through three different types of circumventing the inconvenience of working with the full free-surface Euler system. Starting from depth-averaged models that reduce significantly the complexity by reducing the dimension of the equations, we begin by the simplest yet accurate non-linear model given by *de Saint-Venant* [10]. Known as the *Saint-Venant* system, this model is used to accurately model tsunamis, dam breaks, and floods among many other phenomena that occur in physical situations where the vertical dimension is considerably smaller than the horizontal one. It is also widely used for numerical approximation. Accessing additional information on the vertical variation of the horizontal fluid velocity requires a more complex mathematical model, which introduces the second set of equations tackled in this manuscript: the free-surface hydrostatic Euler equations. Unlike the Saint-Venant system, these equations provide better approximation of fluid phenomena while taking into account additional information related to the vertical variation of the fluid velocity. Despite its significance, this approximation does not overcome the complexity of the time-dependent domain and requires additional tools to access its mathematical properties. Throughout the third chapter, we introduce a transformation of the domain which circumvents its time-dependence while obtaining a nice structure of the mathematical model. The application of the hydrostatic Euler equations extends beyond shallow-water flows but is not applicable in scenarios where dispersive phenomena might occur. Dispersive phenomena are taken into account in the *Boussinesq* approximation [22], which will be the third topic in this thesis. The Boussinesq equations considered in this fourth chapter will deal with small-amplitude long-wave regimes in an irrotational setting. This means a reduction in the dimension of the system which results in a significant simplification of the original Navier-Stokes equations while preserving nonlinear and dispersive properties that are favorable to that of the Saint-Venant system or the hydrostatic Euler equations which allow them to accurately model additional scenarios such as wave interaction near the coastline.

1.1 Structure of the document

The manuscript is arranged as follows. Chapter 2 is dedicated to the Saint-Venant system in the context of implicit kinetic schemes mainly focusing on the one-dimensional case and providing perspective for the extension to the two-dimensional setting. We detail the fully implicit kinetic scheme in the case of flat topography, and we show that an entropy inequality holds in the case of the half-disk Maxwellian. Two steps are required for the implementation of the implicit scheme, the first is inverting the matrix of the system which will enable us to get an analytic expression of the updates and the second is finding an analytic expression of the macroscopic updates. We will prove that the first step is resolved since we are able to calculate the inverse of the system matrix

by hand without further complication. However, the second step requires a step back due to the complex nature of the integrals involved in defining the macroscopic updates. In fact, it requires the computation of integrals of the half-disk Maxwellian against a second degree polynomial and this seems hardly possible with the expression at hand. Instead of integrating the expressions using the classical Maxwellian, we use the index Maxwellian, which proves to be a valuable tool for these calculations. In fact, using quadrature formulae will lead to loss of accuracy in addition to the higher computational cost involved for numerical integration which is why the index Maxwellian is a compromise worth exploring. To ensure well-defined boundary conditions, we adopt the strategy proposed in Bristeau and Coussin's work [26]. Furthermore, we present a strategy to enhance the algorithmic complexity of the implicit scheme, making it more practical for implementation. We also extend the implicit kinetic scheme to the two-dimensional Saint-Venant system. An iterative strategy is also discussed in Chapter 2, where we demonstrate the dissipation of the entropy inequality after some rank. Finally, we conduct numerical tests to assess the performance and validity of the proposed methods.

Chapter 3 is dedicated to the free-surface Hydrostatic Euler system in the view of the transformation given in [91, 95]. The equivalence between the two systems is established under certain regularity conditions. Furthermore, some examples are presented using particular solutions. This enables us to better understand the behaviour of the system and the time evolution of the transformation. Moreover, the spectral analysis of the system is fully detailed and the notion of generalized Riemann invariants is introduced. This is followed by a localisation of the eigenvalues and a discussion of the case of limiting Hölder regularity on the velocity profile. A simple generalization extends the spectral analysis to the case of variable topography. This is complemented with a multi-layer approach following [4] where the spectrum of the discrete system is given. A discussion on the links between the continuous and discrete systems is also presented.

Chapter 4 is a work in progress that focuses on the two-dimensional Boussinesq equations [14,15]. The aim of this project is to find a well-justified derivation of approximate balance laws to the two-dimensional Boussinesq equations. We start by presenting the derivation of the Boussinesq equations using a systematic approach which involves introducing a series of perturbations based on the shallow-water parameter. As a result, we are able to justify the error bounds associated with the approximate balance laws following the methodology applied in [57, 58] for the KdV equations.

Throughout the introduction we will summarize the principle points of the three projects involved and the details are spared for Chapter 2, Chapter 3, and Chapter 4.

1.2 Saint-Venant system in 1d

The Saint-Venant equations [10] were first introduced in 1871 by the French mathematician *de Saint Venant* and they describe the motion of shallow water flows. They are known as the simplest approximation of the water-wave problem for free-surface flows in shallow water and are widely used to analyze and predict the behavior of water in rivers, pipe-flows, tidal currents, and open channels, where the depth of the flow is much smaller compared to the horizontal wave length. The Saint-Venant equations are also used to model physical phenomena such as tsunamis, dam breaks, floods and can contribute in the prevention and forecast of these natural disasters. These non-linear equations are hyperbolic in nature and it is shown that solutions can exhibit discontinuities even for smooth initial data. On the numerical level, the Saint-Venant equations are classically discretized using finite volume schemes. In this thesis, the analysis of the former equations is performed under the context of kinetic schemes. Kinetic schemes offer an efficient, accurate, and stable resolution for the Saint-Venant equations. Moreover, they exhibit good properties such as satisfying a discrete

entropy inequality and preserving the water height. We use an implicit-in-time kinetic interpretation of the Saint-Venant system and we investigate its advantages on the explicit-in-time scheme proposed in [76, 79] for the case of a flat bathymetry and in [5] for variable bathymetry.

The corresponding chapter (Chapter 2) is based on the preprint [44] written in collaboration with *J. Sainte-Marie* and *M. Rigal* [82] and is available on *HAL* (hal-04048832).

The derivation of the Saint-Venant equations involves depth-averaging the full governing equations of fluid motion and neglecting less significant terms under a shallow water regime where the fluid depth-to-wavelength ratio is small. Consider a fluid flow in domain delimited between a fixed topography $z_b(x)$ and a varying free-surface $\eta(t,x) = h(t,x) + z_b(x)$, where the fluid height is denoted $h(t,x) \geq 0$. Denote by $u(t,x)$ the vertically averaged horizontal velocity of the fluid and by g the gravitational acceleration, see Fig. 1.1. The one-dimensional Saint-Venant system is given by:

$$\begin{cases} \dfrac{\partial h}{\partial t} + \dfrac{\partial (hu)}{\partial x} = 0, \\ \dfrac{\partial (hu)}{\partial t} + \dfrac{\partial}{\partial x}\left(hu^2 + \dfrac{gh^2}{2} \right) = -gh\partial_x z_b. \end{cases} \tag{1.1}$$

The Saint-Venant system (1.1) can be rewritten as:

$$\partial_t U + \partial_x F(U) = S(U, z_b), \tag{1.2}$$

where $U = (h, hu)^T$, $F(U) = (hu, hu^2 + gh^2/2)^T$ and $S(U, z_b) = (0, -gh\partial_x z_b)^T$. The Saint-Venant system exhibits important properties. The system is hyperbolic and admits two real eigenvalues $u \pm \sqrt{gh}$. The presence of non-linearity in the system accounts for interactions between the variables of the flow which can lead to complex flow behavior such as the formation of shock waves or hydraulic jumps. Therefore, we have to take into account weak solutions.

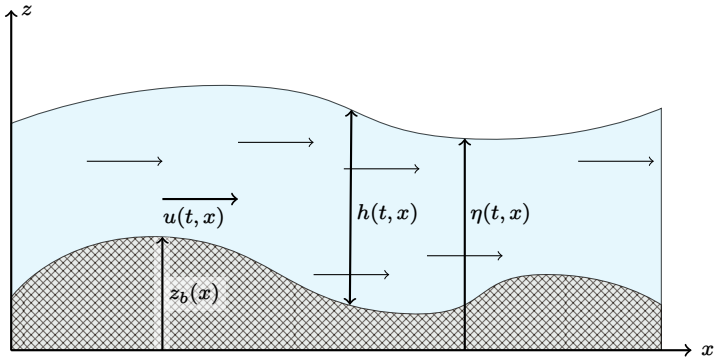

Figure 1.1: Saint-Venant setting

1.2.1 Entropy

In general, for systems of conservation laws with smooth initial data, the classical solution may not exist beyond a certain time due to the crossing of characteristics and thus we have to consider weak solutions. Weak solutions are not necessarily unique but we are interested in weak solutions that are

"physically relevant" in the sense that the solution should satisfy the following entropy inequality:

$$\partial_t \mathcal{E}(U) + \partial_x G_{\mathcal{E}}(U) \leq 0, \tag{1.3}$$

for all pairs $(\mathcal{E}, G_{\mathcal{E}})$ where $\mathcal{E}$ is a convex entropy and $G_{\mathcal{E}}$ is the entropy flux satisfying the following relation:

$$G'_{\mathcal{E}}(U) = F'(U)\mathcal{E}'(U). \tag{1.4}$$

Note that relation (1.4) ensures that any classical solution of (1.1) satisfies the entropy inequality (1.3). In particular, the Saint-Venant system (1.1) satisfies the following entropy (energy) inequality:

$$\partial_t E(U) + \partial_x G(U) \leq 0,$$

where the energy of the system (1.1) and its associated flux satisfy relations (1.3) and (1.4) and are given by:

$$E(U) = \frac{hu^2}{2} + \frac{gh^2}{2} + ghz_b, \qquad G(U) = \left(\frac{hu^2}{2} + gh^2 + ghz_b\right) u. \tag{1.5}$$

Any solution of the Saint-Venant system (1.1) satisfying (1.3) and (1.4) for $\mathcal{E} = E$ is referred to as entropy solution. The existence of entropy solutions in the case of flat bathymetry is given by Lions, Perthame, and Souganidis [71] by the means of a *kinetic formulation* [80] which accounts for all existing entropies of the system, not only the energy. In the case of a single entropy, in this case the energy, we obtain the less precise *kinetic representation*. We are interested in the construction of *well-balanced* schemes which preserve the lake-at-rest steady states presented in the following section.

1.2.2 Stationary solutions

A steady state refers to a solution that is constant in time. In other words, the steady states of the Saint-Venant are obtained by setting $\partial_t h = 0$, $\partial_t u = 0$ in system (1.1):

$$\begin{cases} \partial_x(hu) = 0, \\ \partial_x(hu^2 + \dfrac{gh^2}{2}) = -gh\partial_x z_b, \end{cases}$$

which gives the following relation for steady state solutions:

$$\partial_x \left(\frac{u^2}{2} + g(h + z_b)\right) = 0.$$

The stationary solutions correspond to steady state solutions when $u = 0$. This leads to a balance between the pressure term and the source term and gives $h + z_b = cst$ corresponding to a flat surface. This state is what we refer to as *hydrostatic equilibrium* or often called lake-at-rest and it is important to preserve this equilibrium at the discrete level to guarantee better numerical results. Numerical schemes that preserve the hydrostatic equilibrium are referred to as *well-balanced* schemes.

1.2.3 Kinetic representation

Kinetic schemes offer a unique perspective in the numerical resolution of the Saint-Venant system. The kinetic description of the flow views the system as a collection of particles following a density distribution $f(t, x, \xi) \geq 0$ at time $t > 0$ and position $x \in \mathbb{R}$ traveling with velocity $\xi \in \mathbb{R}$. Collisions

between particles are also accounted for by the means of a collision operator $Q[f]$ which captures the interactions between the particles and plays an important role in describing the evolution of the system. The kinetic description offers insight to the behavior of the system at the mesoscopic level through the Boltzmann kinetic equations,

$$\partial_t f(t,x,\xi) + \xi \partial_x f(t,x,\xi) - g(\partial_x z_b)\partial_\xi f(t,x,\xi) = \frac{1}{\varepsilon} Q[f](t,x,\xi), \tag{1.6}$$

where $\dfrac{1}{\varepsilon}$ is the frequency of collisions. The collision operator $Q[f]$ should satisfy the additional constraints:

$$\int_{\mathbb{R}} \begin{pmatrix} 1 \\ \xi \end{pmatrix} Q[f](t,x,\xi)\, d\xi = 0 \quad \text{for a.e. } (t,x), \tag{1.7}$$

which ensures that the total mass and momentum of the system is conserved despite the collisions. Several choices for the collision operator satisfying (1.7) exist in the literature, we consider one of the simplest choices for the collision operator given by *Bhatnagar, Gross,* and *Krook* [13] and known in the literature as the BGK operator,

$$Q_{BGK}[f](t,x,\xi) := M(U_f(t,x),\xi) - f(t,x,\xi), \tag{1.8}$$

where

$$U_f = \int_{\mathbb{R}} \begin{pmatrix} 1 \\ \xi \end{pmatrix} f(t,x,\xi)\, d\xi\,.$$

The collision operator is responsible for the relaxation of the distribution towards a state of hydrodynamic equilibrium (as $\varepsilon \to 0$) known as the Maxwellian distribution which will be denoted by $M(U,\xi)$ throughout the manuscript. We perform a BGK splitting which divides the collision into two steps:

- Transport step: $\partial_t f + \xi \partial_x f - g(\partial_x z_b)\partial_\xi f = 0$ where the particles move freely without any collisions.
- Relaxation step: $\partial_t f = \dfrac{M(U_f(t,x),\xi) - f(t,x,\xi)}{\varepsilon}$ where the distribution function $f(t,x,\xi)$ is adjusted to reach an equilibrium state $M(U_f(t,x),\xi)$ as $\varepsilon \to 0$ [18].

While the kinetic representation (1.6) offers a mesoscopic view of the system, the hydrodynamic limit $\varepsilon \to 0$ allows us to recover the macroscopic view of the system by assuming permanent collisions between the particles provided that the operator $Q[f]$ satisfies for a.e. (t,x):

$$Q[f] \equiv 0 \iff f(\xi) = M(U_f,\xi).$$

In order to establish a link between the kinetic equations and the Saint-Venant system, we will choose the Maxwellian that satisfies the following relations for any solution $U \in \mathbb{R}^+ \times \mathbb{R}$ of (1.1):

$$\int_{\mathbb{R}} \begin{pmatrix} 1 \\ \xi \\ \xi^2 \end{pmatrix} M(U,\xi)\, d\xi = \begin{pmatrix} h \\ hu \\ hu^2 + gh^2/2 \end{pmatrix} \quad \forall U \in \mathbb{R}^+ \times \mathbb{R}. \tag{1.9}$$

The Maxwellian plays an essential role in recovering the macroscopic variables, indeed by integrating the Maxwellian against the vector $(1,\xi,\xi^2)^T$ one can recover equations (1.1). This is not possible for a generic distribution in the following sense: although any density $f(t,x,\xi)$ can be used to define the macroscopic variable U_f, the integral of f against ξ^2 does not necessarily yield the second

component of $F(U_f)$ which is why a Maxwellian is needed. Following [5] we introduce the classical kinetic Maxwellian

$$M(U,\xi) = \frac{1}{g\pi}\left(2gh - (\xi-u)^2\right)_+^{1/2}, \tag{1.10}$$

where $x_+ \equiv \max(0,x)$ for all $x \in \mathbb{R}$. The Maxwellian (1.10) satisfies the moment relations (1.9). In addition, it is of particular importance since it minimizes the entropy, more on that can be found in [78] and in Chapter 2. A family of Maxwellians satisfying (1.9) can be defined following the work in [78].

The conditions (1.9) allow us to obtain a *kinetic representation* [17,18,33,76] of the Saint-Venant system (1.1) through Lemma 1 by taking the limit $\varepsilon \to 0$ (formally) in equation (1.6) using the BGK operator (1.8). This notion differs from the so-called *kinetic formulations* which account for all the entropies of the system [70, 80].

Lemma 1. *If the topography $z_b(x)$ is Lipschitz continuous, the pair of functions $U = (h, hu)$ is a weak solution to the Saint-Venant system (1.1) if and only if $M(U,\xi)$ satisfies (1.9) and the kinetic equation*

$$\partial_t M(U,\xi) + \xi\partial_x M(U,\xi) - g(\partial_x z_b)\partial_\xi M(U,\xi) = Q,$$

for some "collision term" $Q(t,x,\xi)$ that satisfies (1.7) for a.e. (t,x).

In [17], *F. Bouchut* introduces a comprehensive framework for the construction of BGK models based on the notion of kinetic entropy which can be considered as an energy distribution taken at the kinetic level. Precisely, it follows from the fact that in the case of flat topography, the kinetic entropy allows us to recover the entropy inequality as $\varepsilon \to 0$. We recall that the half-disk Maxwellian (1.10) is associated to the following kinetic entropy,

$$H(f,\xi,z) = \frac{\xi^2}{2}f + \frac{g^2\pi^2}{6}f^3 + gz_b f.$$

Then one can check that the integral of $M(U,\xi)$ against the vector $(1,\xi)^T$ gives the entropy-entropy flux pair (1.5)

$$\int_{\mathbb{R}} \begin{pmatrix} 1 \\ \xi \end{pmatrix} H\big(M(U,\xi),\xi,z_b\big)\, d\xi = \begin{pmatrix} E(U) \\ G(U) \end{pmatrix}. \tag{1.11}$$

The choice (1.10) for the Maxwellian is essential for the derivation of the entropy inequality on the kinetic level and consequently on the macroscopic level through relations (1.11), see [80]. Indeed in the case of a single entropy, as the case at hand, the construction of a BGK model requires that the Maxwellian is chosen to be the distribution which minimizes the chosen convex entropy while preserving the moment relations (1.11). It turns out that (1.10) is the only solution to this constrained minimization problem, more about this point in Lemma 3 in Chapter 2. Although a classical Maxwellian of the form (1.10) is linked to a kinetic entropy, in practice it is replaced by other expressions which allow for more efficient computation of the macroscopic updates.

1.2.4 A quick reminder of the explicit kinetic scheme

An explicit kinetic scheme was proposed in [5] accounting for varying topography where the topography was treated with the hydrostatic reconstruction technique [8]. This scheme uses the BGK splitting introduced in Section 1.2.3 and it was shown that the explicit-in-time discretization preserves the positivity of the distribution function under a CFL condition. This is then used to prove that the scheme preserves the positivity of the water height under a CFL constraint. Using the notion of kinetic entropy [17], the authors in [5] are able to prove that in the case of flat topography,

the explicit scheme enters the framework of *flux vector splitting* schemes [18] and satisfies a fully discrete entropy inequality under a CFL condition. Indeed, the CFL restriction allows us to write the microscopic updates as a convex combination of Maxwellians and use the convexity of the entropy to obtain the desired entropy inequality. This allows to deduce the discrete entropy inequality satisfied by the energy given by (1.5) with $z_b = 0$. However when the topography variation is taken into account, this becomes less evident due to the additional terms associated to the topography that are treated using a kinetic interpretation of the hydrostatic reconstruction scheme. The entropy inequality in the time explicit setting contains positive terms that are difficult to control therefore the analysis becomes less obvious and we are left with the following proposition.

Proposition 1. *(Audusse et al. [5]) The explicit scheme with hydrostatic reconstruction does not satisfy a discrete entropy inequality without a quadratic error term no matter what additional restriction on the CFL is considered.*

The above proposition is the main motivation for the following section.

1.2.5 Fully implicit kinetic scheme in the case of flat topography

The advantage behind using a kinetic scheme is highlighted by kinetic representation (1.6). Instead of discretizing the non-linear Saint-Venant system (1.1), we now discretize the equation (1.6) alternating between a linear transport step and a relaxation step through the BGK operator (1.8). To recover the macroscopic updates and the entropy inequaltiy at the macroscopic level, we integrate the Maxwellian (resp. the associated kinetic entropy) against the vector $(1, \xi)^T$. Starting from the case of flat topography, we will use an implicit upwind scheme to discretize the transport step with initial data $f(t^n, x, \xi) = M(U^n(x), \xi)$. The implicit update is the following:

$$f_i^{n+1-} = M_i - \sigma\xi \left(\mathbb{1}_{\xi<0} f_{i+1}^{n+1-} + \mathbb{1}_{\xi>0} f_i^{n+1-} - \mathbb{1}_{\xi<0} f_i^{n+1-} - \mathbb{1}_{\xi>0} f_{i-1}^{n+1-} \right), \tag{1.12}$$

with $\sigma = \Delta t^n / \Delta x$ and $1 \leq i \leq P$ where P is the number of interior cells. This consists in finding $f^{n+1} = \{f_i^{n+1-}\}_{i \in \{1,\dots,P\}}$ satisfying

$$(\mathbf{I} + \sigma \mathbf{L}) f^{n+1} = M + \sigma B^{n+1}, \tag{1.13}$$

where $\mathbf{I}$ is the identity matrix of length P and $\mathbf{L} \in \mathbb{R}^{P \times P}$ is given by

$$\mathbf{L} = \begin{pmatrix} |\xi| & \xi \mathbb{1}_{\xi<0} & 0 & \dots & 0 \\ -\xi \mathbb{1}_{\xi>0} & |\xi| & \xi \mathbb{1}_{\xi<0} & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -\xi \mathbb{1}_{\xi>0} & |\xi| & \xi \mathbb{1}_{\xi<0} \\ 0 & \dots & 0 & -\xi \mathbb{1}_{\xi>0} & |\xi| \end{pmatrix}.$$

The three vectors f^{n+1}, M and B^{n+1} of $\mathbb{R}^P$ are defined by

$$f^{n+1} = \begin{pmatrix} f_1^{n+1-} \\ \vdots \\ f_i^{n+1-} \\ \vdots \\ f_P^{n+1-} \end{pmatrix}, \quad M = \begin{pmatrix} M_1 \\ \vdots \\ M_i \\ \vdots \\ M_P \end{pmatrix} \quad \text{and} \quad B^{n+1} = \begin{pmatrix} \mathbb{1}_{\xi>0} \xi M_0^{n+1}, \\ 0 \\ \vdots \\ 0 \\ -\mathbb{1}_{\xi<0} \xi M_{P+1}^{n+1}, \end{pmatrix}.$$

The indices 0 and $P+1$ refer to the ghost cells. Note that the notation $\{f_i^{n+1-}\}_{i\in\{1,\dots,P\}}$ refers to the updates corresponding to the transport step and before the projection step, which means that this quantity is not a Maxwellian. Working with an implicit scheme often requires inverting an operator, here $(\mathbf{I}+\sigma\mathbf{L})$, to obtain an explicit expression of the vector of updates $f^{n+1}(\xi)$. However, due to the particular nature of the matrix $(\mathbf{I}+\sigma\mathbf{L})$ and due to the upwinding of fluxes considered, we are able to compute its inverse $(\mathbf{I}+\sigma\mathbf{L})^{-1}$ and prove that all its coefficients are positive. The macroscopic updates are given by:

$$h^{n+1}=\int_{\mathbb{R}}(\mathbf{I}+\sigma\mathbf{L})^{-1}(M+\sigma B^{n+1})d\xi,\quad (hu)^{n+1}=\int_{\mathbb{R}}\xi(\mathbf{I}+\sigma\mathbf{L})^{-1}(M+\sigma B^{n+1})d\xi. \tag{1.14}$$

The following Proposition holds.

Proposition 2. *The update (1.14) is consistent with the Saint-Venant system (1.1) with flat topography $z_b=0$, preserves the positivity of the water height, admits a unique non-negative solution, and satisfies a discrete entropy inequality provided that M is a vector of half-disk Maxwellians of the form (1.10).*

The existence of a discrete entropy inequality is strongly linked to the implicit nature of the scheme, this is illustrated by the following simple example.

Example 1 (The linear transport equation)**.** *Let $a>0$ be a positive constant and consider the following linear transport equation with periodic boundary conditions:*

$$\begin{cases}\partial_t u+a\partial_x u=0, & t>0, x\in[-1,1]\\ u(0,x)=u_0(x),\\ u(t,-1)=u(t,1).\end{cases} \tag{1.15}$$

We will consider the solutions of (1.15) in $L^2[-1,1]$. The energy of the system is given by $E(u)=\dfrac{u^2}{2}$ and by multiplying the first equation in (1.15) by $u\in L^2[-1,1]$, we deduce that the total energy of the system is constant in time, in particular no increase of energy appears. This is an important stability property that we would like to preserve on the discrete level through an appropriate numerical scheme. Let us demonstrate the effect of the explicit (respectively implicit) upwind scheme on the energy dissipation at the discrete level. Consider the following explicit-in-time upwind scheme for system (1.15)

$$u_i^{n+1}-u_i^n=-a\frac{\Delta t}{\Delta x}(u_i^n-u_{i-1}^n). \tag{1.16}$$

The discrete energy is obtained (in analogy to the continuous case) after multiplying the scheme by u_i^n. The upwind scheme in a time-explicit setting satisfies:

$$\frac{E(u_i^{n+1})-E(u_i^n)}{\Delta t}+a\frac{E(u_i^n)-E(u_{i-1}^n)}{\Delta x}=\frac{1}{2}\left(u_i^{n+1}-u_i^n\right)^2-\frac{a}{2\Delta x}\left(u_i^n-u_{i-1}^n\right)^2. \tag{1.17}$$

Replacing the difference $u_i^{n+1}-u_i^n$ by the scheme (1.16), one gets

$$\frac{1}{2}\left(u_i^{n+1}-u_i^n\right)^2-\frac{a}{2\Delta x}\left(u_i^n-u_{i-1}^n\right)^2=\frac{a}{2\Delta x}(-\Delta x+a\Delta t)\left(u_i^n-u_{i-1}^n\right)^2,$$

where the right-hand side is positive unless the CFL condition $a\Delta t<\Delta x$ is satisfied. In fact, this is due to the positive term on the right-hand side of (1.17) resulting from the time discretization and

which is not always dominated by the second dissipating term enforced by the upwinding. Whereas in an implicit-in-time upwind scheme, the discrete energy satisfies:

$$\frac{E(u_i^{n+1})-E(u_i^n)}{\Delta t}+a\frac{E(u_i^{n+1})-E(u_{i-1}^{n+1})}{\Delta x}=-\frac{1}{2}\left(u_i^{n+1}-u_i^n\right)^2-\frac{a}{2\Delta x}\left(u_i^{n+1}-u_{i-1}^{n+1}\right)^2,$$

where the right-hand side has always a negative sign without any CFL constraint. This is obtained by multiplying the scheme by u_i^{n+1}. One can also replace the difference $u_i^{n+1}-u_i^n$ by the implicit version of the scheme (1.16),

$$-\frac{1}{2}\left(u_i^{n+1}-u_i^n\right)^2-\frac{a}{2\Delta x}\left(u_i^{n+1}-u_{i-1}^{n+1}\right)^2=-\frac{a}{2\Delta x}(\Delta x+a\Delta t)\left(u_i^{n+1}-u_{i-1}^{n+1}\right)^2.$$

Hence the implicit-in-time setting can guarantee that the upwind scheme dissipates the energy without any CFL constraint. As we will see in Chapter 2 we will prove that the implicit scheme is favorable to the explicit one on the level of entropy dissipation. This is done using the notion of kinetic entropy, see [17].

Using the explicit expression of the inverse matrix $(\mathbf{I}+\sigma\mathbf{L})^{-1}$, the computation of the macroscopic update associated with (1.12) can be expressed explicitly. This involves evaluating the integral of

$$\mathbb{1}_{\pm\xi>0}\frac{(\pm\sigma\xi)^k}{(1\pm\sigma\xi)^{k+1}}M(U,\xi) \tag{1.18}$$

against 1, ξ and ξ^2 for $0\leq k\leq P-1$. Given the Maxwellian (1.10), this seems hardly possible and instead we will use the simpler equilibrium function given by

$$M(U,\xi)=\frac{h}{2\sqrt{3}c}\mathbb{1}_{|\xi-u|\leq\sqrt{3}c},\quad c=\sqrt{\frac{gh}{2}}, \tag{1.19}$$

and referred to as the index Maxwellian. The index Maxwellian (1.19) satisfies the relations (1.9) and allows us to find the expressions of the integral of the quantities (1.18) against 1, ξ and ξ^2 for $0\leq k\leq P-1$. Although we are unable to prove that it satisfies a discrete entropy inequality as previously noted in Proposition 2 in the case of (1.10), the index Maxwellian remains a better option than using a quadrature formula to approximate the integrals involving the half-disk Maxwellian. Moreover, the implicit kinetic scheme obtained using the index Maxwellian (1.19) will also be positive and will enable us to find explicit expressions of the macroscopic updates which an important advantage of the fully implicit scheme describing a non-linear system in the first place.

Hence it becomes clear that a kinetic solver can potentially offer a reduced computational cost in comparison to other finite-volume solvers due to its ability to explicitly define the macroscopic updates. Although the implicit scheme is not restricted by a CFL condition which is required in explicit schemes, it is important to assess the computational cost and efficiency of the implicit scheme in comparison to the explicit one [5] in the context of kinetic schemes. Let Δt^n and Δt^n_{imp} denote the time-steps associated to the explicit scheme and implicit scheme (1.12) respectively, then the implicit scheme has a lower computational cost when

$$\frac{\Delta t^n_{imp}}{\Delta t^n}\gg P.$$

When it comes to the efficiency of the implicit scheme; the relation between the error and the computational time, in cases such as a low-Froude regime where the fast dynamics do not play an important role it can be better to consider larger time steps thus the use of an implicit kinetic scheme

is more advantageous. Nevertheless, taking large time steps results in a very coarse resolution and might lead to less accurate results and therefore explicit schemes remain preferable unless one takes into account the greater stability obtained by the discrete entropy inequality which is provided by the implicit scheme.

1.2.6 Iterative resolution

When using the half-disk Maxwellian (1.10), it becomes very difficult to compute the explicit expression of the macroscopic updates, see (1.11). Indeed, the macroscopic updates are recovered by integrating f^{n+1} against the vector $(1,\xi)^T$, see relations (1.9). An alternative approach to the resolution of the implicit scheme (1.13) where M is the vector of half-disk Maxwellians of the form (1.10) is given by an iterative process. Using an iterative process circumvents the difficulty presented in Section 1.2.5 where integrating the expression (1.18) in the case of the half-disk Maxwellian is not achievable. It is also computationally less expensive than numerical integration and will be essential to treat the non-linear terms that will appear later when the scheme is extended to account for topography variations. We use the following iterative process based on a *Gauss-Jacobi* decomposition as an approximation to the solution of (1.13):

$$\left\{\begin{array}{l} f^{n+1,0}(\xi) = M \\ (1+\alpha) f^{n+1,k+1}(\xi) = (\alpha \mathbf{I} - \sigma^k \mathbf{L}) M^{n+1,k} + M + \sigma^k B^{n+1,k} \\ M^{n+1,k+1} = f^{n+1,k+1} + \Delta t^k Q^{n+1,k+1} \end{array}\right. , \tag{1.20}$$

where k is the index of iteration and $\alpha \geq 0$ is a relaxation parameter. Note that the superscript $\square^{n+1-}$ that was previously used (see Section 1.2.5) to indicate that the distribution does not correspond to a Maxwellian at time Δt, is dropped for the sake of simplicity. The Gauss-Jacobi type decomposition we consider consists of writing

$$\mathbf{I} + \sigma \mathbf{L} = \mathbf{D} - \mathbf{N},$$

where $\mathbf{D}, \mathbf{N} \in \mathbb{R}^{P \times P}$ are chosen as follows

$$\mathbf{D} = (1+\alpha)\mathbf{I}, \quad \text{and} \quad \mathbf{N} = \alpha \mathbf{I} - \sigma \mathbf{L}.$$

In fact, one can consider different choices for $\mathbf{D}$ and $\mathbf{N}$ in $\mathbb{R}^{P \times P}$ as long as $\mathbf{D}$ is invertible but we do not investigate further this direction in this manuscript due to the time constraints, and we hope to explore other choices in the future. No matrix inversion is required in the iterative process and the macroscopic updates associated to the half-disk Maxwellian can be explicitly expressed through the direct integration of expression (1.18) against $(1,\xi)^T$. The importance of the iterative resolution (1.20) in the case of flat topography is that it enables us to use the Maxwellian (1.10) which is associated with a discrete entropy inequality as stated in Proposition 2. The iterative scheme (1.20) dissipates the entropy at each rank $k \in \mathbb{N}$ provided some CFL condition is satisfied. The following Proposition holds.

Proposition 3. *Given that it converges, the iterative process (1.20) satisfies a discrete entropy inequality which is energy dissipating after some rank $k \in \mathbb{N}$.*

The CFL restriction on the iterative process ensures the positivity and the convergence of the method but is viewed as a downside of using an iterative scheme in comparison to a fully-implicit one. However, as we will see below in the case of varying topography, even under a CFL constraint, the iterative process (1.20) of the implicit scheme (1.13) provides an advantage over the explicit scheme through the dissipation of the entropy on the macroscopic level.

1.2.7 Variable topography and hydrostatic reconstruction

In the case of varying topography, additional tools are needed to approximate the source term appearing on the right-hand side of the Saint-Venant system (1.1) which introduces additional terms in the kinetic scheme (1.12). In order to derive a *well-balanced* scheme in the case of varying topography, a hydrostatic reconstruction technique was introduced by *Audusse et al.* in [8]. This allows to preserve the positivity of the water height as well as preserve the lake-at-rest steady states presented in Section 1.2.2. The hydrostatic reconstruction scheme is based on the principle of hydrostatic equilibrium, which states that the pressure distribution in a fluid at rest is determined solely by the depth of the fluid column.

The fully implicit kinetic scheme in Section 1.2.5 no longer yields explicit expressions of the updates due to the non-linearity induced by the topography variation and hence an iterative resolution is needed. Using the *Gauss-Jacobi* decomposition as in Section 1.2.6 for flat topography, we can write the iterative scheme with hydrostatic construction as follows:

$$\begin{cases} f^{n+1,0}(\xi) = M, \\ (1+\alpha) f_i^{n+1,k+1} = M_i + \alpha M_i^{n+1,k} - \sigma^k \xi \Big(\mathbb{1}_{\xi<0} (M_{i+1/2+}^{n+1,k} - M_{i-1/2+}^{n+1,k}) \\ \qquad + \mathbb{1}_{\xi>0} (M_{i+1/2-}^{n+1,k} - M_{i-1/2-}^{n+1,k}) \Big) + \sigma^k (\xi - u_i^{n+1,k}) (M_{i+1/2-}^{n+1,k} - M_i^{n+1,k}) \\ \qquad\qquad - \sigma^k (\xi - u_i^{n+1,k}) (M_{i-1/2+}^{n+1,k} - M_i^{n+1,k}), \end{cases} \tag{1.21}$$

with the following notations

$$M_{\square}^{n+1,k} = M(U_{\square}^{n+1,k}, \xi), \quad h^{n+1,k} = \int_{\mathbb{R}} f^{n+1,k}(\xi)\, \mathrm{d}\xi, \quad (hu)^{n+1,k} = \int_{\mathbb{R}} \xi f^{n+1,k}(\xi)\, \mathrm{d}\xi,$$

where the square symbol "$\square$" in subscript can be replaced by i (centered value) or $i \pm 1/2\mp$ (reconstructed interfacial value). The terms multiplied by $(\xi - u_i^{n+1,k})$ in (1.21) corresponds to the kinetic interpretation of the source term [5] which is done by the means of a hydrostatic reconstruction technique. The iterative kinetic scheme (1.21) dissipates the energy unlike the explicit scheme where the entropy inequality contains positive terms that are not necessarily well controlled. This result is well illustrated in Fig. 1.2. We have the following result.

Proposition 4. *Given that it converges, the iterative scheme (1.21) preserves the positivity of the water height and satisfies a discrete entropy inequality which is entropy dissipating after some rank* $k \in \mathbb{N}$.

In the scope of this project we are able to address some questions that have received a lot of attention in the community. Is it practical to use an implicit scheme instead of an explicit one? What are the advantages and disadvantages of the implicit scheme? Several well-balanced techniques have been proposed to discretize the classical Saint-Venant system for shallow water flows with non-flat topography. Among these approaches, the hydrostatic reconstruction scheme stands out [5, 8, 9, 21]. This technique consists of choosing an arbitrary solver which is entropy satisfying, to yield a semi-discrete entropy inequality in return. The authors in [5] prove that the hydrostatic reconstruction technique coupled with a kinetic solver satisfies a fully discrete entropy inequality with positive error terms that vanish in the case of flat topography and in the semi-discrete limit [8]. However this does not mean that the explicit scheme with varying topography is entropy dissipating, on the contrary, the work in [5] proves otherwise. This is expected due to the less dissipative nature of fully-explicit schemes in comparison to semi-explicit schemes as well as implicit schemes. This is illustrated by the linear transport equation in Example 1 where the time explicit setting results in positive error terms that are not necessarily controlled. The CFL condition, no matter how restrictive, does not

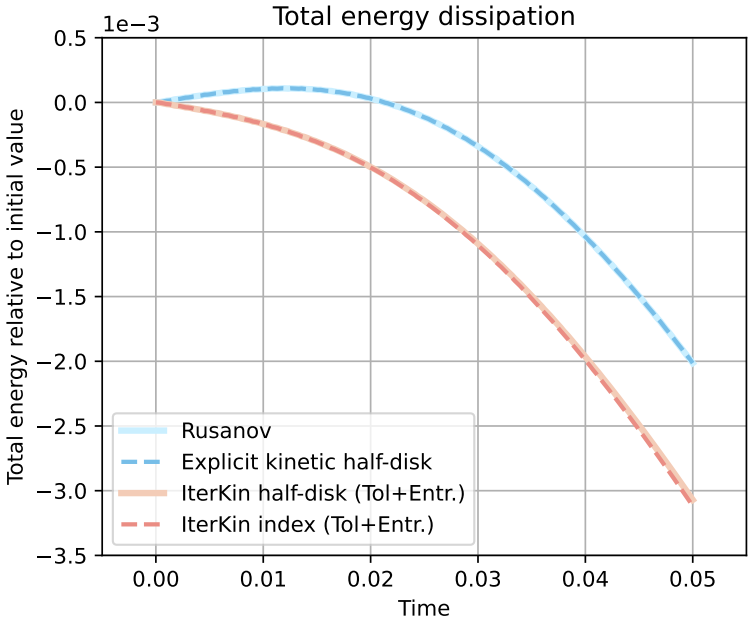

Figure 1.2: Evolution of the relative total energy obtained for various explicit and iterative kinetic schemes.

guarantee that the explicit scheme with hydrostatic reconstruction dissipates the entropy which is why exploring an implicit scheme can be interesting. So if you have wondered if an implicit kinetic scheme for the Saint-Venant system has been considered, the answer is yes, and the reason why it is interesting is the favorable entropy-dissipating property.

This gets us to the second question regarding the practical use of such a scheme. As we will see in Chapter 2, the implicit scheme can be less computationally expensive than the explicit counterpart but it is necessary to include other factors in order to asses its practical use, such as its efficiency. Although taking larger time steps can lead to less accurate results which means that the explicit scheme is still preferable, there are some scenarios where the implicit scheme is a good alternative such as the case of low *Froude* regimes. So the answer to the question is: it depends on the context and we will illustrate this by numerical tests at the end of Chapter 2.

Other directions remain unexplored within the project and provide a perspective for further investigation on the implicit scheme: Experimenting with other iterative strategies in hope of accelerating the convergence rate of the iterative scheme; Improving the convergence hypothesis on the iterative scheme; Studying the two-dimensional iterative scheme for varying topography.

1.3 Hydrostatic Euler system

Despite its ability to accurately describe fluid motion in shallow water, the Saint-Venant system does not take into account the vertical fluid variations which are neglected in this context. On the other hand, the free-surface hydrostatic Euler system accounts for both horizontal and vertical variations of the velocity field which makes it mathematically more accurate in modeling free-surface fluids under hydrostatic pressure. These equations have been extensively studied [23, 24, 74, 81]

and their stability remains a challenging and ongoing research area due to the complex nature of these equations, which unlike the Saint-Venant equations, do not fall under any classical type of partial differential equations namely hyperbolic, parabolic, or elliptic. In this work we rewrite the free-surface hydrostatic Euler system in a way that resembles the classical hyperbolic structure of a non-linear system with the exception of an integral non-local operator. The non-linear system under consideration encompasses certain features that differ from classical hyperbolic systems, notably this system involves no vertical derivatives, which means that the equations do not explicitly account for variations in the vertical direction and the coefficients of the horizontal derivatives involve an integral operator. Furthermore, in order to rewrite the hydrostatic Euler system in this new form, we employ a transformation which maps the time-dependent domain into a fixed fluid domain where the variables are written in a semi-Lagrangian formulation. This formulation was first introduced by V.E. Zakharov [95] to derive an infinite system of conservation laws for shear flows in shallow water.

The following project resulted in a paper [38] in collaboration with *B. Di Martino*, *E. Godlewski*, *J. Guillod*, and *J. Sainte-Marie*. The pre-print can be found on *HAL* (hal-04190892) or *arxiv* (arXiv:2308.15083).

This section views the stability of these equations through the lens of hyperbolic systems of conservation laws. Although *a priori* the hydrostatic Euler system is not hyperbolic in nature, there exists a transformation which bridges the gap between the system and classical hyperbolic-type systems allowing us to borrow some classical tools to build a better understanding of the former set of equations which remain a challenging research topic. We recall that the hydrostatic pressure is a function of the fluid depth and density and is obtained by neglecting the vertical pressure gradients in the vertical momentum equation. This assumption reduces the dimension of the Euler equations. Throughout this project we will assume the following setting:

- The flow is incompressible
- The pressure is hydrostatic
- The fluid domain Ω_t is bounded from below by the bottom topography and from above by a free-surface
- To avoid imposing boundary conditions on the horizontal direction, which is not the interest of this project, we assume that the horizontal coordinates lie in an infinite domain.

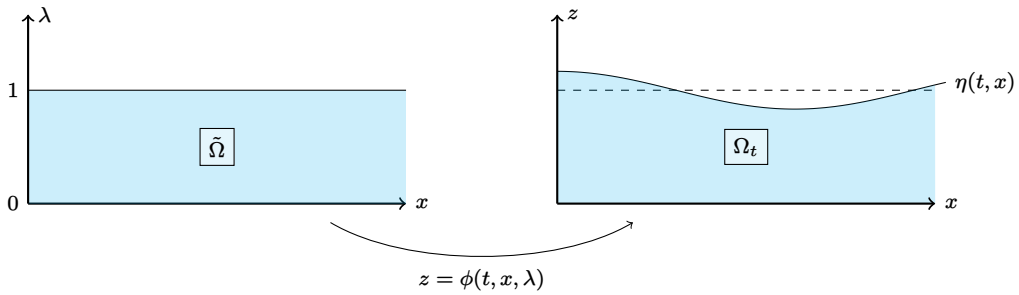

Figure 1.3: The transformation ϕ in a two-dimensional setting

1.3.1 The transformation

Understanding the nature of the hydrostatic Euler equations is essential since the equations provide access to the vertical variation of the velocity which is neglected in the context of the Saint-Venant system. It is also important for the development of appropriate numerical schemes that could be valuable for the simulation and prediction of various natural phenomena occurring in the ocean and at the atmosphere and helps researchers to tackle more complex fluid phenomena that resemble real-world situations under a certain level of accuracy. In the view of [91, 95], the hydrostatic Euler system for incompressible free-surface flows can be mapped onto a fixed domain $\tilde{\Omega}$, as shown in Fig. 1.3, where the new system can be written in a generalized quasi-linear form. We first prove the existence and uniqueness of such a transformation as well as the equivalence between the two systems, this will be further detailed in Chapter 3. The result is summarized as follows.

Under certain regularity conditions imposed on the velocity components and the fluid height associated to the hydrostatic Euler system, there exists a transformation which maps the time-varying domain into a fixed flat domain allowing us to write the system in the following quasi-linear form:

$$\frac{\partial \boldsymbol{U}}{\partial t} + A_1(\boldsymbol{U})\frac{\partial \boldsymbol{U}}{\partial x} + A_2(\boldsymbol{U})\frac{\partial \boldsymbol{U}}{\partial y} = 0, \tag{1.22}$$

where $A_1(\boldsymbol{U})$ and $A_2(\boldsymbol{U})$ are two 3×3 non-local operators acting on the vertical variable $\lambda \in [0, 1]$ and $\boldsymbol{U}$ is the vector composed of the system variables: the fluid variation and the horizontal velocity components. Moreover, the transformation is locally invertible and preserves the orientation. This provides the local (in time) equivalence between the two systems. In general, the well-posedness result is not global. This is due to the non-linearity of the hyperbolic differential equation which governs the evolution of the transformation and might result in the blow-up of the space derivative in finite critical time, see Fig. 1.4. This is not surprising since the original hydrostatic Euler equations are not globally well-posed either. We will not delve further into the specifics in this section, but it is worth noting that this observation will be exemplified and explored in Chapter 3. The equivalence between the two systems in a thin two dimensional layer was presented by *Y. Brenier* in [23]. The analysis in [23] specifically applies to a bounded domain and does not directly extend to free-surface flows. It focuses on smooth solutions of the system where the transformation ϕ remains invertible for a short time interval. The analysis in this project is focused on free-surface flows.

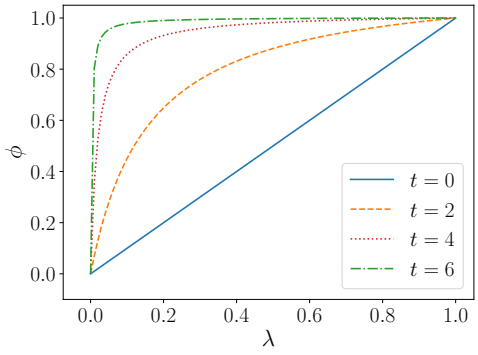

Figure 1.4: Graph of the transformation ϕ for a flow with vorticity where the initial water depth equal to 1 at times $t = 0, 2, 4, 6$.

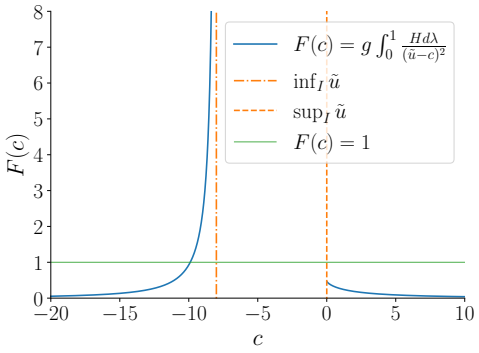

Figure 1.5: The case $\tilde{u} \in C^{0,1/3}(I)$ and $H(\lambda) = 1$ for $g = 10$.

The formulation (1.22) was first introduced by V.E. Zakharov in [95], and has been later on

used in [29, 30, 86, 90, 91]. It allows us to access the spectral elements of the system which was not possible for the free-surface hydrostatic Euler system, hence the advantage of introducing such a transformation. This strategy however, comes at the expense of a non-local operator and new transformed variables where the fluid height is replaced by a function which depends additionally on the vertical coordinate, as we will discuss later in Chapter 3.

1.3.2 The spectrum

In the classical case, with a d-dimensional ($d \geq 1$) square matrix A_0 and a d-dimensional vector $\tilde{U}$, a system of the form

$$\frac{\partial \tilde{U}}{\partial t} + A_0(\tilde{U})\frac{\partial \tilde{U}}{\partial x} = 0, \tag{1.23}$$

is said to be (strictly) *hyperbolic* if for every state $\tilde{U}$ the matrix $A_0(\tilde{U})$ has d real (distinct) eigenvalues and a complete set of eigenvectors. In the case of such classical hyperbolic systems, the eigenvalues of the *Jacobian* matrix $A_0(\tilde{U})$ are real and they correspond to wave speeds. Each eigenvalue is associated with an eigenvector which can allow us to determine the direction of wave propagation in certain cases. Hence the eigenelements play a crucial role in understanding the behavior of the solution. In the case of the system at hand, the notion of *hyperbolicity* becomes more complicated. We consider, for convenience, the system (1.23) with $d = 2$ and

$$A_0(\tilde{U}) = \begin{pmatrix} \tilde{u} & H \\ g\int_0^1 \cdot \, \mathrm{d}\lambda & \tilde{u} \end{pmatrix}, \quad \text{where} \quad \tilde{U} = \begin{pmatrix} H \\ \tilde{u} \end{pmatrix}.$$

Let $I := [0, 1]$ denote the vertical domain. We note that $H = H(t, x, \lambda)$ and $\tilde{u} = \tilde{u}(t, x, \lambda)$ but throughout the spectral analysis, the time variable and the horizontal variables are not essential and therefore will be disregarded. As a result we view the system unknowns as functions of the vertical variable $\lambda \in I$. Note that the function H is no longer the water depth, it is instead a function of λ which when integrated over I gives the water depth h associated to the hydrostatic Euler system. The authors in [32, 88, 91] state that the spectrum consists of two types of values: values that are not in the range of values of the velocity; and the values of the velocity on the interval I. We take this statement a step further to prove a complete decomposition of the spectrum.

Let $C^{0,\alpha}(I)$ be the space of Hölder continuous functions of index α over I. For the sake of simplicity, we will use the notation A_0 to refer to $A_0(\tilde{U})$.

Theorem 1. *If $\tilde{u} \in C^{0,1/4}(I)$, $H \in C^0(I)$ and $H > 0$, then the spectrum of $A_0 : L^2(I) \times L^2(I) \to L^2(I) \times L^2(I)$ is characterized by*

$$\begin{aligned}
\sigma_p(A_0) &= \left\{ c \in \mathbb{C} \setminus \tilde{u}(I) \, : \, \int_0^1 \frac{gH}{(c - \tilde{u})^2} \mathrm{d}\lambda = 1 \right\} \cup \left\{ c \in \tilde{u}(I) \, : \, \text{meas}\left(\tilde{u}^{-1}(c)\right) > 0 \right\}, \\
\sigma_c(A_0) &= \left\{ c \in \tilde{u}(I) \, : \, \text{meas}\left(\tilde{u}^{-1}(c)\right) = 0 \right\}, \\
\sigma_r(A_0) &= \emptyset,
\end{aligned}$$

where the definitions of the point spectrum $\sigma_p(A_0)$, continuous spectrum $\sigma_c(A_0)$, and residual spectrum $\sigma_r(A_0)$ are given in Appendix 3.A.

In classical hyperbolic theory, the spectrum is fully discrete which is not the case for A_0 due to the presence of a continuous spectrum as shown in Theorem 1. Although the values $\tilde{u}(I)$ are elements of the continuous spectrum, they are not part of the *eigenvalues* of the system unless meas $\left(\tilde{u}^{-1}(c)\right) > 0$. The notion of *eigenvalues* will be restricted to the elements of the point spectrum.

1.3.3 Eigenvalues and Riemann Invariants

The new system presents us with several challenges which we will address throughout the manuscript, mainly:

- There is no explicit expression of the eigenvalues.
- The classical notions (characteristic fields, Riemann invariants) of hyperbolic theory are not well-understood in the context of the generalized hyperbolic system (1.23).

Despite having no explicit expression of the eigenvalues, one can still look for alternative ways to determine the nature (real or complex) of the eigenvalues in order to prove the generalized hyperbolicity of system (1.23). The authors in [30,89] derived the following implicit formula for the eigenvalues:

$$\int_0^1 \frac{gH}{(c-\tilde{u})^2}\mathrm{d}\lambda = 1. \tag{1.24}$$

However this has been done without any precision on the regularity of the system variables. This stems from the fact that for regular velocity profiles, in particular strictly monotone velocity profiles [30], the discrete spectrum coincides with the point spectrum and the eigenvalues c of the system (1.23) are given by relation (1.24). The case of non-monotone velocity profiles has been considered in [89]. It is important to recall that the variables are considered as functions of λ and therefore *monotonicity* is associated to the vertical variable only. The intuition behind the importance of this property comes from the following observation.

Observation: Consider the hydrostatic Euler equations in an irrotational flow such that the vertical variation of the horizontal velocity is neglected. This corresponds exactly to the *Saint-Venant* equations (1.1). These equations are known to be hyperbolic and therefore no additional hyperbolicity conditions will be imposed. In fact, the transformation introduced in Section 1.3.1 no longer serves its purpose and therefore is unnecessary. However, it raises the question for rotational flows associated with a non-zero vertical derivative of the horizontal velocity which divides the flow into classes of monotone and non-monotone velocity profiles [30,89].

According to [30,32,64,88], there exist two real distinct eigenvalues, solutions of (1.24) outside the range of values of the velocity. This was done using the limiting values of the expression (1.24) in the upper and lower complex half-planes. We prove in Proposition 5 that this can only be true under a certain Hölder regularity of the velocity profile and if violated it will result in the loss of one or both real eigenvalues. A simple example is given by Fig. 1.5.

Proposition 5. *Assuming that $\tilde{u} \in C^{0,1/2}(I)$, $H \in C^0(I)$ and $H > 0$, there exists exactly two real eigenvalues $c_\pm \in J_\pm$ solutions of (1.24), where the intervals $J_\pm$ are defined as follows:*

$$J_- = \left[\tilde{u}_- - \sqrt{gh}, \tilde{u}_-\right], \qquad J_+ = \left[\tilde{u}_+, \tilde{u}_+ + \sqrt{gh}\right],$$

with

$$\tilde{u}_- = \inf_I \tilde{u}, \qquad \tilde{u}_+ = \sup_I \tilde{u}, \qquad h = \int_0^1 H\mathrm{d}\lambda.$$

One can easily show that the search for real eigenvalues can be reduced to the intervals $J_\pm$ (the proof can be found for Proposition 14 of Chapter 3). Let $c \in J_\pm$ be a real solution of (1.24) and let $F(c) := g\displaystyle\int_0^1 \frac{Hd\lambda}{(c-\tilde{u})^2}$. We will show that the regularity of $\tilde{u}$ plays an important role in determining

the exact number of the real eigenvalues present in $J_{\pm}$. If $\tilde{u}$ is α-Hölder continuous this means that $(c-\tilde{u})^2$ can only be $\alpha/2$-Hölder continuous and hence if $\alpha = 1/2$ then $\dfrac{1}{(c-\tilde{u})}$ can be non-integrable. Indeed for $\alpha = 1/2$, one can show that $F(c) \geq L\log(\delta^2)$ for some positive constant $L > 0$, for all $\delta > 0$ and hence

$$\lim_{c\to\tilde{u}_-, c<\tilde{u}_-} F(c) = \infty \quad \text{and} \quad \lim_{c\to\tilde{u}_+, c>\tilde{u}_+} F(c) = \infty.$$

However, this is not possible for $\alpha = 1/4$ unless the Hölder constant $K > 0$ satisfies the following smallness assumption:

$$K \leq \sqrt{g \inf_I H}.$$

The proof will be detailed later in Chapter 3. In fact one can show that $C^{0,1/2}(I)$ is the critical space, as a (more complicated) smallness assumption on K is required in $C^{0,\alpha}(I)$ for $\frac{1}{4} < \alpha < \frac{1}{2}$.

It is important to note that Proposition 5 guarantees the existence of two real eigenvalues but does not guarantee that complex values do not exist. For this reason, we look for an additional hypothesis which guarantees that all the eigenvalues satisfying (1.24) are real and distinct. If the equation (1.24) admits only real and distinct solutions $c \in \sigma_p(A_0)$, we say that system (1.23) is *"generalized hyperbolic"*. As shown in [30] there are two elements that can guarantee the *generalized hyperbolicity* of the system; the monotonicity of the velocity profile and its convexity. These two hypotheses where assumed on the flow in the domain Ω_t through an equivalent relation to (1.24) in the variable z instead of λ using the change of variables $z = \phi(t,x,\lambda)$:

$$1 = \int_0^1 \frac{gH}{(c-\tilde{u})^2}\mathrm{d}\lambda = \int_{z_b}^{\eta} \frac{g}{(c-u)^2}\mathrm{d}z.$$

We prove that for sufficiently regular solutions with a monotonic velocity profile that satisfies $\partial_\lambda\big(H/\partial_\lambda\tilde{u}\big) \neq 0$ for all $\lambda \in I$, the spectrum $\sigma(A_0)$ is fully real and there exists only two real eigenvalues $c_\pm$ given by Proposition 5. The condition $\partial_\lambda\big(H/\partial_\lambda\tilde{u}\big) \neq 0$ ensures that the velocity profile of the hydrostatic Euler system in the initial domain is convex. This is in accordance with [30, Lemma 1,2,3] where the hypothesis is made on the initial system. We proceed in the analysis to another important and challenging notion for system which is the notion *Riemann invariants*.

The notion of *Riemann invariants* is closely linked to the eigenelements of the system, in fact the Riemann invariants are specific functions of the system variables which are constant along the integral curves of the corresponding eigenvectors (*i.e.* along the trajectories of the corresponding vector fields). A comprehensive view on hyperbolic conservation laws is given in [51]. The classical notions of *characteristic fields* and *Riemann invariants* no longer hold in the generalized case but instead one can define generalized notions which, in analogy to the classical notions, could help us get a step closer into understanding the characteristic behavior of the hydrostatic Euler system. For a regular solution $\tilde{U}$ of (1.23), consider the eigenvalues $c \in \sigma_p(A_0)$. These eigenvalues depend on the time and horizontal space variables t and x respectively, but do not depend on the vertical space variable $\lambda \in I$. Note that the dependence on t and x is given through the variable $\tilde{U}$. The *generalized Riemann invariants* [90] are quantities transported by these velocities for all λ. We shall call a generalized Riemann invariant associated to $c(t,x)$, any function $R(t,x)$ satisfying:

$$\frac{\partial R}{\partial t} + c\,\frac{\partial R}{\partial x} = 0.$$

For a better understanding of this generalized notion, we will use the Saint-Venant system (1.1) as

a reference example.

Example 2 (One-dimensional Saint-Venant system)**.** *Consider the Saint-Venant system (1.1) with a flat topography $z_b = 0$ or equivalently (1.2). For a strictly positive water height $h > 0$, the matrix of the system written in non-conservative variables (h, u), is given by:*

$$A(h,u) := \begin{pmatrix} u & h \\ g & u \end{pmatrix}.$$

The above matrix admits two real distinct eigenvalues $u \pm \sqrt{gh}$ for $h > 0$ which means that the system (1.1) with $z_b = 0$ is hyperbolic. An eigenvector associated to the eigenvalue $c_\pm = u \pm \sqrt{gh}$ is given by $\varphi_\pm = \left(1, \pm\sqrt{gh}/h\right)^T$. Recall from [51, Chapter II, Definition 3.2] that classically the Riemann invariant $r_\pm = r_\pm(h,u)$ is constant along the trajectories of the vector field $\varphi_\pm$, which means that

$$\nabla_{(h,u)} r_\pm \cdot \varphi_\pm = 0.$$

In the classical case of 2×2 systems, the Riemann invariants can also be found by multiplying the equations by an orthogonal basis to the basis of eigenvectors. Due to the particular structure of the 2×2 system, the Riemann invariants $r_\pm$ are not only constant along the trajectories of the vector field $\varphi_\pm$ but also the Riemann invariants $r_\pm$ are transported by the eigenvalues $c_\mp$ respectively:

$$\partial_t r_\pm + c_\mp \partial_x r_\pm = 0.$$

This method, as we will see later, can be generalized to find the Riemann invariants of general 2×2 systems of the form (1.23). Indeed the generalized Riemann invariants will be chosen to be transported by the eigenvalues but it is difficult to say whether these quantities will be preserved along the integral curves of the corresponding eigenvectors as in the classical case. Moreover, a notion of characteristic fields is difficult to construct in the generalized case, this is due to the presence of an integral operator. It introduces an additional difficulty since it is not clear how the system of Riemann invariants can be linked to the original system.

The above analysis is complemented by a vertical discretization of the domain which results in a *multi-layer* model of (1.23).

1.3.4 Multilayer Analysis

E. Audusse in [7] introduced a discretization which gives access to the vertical variation of the horizontal velocity in the Saint-Venant system. This formulation has also been later used in [6,25,46]. We adapt this strategy to the system (1.23) by introducing a piece-wise constant approximation of the velocity profile $\tilde{u}$ in the vertical variable $\lambda \in I$ to obtain an N-layer formulation (the layer α is shown in Fig. 1.6) of the following form:

$$\frac{\partial \tilde{\boldsymbol{\mathcal{U}}}}{\partial t} + A_N(\tilde{\boldsymbol{\mathcal{U}}}) \frac{\partial \tilde{\boldsymbol{\mathcal{U}}}}{\partial x} = \tilde{\boldsymbol{S}}. \tag{1.25}$$

Unlike the continuous system (1.23), the system (1.25) is a classical quasi-linear system with a $2N \times 2N$ matrix $A_N(\tilde{\boldsymbol{\mathcal{U}}})$. The vector $\tilde{\boldsymbol{\mathcal{U}}} = (\boldsymbol{H}_N, \tilde{\boldsymbol{U}}_N)$ corresponds to a vector of $2N$ entries denoting $\mathbb{P}_0-$approximations in λ of H and $\tilde{u}$ respectively:

$$\boldsymbol{H}_N = (H_1, \ldots, H_N)^T, \qquad \tilde{\boldsymbol{U}}_N = (\tilde{u}_1, \ldots, \tilde{u}_N)^T,$$

and $\tilde{S} = (\mathbf{0}_N, -g\partial_x z_b \mathbf{1}_N)$. The derivation of the multilayer model will be detailed in Section 3.5 of Chapter 3. System (1.25) does not include mass-exchange terms unlike the free-surface multilayer shallow water models present in the literature and in contrast to other multilayer schemes, the considered piece-wise constant approximation is exact with zero truncation error. However we do not look for a comparison between the two systems in this manuscript and instead we focus on the system at hand. The objective is to examine the eigenvalues of the multilayer model (1.23).

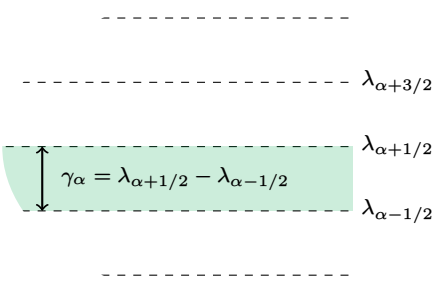

Figure 1.6: The layer α contains the points of coordinates (x, λ) with $\lambda \in \mathcal{L}_\alpha = [\lambda_{\alpha-1/2}, \lambda_{\alpha+1/2}]$ where $0 = \lambda_{1/2} < \lambda_{3/2} < \cdots < \lambda_{N+1} = 1$.

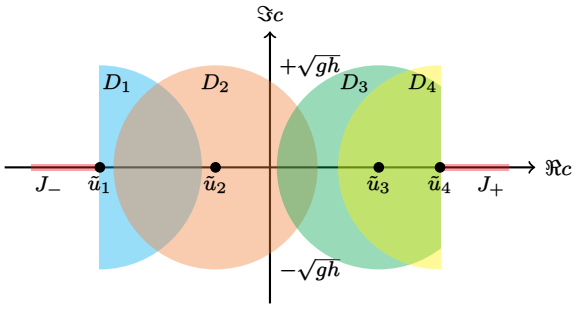

Figure 1.7: Example for $N = 4$: the spectrum $\sigma(A_N)$ is included in the intervals $J_\pm$ and disks D_i. Moreover exactly one eigenvalue $c_\pm$ is in each interval $J_\pm$.

For distinct values $\tilde{u}_i$, we prove that the eigenvalues of system (1.25) satisfy the following relation:

$$\sum_{i=1}^{N} \frac{g\gamma_i H_i}{(\tilde{u}_i - c)^2} = 1. \tag{1.26}$$

Expression (1.26) resembles equation (1.24) for the discrete spectrum. Unlike the case of the continuous system, the eigenvalues of the matrix $A_N(\tilde{\boldsymbol{u}})$ of the discrete system can be determined using classical theory. For an arbitrary number of layers N, we redefine the following notations:

$$\tilde{u}_- = \inf_i \tilde{u}_i, \qquad \tilde{u}_+ = \sup_i \tilde{u}_i.$$

The following proposition resembles Proposition 5 for the continuous system.

Proposition 6. *Let $\boldsymbol{H}_N > 0$, then the intervals*

$$J_- = \left[\tilde{u}_- - \sqrt{gh_N}, \tilde{u}_-\right], \qquad J_+ = \left[\tilde{u}_+, \tilde{u}_+ + \sqrt{gh_N}\right],$$

contain exactly one real eigenvalue each. Moreover, the other eigenvalues are contained in the union of N disks $\{D_i\}_{i\in\{1,\dots,N\}}$ centered at $\tilde{u}_i$ with radius $\sqrt{gh_N}$.

A simple example is given in Fig. 1.7 for $N = 4$. Proposition 6 does not rule out the existence of real eigenvalues within the interval $[\tilde{u}_-, \tilde{u}_+]$ therefore additional assumptions are required. To determine the nature of the eigenvalues c solutions of (1.26), it is helpful to consider a simple example with $N = 2$. By observing Fig. 1.8 and Fig. 1.9, we can expect that the length of the interval $[\tilde{u}_1, \tilde{u}_2]$ influences whether the eigenvalues inside the interval are real or complex, while the two other eigenvalues on the outside are unaffected. We expect that for distinct values of $\tilde{u}_i$, two real eigenvalues c_- and c_+ will exist outside the interval $[\tilde{u}_-, \tilde{u}_+]$ as pointed out in Proposition 6, and the presence of real eigenvalues in $[\tilde{u}_i, \tilde{u}_{i+1}]$ will depend on the smallness of the quantity $|\tilde{u}_i - \tilde{u}_{i+1}|$

or equivalently, the difference $\tilde{u}_+ - \tilde{u}_-$. This turns out to be true in the following sense. If all $\tilde{u}_i$ are distinct, $\boldsymbol{H}_N > 0$, and either $\tilde{u}_+ - \tilde{u}_- < \sqrt{g h_N}$ with $h_N = \sum_{i=1}^{N} \gamma_i H_i$, or $\max_i(|\tilde{u}_i - \tilde{u}_{i+1}|^2) < 8g \min_i(\gamma_i H_i)$, then system (1.25) admits only two real eigenvalues $c_- \in J_-$ and $c_+ \in J_+$ solutions of (1.26) and this rules out the possibility of other real eigenvalues existing inside the interval $[\tilde{u}_-, \tilde{u}_+]$ providing a clear characterization of all possible real eigenvalues of (1.25).

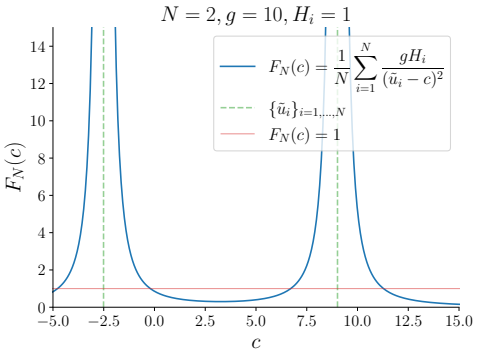

Figure 1.8: $|\tilde{u}_2 - \tilde{u}_1| > \sqrt{\frac{gH_1 + gH_2}{N}}$

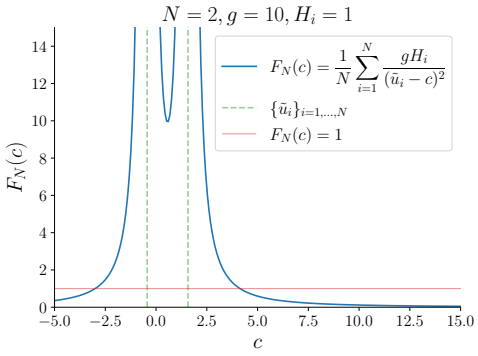

Figure 1.9: $|\tilde{u}_2 - \tilde{u}_1| < \sqrt{\frac{gH_1 + gH_2}{N}}$

This however means that the other eigenvalues might have a non-zero imaginary part. The next step is naturally to look for a link between the discrete system and the continuous system as the number of layers N increases. In other words we are presented with the following question: Does the spectrum of (1.25) converge to the spectrum of (1.23) as $N \to \infty$? This question turns out to be not so trivial and not necessarily true. However, since the monotonicity and convexity of the velocity $\tilde{u}$ in the vertical variable ensures in a way the hyperbolicity of the continuous system in a generalized sense, one can hope to find similar results for the discrete system starting from the same hypothesis. Therefore, assuming that the solutions of system (1.23) are sufficiently regular and that the velocity is strictly monotonic in λ and satisfies $\partial_\lambda \left(H/\partial_\lambda u\right) \neq 0$, we are able to prove that the imaginary part of the remaining eigenvalues tends to zero when N tends to infinity resulting in all real eigenvalues. Therefore the multi-layer formulation can be practical for the numerical approximation of the model at hand since the findings are consistent with the continuous model as the number of layers increases. There are several advantages and disadvantages to this technique over the existing methods that are used in the approximation of hydrostatic models (see [3]). It is important to note that this technique contains derivatives only along the horizontal direction as shown in (1.25), and no exchange terms between the layers appear in the model. Moreover, the estimates on the eigenvalues can eventually provide stability conditions for the time discretization. On the other hand, the downside is that the change of variables can become singular and therefore difficult to invert which will require reinterpolating the variables each time it becomes singular.

The well-posedness of the free-surface hydrostatic Euler equations is a challenging research area and up to our knowledge it remains an open problem. The wall-bounded case has been studied in [23, 74] but the extension to the free-surface case could potentially be invalid. This project could be further developed to encompass other aspects. A comparison between the multi-layer shallow water model [7] and system (1.25) could be interesting especially since the multilayer model (1.25) does not involve exchange terms on contrary to the multilayer shallow water model in [7]. Moreover, based on the assumptions made to establish these properties and supported by numerical experiments, it is evident that certain physical velocity profiles may lead to the presence of complex

eigenvalues. Exploring and analyzing these scenarios could lead to a better understanding of the model.

1.4 Boussinesq System

Despite being able to accurately describe the behavior of fluids in shallow water and accounting for vertical velocity variations under hydrostatic pressure, the Saint-Venant equations and the hydrostatic Euler equations do not account for wave dispersion which is present in real-life situations. The weakly non-linear Boussinesq equations [14, 15] are a generalization of the Saint-Venant equations which take into account the dispersive effects by incorporating higher order terms allowing a more accurate modeling of near-shore and coastal regions. These higher order terms in the Boussinesq system appear due to the non-hydrostatic pressure contribution which is neglected in the Saint-Venant equations. Accounting for dispersive effects in a long-wave, small-amplitude regime allows the Boussinesq equations to capture phenomena such as wave breaking and wave interaction in coastal regions. These equations were first derived by *Boussinesq* in the late 19^{th} century and later on extended to include a whole-class of Boussinesq-type equations. In this part of the manuscript we will consider two notable methods in order to rigorously justify the formally derived approximate mechanical balance laws satisfied by the Boussinesq equations. The first method allows the derivation of the Boussinesq equations by a series of asymptotic expansions under the shallow-water assumption while the latter provides a more systematic approach using a series of perturbations on the shallow-water parameter. We present a common ground between the two methods by carefully regrouping the terms of similar order and bounding the error terms by known expressions. Consequently, this allows us to calculate the error bounds on the approximate balance laws following [57]. This project is in collaboration with *S. Israwi*, *H. Kalisch*, and *D. Mitsotakis*. I will present partial results that will be further improved for publication.

NB. This work started under difficult circumstances in 2019 *in Lebanon; it is worth to mention Covid-*19 *lockdowns, the explosion of the port in Beirut which shut-down the institutions, and lack of financing for my thesis due to the economical crisis. The working conditions were not suitable for collaborative research and furthermore prohibited regular follow-up meetings. This was resolved by moving to France with the help of my supervisors J. Sainte-Marie and B. Di Martino where the two projects (Chapter 2 and Chapter 3) took place and were financially supported by Inria and later on by a full-time ATER (teaching) contract at Sorbonne Université. We look forward to finalize the work in Chapter 4 soon.*

Let us first define the setup of the two-dimensional Euler system.

- The fluid is ideal
- The flow is incompressible
- The flow is irrotational
- The fluid domain Ω_t is bounded from below by the bottom topography and from above by a free-surface
- Boundary conditions are imposed on the free-surface and at the bottom

Note that unlike the case of the Saint-Venant system as well as the hydrostatic Euler system, the pressure in this section is non-hydrostatic. It incorporates terms that account for the variation of the vertical velocity.

1.4.1 Long wave regimes

The Euler equations accurately describe the propagation of free-surface waves in the absence of viscosity. Mass, momentum, and energy are conserved by these equations but their complexity arises both on the mathematical and numerical level which lead to the derivation of simpler models with less accuracy but more applicability. One way to overcome the difficulties imposed by the three-dimensional model is to consider long wave regimes. The long-wave regime is sometimes referred to as the shallow-water regime. Not to be confused with the non-linear shallow-water equations (1.1), otherwise known as the *Saint-Venant* equations. The *Saint-Venant* equations are the simplest depth-averaged model and are obtained at a low precision. The Saint-Venant equations are very useful and widely used to efficiently model wave propagation in shallow water and satisfy exact conservation laws. They do not incorporate dispersive effects, which are typically described by higher order derivatives, we will consider other higher precision models for long-wave approximation. The derivation of such models in the fluid setting described previously, can be done by means of an asymptotic expansion. We will use a finite expansion of the velocity potential, on the long-wave parameter. Recall that the flow is assumed to be irrotational which allows us to view the velocity vector as the gradient of some potential.

The long-wave regimes are divided into two significant types: large amplitude, long wave and small amplitude, long wave. These measurements are relative to the depth of the fluid. Large amplitude, long waves are typically described by the *Serre-Green-Naghdi* equations [55, 83] and the small amplitude, long waves by the *Boussinesq* equations [22]. In the one-dimensional setting for small amplitude long wave regimes, one can also consider the *Korteweg-de Vries* equation which is characterized by its ability to accurately model *solitary waves*.

We will focus on two-dimensional models and mainly on the derivation and justification of approximate mass, momentum, and energy conservation. In fact, for the one-dimensional version of the Serre-Green-Naghdi and Boussinesq equations, appropriate approximations of mass, momentum, and energy balance laws have been derived. These estimates hold up to the same order of accuracy of the equations [1, 61]. We will consider a small amplitude, long wave regime described by two-dimensional Boussinesq-type equations and we will try to extend and justify the derivation of the respective approximate balance laws of mass, momentum, and energy. The Boussinesq-approximation has been extended to include a whole class of models [14, 15, 67]. Different Boussinesq-type models have been derived due to their preferable dispersive properties. To improve linear frequency dispersion, one technique used is substituting the higher order time derivatives on the velocity by high order space derivatives on the free-surface elevation. Introducing other parameters requires another trick. Since the Boussinesq equations can be completely written in terms of the free-surface elevation and the velocity (known as the *Zakharov-Craig-Sulem* formulation), where mainly the averaged velocity is considered, one more degree of freedom can be obtained by changing the choice of the velocity variable considered. This means that instead of writing the system in terms of the averaged velocity, for instance we can choose to write them in terms of the velocity at the bottom or the velocity at another level line in the domain. An asymptotic expansion on the long-wave parameter, followed by the two tricks mentioned above, will allow us to derive a four-parameter family of Boussinesq-type models.

1.4.2 The Zakharov-Craig-Sulem Formulation

Zakharov noted in [94] that given the free-surface elevation and the trace of the velocity potential at the free-surface, one can fully define the flow. The formulation was then given in terms of the Dirichlet-Neumann operator by *Craig* and *Sulem* [34, 35]. This formulation is mostly used to derive

asymptotic models in the long-wave regime. Let $\mathring{\Phi}$ denote the velocity potential of the Euler system described in the above setting (ideal, incompressible, irrotational fluid). The velocities are given as follows:

$$\mathbf{u} = \nabla\mathring{\Phi}, \qquad w = \mathring{\Phi}_z$$

Then the Euler equations with Neumann boundary conditions at the bottom and kinematic boundary conditions at the free-surface can be rewritten in terms of the velocity potential and the free-surface variables. The derivation of asymptotic models relies on appropriate scaling. Let x, y denote the horizontal variables, z denote the vertical variable, t the time, η the free-surface, h_0 the undisturbed water depth, and g the gravitational acceleration. The non-dimensional form of the previous variables are written with an additional tilde, l and A denote a characteristic wavelength and wave amplitude. Let α and β be the measures of non-linearity and frequency dispersion defined as follows

$$\alpha = \frac{A}{h_0}, \qquad \beta = \frac{h_0^2}{l^2} \ . \tag{1.27}$$

Appropriate assumptions on the respective magnitude of the parameters α and β, lead to the derivation of (simpler) asymptotic models from the Euler equations. The Stokes number

$$S = \frac{\alpha}{\beta} \ ,$$

is introduced in order to quantify the applicability of the equation to a particular regime of surface water waves. For the Boussinesq regime, the Stokes number is usually considered to be of order 1. With the above non-dimensionalization, we note that the free surface is given by $\tilde{z} = 1 + \alpha\tilde{\eta}$. It is noted that in the sequel we consider the *Zakharov-Craig-Sulem formulation* of the Euler equations:

$$\begin{cases} \tilde{\eta}_{\tilde{t}} - \dfrac{1}{\beta}\mathcal{G}_\beta[\alpha\tilde{\eta}]\psi = 0 \ , \\ \psi_{\tilde{t}} + \tilde{\eta} + \dfrac{\alpha}{2}|\nabla\psi|^2 - \dfrac{\alpha}{\beta}\dfrac{[\mathcal{G}_\beta[\alpha\tilde{\eta}]\psi + \alpha\beta\nabla\tilde{\eta}\cdot\nabla\psi]^2}{2(1+\alpha^2\beta|\nabla\tilde{\eta}|^2)} = 0 \ . \end{cases} \tag{1.28}$$

Given a solution of this system, we reconstruct the potential $\tilde{\Phi}$ by solving the Laplace equation (1.29) below. More precisely, we introduce the trace of the velocity potential at the free surface, defined as

$$\psi = \tilde{\Phi}_{|_{\tilde{z}=1+\alpha\tilde{\eta}}} \ ,$$

and the Dirichlet-Neumann operator $\mathcal{G}_\beta[\alpha\tilde{\eta}]\cdot$ as

$$\mathcal{G}_\beta[\alpha\tilde{\eta}]\psi = \partial_{\tilde{z}}\tilde{\Phi}_{|_{\tilde{z}=1+\alpha\tilde{\eta}}} \ ,$$

with $\tilde{\Phi}$ solving the boundary value problem

$$\begin{cases} \beta\partial_{\tilde{x}}^2\tilde{\Phi} + \beta\partial_{\tilde{y}}^2\tilde{\Phi} + \partial_{\tilde{z}}^2\tilde{\Phi} = 0 \ , \\ \partial_n\tilde{\Phi}_{|_{\tilde{z}=0}} = 0 \ , \\ \tilde{\Phi}_{|_{\tilde{z}=1+\alpha\tilde{\eta}}} = \psi \ . \end{cases} \tag{1.29}$$

We detail in the next section some parts of the derivation of a four-parameter Boussinesq system as they will be relevant for Section 1.4.4. The derivation consists of using an approximation on the velocity potential, we will adopt the same technique used in [67].

1.4.3 Asymptotic expansion

Consider the following finite expansion of the velocity potential in the shallowness parameter β:

$$\tilde{\Phi}^{app} = \sum_{j=0}^{N} \beta^j \tilde{\Phi}_j. \tag{1.30}$$

The functions $\{\tilde{\Phi}_j\}_{j\in\{1,\dots,N\}}$ can be calculated by substituting (1.30) into (1.29) and canceling the residual up to the order $\mathcal{O}(\beta^{N+1})$:

$$\begin{aligned}
\tilde{\Phi}_0 &= \psi \ , \\
\tilde{\Phi}_1 &= -\frac{1}{2}\tilde{z}^2\Delta\psi + \frac{1}{2}\Delta\psi + \alpha\tilde{\eta}\Delta\psi + \frac{1}{2}\alpha^2\tilde{\eta}^2\Delta\psi \ , \\
\tilde{\Phi}_2 &= \frac{1}{24}\tilde{z}^4\Delta^2\psi - \frac{1}{4}\tilde{z}^2\Delta^2\psi + \frac{5}{24}\Delta^2\psi + \frac{5}{6}\alpha\tilde{\eta}\Delta^2\psi - \frac{1}{2}\alpha\tilde{z}^2\Delta\tilde{\eta}\Delta\psi + \frac{1}{2}\alpha\Delta\tilde{\eta}\Delta\psi \\
&\quad - \alpha\tilde{z}^2\nabla\tilde{\eta}\cdot\nabla\Delta\psi + \alpha\nabla\tilde{\eta}\cdot\nabla\Delta\psi - \frac{1}{2}\alpha\tilde{z}^2\tilde{\eta}\Delta^2\psi + \mathcal{O}(\alpha^2) \ .
\end{aligned}$$

The explicit form of the terms $\{\tilde{\Phi}_j\}_{j\in\{2,\dots,N\}}$ will not be relevant in the Boussinesq regime. Instead of using the averaged velocity or the velocity at the bottom, we will use the velocity at some level line of the domain. Let $\tilde{U},\tilde{V}$ be the dimensionless velocities at a dimensionless height θ $(0\leq\theta\leq 1)$ in the fluid column:

$$\tilde{U} = \tilde{\Phi}^{app}_{\tilde{x}}\big|_{\tilde{z}=\theta} = \psi_{\tilde{x}} - \frac{\beta}{2}(\theta^2-1)\Delta\psi_{\tilde{x}} + \mathcal{O}(\alpha\beta,\beta^2), \tag{1.31}$$

$$\tilde{V} = \tilde{\Phi}^{app}_{\tilde{y}}\big|_{\tilde{z}=\theta} = \psi_{\tilde{y}} - \frac{\beta}{2}(\theta^2-1)\Delta\psi_{\tilde{y}} + \mathcal{O}(\alpha\beta,\beta^2). \tag{1.32}$$

Replacing $\tilde{\Phi}$ by $\tilde{\Phi}^{app}$ means passing from equations on $\tilde{\Phi}$ to equations on ψ, but in order to use the velocity variables $\tilde{U}$ and $\tilde{V}$, one should write ψ in terms of $\tilde{\Phi}^{app}\big|_{z=\theta}$. In fact, using (4.9) one can formally write:

$$\psi = \left(1 - \frac{\beta}{2}(z^2-1)\right)^{-1}\tilde{\Phi}^{app} + \mathcal{O}(\alpha\beta,\beta^2),$$

or equivalently

$$\psi = \tilde{\Phi}^{app}\big|_{\tilde{z}=\theta} + \frac{\beta}{2}\left(\theta^2-1\right)\Delta\,\tilde{\Phi}^{app}\big|_{\tilde{z}=\theta} + \mathcal{O}(\alpha\beta,\beta^2). \tag{1.33}$$

Details are provided in Chapter 4. We consider the following Boussinesq system:

$$\tilde{U}_{\tilde{t}} + \tilde{\eta}_{\tilde{x}} + \frac{\beta}{2}(\theta^2-1)\Delta\tilde{U}_{\tilde{t}} + \alpha\left(\tilde{U}\tilde{U}_{\tilde{x}} + \tilde{V}\tilde{V}_{\tilde{x}}\right) = \mathcal{O}(\alpha\beta,\beta^2) \ , \tag{1.34a}$$

$$\tilde{V}_{\tilde{t}} + \tilde{\eta}_{\tilde{y}} + \frac{\beta}{2}(\theta^2-1)\Delta\tilde{V}_{\tilde{t}} + \alpha\left(\tilde{U}\tilde{U}_{\tilde{y}} + \tilde{V}\tilde{V}_{\tilde{y}}\right) = \mathcal{O}(\alpha\beta,\beta^2) \ , \tag{1.34b}$$

$$\tilde{\eta}_{\tilde{t}} + \tilde{U}_{\tilde{x}} + \tilde{V}_{\tilde{y}} + \frac{\beta}{2}\left[\theta^2 - \frac{1}{3}\right]\left(\Delta\tilde{U}_{\tilde{x}} + \Delta\tilde{V}_{\tilde{y}}\right) + \alpha\left((\tilde{\eta}\tilde{U})_{\tilde{x}} + (\tilde{\eta}\tilde{V})_{\tilde{y}}\right) = \mathcal{O}(\alpha\beta,\beta^2) \ . \tag{1.34c}$$

One can observe that:

$$\tilde{U}_{\tilde{t}} + \tilde{\eta}_{\tilde{x}} = \mathcal{O}(\alpha, \beta) \ , \tag{1.35}$$

$$\tilde{V}_{\tilde{t}} + \tilde{\eta}_{\tilde{y}} = \mathcal{O}(\alpha, \beta) \ , \tag{1.36}$$

$$\tilde{\eta}_{\tilde{t}} + \tilde{U}_{\tilde{x}} + \tilde{V}_{\tilde{y}} = \mathcal{O}(\alpha, \beta) \ , \tag{1.37}$$

This means that one can introduce parameters that compensate a certain weight of the time derivatives of the velocity with space derivatives on the free-surface in the higher order terms of (1.34a)-(1.34c). More precisely for $\mu \in \mathbb{R}$:

$$\begin{aligned}\Delta\tilde{U}_{\tilde{t}} &= \mu\Delta\tilde{U}_{\tilde{t}} + (1-\mu)\Delta\tilde{U}_{\tilde{t}} \\ &= \mu\Delta\tilde{U}_{\tilde{t}} - (1-\mu)\Delta\tilde{\eta}_{\tilde{x}} + \mathcal{O}(\alpha, \beta).\end{aligned}$$

Similar approximations hold for $\tilde{V}$ and $\tilde{\eta}$. The full equations will be given in Chapter 4. This technique allows a certain degree of freedom in the asymptotic expansion for improving the dispersive properties of the Boussinesq equations.

The derivation of the $a-b-c-d$ family of Boussinesq equations of [15] is classically performed using (formally) an infinite asymptotic expansion on the vertical variable $\tilde{z}$ instead of the finite series (1.30). We will see the details of the derivation later in Chapter 4. Since we aim in finding a rigorous derivation of the mechanical balance laws for two-dimensional family of Boussinesq equations, we must first estimate approximations between the velocities $\nabla_{x,z}\tilde{\Phi}^{app}$ and $\nabla_{x,z}\tilde{\Phi}$, where $\nabla_{x,z} = (\nabla, \partial_z)^T$. It is also essential to define the pressure in the Boussinesq regime and justify that it approximates the pressure term in the Euler equations up to the same order as that of the system. Let $\omega = \tilde{\Phi} - \tilde{\Phi}^{app}$, then ω satisfies the following boundary-value problem

$$\begin{cases} \beta\partial_{\tilde{x}}^2\omega + \beta\partial_{\tilde{y}}^2\omega + \partial_{\tilde{z}}^2\omega = r \text{ in } \Omega_t, \\ \omega|_{\tilde{z}=1+\alpha\tilde{\eta}} = 0, \\ \partial_n\omega|_{\tilde{z}=0} = 0, \end{cases} \tag{1.38}$$

where $r = \alpha\beta^2 r_1 + \beta^3 r_2$ is a regular function in terms of $\tilde{z}$ and the derivatives of ψ. In order to find appropriate bounds on ω in the domain Ω_t, we will use the classical so-called Σ-transformation. This transformation maps the time-dependent domain Ω_t onto a flat strip S allowing us to recover L^∞ estimates in the original domain. The following estimates are an adaptation of the work done by *D. Lannes* in [67, Chapter 2].

1.4.4 Important estimates

Let $(\eta^{Euler}, \tilde{\Phi})$ be a regular solution of the Euler system such that $(\eta^{Euler}, \nabla\psi) \in H^s(\mathbb{R}^2) \times H^s(\mathbb{R}^2)$ with s large enough. Assume that the total water depth satisfies:

$$\exists h_{min} > 0, \qquad \forall X = (x, y) \in \mathbb{R}^2, \qquad 1 + \alpha\tilde{\eta} \geq h_{min}. \tag{1.39}$$

Then, for $0 < \tilde{t} < T/\beta$ we have,

$$|\nabla_{x,z}\tilde{\Phi}^{app} - \nabla_{x,z}\tilde{\Phi}|_{L^\infty(\Omega_t)} \lesssim \alpha\beta + \beta^2 \ .$$

This estimate also holds for higher accuracy $\mathcal{O}(\alpha^i\beta^j, \alpha^j\beta^i)$ with $i, j \geq 1$ which might be useful for higher-order models, but as far as we are concerned, $\mathcal{O}(\alpha\beta, \beta^2)$ is sufficient in the Boussinesq regime.

The momentum balance requires an additional approximation of the pressure term which is given by the Bernoulli equation:

$$\tilde{P}' = -\tilde{\Phi}_{\tilde{t}} - \frac{1}{2}\alpha(\tilde{\Phi}_{\tilde{x}}^2 + \tilde{\Phi}_{\tilde{y}}^2) - \frac{1}{2}\frac{\alpha}{\beta}(\tilde{\Phi}_{\tilde{z}}^2).$$

We define the approximation of the pressure by replacing $\tilde{\Phi}$ by $\tilde{\Phi}^{app}$, however an expression of $\tilde{\Phi}_{\tilde{t}}^{app}$ in terms of the velocities $\tilde{U}$ and $\tilde{V}$ is needed. In the Boussinesq regime the approximate potential satisfies:

$$\tilde{\Phi}_{\tilde{t}}^{app} + \frac{\alpha}{2}\left(\tilde{\Phi}_{\tilde{x}}^{app,2} + \tilde{\Phi}_{\tilde{y}}^{app,2} + \frac{1}{\beta}\tilde{\Phi}_{\tilde{z}}^{app,2}\right) + \tilde{\eta} = \mathcal{O}(\alpha\beta, \beta^2),$$

hence using the following identity:

$$-\frac{\alpha\beta}{2}\Delta\tilde{\Phi}_{\tilde{t}}^{app} = -\frac{\alpha\beta}{2}\partial_{\tilde{t}}\partial_{\tilde{x}}\tilde{\Phi}_{\tilde{x}}^{app} + \partial_{\tilde{t}}\partial_{\tilde{y}}\tilde{\Phi}_{\tilde{y}}^{app},$$

one can obtain an expression of the approximate pressure term:

$$\tilde{Q} = \tilde{\eta} + \frac{1}{2}\beta(\tilde{z}^2 - 1)[\tilde{U}_{\tilde{x}\tilde{t}} + \tilde{V}_{\tilde{y}\tilde{t}}].$$

In order to justify the approximation of the pressure given by $\tilde{Q}$, we begin with a useful result derived from the Zakharov-Craig-Sulem equations, cf. [66]. Let $\tilde{\eta} \in H^{s+1/2}(\mathbb{R}^2) \cap H^{t_0+2}(\mathbb{R}^2)$ with $s \geq 0$, $t_0 > 1$ satisfying (1.39). Then, the following mappings are continuous:

$$\begin{aligned} \mathcal{G}_\beta[\alpha\tilde{\eta}] \colon \dot{H}^{s+1}(\mathbb{R}^2) &\to H^{s-1/2}(\mathbb{R}^2) \\ \psi &\mapsto \mathcal{G}_\beta[\alpha\tilde{\eta}]\psi \end{aligned} \tag{1.40}$$

$$\begin{aligned} \nu[\beta\tilde{\eta}] \colon \dot{H}^{s+1/2}(\mathbb{R}^2) &\to H^{s-1/2}(\mathbb{R}^2) \\ \psi &\mapsto \frac{[\mathcal{G}_\beta[\alpha\tilde{\eta}]\psi + \alpha\beta\nabla\tilde{\eta}\cdot\nabla\psi]^2}{2(1+\alpha^2\beta|\nabla\tilde{\eta}|^2)} \end{aligned} \tag{1.41}$$

Then, for $0 < \tilde{t} < T/\beta$ we have,

$$|\tilde{\Phi}_{\tilde{t}}^{app} - \tilde{\Phi}_{\tilde{t}}|_{L^\infty(\Omega_t)} \lesssim \alpha\beta + \beta^2\,,$$

Due to the approximation involved in deriving the Boussinesq system, the mechanical balance laws are no longer exact in the Boussinesq regime. Instead, we look for approximate balance laws that hold up to the same order as that of the derived system.

1.4.5 Mass, momentum, and energy balance

The incompressibility condition combined with the kinematic boundary conditions at the free-surface and at the bottom leads to the following non-dimensional mass balance satisfied by the Euler equations:

$$\frac{\partial}{\partial\tilde{t}}\,(1+\alpha\tilde{\eta}) + \frac{\partial}{\partial\tilde{x}}\int_{\tilde{z}=0}^{1+\alpha\tilde{\eta}} \alpha\tilde{\Phi}_{\tilde{x}}\,d\tilde{z} + \frac{\partial}{\partial\tilde{y}}\int_{\tilde{z}=0}^{1+\alpha\tilde{\eta}} \alpha\tilde{\Phi}_{\tilde{y}}\,d\tilde{z} = 0\ . \tag{1.42}$$

The approximate mass and approximate mass flux will be defined such that equation (1.42) holds up to order $\mathcal{O}(\alpha\beta^2, \alpha^2\beta)$ or equivalently $\mathcal{O}(\beta^3)$. This requires an explicit expression of the integrals involved while replacing $\tilde{\Phi}$ by $\tilde{\Phi}^{app}$. Note that the mass is given by $\tilde{M} = 1 + \alpha\tilde{\eta}$. We demonstrate

the choice of the mass flux using the first integral:

$$\begin{aligned}\frac{\partial}{\partial\tilde{x}}\int_{\tilde{z}=0}^{1+\alpha\tilde{\eta}}\alpha\tilde{\Phi}_{\tilde{x}}^{app}\,d\tilde{z} &= \frac{\partial}{\partial\tilde{x}}\int_{\tilde{z}=0}^{1+\alpha\tilde{\eta}}\left\{\alpha\tilde{\Phi}_{0,\tilde{x}}+\alpha\beta\tilde{\Phi}_{1,\tilde{x}}\right\}\,d\tilde{z}+\mathcal{O}(\beta^3)\\ &= \frac{\partial}{\partial\tilde{x}}\int_{\tilde{z}=0}^{1+\alpha\tilde{\eta}}\left\{\alpha\psi_{\tilde{x}}-\frac{\alpha\beta}{2}\tilde{z}^2\Delta\psi_{\tilde{x}}+\frac{\alpha\beta}{2}\Delta\psi_{\tilde{x}}\right\}\,d\tilde{z}+\mathcal{O}(\alpha\beta^2)\\ &= \frac{\partial}{\partial\tilde{x}}\left\{\alpha(1+\alpha\tilde{\eta})\psi_{\tilde{x}}-\frac{\alpha\beta}{6}\Delta\psi_{\tilde{x}}+\frac{\alpha\beta}{2}\Delta\psi_{\tilde{x}}\right\}+\mathcal{O}(\alpha\beta^2).\end{aligned}$$

We can then find the approximate mass flux by replacing $\psi_{\tilde{x}}$ by its expression in terms of $\tilde{\Phi}_{\tilde{x}}^{app}$ at the height $0\leq\theta\leq 1$ given by (1.33)

$$\psi_{\tilde{x}}=\tilde{\Phi}_{\tilde{x}}^{app}+\frac{\beta}{2}(\theta^2-1)\Delta\tilde{\Phi}_{\tilde{x}}^{app}+\mathcal{O}(\beta^3),$$

where $\tilde{U}=\tilde{\Phi}_{\tilde{x}}^{app}$ is the approximate velocity taken as the velocity in the Boussinesq approximation. Therefore we define the mass flux as:

$$\tilde{q}_{m_x}=\alpha(1+\alpha\tilde{\eta})\tilde{U}+\frac{\alpha\beta}{2}(\theta^2-1/3)\Delta\tilde{U}.$$

A similar definition holds for $\tilde{q}_{m_y}$. Finally we can bound all the error terms as follows. Let $(\eta^{Euler},\tilde{\Phi})$ be a regular solution of the Euler system such that $(\eta^{Euler},\nabla\psi)\in H^s(\mathbb{R}^2)\times H^s(\mathbb{R}^2)$ with s large enough. Then, there exists a constant C independent of β such that:

$$\left|\frac{\partial}{\partial\tilde{t}}\tilde{M}+\frac{\partial}{\partial\tilde{x}}q_{m_x}+\frac{\partial}{\partial\tilde{y}}q_{m_y}\right|_{L^\infty(\mathbb{R}^2)}\leq C(\alpha^2\beta+\alpha\beta^2).$$

The momentum balance reads as follows:

$$\begin{aligned}&\alpha\frac{\partial}{\partial\tilde{t}}\int_{\tilde{z}=0}^{1+\alpha\tilde{\eta}}\tilde{\Phi}_{\tilde{x}}\,d\tilde{z}+\frac{\partial}{\partial\tilde{x}}\int_{\tilde{z}=0}^{1+\alpha\tilde{\eta}}\left\{\alpha^2(\tilde{\Phi}_{\tilde{x}}^2)+\alpha\tilde{P}'-(\tilde{z}-1)\right\}\tilde{z}+\alpha^2\frac{\partial}{\partial\tilde{y}}\int_{\tilde{z}=0}^{1+\alpha\tilde{\eta}}\tilde{\Phi}_{\tilde{x}}\tilde{\Phi}_{\tilde{y}}\,d\tilde{z}=0\ ,\\ &\alpha\frac{\partial}{\partial\tilde{t}}\int_{\tilde{z}=0}^{1+\alpha\tilde{\eta}}\tilde{\Phi}_{\tilde{y}}\,d\tilde{z}+\frac{\partial}{\partial\tilde{y}}\int_{\tilde{z}=0}^{1+\alpha\tilde{\eta}}\left\{\alpha^2(\tilde{\Phi}_{\tilde{y}}^2)+\alpha\tilde{P}'-(\tilde{z}-1)\right\}\tilde{z}+\alpha^2\frac{\partial}{\partial\tilde{x}}\int_{\tilde{z}=0}^{1+\alpha\tilde{\eta}}\tilde{\Phi}_{\tilde{x}}\tilde{\Phi}_{\tilde{y}}\,d\tilde{z}=0\ .\end{aligned}$$

Note that as mentioned previously in Section 1.4.4, the pressure estimate plays a role in the justification of the approximate momentum balance. Hence, we are able to replace $\tilde{\Phi}$ by $\tilde{\Phi}^{app}$ and integrate in a similar way to obtain the following estimates:

$$\begin{aligned}&\left|\frac{\partial}{\partial\tilde{t}}\left\{(1+\alpha\tilde{\eta})\tilde{U}+\frac{\beta}{2}\left(\theta^2-\frac{1}{3}\right)\Delta\tilde{U}\right\}\right.\\ &\qquad\left.+\frac{\partial}{\partial\tilde{x}}\left\{\tilde{\eta}+\alpha\tilde{U}^2+\frac{\alpha}{2}\tilde{\eta}^2-\frac{1}{3}\beta(\tilde{U}_{\tilde{x}\tilde{t}}+\tilde{V}_{\tilde{y}\tilde{t}})+\frac{1}{2}\right\}+\frac{\partial}{\partial\tilde{y}}(\alpha\tilde{U}\tilde{V})\right|_{L^\infty(\mathbb{R}^2)}\leq C_2(\alpha\beta+\beta^2),\end{aligned}$$

$$\begin{aligned}&\left|\frac{\partial}{\partial\tilde{t}}\left\{(1+\alpha\tilde{\eta})\tilde{V}+\frac{\beta}{2}\left(\theta^2-\frac{1}{3}\right)\Delta\tilde{V}\right\}\right.\\ &\qquad\left.+\frac{\partial}{\partial\tilde{y}}\left\{\tilde{\eta}+\alpha\tilde{V}^2+\frac{\alpha}{2}\tilde{\eta}^2-\frac{1}{3}\beta(\tilde{U}_{\tilde{x}\tilde{t}}+\tilde{V}_{\tilde{y}\tilde{t}})+\frac{1}{2}\right\}+\frac{\partial}{\partial\tilde{x}}(\alpha\tilde{U}\tilde{V})\right|_{L^\infty(\mathbb{R}^2)}\leq C_3(\alpha\beta+\beta^2).\end{aligned}$$

A similar estimate holds for the energy. The details can be found in Theorem 10, Theorem 11, and Theorem 12 of Chapter 4.

To my knowledge, these approximation results are new, however, as already mentioned, this work is not complete. It would be important to study more closely the existing literature and compare with related studies. Also, it would be interesting to develop some appropriate numerical approximation to test the model on practical real-life situations.

Chapter 2

Implicit kinetic schemes for the Saint-Venant system

Outline of the current chapter

This chapter is written in collaboration with *Jacques Sainte-Marie* and *Mathieu Rigal*. It can be found as a pre-print [44] on *HAL* (hal-04048832).

Explicit kinetic schemes applied to the nonlinear shallow water equations have been extensively studied in the past. The novelty of this paper is to investigate an implicit version of such methods in order to improve their stability properties. In the case of a flat bathymetry we obtain a fully implicit kinetic solver satisfying a discrete entropy inequality and keeping the water height non negative

without any restriction on the time step. Remarkably, a simplified version of this nonlinear implicit scheme allows to express the update explicitly which we implement in practice. An extension to the two-dimensional case is also discussed. The case of varying bottoms is then dealt with through an iterative solver combined with the hydrostatic reconstruction technique. We show that this scheme preserves the water height positivity under a CFL condition and satisfies a discrete entropy inequality without error term, which is an improvement over its explicit version. Finally we perform some numerical validations underlining the advantages and the computational cost of our strategy.

2.1 Introduction

Mathematical models for free surface flows are widely studied but their analysis and numerical approximation remain a challenging issue. The incompressible Navier-Stokes system with free surface being very difficult to study, it is often replaced by the classical Saint-Venant system [10, 48] that is a hyperbolic system of conservation laws approximating various geophysical flows, such as rivers, coastal areas, and oceans when completed with a Coriolis term, and granular flows when completed with friction terms.

The derivation of an efficient, robust and stable numerical scheme for the Saint-Venant system has received an extensive coverage, we refer the reader to [20,52,60,93] and references therein. One of the challenges involves the construction of a well-balanced scheme i.e. preserving some characteristic stationary solutions. In a recent work, some of the authors have proposed a numerical scheme for the Saint-Venant system with topography (2.1) satisfying a fully discrete entropy inequality [5]. The proposed numerical scheme is based on a kinetic solver [9, 17, 18, 27, 54, 78] coupled with the hydrostatic reconstruction technique introduced in [8] for the numerical treatment of the topography source term. Based on the results obtained in [5], Bouchut and Lhebrard have proved the convergence of the the kinetic hydrostatic reconstruction scheme for the Saint-Venant system (2.1), see [21].

Finite volume approaches for the approximation of conservation laws have to deal with a CFL constraint that can be very restrictive for some applications where large time scales and significant wave velocities have to be considered. This is for instance the case in the low Froude regime, where the surface gravity waves travel at a much larger velocity than the fluid particles. Moreover, the explicit in time discretization induces a non-negative term in the discrete energy balance that cannot be always controlled by the dissipation coming from the upwinding of the numerical fluxes, see [5]. The novelties of this paper are:

- to propose a fully implicit kinetic scheme in 1d and 2d for the Saint-Venant system satisfying a fully discrete entropy inequality without any restriction on the time step,
- to evaluate the practical interest of implicit schemes for the Saint-Venant system in the sense that the CFL constraint for explicit schemes is replaced in the context of implicit schemes by the computational costs due to the computation of the numerical fluxes. Indeed in the explicit setting, the numerical fluxes at an interface depend on the value of the variables at the two neighbouring vertices whereas in the implicit context, the numerical fluxes depend on the value of the variables at all vertices (the stencil encompasses the whole computational domain).

Notice that an implicit scheme often requires to invert an operator – here a matrix – at each time step. However, the kinetic scheme gives a very favorable context in which we have an explicit expression of the inverse of the operator. Hence, one can hardly imagine a truly implicit scheme for the Saint-Venant system with a lower computational cost than a kinetic solver.

The aim of this paper is to propose an implicit – in time – version of the kinetic scheme given in [5] and to study its properties. More precisely, we prove some stability properties and most importantly, we derive a fully discrete entropy inequality without any error term.

This paper is organized as follows. First, we recall the formulation of the Saint-Venant system, its kinetic description and the framework of its numerical approximation in the context of a kinetic solver. Then, the implicit kinetic scheme for the Saint-Venant system without topography is proposed and studied in 1d and in the 2d case. An iterative version of the implicit scheme is proposed in Section 2.5 where the topography can be taken into account through the hydrostatic reconstruction technique [9]. Finally, numerical examples are given to evaluate the interest of our approach.

2.2 The Saint-Venant system and its kinetic interpretation

2.2.1 The Saint-Venant system

The classical Saint Venant system for shallow water describes the height of water $h(t,x) \geq 0$, and the water velocity $u(t,x) \in \mathbb{R}$ (x denotes a coordinate in the horizontal direction) in the direction parallel to the bottom. It assumes a slowly varying topography $z_b(x)$, and reads

$$\begin{aligned} &\partial_t h + \partial_x(hu) = 0, \\ &\partial_t(hu) + \partial_x(hu^2 + g\frac{h^2}{2}) + gh\partial_x z_b = 0, \end{aligned} \tag{2.1}$$

where $g > 0$ is the gravity constant. This system is completed with an entropy (energy) inequality

$$\partial_t\left(h\frac{u^2}{2} + g\frac{h^2}{2} + ghz_b\right) + \partial_x\left((h\frac{u^2}{2} + gh^2 + ghz_b)u\right) \leq 0.$$

We shall denote $U = (h, hu)^T$ and

$$\eta(U) = h\frac{u^2}{2} + g\frac{h^2}{2}, \qquad G(U) = (h\frac{u^2}{2} + gh^2)u,$$

the entropy and entropy fluxes without topography.

2.2.2 Kinetic interpretation of the Saint-Venant system

The reader can refer to [5] and references therein for a complete presentation of the description of the Saint-Venant system.

The classical kinetic Maxwellian (see e.g. [78]) is given by

$$M(U,\xi) = \frac{1}{g\pi}\Big(2gh - (\xi - u)^2\Big)_+^{1/2}, \tag{2.2}$$

where $\xi \in \mathbb{R}$ and $x_+ \equiv \max(0,x)$ for any $x \in \mathbb{R}$. It satisfies the following moment relations,

$$\begin{aligned} \int_{\mathbb{R}} M(U,\xi)\,d\xi = h, \qquad \int_{\mathbb{R}} \xi M(U,\xi)\,d\xi = hu, \\ \int_{\mathbb{R}} \xi^2 M(U,\xi)\,d\xi = hu^2 + g\frac{h^2}{2}. \end{aligned} \tag{2.3}$$

These definitions allow us to obtain a *kinetic representation* of the Saint-Venant system.

Lemma 2. *If the topography $z_b(x)$ is Lipschitz continuous, the pair of functions (h, hu) is a weak*

solution to the Saint-Venant system (2.1) if and only if $M(U,\xi)$ satisfies the kinetic equation

$$\partial_t M + \xi \partial_x M - g(\partial_x z_b)\partial_\xi M = Q, \tag{2.4}$$

for some "collision term" $Q(t,x,\xi)$ that satisfies, for a.e. (t,x),

$$\int_{\mathbb{R}} Q d\xi = \int_{\mathbb{R}} \xi Q d\xi = 0. \tag{2.5}$$

Proof. If (2.4) and (2.5) are satisfied, we can multiply (2.4) by $(1,\xi)^T$, and integrate with respect to ξ. Using (2.3) and (2.5) and integrating by parts the term in $\partial_\xi M$, we obtain (2.1). Conversely, if (h,hu) is a weak solution to (2.1), just define Q by (2.4); it will satisfy (2.5) according to the same computations. □

The standard way to use Lemma 2 is to write a kinetic relaxation equation [17,18,33,76,77], like

$$\partial_t f + \xi \partial_x f - g(\partial_x z_b)\partial_\xi f = \frac{M-f}{\epsilon}, \tag{2.6}$$

where the distribution $f(t,x,\xi)$ is positive, where $M = M(U,\xi)$ with $U(t,x) = \int (1,\xi)^T f(t,x,\xi) d\xi$, and where $\epsilon > 0$ is a relaxation time. In the limit $\epsilon \to 0$ we recover formally the formulation (2.4), (2.5). We refer to [17] for general considerations on such kinetic relaxation models without topography, the case with topography being introduced in [78]. Note that the notion of *kinetic representation* as (2.4), (2.5) differs from the so called *kinetic formulations* where a large set of entropies is involved, see [80]. For systems of conservation laws, these kinetic formulations include non-advective terms that prevent from writing down simple approximations. In general, kinetic relaxation approximations can be compatible with just a single entropy. Nevertheless this is enough for proving the convergence as $\epsilon \to 0$, see [12].

The interest of the particular form (2.2) lies in its link with a kinetic entropy. Consider the kinetic entropy,

$$H(f,\xi,z_b) = \frac{\xi^2}{2} f + \frac{g^2\pi^2}{6} f^3 + g z_b f,$$

where $f \geq 0$, $\xi \in \mathbb{R}$ and $z_b \in \mathbb{R}$, and its version without topography

$$H_0(f,\xi) = \frac{\xi^2}{2} f + \frac{g^2\pi^2}{6} f^3.$$

Then one can check the relations

$$\int_{\mathbb{R}} H\big(M(U,\xi),\xi,z_b\big)\, d\xi = \eta(U) + ghz_b,$$

$$\int_{\mathbb{R}} \xi H\big(M(U,\xi),\xi,z_b\big)\, d\xi = G(U) + ghz_b u.$$

One has the following entropy relations.

Lemma 3. *Let $f(\xi) \geq 0$ satisfy $\int f(\xi)d\xi = h$ and $\int \xi f(\xi) d\xi = hu$ (assumed finite). The half-disk Maxwellian (2.2) satisfies the three properties below.*

(i) For any $\xi \in \mathbb{R}$ one has

$$H_0(f,\xi) \geq H_0(M(U,\xi),\xi) + \eta'(U)\cdot\begin{pmatrix}1\\ \xi\end{pmatrix}(f - M(U,\xi)). \tag{2.7}$$

(ii) For $U = (h,hu)^T$ there holds the inequality

$$\eta(U) = \int_{\mathbb{R}} H_0\big(M(U,\xi),\xi\big)\,d\xi \leq \int_{\mathbb{R}} H_0\big(f(\xi),\xi\big)\,d\xi. \tag{2.8}$$

(iii) The kinetic entropy equality

$$\partial_t H(M,\xi,z_b) + \partial_x\big(\xi H(M,\xi,z_b)\big) - g(\partial_x z_b)\partial_\xi H(M,\xi,z_b) = \partial_f H(M,\xi,z_b)Q,$$

holds, leading to the macroscopic inequality

$$\partial_t \int_{\mathbb{R}} H(M,\xi,z_b)d\xi + \partial_x \int_{\mathbb{R}} \xi H(M,\xi,z_b)d\xi \leq 0.$$

Proof of Lemma 3. The property *(ii)* was proved by Perthame and Simeoni in [78]. It is simply recovered from *(i)* by integrating (2.7) over $\xi \in \mathbb{R}$ and using the fact that $f - M(U,\xi)$ is a collision term. A proof of inequality (2.7) is given in [5] and relies on the following relation

$$\partial_f H_0(M(U,\xi),\xi) = \begin{cases} \eta'(U)\cdot\begin{pmatrix}1\\ \xi\end{pmatrix} & \text{if } \xi \in \operatorname{supp} M(U,\cdot) \\ \geq \eta'(U)\cdot\begin{pmatrix}1\\ \xi\end{pmatrix} & \text{otherwise} \end{cases}, \tag{2.9}$$

which we will use in Section 2.5 to obtain discrete entropy inequalities. For proving *(iii)*, we multiply (2.4) by $\partial_f H(M,\xi,z_b)$ and an integration in ξ of the obtained equation gives the result. □

2.2.3 Kinetic scheme for the Saint-Venant system

For numerical purposes it is usual to replace the right-hand side in the kinetic relaxation equation (2.6) by a time discrete projection to the Maxwellian state. When space discretization is present it leads to flux-vector splitting schemes, see [18] for the case without topography, [78] for the case with topography, and [9] for the 2d case on unstructured meshes.

We would like to approximate the solution $U(t,x)$, $x \in \mathbb{R}$, $t \geq 0$ of the system (2.1) by discrete values U_i^n, $i \in \mathbb{Z}$, $n \in \mathbb{N}$. In order to do so, we consider a grid of points $x_{i+1/2}$, $i \in \mathbb{Z}$,

$$\ldots < x_{i-1/2} < x_{i+1/2} < x_{i+3/2} < \ldots,$$

and we define the cells (or finite volumes) and their lengths

$$C_i =]x_{i-1/2}, x_{i+1/2}[, \qquad \Delta x_i = x_{i+1/2} - x_{i-1/2}.$$

We consider discrete times t^n with $t^{n+1} = t^n + \Delta t^n$, and we define the piecewise constant functions $U^n(x)$ corresponding to time t^n and $z_b(x)$ as

$$U^n(x) = U_i^n, \quad z_b(x) = z_i, \quad \text{for } x_{i-1/2} < x < x_{i+1/2}.$$

A finite volume scheme for solving (2.1) is a formula of the form

$$U_i^{n+1} = U_i^n - \sigma_i(F_{i+1/2-} - F_{i-1/2+}), \tag{2.10}$$

where $\sigma_i = \Delta t^n/\Delta x_i$, telling how to compute the values U_i^{n+1} knowing U_i^n and discretized values z_i of the topography. Here we consider first-order three points schemes where

$$F_{i+1/2-} = \mathcal{F}_l(U_i^{n+p}, U_{i+1}^{n+p}, z_{i+1} - z_i), \qquad F_{i+1/2+} = \mathcal{F}_r(U_i^{n+p}, U_{i+1}^{n+p}, z_{i+1} - z_i), \tag{2.11}$$

with $p = 0, 1$. The value $p = 0$ classically corresponds to a first order explicit time scheme for solving (2.1) whereas $p = 1$ means an implicit time scheme. In this paper, we focus on the case $p = 1$. The functions $\mathcal{F}_{l/r}(U_l, U_r, \Delta z) \in \mathbb{R}^2$ are the numerical fluxes, see [20].

Indeed the method used in [78] in order to solve (2.1) can be viewed as solving

$$\partial_t f + \xi \partial_x f - g(\partial_x z_b)\partial_\xi f = 0 \tag{2.12}$$

for the unknown $f(t, x, \xi)$, over the time interval (t^n, t^{n+1}), with initial data

$$f(t^n, x, \xi) = M(U^n(x), \xi). \tag{2.13}$$

Defining the update as

$$U_i^{n+1} = \frac{1}{\Delta x_i} \int_{x_{i-1/2}}^{x_{i+1/2}} \int_{\mathbb{R}} \begin{pmatrix} 1 \\ \xi \end{pmatrix} f(t^{n+1-}, x, \xi)\, dx d\xi, \tag{2.14}$$

and

$$f_i^{n+1}(\xi) = \frac{1}{\Delta x_i} \int_{x_{i-1/2}}^{x_{i+1/2}} f(t^{n+1-}, x, \xi)\, dx, \tag{2.15}$$

the formula (2.14) can then be written

$$U_i^{n+1} = \int_{\mathbb{R}} \begin{pmatrix} 1 \\ \xi \end{pmatrix} f_i^{n+1}(\xi)\, d\xi. \tag{2.16}$$

This formula can in fact be written under the form (2.10), (2.11) for some numerical fluxes $\mathcal{F}_{l/r}$.

2.3 An implicit kinetic scheme

In this section we consider the problem (2.1) without topography, and the kinetic scheme (2.12)-(2.16). First we present and study the discrete implicit kinetic then we detail the macroscopic scheme obtained from the kinetic discretization.

2.3.1 Implicit scheme without topography

Without topography, the kinetic scheme is a *flux vector splitting* scheme [18]. The update (2.15) of the solution of (2.12),(2.13) simplifies to the discrete kinetic scheme

$$f_i^{n+1-} = M_i - \sigma\xi \left(\mathbb{1}_{\xi<0} f_{i+1}^{n+1-} + \mathbb{1}_{\xi>0} f_i^{n+1-} - \mathbb{1}_{\xi<0} f_i^{n+1-} - \mathbb{1}_{\xi>0} f_{i-1}^{n+1-} \right), \tag{2.17}$$

with $\sigma = \Delta t^n/\Delta x$. We rewrite the previous equations under the form

$$\left\{\begin{array}{l} -\sigma \mathbb{1}_{\xi>0}\xi f_{i-1}^{n+1-} + (1+\sigma|\xi|) f_i^{n+1-} + \sigma \mathbb{1}_{\xi<0}\xi f_{i+1}^{n+1-} = M_i \\ (1+\sigma|\xi|) f_1^{n+1-} + \sigma \mathbb{1}_{\xi<0}\xi f_2^{n+1-} = M_1 + \sigma \mathbb{1}_{\xi>0}\xi M_0^{n+1}, \\ -\sigma \mathbb{1}_{\xi>0}\xi f_{P-1}^{n+1-} + (1+\sigma|\xi|) f_P^{n+1-} = M_P - \sigma \mathbb{1}_{\xi<0}\xi M_{P+1}^{n+1}, \end{array}\right. \tag{2.18}$$

The quantities $M_0^{n+1} = M(U_0^{n+1},\xi)$ and $M_{P+1}^{n+1} = M(U_{P+1}^{n+1},\xi)$ appearing in the last two lines of (2.18) account for the imposed boundary conditions. In a first step, we assume that M_0^{n+1} and M_{P+1}^{n+1} are two known kinetic Maxwellian, their expressions will be discussed in more details in the paragraph devoted to the practical computation of the implicit variables, see Section 2.3.4.

With obvious notations, the system (2.18) consists in finding $f^{n+1} = \{f_i^{n+1-}\}_{i\in\{1,\dots,P\}}$ satisfying

$$(\mathbf{I} + \sigma \mathbf{L}) f^{n+1} = M + \sigma B^{n+1}, \tag{2.19}$$

where $\mathbf{I}$ is the identity matrix of length P and $\mathbf{L} \in \mathbb{R}^{P\times P}$ is given by

$$\begin{pmatrix} |\xi| & \xi\mathbb{1}_{\xi<0} & 0 & \dots & 0 \\ -\xi\mathbb{1}_{\xi>0} & |\xi| & \xi\mathbb{1}_{\xi<0} & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -\xi\mathbb{1}_{\xi>0} & |\xi| & \xi\mathbb{1}_{\xi<0} \\ 0 & \dots & 0 & -\xi\mathbb{1}_{\xi>0} & |\xi| \end{pmatrix}.$$

The three vectors f^{n+1}, M and B^{n+1} of $\mathbb{R}^P$ are defined by

$$f^{n+1} = \begin{pmatrix} f_1^{n+1-} \\ \vdots \\ f_i^{n+1-} \\ \vdots \\ f_P^{n+1-} \end{pmatrix}, \quad M = \begin{pmatrix} M_1 \\ \vdots \\ M_i \\ \vdots \\ M_P \end{pmatrix} \quad \text{and} \quad B^{n+1} = \begin{pmatrix} \mathbb{1}_{\xi>0}\xi M_0^{n+1}, \\ 0 \\ \vdots \\ 0 \\ -\mathbb{1}_{\xi<0}\xi M_{P+1}^{n+1}, \end{pmatrix}. \tag{2.20}$$

The practical computation of the densities vector f^{n+1} will be discussed in Section 2.3.2. Hereafter, we focus on the properties of the numerical scheme (2.18) and the two following results hold.

Lemma 4. *The matrix $\mathbf{I} + \sigma\mathbf{L}$ defined by (2.19)*

(i) is invertible for any σ and ξ,

(ii) its inverse $(\mathbf{I} + \sigma\mathbf{L})^{-1}$ has only positive coefficients.

Proposition 7. *The numerical scheme (2.18) satisfies the following properties*

(i) the discretization (2.18) is consistent at the macroscopic level with (2.1),

(ii) the system (2.18) – or equivalently the system (2.19) – admits an unique solution and the solution satisfies

$$f_i^{n+1-} = f_i^{n+1-}(\xi) \geq 0, \quad \forall 1 \leq i \leq P, \ \forall \xi \in \mathbb{R}.$$

Since the system (2.19) admits a unique solution of positive quantities, it defines an implicit kinetic scheme.

Proposition 8. *The numerical scheme defined by (2.19) satisfies the fully discrete entropy equality*

$$H_0(f_i^{n+1-}) = H_0(M_i) - \sigma\Big(H_{0,i+1/2}^{n+1-} - H_{0,i-1/2}^{n+1-}\Big) - \Psi(f_i^{n+1-}, M_i) \\ + \sigma\xi\Big(\mathbb{1}_{\xi<0}\Psi(f_i^{n+1-}, f_{i+1}^{n+1-}) - \mathbb{1}_{\xi>0}\Psi(f_i^{n+1-}, f_{i-1}^{n+1-})\Big) \tag{2.21}$$

where $H_{0,i+1/2}^{n+1-}$, $H_{0,i-1/2}^{n+1-}$ *are given by*

$$H_{0,i+1/2}^{n+1-} = \xi\mathbb{1}_{\xi<0}H_0(f_{i+1}^{n+1-}) + \xi\mathbb{1}_{\xi>0}H_0(f_i^{n+1-}),$$

$$H_{0,i-1/2}^{n+1-} = \xi\mathbb{1}_{\xi<0}H_0(f_i^{n+1-}) + \xi\mathbb{1}_{\xi>0}H_0(f_{i-1}^{n+1-}),$$

and where the function Ψ *is defined by*

$$\Psi : \mathbb{R}^2 \ni (a,b) \longmapsto \frac{g^2\pi^2}{6}(b+2a)(b-a)^2. \tag{2.22}$$

Since Ψ *is positive on* $\mathbb{R}_+^2$*, the last two terms of equality (2.21) define a nonpositive dissipative term.*

Notice that the results obtained in Proposition 7 and Proposition 8 do not require any CFL condition. A consequence of Proposition 8 is that, when using the classical Maxwellian (2.2), the macroscopic scheme associated to (2.17) will satisfy a discrete entropy inequality that always dissipates the energy. In fact since the Maxwellian (2.2) minimizes the functional (2.8) we have the following upper bound on the macroscopic entropy $\eta(U_i^{n+1})$

$$\eta(U_i^{n+1}) = \int_{\mathbb{R}} H_0(M(U_i^{n+1},\xi),\xi)d\xi \leq \int_{\mathbb{R}} H_0(f_i^{n+1-}(\xi),\xi)d\xi.$$

We then use equality (2.21) yielding

$$\eta(U_i^{n+1}) \leq \eta(U_i^n) - \sigma\bigg(\int_{\mathbb{R}} H_{0,i+1/2}^{n+1-}(\xi)d\xi - \int_{\mathbb{R}} H_{0,i-1/2}^{n+1-}(\xi)d\xi\bigg) + \int_{\mathbb{R}} D_i(\xi)d\xi, \tag{2.23}$$

where D_i is the negative dissipation term corresponding to the last three lines of (2.21).

Proof of Lemma 4. The matrix $\mathbf{I}+\sigma\mathbf{L}$ writes

$$\mathbf{I}+\sigma\mathbf{L} = \begin{pmatrix} 1+\sigma|\xi| & \sigma\xi\mathbb{1}_{\xi<0} & 0 & \dots & 0 \\ -\sigma\xi\mathbb{1}_{\xi>0} & 1+\sigma|\xi| & \sigma\xi\mathbb{1}_{\xi<0} & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -\sigma\xi\mathbb{1}_{\xi>0} & 1+\sigma|\xi| & \sigma\xi\mathbb{1}_{\xi<0} \\ 0 & \dots & 0 & -\sigma\xi\mathbb{1}_{\xi>0} & 1+\sigma|\xi| \end{pmatrix},$$

and it is easy to see that the matrix $\mathbf{I}+\sigma\mathbf{L}$ is strictly diagonally dominant and hence invertible. Moreover the matrix $\mathbf{\Lambda} = \mathbf{I}+\sigma\mathbf{L}$ is such that

$$\mathbf{\Lambda}_{i,i} > 0, \quad \text{and} \quad \mathbf{\Lambda}_{i,j} \leq 0, \text{ when } i \neq j,$$

meaning $\mathbf{I}+\sigma\mathbf{L}$ is a monotone matrix and hence the solution of (2.19) satisfies

$$f_i^{n+1-} = \left((\mathbf{I}+\sigma\mathbf{L})^{-1}(M+\sigma B^{n+1,k})\right)_i \geq 0, \quad \forall i,$$

proving the result.

Denoting $\mathbf{L}^d$ (resp. $\mathbf{L}^{nd}$) the diagonal (resp. non diagonal) part of $\mathbf{L}$ we can write

$$\mathbf{I}+\sigma\mathbf{L} = (\mathbf{I}+\sigma\mathbf{L}^d)\left(\mathbf{I}-(\mathbf{I}+\sigma\mathbf{L}^d)^{-1}(-\sigma\mathbf{L}^{nd})\right),$$

where all the entries of the matrix

$$\mathbf{J} = (\mathbf{I}+\sigma\mathbf{L}^d)^{-1}(-\sigma\mathbf{L}^{nd}),$$

are non negative and less than 1. And hence, we can write

$$(\mathbf{I}+\sigma\mathbf{L})^{-1} = (\mathbf{I}-\mathbf{J})^{-1}(\mathbf{I}+\sigma\mathbf{L}^d)^{-1} = \sum_{k=0}^{\infty}\mathbf{J}^k(\mathbf{I}+\sigma\mathbf{L}^d)^{-1},$$

proving all the entries of $(\mathbf{I}+\sigma\mathbf{L})^{-1}$ are non negative. □

Proof of Proposition 7. (i) The four terms in parentheses in (2.17) are conservative, and are classically consistent with $\xi\partial_x f$ in (2.12).
(ii) This is a direct consequence of Lemma 4. □

The proof of Proposition 8 makes use of the following Lemma which will also be useful later.

Lemma 5. *The following identity holds for any real pair (a,b) and for any $\xi\in\mathbb{R}$*

$$H_0(b,\xi) = H_0(a,\xi) + \partial_f H_0(a,\xi)(b-a) + \Psi(a,b), \tag{2.24}$$

with the function Ψ defined in (2.22). Especially, we recover the convexity of $H_0(\cdot,\xi)$ on $\mathbb{R}_+$ thanks to the positivity of Ψ on $\mathbb{R}_+^2$. Equality (2.24) remains satisfied if we replace H_0 by H.

Proof of Lemma 5. For any (a,b) in $\mathbb{R}^2$ there holds

$$\begin{aligned}
\partial_f H_0(a)(b-a) &= \frac{\xi^2}{2}b + \frac{g^2\pi^2}{2}a^2b - \frac{\xi^2}{2}a - \frac{g^2\pi^2}{2}a^3 \\
&= H_0(b) + \frac{g^2\pi^2}{2}a^2b - \frac{g^2\pi^2}{6}b^3 - H_0(a) - \frac{g^2\pi^2}{2}a^3 + \frac{g^2\pi^2}{6}a^3 \\
&= H_0(b) - H_0(a) - \frac{g^2\pi^2}{6}\Big(b^3 - a^3 - 3a^2(b-a)\Big),
\end{aligned}$$

and equality (2.24) is recovered using the formula

$$b^3 - a^3 - 3a^2(b-a) = (b+2a)(b-a)^2.$$

This result is extended to the kinetic entropy H owing to the relation $H(f,\xi) = H_0(f,\xi) + gz_b f$. □

Proof of Proposition 8. The proof follows similar lines as what was done in the case of the fully explicit version of the kinetic scheme in [5]. Instead of multiplying Equation (2.17) by $\partial_f H_0(f_i^n)$,

we multiply it with $\partial_f H_0(f_i^{n+1-})$, which leads to

$$\begin{aligned}\partial_f H_0(f_i^{n+1-})(f_i^{n+1-} - M_i) = & -\sigma\xi \mathbb{1}_{\xi<0}\partial_f H_0(f_i^{n+1-})(f_{i+1}^{n+1-} - f_i^{n+1-}) \\ & + \sigma\xi \mathbb{1}_{\xi>0}\partial_f H_0(f_i^{n+1-})(f_{i-1}^{n+1-} - f_i^{n+1-}) \ .\end{aligned} \tag{2.25}$$

In equation (2.25), we recognize three terms of the form $\partial_f H(a)(b-a)$ with $a = f_i^{n+1-}$ and $b \in \{f_{i-1}^{n+1-}, M_i, f_{i+1}^{n+1-}\}$. Taking advantage of Lemma 5 we can write

$$\begin{aligned}H_0(f_i^{n+1-}) - H_0(M_i) + \Psi(f_i^{n+1-}, M_i) = & -\sigma\xi \mathbb{1}_{\xi<0}\Big(H_0(f_{i+1}^{n+1-}) - H_0(f_i^{n+1-}) - \Psi(f_i^{n+1-}, f_{i+1}^{n+1-})\Big) \\ & + \sigma\xi \mathbb{1}_{\xi>0}\Big(H_0(f_{i-1}^{n+1-}) - H_0(f_i^{n+1-}) - \Psi(f_i^{n+1-}, f_{i-1}^{n+1-})\Big),\end{aligned}$$

and we conclude by grouping the expressions. □

2.3.2 Practical computation of the implicit update

When dealing with implicit schemes, one has often to invert an operator and the key point of the numerical scheme (2.19) is the computation of the inverse of the matrix $\mathbf{I} + \sigma\mathbf{L}$. In our case, it will be possible to compute analytically this inverse thanks to the triangular structure of the matrix, which is due to the upwinding of the fluxes in (2.17). More precisely we decompose $(\mathbf{I} + \sigma\mathbf{L})^{-1}$ as the contributions of the left- and right-going information, which gives

$$(\mathbf{I} + \sigma\mathbf{L})^{-1} = (\mathbf{I} + \sigma\mathbf{L}^+)^{-1}\mathbb{1}_{\xi<0} + (\mathbf{I} + \sigma\mathbf{L}^-)^{-1}\mathbb{1}_{\xi>0}, \tag{2.26}$$

with the upwinding matrices $\mathbf{L}^+$ and $\mathbf{L}^-$ corresponding to

$$\mathbf{L}^+ = \begin{pmatrix} -\xi & \xi & 0 & \dots & 0 \\ 0 & -\xi & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \xi \\ 0 & \dots & & 0 & -\xi \end{pmatrix}, \quad \mathbf{L}^- = \begin{pmatrix} \xi & 0 & & \dots & 0 \\ -\xi & \ddots & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \dots & 0 & -\xi & \xi \end{pmatrix}.$$

Introducing $\mathbf{J}^+$ and $\mathbf{J}^-$ the matrices of $\mathbb{R}^{P\times P}$ defined as

$$(\mathbf{J}^+)_{i,j} = \begin{cases} 1 & \text{if } i = j-1 \\ 0 & \text{otherwise} \end{cases}, \quad (\mathbf{J}^-)_{i,j} = \begin{cases} 1 & \text{if } i = j+1 \\ 0 & \text{otherwise} \end{cases},$$

we can write

$$\begin{cases} (\mathbf{I} + \sigma\mathbf{L}^-)^{-1} = \big((1+\sigma\xi)\mathbf{I} - \sigma\xi\mathbf{J}^-\big)^{-1} = \dfrac{1}{1+\sigma\xi}\left(\mathbf{I} - \dfrac{\sigma\xi}{1+\sigma\xi}\mathbf{J}^-\right)^{-1} \\ (\mathbf{I} + \sigma\mathbf{L}^+)^{-1} = \big((1-\sigma\xi)\mathbf{I} + \sigma\xi\mathbf{J}^+\big)^{-1} = \dfrac{1}{1-\sigma\xi}\left(\mathbf{I} + \dfrac{\sigma\xi}{1-\sigma\xi}\mathbf{J}^+\right)^{-1} \end{cases}.$$

The above inverses can be computed through geometric sums since $\mathbf{J}_P^+$ and $\mathbf{J}_P^-$ have a spectral radius equal to zero. More specifically these two matrices are nilpotent, which implies that the geometric

sums in question contain a finite number of nonzero terms and are given below

$$(\mathbf{I}+\sigma\mathbf{L}^-)^{-1}=\sum_{k=0}^{P}\frac{(\sigma\xi)^k}{(1+\sigma\xi)^{k+1}}(\mathbf{J}^-)^k,$$
$$(\mathbf{I}+\sigma\mathbf{L}^+)^{-1}=\sum_{k=0}^{P}\frac{(-\sigma\xi)^k}{(1-\sigma\xi)^{k+1}}(\mathbf{J}^+)^k.$$

To conclude we give the analytic expression of the inverse:

$$(\mathbf{I}+\sigma\mathbf{L}^-)^{-1}_{i,j}=\begin{cases}\dfrac{(\sigma\xi)^{i-j}}{(1+\sigma\xi)^{i-j+1}} & \text{if } i\geq j\\ 0 & \text{else}\end{cases}, \tag{2.27}$$

$$(\mathbf{I}+\sigma\mathbf{L}^+)^{-1}_{i,j}=\begin{cases}\dfrac{(-\sigma\xi)^{j-i}}{(1-\sigma\xi)^{j-i+1}} & \text{if } i\leq j\\ 0 & \text{else}\end{cases}. \tag{2.28}$$

Especially we recover the properties enumerated in Lemma 4, since we see that all the coefficients of the inverse (2.26) are comprised between zero and one respectively when $\xi\geq 0$ and $\xi\leq 0$.

2.3.3 Macroscopic implicit scheme

We now turn towards obtaining an explicit writing of the macroscopic update associated to (2.17). Since the right hand side of (2.19) is made of Maxwellians, we will see that this amounts to compute the integral of

$$\mathbb{1}_{\pm\xi>0}\frac{(\pm\sigma\xi)^k}{(1\pm\sigma\xi)^{k+1}}M(U,\xi)$$

against 1, ξ and ξ^2 for $0\leq k\leq P-1$. This hardly seems possible with the classical Maxwellian proposed in (2.2). Instead in this section we will use the simpler equilibrium function given by

$$M(U,\xi)=\frac{h}{2\sqrt{3}c}\mathbb{1}_{|\xi-u|\leq\sqrt{3}c},\quad c=\sqrt{\frac{gh}{2}}, \tag{2.29}$$

and referred to as the index Maxwellian. This is the simplest choice we can make, and it will enable us to obtain analytic expressions for the aforementioned integrals. Furthermore it satisfies all the moment relations (2.3), and we make the following remark.

Remark 1. *We recall that the half-disk Maxwellian $M(U,\xi)$ defined by (2.2) has some optimal properties presented in Lemma 3, which allow to obtain the discrete entropy inequality (2.23) at the macroscopic scale. Other choices of Maxwellian are possible such as (2.29), but the previous discrete entropy inequality is not granted to hold anymore. A general possibility is to choose $M(U,\xi)$ of the form*

$$M(U,\xi)=\frac{h}{c}\chi\Big(\frac{\xi-u}{c}\Big).$$

To satisfy the moment relations (2.3), it is then sufficient for χ to be an even function verifying

$$\int_{\mathbb{R}}\chi(z)dz=\int_{\mathbb{R}}z^2\chi(z)dz=1.$$

Furthermore we ask χ to be nonnegative with compact support, and possible choices are for instance

$$\chi_1(z) = \frac{1}{2\sqrt{3}} \mathbb{1}_{|z|\leq\sqrt{3}}, \quad or \quad \chi_2(z) = \frac{3}{20\sqrt{5}} z^2 + \frac{3}{4\sqrt{5}} \mathbb{1}_{|z|\leq\sqrt{5}}. \tag{2.30}$$

The definition (2.29) of the index Maxwellian corresponds to the first shape function χ_1 in (2.30), whereas the definition (2.2) of the half-disk Maxwellian corresponds to the third choice below

$$\chi_3(z) = \frac{1}{\pi} \sqrt{1 - \frac{z^2}{4}} \mathbb{1}_{|z|\leq 2}.$$

We proceed in two steps to compute our scheme with boundary conditions. The strategy consists to dissociate the contribution of the information coming from the interior of the computational domain, and the one coming from the exterior as below

$$\begin{cases} U^{\text{int}} = \displaystyle\int_{\mathbb{R}} \begin{pmatrix} 1 \\ \xi \end{pmatrix} (\mathbf{I} + \sigma \mathbf{L})^{-1} M \, d\xi \\ U^{\text{ext}} = \displaystyle\int_{\mathbb{R}} \begin{pmatrix} 1 \\ \xi \end{pmatrix} (\mathbf{I} + \sigma \mathbf{L})^{-1} \sigma B^{n+1} \, d\xi \end{cases}. \tag{2.31}$$

The final update is then set as $U^{n+1} = U^{\text{int}} + U^{\text{ext}}$, which coincides with definition (2.16). We postpone the details about the computation of B^{n+1} to the next section, and assume that it is known for now. First for U^{int}, we have

$$U^{\text{int}} = \int_{\mathbb{R}} \mathbb{1}_{\xi\leq 0} \begin{pmatrix} 1 \\ \xi \end{pmatrix} (\mathbf{I} + \sigma \mathbf{L}^+)^{-1} M d\xi + \int_{\mathbb{R}} \mathbb{1}_{\xi\geq 0} \begin{pmatrix} 1 \\ \xi \end{pmatrix} (\mathbf{I} + \sigma \mathbf{L}^-)^{-1} M d\xi.$$

Plugging the analytic expressions (2.27) and (2.28) in the above integrals, we can express the i-th component of U^{int} as

$$\begin{aligned} U_i^{\text{int}} &= \int_{\xi<0} \sum_{j=1}^{P} \begin{pmatrix} 1 \\ \xi \end{pmatrix} (\mathbf{I} + \sigma \mathbf{L}^+)^{-1}_{i,j} M_j d\xi + \int_{\xi>0} \sum_{j=1}^{P} \begin{pmatrix} 1 \\ \xi \end{pmatrix} (\mathbf{I} + \sigma \mathbf{L}^-)^{-1}_{i,j} M_j d\xi \\ &= \int_{\xi<0} \sum_{j=i}^{P} \begin{pmatrix} 1 \\ \xi \end{pmatrix} \frac{(-\sigma\xi)^{j-i}}{(1-\sigma\xi)^{j-i+1}} M_j d\xi + \int_{\xi>0} \sum_{j=1}^{i} \begin{pmatrix} 1 \\ \xi \end{pmatrix} \frac{(\sigma\xi)^{i-j}}{(1+\sigma\xi)^{i-j+1}} M_j d\xi. \end{aligned} \tag{2.32}$$

A detailed expression of the quantities appearing in relation (2.32) is given in Appendix A. Similarly for the exterior contribution we have

$$U^{\text{ext}} = \sigma \int_{\xi<0} \begin{pmatrix} 1 \\ \xi \end{pmatrix} (\mathbf{I} + \sigma \mathbf{L}^+)^{-1} B^{n+1} d\xi + \sigma \int_{\xi>0} \begin{pmatrix} 1 \\ \xi \end{pmatrix} (\mathbf{I} + \sigma \mathbf{L}^-)^{-1} B^{n+1} d\xi.$$

Using definition (2.20) and equalities (2.27)–(2.28), the i-th component of U^{ext} is

$$U_i^{\text{ext}} = \int_{\xi<0} \begin{pmatrix} 1 \\ \xi \end{pmatrix} \frac{(-\sigma\xi)^{P-i+1}}{(1-\sigma\xi)^{P-i+1}} M_{P+1}^{n+1} d\xi + \int_{\xi>0} \begin{pmatrix} 1 \\ \xi \end{pmatrix} \frac{(\sigma\xi)^{i}}{(1+\sigma\xi)^{i}} M_0^{n+1} d\xi.$$

We can reuse the primitive obtained in Appendix 2.A to express the water height h_i^{ext}. However for

the flux $(hu)_i^{\text{ext}}$ we need to write a primitive for

$$\int_{\pm\xi>0} \xi \frac{(\pm\sigma\xi)^k}{(1\pm\sigma\xi)^k} M(U,\xi)d\xi,$$

which is given in Appendix 2.B. In the end we obtain a fully explicit writing of the implicit update at the macroscopic level.

2.3.4 Boundary conditions

In this paragraph we discuss how to enforce the boundary conditions. These are represented by the exterior contribution U^{ext} introduced in (2.31), and accordingly we need to specify the ghost values U_0^{n+1} and U_{P+1}^{n+1} appearing in the definition (2.20) of B^{n+1}. The problem we are facing is that these ghost quantities depend on the neighboring values in cells C_1 and C_P at time t^{n+1}, and which are themselves unknown. Hence we have an implicit problem where the relation between the ghost and border terms can be nonlinear depending on the type of boundary conditions. In practice we will avoid this issue by substituting B^{n+1} with B^n in the definition of U^{ext}. Doing so can be interpreted as a first order approximation in time since we have

$$U_0^{n+1} = U_0^n + O(\Delta t), \quad U_{P+1}^{n+1} = U_{P+1}^n + O(\Delta t).$$

The benefit is that we can more easily determine the ghost quantities U_0^n, U_{P+1}^n at time t^n based on U_1^n, U_P^n following the procedure described hereafter and similar to that of Bristeau and Coussin in [26]. We will focus on fluvial flows where the material velocity of particles $|u|$ is smaller than the celerity of surface gravity waves $\sqrt{gh}$. In particular low Froude flows enter this regime. Since in this case the eigenvalues $u-\sqrt{gh}$ and $u+\sqrt{gh}$ have opposite sign, at each boundary we have exactly one wave entering the domain and one wave leaving it. Hence we dispose of a single degree of freedom to set the ghost values, which generally consists in enforcing either a given water height or a discharge. The ghost state is then fully determined by asking the outward-going Riemann invariant to remain constant through the interface.

Given water height. First we treat the case where the water height is enforced at the boundary of the domain. We denote by $h_{g,l}$ the value attributed to the left ghost cell, and $h_{g,r}$ the one attributed to the right ghost cell. Together with the condition on the outgoing Riemann invariant, we get the following nonlinear systems

$$\begin{cases} h_0^n = h_{g,l} \\ u_0^n - 2\sqrt{gh_0^n} = u_1^n - 2\sqrt{gh_1^n} \end{cases}, \quad \begin{cases} h_{P+1}^n = h_{g,r} \\ u_{P+1}^n + 2\sqrt{gh_{P+1}^n} = u_P^n + 2\sqrt{gh_P^n} \end{cases}.$$

They can be solved explicitly and we get

$$U_0^n = h_{g,l}\begin{pmatrix} 1 \\ u_1^n - 2(\sqrt{gh_1^n} - \sqrt{gh_{g,l}}) \end{pmatrix},$$
$$U_{P+1}^n = h_{g,r}\begin{pmatrix} 1 \\ u_P^n + 2(\sqrt{gh_P^n} - \sqrt{gh_{g,r}}) \end{pmatrix}.$$

Given flux. Another possibility is to enforce the discharge at the boundary, and we denote by $q_{g,l}$ and $q_{g,r}$ the left and right ghost values. This time around, the constraint on the Riemann invariant

will enable to determine the ghost water height. Indeed we have the systems

$$\left\{\begin{array}{l} q_0^n = q_{g,l} \\ u_0^n - 2\sqrt{gh_0^n} = u_1^n - 2\sqrt{gh_1^n} \end{array}\right. , \quad \left\{\begin{array}{l} q_{P+1}^n = q_{g,r} \\ u_{P+1}^n + 2\sqrt{gh_{P+1}^n} = u_P^n + 2\sqrt{gh_P^n} \end{array}\right. , \tag{2.33}$$

and the equalities satisfied by the Riemann invariants amount to finding the real roots of the third order polynomials in $\sqrt{h_0^n}$ and $\sqrt{h_{P+1}^n}$ below

$$\begin{aligned} -2\sqrt{g}(h_0^n)^{3/2} - (u_1^n - 2\sqrt{gh_1^n})h_0^n + q_{g,l} = 0, \\ 2\sqrt{g}(h_{P+1}^n)^{3/2} - (u_P^n + 2\sqrt{gh_P^n})h_{P+1}^n + q_{g,r} = 0. \end{aligned}$$

Or equivalently,

$$(h_0^n)^{3/2} + \frac{u_1^n - 2\sqrt{gh_1^n}}{2\sqrt{g}} h_0^n - \frac{q_{g,l}}{2\sqrt{g}} = 0, \tag{2.34}$$

$$(h_{P+1}^n)^{3/2} - \frac{u_P^n + 2\sqrt{gh_P^n}}{2\sqrt{g}} h_{P+1}^n + \frac{q_{g,r}}{2\sqrt{g}} = 0. \tag{2.35}$$

and the discriminants of (2.34),(2.35) are given respectively by

$$\begin{aligned} \Delta_0 = -\frac{q_{g,l}}{432g^2}\left((u_1^n - 2\sqrt{gh_1^n})^3 - 27gq_{g,l}\right), \\ \Delta_P = -\frac{q_{g,r}}{432g^2}\left((u_P^n + 2\sqrt{gh_P^n})^3 + 27gq_{g,r}\right). \end{aligned}$$

We consider that in particular $|u_1^n| \leq 2\sqrt{gh_0^n}$ and $|u_P^n| \leq 2\sqrt{gh_P^n}$. It is clear that the sign of $q_{g,l}$ has an impact on the sign of the discriminant and hence on the number of possible roots as well as on the sign of the roots as we demonstrate below. Consider the following polynomial,

$$\mathcal{P}(x) = x^3 + \frac{u_1^n - 2\sqrt{gh_1^n}}{2\sqrt{g}} x^2 - \frac{q_{g,l}}{2\sqrt{g}},$$

then $\mathcal{P}'$ admits two real distinct roots $0, -\frac{u_1^n - 2\sqrt{gh_1^n}}{3\sqrt{g}} = x_1 > 0$ with $\mathcal{P}(x_1) = \frac{(u_1^n - 2\sqrt{gh_1^n})^3}{54g\sqrt{g}} - \frac{q_{g,l}}{2\sqrt{g}}$. Hence, the equation $\mathcal{P}(x) = 0$ admits at least one real non-negative root in the case $q_{g,l} \in [\frac{(u_1^n - 2\sqrt{gh_1^n})^3}{27g}, +\infty]$. In the case $q_{g,l} < \frac{(u_1^n - 2\sqrt{gh_1^n})^3}{27g}$, equations (2.33) are no longer valid and one can choose to impose h_0^n instead. A similar argument holds for the right ghost cell. Consider the polynomial,

$$\mathcal{Q}(x) = x^3 - \frac{u_P^n + 2\sqrt{gh_P^n}}{2\sqrt{g}} x^2 + \frac{q_{g,r}}{2\sqrt{g}},$$

then $\mathcal{Q}'$ admits two real distinct roots $0, \frac{u_P^n + 2\sqrt{gh_P^n}}{3\sqrt{g}} = x_2 > 0$ with $\mathcal{Q}(x_2) = -\frac{(u_P^n + 2\sqrt{gh_P^n})^3}{54g\sqrt{g}} + \frac{q_{g,r}}{2\sqrt{g}}$. Hence the equation $\mathcal{Q}(x) = 0$ admits at least one real non-negative root in the case $q_{g,r} \in [-\infty, \frac{(u_P^n + 2\sqrt{gh_P^n})^3}{27g}]$. In the case $q_{g,r} > \frac{(u_P^n + 2\sqrt{gh_P^n})^3}{27g}$, equations (2.33) are no longer valid and one can choose to impose h_0^n and/or h_P^n instead.

Note that in this case, our approach differs from that of Bristeau and Coussin in [26], where the ghost value is chosen such that the resulting numerical flux at the interface coincides with the

boundary discharge. Instead we do not enforce any value at the interface but directly in the ghost cell, which can be seen as a first order simplification in space. We also comment on the fact that nothing prevents us from mixing the boundary conditions, for instance we can enforce a water height on the left boundary, and a discharge on the right. A common practice for channel flows is to enforce the water height at the inlet and the flux at the outlet.

Remark 2. *When substituting B^{n+1} with B^n in the implicit kinetic scheme (2.19), the corresponding update can be reformulated as $(\overline{\mathbf{I}+\sigma\mathbf{L}})\overline{f}^{n+1} = \overline{M}^n$ with*

$$\overline{\mathbf{I}+\sigma\mathbf{L}} = \left(\begin{array}{c|c|c} 1 & & \\ \hline -\sigma\xi\mathbb{1}_{\xi>0} & \multirow{3}{*}{$\mathbf{I}+\sigma\mathbf{L}$} & \vdots \\ 0 & & 0 \\ \vdots & & \sigma\xi\mathbb{1}_{\xi<0} \\ \hline & & 1 \end{array}\right), \overline{f}^{n+1} = \begin{pmatrix} f_0^{n+1} \\ f^{n+1} \\ f_{P+1}^{n+1} \end{pmatrix}, \overline{M}^n = \begin{pmatrix} M_0^n \\ M^n \\ M_{P+1}^n \end{pmatrix}$$

As a consequence the maximum principle $\|\overline{f}^{n+1}(\xi)\|_\infty \leq \|\overline{M}^n(\xi)\|_\infty$ holds for any ξ in $\mathbb{R}$ during the transport step. In fact we can verify that matrix $(\overline{\mathbf{I}+\sigma\mathbf{L}})$ is monotone, and following the argument involved in Lemma 5.1 from [3] we can write

$$0 \leq \overline{f}^{n+1} = (\overline{\mathbf{I}+\sigma\mathbf{L}})^{-1}\overline{M}^n \leq (\overline{\mathbf{I}+\sigma\mathbf{L}})^{-1}(\|\overline{M}^n\|_\infty \mathbf{1}),$$

with $\mathbf{1}$ the vector from $\mathbb{R}^{P+2}$ whose entries are all equal to one. Using equality $(\overline{\mathbf{I}+\sigma\mathbf{L}})^{-1}\mathbf{1} = \mathbf{1}$ allows to conclude. Note however that there is no such principle at the macroscopic scale, similarly to the continuous Saint-Venant system.

2.3.5 Implementation and computational costs

It is important to try and keep a reasonable algorithmic complexity so that the implicit method presented in the previous lines remains usable in practice. We discuss here how to improve its computational cost by a substantial margin. In Appendix 2.A, we show that the i-th component of vectors h^{int} and $(hu)^{\text{int}}$ have the form

$$\left\{\begin{array}{l} h_i^{\text{int}} = \dfrac{1}{2\sigma\sqrt{3}}\left(\displaystyle\sum_{j=i}^{P}\sqrt{\frac{2h_j^n}{g}}(Ah)_{i,j} + \sum_{j=1}^{i}\sqrt{\frac{2h_j^n}{g}}(Bh)_{i,j}\right) \\ (hu)_i^{\text{int}} = \dfrac{1}{2\sigma^2\sqrt{3}}\left(-\displaystyle\sum_{j=i}^{P}\sqrt{\frac{2h_j^n}{g}}(Ahu)_{i,j} + \sum_{j=1}^{i}\sqrt{\frac{2h_j^n}{g}}(Bhu)_{i,j}\right) \end{array}\right., \tag{2.36}$$

where Ah, Ahu are dense upper triangular matrices, and Bh, Bhu are dense lower triangular matrices. Therefore computing h^{int} and $(hu)^{\text{int}}$ through (2.36) is analog to performing a matrix-vector product which has a quadratic complexity $O(P^2)$, and we cannot hope to do better than that. However the coefficients (2.77)–(2.80) of the above matrices involve a summation, and at a first glance the cost to assemble them is seemingly cubic. This is quite expensive and can render the method pretty much inefficient. However this complexity can be reduced to a quadratic cost by computing the coefficients in the correct order. More specifically we show that all the matrices above can be defined through a recurrence relation allowing to compute each coefficient from a previous one in

$O(1)$ operation. In fact, denoting $y = x/(1+x)$ and $z = \ln|1+x|$, the matrix Ah is given by

$$\begin{pmatrix} \big[z\big]_{-\min(0,b_1)\sigma}^{-\min(0,a_1)\sigma} & \big[z-y\big]_{-\min(0,b_2)\sigma}^{-\min(0,a_2)\sigma} & \cdots & \cdots & \big[z-\sum_{l=1}^{P-1} y^l/l\big]_{-\min(0,b_P)\sigma}^{-\min(0,a_P)\sigma} \\ 0 & \big[z\big]_{-\min(0,b_2)\sigma}^{-\min(0,a_2)\sigma} & & & \vdots \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \big[z-y\big]_{-\min(0,b_P)\sigma}^{-\min(0,a_P)\sigma} \\ 0 & \cdots & \cdots & 0 & \big[z\big]_{-\min(0,b_P)\sigma}^{-\min(0,a_P)\sigma} \end{pmatrix},$$

where $a_j^n = u_j^n - \sqrt{3}\,c_j^n$ and $b_j^n = u_j^n + \sqrt{3}\,c_j^n$. This corresponds to the recursive definition below

$$(Ah)_{i,j} = \begin{cases} 0 & \text{if } j < i \\ \big[\ln(|1+x|)\big]_{-\min(0,b_j)\sigma}^{-\min(0,a_j)\sigma} & \text{if } i = j \\ (Ah)_{i+1,j} - \frac{1}{j-i}\big[y^{j-i}\big]_{-\min(0,b_j)\sigma}^{-\min(0,a_j)\sigma} & \text{if } j > i \end{cases}.$$

Likewise, the lower triangular matrix Bh is given by

$$\begin{pmatrix} \big[z\big]_{\max(0,a_1^n)\sigma}^{\max(0,b_1^n)\sigma} & 0 & \cdots & \cdots & 0 \\ \big[z-y\big]_{\max(0,a_1^n)\sigma}^{\max(0,b_1^n)\sigma} & \ddots & & & \\ \vdots & & \ddots & & \\ \vdots & & & \big[z\big]_{\max(0,a_{P-1}^n)\sigma}^{\max(0,b_{P-1}^n)\sigma} & 0 \\ \big[z-\sum_{l=1}^{P-1} y^l/l\big]_{\max(0,a_1^n)\sigma}^{\max(0,b_1^n)\sigma} & \cdots & \cdots & \big[z-y\big]_{\max(0,a_{P-1}^n)\sigma}^{\max(0,b_{P-1}^n)\sigma} & \big[z\big]_{\max(0,a_P^n)\sigma}^{\max(0,b_P^n)\sigma} \end{pmatrix},$$

and can be defined by the following recurrence formula

$$(Bh)_{i,j} = \begin{cases} 0 & \text{if } i < j \\ \big[\ln(|1+x|)\big]_{\max(0,a_j)\sigma}^{\max(0,b_j)\sigma} & \text{if } i = j \\ (Bh)_{i-1,j} - \frac{1}{i-j}\big[y^{i-j}\big]_{\max(0,a_j)\sigma}^{\max(0,b_j)\sigma} & \text{if } i > j \end{cases}.$$

Hence it is more efficient to assemble matrices Ah and Bh column wise, starting from the diagonal coefficient and moving towards the first or last row. This way we only have to subtract one term to the previous coefficient so as to get the next one, and the cost of this operation is in $O(1)$. Since there are $P(P+1)/2$ coefficients to compute in total, the assembly of Ah and Bh following this strategy requires $O(P^2)$ steps. A similar conclusion is achieved for Ahu and Bhu, although the recurrence relation is less straightforward to obtain. We first remark that, introducing $(l)_{i,j} = i - j + 1$ the

relations (2.79) and (2.80) become

$$(Ahu)_{i,j} = \mathbb{1}_{j\geq i}\Big[-(l)_{j,i}\ln|1+x| + x + \sum_{k=1}^{j-i} k\,\frac{y^{(l)_{j,i}-k}}{(l)_{j,i}-k}\Big]_{-\min(0,b_j)\sigma}^{-\min(0,a_j)\sigma},$$
$$(Bhu)_{i,j} = \mathbb{1}_{i\geq j}\Big[-(l)_{i,j}\ln|1+x| + x + \sum_{k=1}^{i-j} k\,\frac{y^{(l)_{i,j}-k}}{(l)_{i,j}-k}\Big]_{\max(0,a_j)\sigma}^{\max(0,b_j)\sigma}.$$

Performing the change of index $r=(l)_{j,i}-k$ for matrix Ahu and $s=(l)_{i,j}-k$ for matrix Bhu we find

$$(Ahu)_{i,j} = \Big[-(l)_{i,j}\ln|1+x| + x + (l)_{i,j}\sum_{r=1}^{j-i}\frac{y^r}{r} - \sum_{r=1}^{j-i} y^r\Big]_{-\min(0,b_j)\sigma}^{-\min(0,a_j)\sigma},$$
$$(Bhu)_{i,j} = \Big[-(l)_{i,j}\ln|1+x| + x + (l)_{i,j}\sum_{s=1}^{i-j}\frac{y^s}{s} - \sum_{s=1}^{i-j} y^s\Big]_{\max(0,a_j)\sigma}^{\max(0,b_j)\sigma}.$$

Next we introduce the matrices defined column wise in a recursive manner

$$(UA)_{i,j} = \begin{cases} 0 & \text{if } j\leq i \\ (UA)_{i+1,j} + \big[y^{j-i}\big]_{-\min(0,b_j)\sigma}^{-\min(0,a_j)\sigma} & \text{if } j>i \end{cases},$$

$$(VA)_{i,j} = \begin{cases} 0 & \text{if } j\leq i \\ (VA)_{i+1,j} + \Big[\frac{y^{j-i}}{j-i}\Big]_{-\min(0,b_j)\sigma}^{-\min(0,a_j)\sigma} & \text{if } j>i \end{cases}.$$

Then we can write that

$$(Ahu)_{i,j} = \begin{cases} 0 & j<i \\ \big[x-\ln|1+x|\big]_{-\min(0,b_j)\sigma}^{-\min(0,a_j)\sigma} & j=i \\ (l)_{j,i}(VA)_{i,j} - (UA)_{i,j} + \big[x-(l)_{j,i}\ln|1+x|\big]_{-\min(0,b_j)\sigma}^{-\min(0,a_j)\sigma} & j>i \end{cases}. \tag{2.37}$$

Similarly we introduce

$$(UB)_{i,j} = \begin{cases} 0 & \text{if } i\leq j \\ (UB)_{i-1,j} + \big[y^{i-j}\big]_{\max(0,a_j)\sigma}^{\max(0,b_j)\sigma} & \text{if } i>j \end{cases},$$

$$(VB)_{i,j} = \begin{cases} 0 & \text{if } i\leq j \\ (VB)_{i-1,j} + \Big[\frac{y^{i-j}}{i-j}\Big]_{\max(0,a_j)\sigma}^{\max(0,b_j)\sigma} & \text{if } i>j \end{cases},$$

so that we have

$$(Bhu)_{i,j} = \begin{cases} 0 & i<j \\ \big[x-\ln|1+x|\big]_{\max(0,a_j)\sigma}^{\max(0,b_j)\sigma} & i=j \\ (l)_{i,j}(VB)_{i,j} - (UB)_{i,j} + \big[x-(l)_{i,j}\ln|1+x|\big]_{\max(0,a_j)\sigma}^{\max(0,b_j)\sigma} & i>j \end{cases}. \tag{2.38}$$

To conclude, through relations (2.37) and (2.38) we are also able to assemble matrices Ahu and Bhu with a quadratic cost with respect to the number of cells, which means that the overall method has a $O(P^2)$ complexity.

We have considered here the specific case of a kinetic solver and one can imagine that an implicit scheme for another finite volume solver can lead to reduced numerical costs. But it is worth noticing that since the explicit expression of the inverse of the matrix $\mathbf{I} + \sigma\mathbf{L}$ is accessible in the kinetic context, one can hardly find a more efficient implicit technique.

Obviously the proposed implicit scheme is not constrained by any CFL condition associated with an explicit scheme, nevertheless it is important to compare the computational costs of the explicit and implicit strategies in the context of a kinetic solver.

Explicit scheme. Let Δt^n be the time step allowing to satisfy the CFL constraint. In order to obtain the expression of U^{n+1} from U^n, approximately $4P$ numerical fluxes have to be computed (2 numerical fluxes at each interface for each variable h and hu). The explicit kinetic scheme is fully detailed in [5, 9].

Implicit scheme. The CFL constraint being relaxed, we can consider a time step $\Delta t^n_{imp} \gg \Delta t^n$. The results obtained in this paragraph shows that the update from U^{n+1} from U^n requires approximately P^2 numerical fluxes to compute.

We conclude that the implicit strategy is less expensive when

$$\frac{\Delta t^n_{imp}}{\Delta t^n} \gg \frac{P^2}{P} = P. \tag{2.39}$$

Note however that the computational cost is not the only factor to account for, and one should also consider the efficiency of the scheme, that is to say the relation between the error and the computational time. Generally, taking a very coarse resolution in time results in poorly accurate results, in which case it is not desirable to have (2.39). Nevertheless there are some cases where the fast dynamics do not play an important role such as in the low Froude regime. Then it might be advantageous to consider large time steps. We will see through the upcoming numerical results from Section 2.6 that the interest of the implicit kinetic scheme is rather limited when it comes to efficiency, at least for the considered test cases. Hence the explicit strategy is preferable to the implicit one, unless we account for the greater stability offered by the latter in terms of discrete entropy inequality.

2.4 The two-dimensional Saint-Venant system

With obvious notations, we consider the 2d Saint-Venant system written under the form

$$\frac{\partial h}{\partial t} + \nabla_{x,y} \cdot (h\mathbf{u}) = 0, \tag{2.40}$$

$$\frac{\partial (h\mathbf{u})}{\partial t} + \nabla_{x,y} \cdot (h\mathbf{u} \otimes \mathbf{u}) + \nabla_{x,y}\big(\frac{g}{2}h^2\big) = -gh\nabla_{x,y} z_{2d}, \tag{2.41}$$

with $\mathbf{u} = (u, v)^T$. The kinetic interpretation of the 2d Saint-Venant system (2.40)-(2.41) is a straightforward extension of Lemma 2 and has been studied in [3, 9].

To build the 2d Gibbs equilibrium, we define the function

$$\chi_{2d}(z_1, z_2) = \frac{1}{4\pi} \mathbb{1}_{z_1^2 + z_2^2 \leq 4}. \tag{2.42}$$

This choice corresponds to the 2d version of the kinetic Maxwellian used in 1d (see [3, Remark 4.2]) and we have

$$M_{2d} = M(U_{2d}, \xi, \gamma) = \frac{h}{c^2}\chi_{2d}\left(\frac{\xi - u}{c}, \frac{\gamma - v}{c}\right), \tag{2.43}$$

with $c = \sqrt{\frac{g}{2}h}$, $(\xi, \gamma) \in \mathbb{R}^2$ and

$$U_{2d} = (h, hu, hv)^T. \tag{2.44}$$

In other words, we have $M_{2d} = \frac{1}{2g\pi}\mathbb{1}_{(\xi-u)^2+(\gamma-v)^2 \leq 2gh}$ and the following lemma holds.

Lemma 6. *If the topography $z_{2d}(x, y)$ is Lipschitz continuous, the pair of functions $(h, h\mathbf{u})$ is a weak solution to the Saint-Venant system (2.40)-(2.41) if and only if $M_{2d}(U, \xi)$ satisfies the kinetic equation*

$$\partial_t M_{2d} + \begin{pmatrix}\xi \\ \gamma\end{pmatrix} \cdot \nabla_{x,y} M_{2d} - g\nabla_{x,y} z_{2d} \cdot \nabla_{\xi,\gamma} M_{2d} = Q_{2d}, \tag{2.45}$$

for some "collision term" $Q_{2d}(t, x, y, \xi, \gamma)$ that satisfies, for a.e. (t, x, y),

$$\int_{\mathbb{R}^2} Q_{2d} d\xi d\gamma = \int_{\mathbb{R}^2} \xi Q_{2d} d\xi d\gamma = \int_{\mathbb{R}^2} \gamma Q_{2d} d\xi d\gamma = 0.$$

Proof of Lemma 6. The proof relies on simple computations. Classically, the integral of Eq. (2.45) over $\mathbb{R}^2$ gives Eq. (2.40) whereas the integral over $\mathbb{R}^2$ of Eq. (2.45) multiplied by $(\xi, \gamma)^T$ gives Eq. (2.41). □

Let us consider a cartesian mesh of a 2d domain $\Omega = (0, L_x) \times (O, L_y)$, the vertices are denoted $P_{i,j}$ for $0 \leq i \leq P+1$, $0 \leq j \leq L+1$. The coordinates of $P_{i,j}$ are $(x_i, y_j)^T$ with

$$x_i = i\Delta x, \qquad y_j = j\Delta y,$$

and $\Delta_x = L_x/(P+1)$, $\Delta_y = L_y/(L+1)$. Without loss of generality, we consider $L_x = L_y$ and $P = L$ hence $\Delta_x = \Delta_y$. We use the following notations (see Fig. 2.1):

- $K_{i,j}$, set of subscripts of nodes $P_{k,l}$ surrounding $P_{i,j}$,
- $|C_{i,j}|$, area of $C_{i,j}$,
- $\partial C_{i,j}$, boundary of $C_{i,j}$

We define the piecewise constant functions $U^n(x, y)$ and $z_{2d}(x, y)$ on cells $C_{i,j}$ corresponding to time t^n as

$$U^n(x, y) = U^n_{i,j}, \quad z_{2d}(x, y) = z_{2d,i,j}, \quad \text{for } (x, y) \in C_{i,j},$$

with $U^n_{2d,i,j} = (h^n_{i,j}, q^n_{x,i,j}, q^n_{y,i,j})^T$ i.e.

$$U^n_{2d,i,j} \approx \frac{1}{|C_{i,j}|}\int_{C_{i,j}} U_{2d}(t^n, x, y) dxdy, \quad z_{2d,i,j} \approx \frac{1}{|C_{i,j}|}\int_{C_{i,j}} z_{2d}(x, y) dxdy,$$

with U_{2d} defined by (2.44).

Let $C_{i,j}$ be a dual cell of the structured mesh defined by the vertices $\{P_{i,j}\}$, see Fig. 2.1. In the case of a flat topography, the integral over $C_{i,j}$ of the convective part of the kinetic equation (2.4)

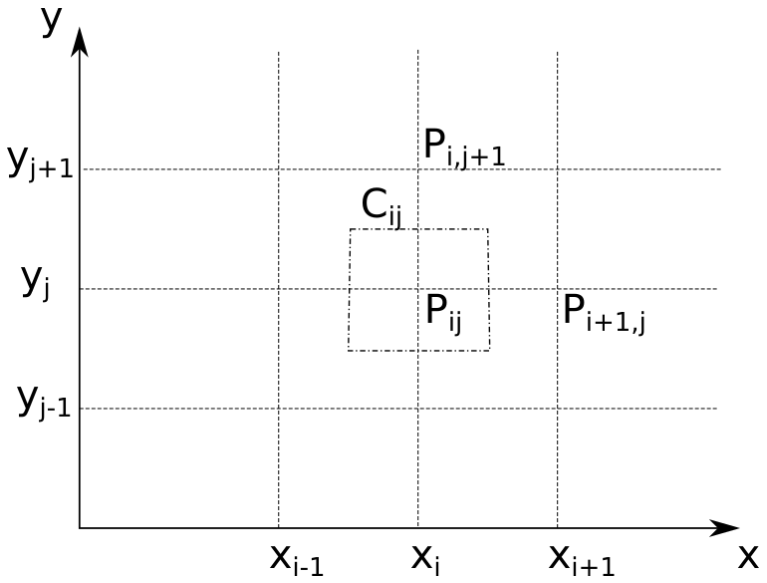

Figure 2.1: The vertices $\{P_{i,j}\}$ and the dual cell $C_{i,j}$.

gives

$$\int_{C_{i,j}} \left(\frac{\partial M_{2d}}{\partial t} + \begin{pmatrix} \xi \\ \gamma \end{pmatrix} . \nabla_{x,y} M_{2d} \right) dxdy \approx |C_{i,j}| \frac{\partial M_{2d,i,j}}{\partial t} + \sum_{(k,l) \in K_{i,j}} \int_{\partial C_{i,j}} M_{i,j,k,l} dl, \tag{2.46}$$

with $M_{2d,i,j} = M(U_{2d,i,j}, \xi, \gamma)$ and the upwinding formula

$$M_{i,j,k,l} = M_{2d,i,j} \zeta_{k,l} \mathbb{1}_{\zeta_{k,l} \geq 0} + M_{2d,k,l} \zeta_{k,l} \mathbb{1}_{\zeta_{k,l} \leq 0},$$

where $\zeta_{k,l} = (\xi, \gamma)^T \cdot \mathbf{n}_{k,l}$, $\mathbf{n}_{k,l}$ for $(k,l) \in K_{i,j}$ being the outward normal to the contour $\partial C_{i,j}$. The implicit Euler scheme applied to the kinetic interpretation (2.46) gives the kinetic scheme

$$f_{2d,i,j}^{n+1} = M_{2d,i,j}^n - \frac{\Delta t^n}{\Delta x} \sum_{(k,l) \in K_{i,j}} \left(f_{2d,i,j}^{n+1} \zeta_{k,l} \mathbb{1}_{\zeta_{k,l} \geq 0} + f_{2d,k,l}^{n+1} \zeta_{k,l} \mathbb{1}_{\zeta_{k,l} \leq 0} \right). \tag{2.47}$$

Denoting

$$f_{2d} = (f_{2d,1,1}, f_{2d,2,1}, \ldots, f_{2d,P,1}, f_{2d,1,2}, \ldots)^T,$$

the kinetic scheme (2.47) also writes

$$\left(\mathbf{I}_{P^2} + \frac{\Delta t^n}{\Delta x} \mathbf{L}_{P^2} \right) f_{2d}^{n+1} = M_{2d} + \frac{\Delta t^n}{\Delta x} B_{2d}^{n+1},$$

where we have used the particular geometry of the mesh and with $\mathbf{I}_{P^2}$ is the identity matrix of

length P^2, B_{2d}^{n+1} accounts for the boundary conditions and the block matrix $\mathbf{L}_{P^2}$ is defined by

$$\mathbf{L}_{P^2} = \begin{pmatrix} \mathbf{D}_{\xi,\gamma} & \mathbf{N}_\gamma^+ & 0 & \dots & \dots \\ \mathbf{N}_\gamma^- & \mathbf{D}_{\xi,\gamma} & \mathbf{N}_\gamma^+ & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \mathbf{N}_\gamma^- & \mathbf{D}_{\xi,\gamma} & \mathbf{N}_\gamma^+ \\ \dots & \dots & 0 & \mathbf{N}_\gamma^- & \mathbf{D}_{\xi,\gamma} \end{pmatrix},$$

where $\mathbf{D}_{\xi,\gamma}, \mathbf{N}_\gamma^\pm$ are $P \times P$ matrices defined by $\mathbf{N}_\gamma^+ = -\gamma \mathbb{1}_{\gamma \geq 0} \mathbf{I}_P$, $\mathbf{N}_\gamma^- = \gamma \mathbb{1}_{\gamma \leq 0} \mathbf{I}_P$ and

$$\mathbf{D}_{\xi,\gamma} = \begin{pmatrix} |\xi| + |\gamma| & \xi \mathbb{1}_{\xi \leq 0} & 0 & \dots & \dots \\ -\xi \mathbb{1}_{\xi \geq 0} & |\xi| + |\gamma| & \xi \mathbb{1}_{\xi \leq 0} & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & -\xi \mathbb{1}_{\xi \geq 0} & |\xi| + |\gamma| & \xi \mathbb{1}_{\xi \leq 0} \\ \dots & \dots & 0 & -\xi \mathbb{1}_{\xi \geq 0} & |\xi| + |\gamma| \end{pmatrix}.$$

Since the matrix

$$\mathbf{I}_{P^2} + \frac{\Delta t^n}{\Delta x} \mathbf{L}_{P^2},$$

has the same structure as the matrix $\mathbf{I} + \sigma \mathbf{L}$ studied in Lemma 4, the results of Proposition 7 and Proposition 8 are valid.

We do not give the explicit formula neither for the inverse of the matrix $\mathbf{I}_{P^2} + \frac{\Delta t^n}{\Delta x} \mathbf{L}_{P^2}$ nor for the numerical fluxes at the macroscopic level.

2.5 An iterative resolution scheme

The kinetic scheme (2.19) requires to solve a linear system and in the previous section, we have seen that it was possible to have an analytic expression for the inverse of the matrix

$$\mathbf{I} + \sigma \mathbf{L}.$$

For the numerical approximation of PDEs e.g. in finite elements methods when the linear system to solve is large an iterative strategy is singled out compared to a direct inversion of the matrix. We propose to follow the same idea here, with mainly two benefits. First it will allow us to use the half disk Maxwellian (2.2), for which we recall the integrals (2.31) could not be computed analytically in the case of the fully implicit kinetic scheme. This is important as it will enable to prove some discrete entropy inequality at the macroscopic scale thanks to (2.8), while having an explicit writting of the update. The second advantage lies in the possibility to couple the iterative strategy with the hydrostatic reconstruction to obtain a well balanced treatment for varying bottoms, which we will discuss in the next section.

More precisely using a Gauss-Jacobi type decomposition, let us rewrite

$$\mathbf{I} + \sigma \mathbf{L} = \mathbf{D} - \mathbf{N},$$

where $\mathbf{D}$ and $\mathbf{N}$ are two matrices from $\mathbb{R}^{P \times P}$ with $\mathbf{D}$ is invertible. Then the scheme (2.19) also

writes

$$f^{n+1} = \mathbf{D}^{-1}\mathbf{N}f^{n+1} + \mathbf{D}^{-1}(M + \sigma B^{n+1}),$$

and if it converges, the sequence $\{f^{n+1,k}\}_{k\in\mathbb{N}}$ defined by

$$\mathbf{D}f^{n+1,k+1} = \mathbf{N}f^{n+1,k} + M + \sigma B^{n+1},$$

converges towards the solution of (2.19).

2.5.1 Case without topography

In this section, we study this iterative strategy with the particular choice

$$\mathbf{D} = (1+\alpha)\mathbf{I}, \quad \text{and} \quad \mathbf{N} = \alpha\mathbf{I} - \sigma\mathbf{L},$$

when the bathymetry is flat and where $\alpha \in \mathbb{R}_+$ is a relaxation parameter. When developped, this iterative process reads:

$$\begin{cases} f^{n+1,0} = M \\ (1+\alpha)f^{n+1,k+1} = (\alpha\mathbf{I} - \sigma\mathbf{L})f^{n+1,k} + M + \sigma B^{n+1,k} \\ \forall 1 \leq i \leq P,\ U_i^{n+1,k} = \displaystyle\int_{\mathbb{R}} \begin{pmatrix} 1 \\ \xi \end{pmatrix} f_i^{n+1,k}(\xi)\,\mathrm{d}\xi \end{cases}, \tag{2.48}$$

with $B^{n+1,k}$ the boundary condition associated with the macroscopic state $U^{n+1,k}$ as explained in Section 2.3.4. The following Proposition highlights the main compromise linked with such an iterative approach, which is the requirement for a CFL condition in order for the method to converge.

Proposition 9. *Assume that $B^{n+1,k}$ remains constant equal to B^n for any k in $\mathbb{N}$. Then (2.48) defines an arithmetico-geometric sequence which converges if the CFL condition $\sigma|\xi| < 1 + 2\alpha$ holds for all ξ belonging to* $\operatorname{supp} M \cup \operatorname{supp} B^n$.

Proof. By recurrence, we can show that for any $k \in \mathbb{N}$ the support of $f^{n+1,k}$ is included in $\operatorname{supp} M \cup \operatorname{supp} B^n$, which is why we restrict to velocities ξ belonging to this set. Consider f the solution of

$$f = \mathbf{D}^{-1}\mathbf{N}f + \mathbf{D}^{-1}(M + \sigma B^n)\ .$$

The sequence $(g^k)_k$ defined by $g^k = f^{n+1,k} - f$ satisfies $g^{k+1} = \mathbf{D}^{-1}\mathbf{N}g^k$ and converges to zero as soon as the spectral radius of $\mathbf{D}^{-1}\mathbf{N}$ is strictly less than one. Since $\mathbf{D}^{-1}\mathbf{N}$ is a triangular matrix, its eigenvalues are given by its diagonal coefficients, all equal to $(1+\alpha)^{-1}(\alpha - \sigma|\xi|)$. Under the assumption $\sigma|\xi| < 1 + 2\alpha$, this quantity is strictly less than one in absolute value, which concludes the proof. □

Remark 3. *As we did in Section 2.3.4 for the fully implicit scheme, we can replace $B^{n+1,k}$ by B^n in the iterative process (2.48). In fact this constitutes a first order approximation in time since we have $f^{n+1,k} = M + \mathcal{O}(\Delta t)$. Under this simplification, one can drop the assumption $B^{n+1,k} = B^n$ from Proposition 9.*

In practice, we wish to apply an iterative method directly at the macroscopic level. An issue with (2.48) is that the distribution involved in the kinetic flux (i.e. the term in factor of $\sigma\mathbf{L}$) is not a vector of Maxwellians, which prevents us to write the recurrence relation at the macroscopic level since there is no general expression for the numerical flux. To bypass this issue, we propose the

following modification of (2.48), where we replace all occurrences of $f^{n+1,k}$ on the right hand side by a vector of Maxwellians $M^{n+1,k}$, which defines a new sequence $(g^{n+1,k}(\xi))_{k\in\mathbb{N}}$ as

$$\left\{\begin{array}{l} g^{n+1,0}(\xi) = M \\ (1+\alpha)g^{n+1,k+1}(\xi) = (\alpha\mathbf{I} - \sigma^k\mathbf{L})M^{n+1,k} + M + \sigma^k B^{n+1,k} \\ M^{n+1,k+1} = g^{n+1,k+1} + \Delta t^k Q^{n+1,k+1} \end{array}\right. . \tag{2.49}$$

This new iterative process is alternating two stages, the first one being the usual transport step, while the second one is a projection step onto the set of Maxwellians yielding $M^{n+1,k+1}$. In this sense (2.49) is an iterative BGK splitting approach where the projection step doesn't modify the macroscopic quantities of interest since the term $Q^{n+1,k}$ is a vector of collision operators each one satisfying the conservation constraints (2.5). Note that the time stepping Δt^k is made dependent on k as the support of $M^{n+1,k}$ can now change from iteration to iteration. It is important to remark that this iterative scheme differs from (2.48) and we cannot apply the result of Proposition 9. The practical implementation of scheme (2.49) is based on its macroscopic version given for all $1 \leq i \leq P$ by

$$(1+\alpha)U_i^{n+1,k+1} = \alpha U_i^{n+1,k} + U_i - \sigma\Big(\mathcal{F}(U_i^{n+1,k}, U_{i+1}^{n+1,k}) - \mathcal{F}(U_{i-1}^{n+1,k}, U_i^{n+1,k})\Big), \tag{2.50}$$

where the numerical flux F is defined as

$$\mathcal{F}(U_L, U_R) = \int_{\mathbb{R}} \xi \begin{pmatrix} 1 \\ \xi \end{pmatrix} \Big(\mathbb{1}_{\xi>0} M(U_L,\xi) + \mathbb{1}_{\xi<0} M(U_R,\xi)\Big)\,\mathrm{d}\xi, \tag{2.51}$$

and where the vectors $U_0^{n+1,k}, U_{P+1}^{n+1,k}$ appearing for $i \in \{1, P\}$ are respectively functions of $U_1^{n+1,k}$ and $U_P^{n+1,k}$ since the boundary conditions are imposed though a ghost cell strategy fully described in [3,26]. Notice that if the sequence $\{U^{n+1,k}\}_{k\in\mathbb{N}} \subset (\mathbb{R}^2)^P$ from (2.50) converges in $(\mathbb{R}^2)^P$, its limit U^{n+1} then satisfies

$$\forall 1 \leq i \leq P, \quad U_i^{n+1} = U_i^n - \sigma\Big(\mathcal{F}(U_i^{n+1}, U_{i+1}^{n+1}) - \mathcal{F}(U_{i-1}^{n+1}, U_i^{n+1})\Big)$$

by continuity of the numerical flux (2.51). Besides we want to remark that the iterative method described here could have been applied with any other numerical flux at the macroscopic level. However, using a numerical flux different from the kinetic one would have made it very difficult (if possible at all) to prove the forthcoming properties, whereas using (2.51) gives us a favorable setting to perform the proofs. These properties include the preservation of the water height positivity under a CFL condition, and the existence of a discrete entropy equality with dissipation.

Proposition 10. *Assume that the water height vectors h^n and $h^{n+1,k}$ are positive. Then the update $g^{n+1,k+1}$ defined in the iterative scheme (2.49) is positive if for all $1 \leq i \leq P$ the CFL condition $\sigma^k|\xi| \leq \alpha + M_i/M_i^{n+1,k}$ holds for any ξ belonging to $\operatorname{supp} M^{n+1,k}$. As a direct consequence, the water height vector $h^{n+1,k+1}$ from scheme (2.50) is positive under these assumptions.*

We postpone the proof of Proposition 10 to the next section, where it is generalized to the case with varying bottom in Proposition 12.

Proposition 11. *Let us denote $\Xi = \operatorname{supp} M^{n+1,k}$, with M the half-disk Maxwellian (2.2). The*

kinetic entropy of the iterative scheme (2.49) satisfies the following inequality

$$H_0(M_i^{n+1,k+1}) \leq \frac{H_0(M_i) + \alpha H_0(M_i^{n+1,k})}{1+\alpha} - \frac{\sigma^k \xi}{1+\alpha}\Big(H_{0,i+1/2}^{n+1,k} - H_{0,i-1/2}^{n+1,k}\Big) + \Delta t^k \eta'(U_i^{n+1,k+1}) \cdot \begin{pmatrix} 1 \\ \xi \end{pmatrix} Q_i^{n+1,k+1} + D_i^{n+1,k+1}, \tag{2.52}$$

with $Q_i^{n+1,k+1} = (M_i^{n+1,k+1} - g_i^{n+1,k+1})/\Delta t^k$ a collision operator verifying the conservation constraints (2.5). The interfacial kinetic entropies $H_{0,i\pm1/2}^{n+1,k}$ are given by

$$H_{0,i-1/2}^{n+1,k} = \mathbb{1}_{\xi>0} H_0(M_{i-1}^{n+1,k},\xi) + \mathbb{1}_{\xi<0} H_0(M_i^{n+1,k},\xi),$$
$$H_{0,i+1/2}^{n+1,k} = \mathbb{1}_{\xi>0} H_0(M_i^{n+1,k},\xi) + \mathbb{1}_{\xi<0} H_0(M_{i+1}^{n+1,k},\xi),$$

and the term $D_i^{n+1,k+1}$ is given by

$$D_i^{n+1,k+1} = -\frac{1}{1+\alpha}\Psi(M_i^{n+1,k+1}, M_i) - \frac{\alpha - \sigma^k|\xi|\mathbb{1}_\Xi}{1+\alpha}\Psi(M_i^{n+1,k+1}, M_i^{n+1,k}) - \frac{\sigma^k|\xi|\mathbb{1}_\Xi}{1+\alpha}\Psi(M_i^{n+1,k+1}, M_{i\pm1}^{n+1,k}),$$

where we recall that the function Ψ defined in (2.22) is positive on $\mathbb{R}_+^2$ and where $i \pm 1 = i - \operatorname{sgn}\xi$. As a consequence, if for any integer k the CFL condition

$$\forall \xi \in \operatorname{supp} M^{n+1,k},\ \sigma^k|\xi| \leq \alpha \tag{2.53}$$

holds, then $D_i^{n+1,k+1}$ is a dissipation term with negative sign and at each iteration the kinetic entropy is dissipated up to terms that are macroscopically zero, that is to say there exists a kinetic entropy flux $\widetilde{H}_{0,i+1/2}^{n+1,k}$, a negative dissipation $\widetilde{D}_i^{n+1,k+1}$ and a term $\widetilde{Z}_i^{n+1,k+1}(\xi)$ whose integral over $\xi \in \mathbb{R}$ is zero such that

$$H_0(M_i^{n+1,k+1},\xi) \leq H_0(M_i,\xi) - \sigma^k\xi\Big(\widetilde{H}_{0,i+1/2}^{n+1,k} - \widetilde{H}_{0,i-1/2}^{n+1,k}\Big) + \widetilde{D}_i^{n+1,k+1} + \widetilde{Z}_i^{n+1,k+1}. \tag{2.54}$$

Before giving the proof we have the remark below.

Remark 4. *Even when the CFL condition (2.53) is not satisfied, we can ensure that the scheme (2.49) satisfies a discrete entropy inequality from some rank k assuming the convergence of the method. In fact, multiplying inequality (2.52) by $1+\alpha$ it is possible to write*

$$\begin{aligned} H_0(M_i^{n+1,k+1},\xi) \leq\ & \\ & H_0(M_i,\xi) - \sigma^k\xi\Big(H_{0,i+1/2}^{n+1,k} - H_{0,i-1/2}^{n+1,k}\Big) + (1+\alpha)\Delta t^k \eta'(U_i^{n+1,k+1}) \cdot \begin{pmatrix} 1 \\ \xi \end{pmatrix} Q_i^{n+1,k+1} \\ & - \Psi(M_i^{n+1,k+1}, M_i) - \sigma^k|\xi|\mathbb{1}_\Xi \Psi(M_i^{n+1,k+1}, M_{i\pm1}^{n+1,k}) \\ & + \alpha\Big(H_0(M_i^{n+1,k},\xi) - H_0(M_i^{n+1,k+1},\xi)\Big) - (\alpha - \sigma^k|\xi|\mathbb{1}_\Xi)\Psi(M_i^{n+1,k+1}, M_i^{n+1,k}). \end{aligned} \tag{2.55}$$

On the right hand side of (2.55), the quantity

$$(1+\alpha)\Delta t^k \eta'(U_i^{n+1,k+1}) \cdot \begin{pmatrix} 1 \\ \xi \end{pmatrix} Q_i^{n+1,k+1}$$

does not cause any issue as it vanishes upon integration over $\xi \in \mathbb{R}$. In fact we recall that the collision term $Q_i^{n+1,k+1}$ satisfies the conservation constraints (2.5), meaning that its integral against $(1,\xi)^T$ vanishes. Therefore in (2.55) the only problematic terms are contained in the last line, as their sign can be positive since we do not assume $\sigma^k|\xi|\mathbb{1}_\Xi \leq \alpha$ anymore. Nevertheless, by regularity of $H_0(\cdot,\xi)$ and by definition (2.22) of Ψ, these terms write as a $\mathcal{O}(M_i^{n+1,k+1} - M_i^{n+1,k})$ and vanish as $k \to \infty$ assuming the method converges. As a consequence, from some rank k these two terms become negligible compared to $-\Psi(M_i^{n+1,k+1}, M_i) < 0$ which remains bounded away from zero, and we recover a dissipation with negative sign. (Note that if $M^{n+1,k}$ was converging to M as $k \to \infty$, it would imply that M solves the fixed point problem and thus $M^{n+1,k} = M$ for all k; putting aside this trivial case, this is why $\Psi(M_i^{n+1,k+1}, M_i)$ remains bounded away from zero).

Proof of Proposition 11. First, we remark that for any ξ there holds

$$\partial_f H_0(M_i^{n+1,k+1},\xi) Q_i^{n+1,k+1} \leq \eta'(U_i^{n+1,k+1}) \cdot \begin{pmatrix} 1 \\ \xi \end{pmatrix} Q_i^{n+1,k+1}, \tag{2.56}$$

which is implied by the relation (2.9) and by the fact that $Q_i^{n+1,k+1} = (M_i^{n+1,k+1} - g_i^{n+1,k+1})/\Delta t^k$ is negative for any $\xi \notin \operatorname{supp} M_i^{n+1,k+1}$. As a consequence of (2.56), inequality (2.52) (and equivalently (2.55)) is verified if there holds

$$\begin{aligned} H_0(M_i^{n+1,k+1}) = \frac{H_0(M_i) + \alpha H_0(M_i^{n+1,k})}{1+\alpha} - \frac{\sigma^k \xi}{1+\alpha}\Big(H_{0,i+1/2}^{n+1,k} - H_{0,i-1/2}^{n+1,k}\Big) \\ + \Delta t^k \partial_f H_0(M_i^{n+1,k+1}) Q_i^{n+1,k+1} + D_i^{n+1,k+1}, \end{aligned} \tag{2.57}$$

To prove equality (2.57) we write the subiteration (2.49) as

$$M_i^{n+1,k+1} = \frac{1}{1+\alpha}\Big(M_i + (\alpha - \sigma^k|\xi|\mathbb{1}_\Xi) M_i^{n+1,k} + \sigma^k|\xi|\mathbb{1}_\Xi M_{i\pm1}^{n+1,k}\Big) + \Delta t^k Q_i^{n+1,k+1}, \tag{2.58}$$

with $i \pm 1 = i - \operatorname{sign}\xi$. Applying Lemma 5 for $a = M_i^{n+1,k+1}$ and $b = M_{i\pm1}^{n+1,k}, M_i, M_i^{n+1,k}$, we respectively get:

$$H_0(M_{i\pm1}^{n+1,k}) = H_0(M_i^{n+1,k+1}) + \partial_f H_0(M_i^{n+1,k+1})(M_{i\pm1}^{n+1,k} - M_i^{n+1,k+1}) + \Psi(M_i^{n+1,k+1}, M_{i\pm1}^{n+1,k}) \tag{2.59}$$

$$H_0(M_i^{n+1,k}) = H_0(M_i^{n+1,k+1}) + \partial_f H_0(M_i^{n+1,k+1})(M_i^{n+1,k} - M_i^{n+1,k+1}) + \Psi(M_i^{n+1,k+1}, M_i^{n+1,k}) \tag{2.60}$$

$$H_0(M_i) = H_0(M_i^{n+1,k+1}) + \partial_f H_0(M_i^{n+1,k+1})(M_i - M_i^{n+1,k+1}) + \Psi(M_i^{n+1,k+1}, M_i) \tag{2.61}$$

Performing the linear combination

$$\frac{1}{1+\alpha}\Big((2.61) + (\alpha - \sigma^k|\xi|\mathbb{1}_\Xi)(2.60)) + \sigma^k|\xi|\mathbb{1}_\Xi (2.59)\Big)$$

and using (2.58) we obtain

$$\frac{1}{1+\alpha}\Big(H_0(M_i)+(\alpha-\sigma^k|\xi|\mathbb{1}_\Xi)H_0(M_i^{n+1,k})+\sigma^k|\xi|\mathbb{1}_\Xi H_0(M_{i\pm1}^{n+1,k})\Big)=$$
$$H_0(M_i^{n+1,k+1})-\Delta t^k\partial_f H_0(M_i^{n+1,k+1})Q_i^{n+1,k+1}+\frac{1}{1+\alpha}\Psi(M_i^{n+1,k+1},M_i)$$
$$+\frac{\alpha-\sigma^k|\xi|\mathbb{1}_\Xi}{1+\alpha}\Psi(M_i^{n+1,k+1},M_i^{n+1,k})+\frac{\sigma^k|\xi|\mathbb{1}_\Xi}{1+\alpha}\Psi(M_i^{n+1,k+1},M_{i\pm1}^{n+1,k})$$

which corresponds to equality (2.57) after rearranging the terms.

Next we proceed by induction to show that the kinetic entropy is dissipated at every iteration assuming the CFL condition (2.53) holds for any integer k. The key argument is that under this CFL condition, the term $D_i^{n+1,k+1}$ defines a convex combination of negative quantities, and is thus negative. The initialization is obvious since we have $M_i^{n+1,0}=M_i$, so we focus on the recurrence. We want to show that (2.54) holds at some rank $k\geq 1$ assuming that it is satisfied at rank $k-1$. Under this assumption we can develop (2.52) as

$$H_0(M_i^{n+1,k+1})\leq$$
$$\frac{1}{1+\alpha}\Big(H_0(M_i)+\alpha\Big(H_0(M_i)-\sigma^k\xi\Big(\widetilde{H}_{0,i+1/2}^{n+1,k-1}-\widetilde{H}_{0,i-1/2}^{n+1,k-1}\Big)+\widetilde{D}_i^{n+1,k}+\widetilde{Z}_i^{n+1,k}\Big)\Big)$$
$$-\frac{\sigma^k\xi}{1+\alpha}\Big(H_{0,i+1/2}^{n+1,k}-H_{0,i-1/2}^{n+1,k}\Big)+\Delta t^k\eta'(U_i^{n+1,k+1})\cdot\begin{pmatrix}1\\ \xi\end{pmatrix}Q_i^{n+1,k+1}+D_i^{n+1,k+1},$$

with $\widetilde{Z}_i^{n+1,k}$ and $\eta'(U_i^{n+1,k+1})\cdot\begin{pmatrix}1\\ \xi\end{pmatrix}Q_i^{n+1,k+1}$ macroscopically zero as per Remark 4. Therefore we have

$$H_0(M_i^{n+1,k+1})\leq H_0(M_i)-\frac{\sigma^k\xi}{1+\alpha}\Big(H_{0,i+1/2}^{n+1,k}+\alpha\widetilde{H}_{0,i+1/2}^{n+1,k-1}-H_{0,i-1/2}^{n+1,k}-\alpha\widetilde{H}_{0,i-1/2}^{n+1,k-1}\Big)$$
$$+\frac{\alpha}{1+\alpha}\widetilde{Z}_i^{n+1,k}+\Delta t^k\eta'(U_i^{n+1,k+1})\cdot\begin{pmatrix}1\\ \xi\end{pmatrix}Q_i^{n+1,k+1}+\frac{\alpha}{1+\alpha}\widetilde{D}_i^{n+1,k}+D_i^{n+1,k+1}$$

and the proof is complete by setting

$$\widetilde{H}_{0,i+1/2}^{n+1,k}=\frac{1}{1+\alpha}\Big(H_{0,i+1/2}^{n+1,k}+\alpha\widetilde{H}_{0,i+1/2}^{n+1,k-1}\Big),\quad \widetilde{D}_i^{n+1,k+1}=\frac{\alpha}{1+\alpha}\widetilde{D}_i^{n+1,k}+D_i^{n+1,k+1},$$
$$\widetilde{Z}_i^{n+1,k+1}=\frac{\alpha}{1+\alpha}\widetilde{Z}_i^{n+1,k}+\Delta t^k\eta'(U_i^{n+1,k+1})\cdot\begin{pmatrix}1\\ \xi\end{pmatrix}Q_i^{n+1,k+1}.$$

□

2.5.2 Case with topography

To deal with varying bathymetries in a well balanced way, the hydrostatic reconstruction technique introduced by Audusse et al. in [8] can be used. It is based on the reconstruction of the water height according to a procedure that we briefly recall. Let $U_i=(h_i,h_iu_i)^T\in\mathbb{R}^2$ denote the vector of quantities of interest over cell $1\leq i\leq P$, with P the number of interior cells and with ghost cells corresponding to indices 0 and $P+1$. The reconstructed states are vectors from $\mathbb{R}^2$ defined on the

left and right neighborhoods of each cell interface as follows:

$$\forall 1 \leq i \leq P, \quad U_{i+1/2-} = \begin{pmatrix} h_{i+1/2-} \\ h_{i+1/2-}u_i \end{pmatrix}, \quad U_{i-1/2+} = \begin{pmatrix} h_{i-1/2+} \\ h_{i-1/2+}u_i \end{pmatrix} .$$

The reconstructed interfacial water heights are given by

$$h_{i-1/2+} = (h_i + z_i - z_{i-1/2})_+, \quad h_{i+1/2-} = (h_i + z_i - z_{i+1/2})_+, \tag{2.62}$$

with the interfacial bathymetry $z_{i+1/2} = \max(z_i, z_{i+1})$. The truly implicit kinetic scheme we are considering reads as below

$$U_i^{n+1} = U_i^n - \sigma(F_{i+1/2-}^{n+1} - F_{i-1/2+}^{n+1}), \tag{2.63}$$

with $\sigma = \Delta t/\Delta x$ and the numerical fluxes decomposed as:

$$F_{i+1/2-}^{n+1} = \mathcal{F}(U_{i+1/2-}^{n+1}, U_{i+1/2+}^{n+1}) + \frac{g}{2}\begin{pmatrix} 0 \\ (h_i^{n+1})^2 - (h_{i+1/2-}^{n+1})^2 \end{pmatrix}$$
$$F_{i-1/2+}^{n+1} = \mathcal{F}(U_{i-1/2-}^{n+1}, U_{i-1/2+}^{n+1}) + \frac{g}{2}\begin{pmatrix} 0 \\ (h_i^{n+1})^2 - (h_{i-1/2+}^{n+1})^2 \end{pmatrix}$$

We recall that in our case the upwinding of the numerical flux $\mathcal{F}$ is induced at the kinetic level according to definition (2.51).

Because the update (2.63) is nonlinear, it is not possible to solve it analytically. Instead we will consider an iterative process with a relaxation parameter $\alpha > 0$ similarly to the one from Section 2.5.1. At the kinetic level, this process consists in introducing for any real ξ the sequence $(f^{n+1,k}(\xi))_{k\in\mathbb{N}} \subset \mathbb{R}_+^P$ initialized with $f^{n+1,0}(\xi) = M(U^n, \xi)$ and defined recursively as:

$$\begin{aligned} &(1+\alpha) f_i^{n+1,k+1} = \\ &\quad M_i + \alpha M_i^{n+1,k} - \sigma^k \xi \Big(\mathbb{1}_{\xi<0}(M_{i+1/2+}^{n+1,k} - M_{i-1/2+}^{n+1,k}) + \mathbb{1}_{\xi>0}(M_{i+1/2-}^{n+1,k} - M_{i-1/2-}^{n+1,k}) \Big) \\ &\quad + \sigma^k(\xi - u_i^{n+1,k})(M_{i+1/2-}^{n+1,k} - M_i^{n+1,k}) - \sigma^k(\xi - u_i^{n+1,k})(M_{i-1/2+}^{n+1,k} - M_i^{n+1,k}), \end{aligned} \tag{2.64}$$

where the last line of (2.64) corresponds to the kinetic interpretation of the topography source term, see [5]. In the above we used the notation

$$M_\square^{n+1,k} = M(U_\square^{n+1,k}, \xi), \quad h^{n+1,k} = \int_{\mathbb{R}} f^{n+1,k}(\xi)\,\mathrm{d}\xi, \quad (hu)^{n+1,k} = \int_{\mathbb{R}} \xi\, f^{n+1,k}(\xi)\,\mathrm{d}\xi,$$

where the square symbol "$\square$" in subscript can be replaced by i (centered value) or $i \pm 1/2\mp$ (reconstructed interfacial value). Making use of the relations

$$\int_{\mathbb{R}} \begin{pmatrix} 1 \\ \xi \end{pmatrix} (\xi - u_i)(M_i - M_{i+1/2-})\,\mathrm{d}\xi = \begin{pmatrix} 0 \\ \frac{g}{2}(h_i^2 - h_{i+1/2-}^2) \end{pmatrix}$$
$$\int_{\mathbb{R}} \begin{pmatrix} 1 \\ \xi \end{pmatrix} (\xi - u_i)(M_i - M_{i-1/2+})\,\mathrm{d}\xi = \begin{pmatrix} 0 \\ \frac{g}{2}(h_i^2 - h_{i-1/2+}^2) \end{pmatrix}$$

given in [5], the macroscopic version of scheme (2.64) obtained by integrating the update against

the vector $(1,\xi)^T$ reads

$$(1+\alpha)U_i^{n+1,k+1} = U_i + \alpha U_i^{n+1,k} - \sigma^k(F_{i+1/2-}^{n+1,k} - F_{i-1/2+}^{n+1,k}). \tag{2.65}$$

If the iterative process (2.65) converges, we recover the implicit scheme (2.63) by setting the macroscopic update as $U^{n+1} = \lim_{k\to\infty} U^{n+1,k}$ and $\sigma = \lim_{k\to\infty}\sigma^k$. In practice we will not be able to compute this limit, hence we will set $U^{n+1} = U^{n+1,k}$ and $\sigma = \sigma^k$ for k large enough, hoping that $U^{n+1,k} \approx U^{n+1,\infty}$. One should also notice that when the bathymetry is flat the hydrostatic reconstruction becomes transparent, and the scheme (2.65) coincides with (2.50).

Finally we propose an estimate for the entropy associated with the scheme (2.64), as well as a CFL condition to ensure its positivity.

Proposition 12. *The following properties are satisfied by the scheme (2.64):*

(i) Assume that the water height vectors h^n and $h^{n+1,k}$ are positive. Then the update $h^{n+1,k+1}$ defined in the iterative scheme (2.65) is positive if for all $1 \le i \le P$ the CFL condition $\sigma^k|\xi| \le \alpha + M_i/M_i^{n+1,k}$ holds for any ξ belonging to $\operatorname{supp} M^{n+1,k}$.

(ii) The kinetic entropy of the iterative process (2.64) verifies the following kinetic entropy inequality

$$\begin{aligned} &H(M_i^{n+1,k+1}, z_i) \le \\ &\quad H(M_i, z_i) - \sigma^k(\widetilde{G}_{i+1/2-}^{n+1,k} - \widetilde{G}_{i-1/2+}^{n+1,k}) + (1+\alpha)\Delta t^k\Big(\eta'(U_i^{n+1,k+1})\cdot\begin{pmatrix}1\\ \xi\end{pmatrix} + gz_i\Big)Q_i^{n+1,k+1} \\ &\quad + \alpha(H(M_i^{n+1,k}, z_i) - H(M_i^{n+1,k+1}, z_i)) \\ &\quad + (1+\alpha)(\Psi(M_i^{n+1,k}, f_i^{n+1,k+1}) - \Psi(M_i^{n+1,k+1}, f_i^{n+1,k+1})) - \Psi(M_i^{n+1,k}, M_i). \end{aligned} \tag{2.66}$$

with $Q_i^{n+1,k+1} = (M_i^{n+1,k+1} - f_i^{n+1,k+1})/\Delta t^k$ a collision term satisfying the conservation constraints (2.5), and where

$$\begin{aligned} \widetilde{G}_{i+1/2-}^{n+1,k} &= \xi\mathbb{1}_{\xi<0}H(M_{i+1/2+}^{n+1,k}, z_{i+1/2}) + \xi\mathbb{1}_{\xi>0}H(M_{i+1/2-}^{n+1,k}, z_{i+1/2}) + \xi H(M_i^{n+1,k}, z_i) \\ &\quad - \xi H(M_{i+1/2-}^{n+1,k}, z_{i+1/2}) + \Big(\eta'(U_i^{n+1,k})\cdot\begin{pmatrix}1\\ \xi\end{pmatrix} + gz_i\Big)\Big(\xi M_{i+1/2-}^{n+1,k} - \xi M_i^{n+1,k} \\ &\quad + (\xi - u_i^{n+1,k})(M_i^{n+1,k} - M_{i+1/2-}^{n+1,k})\Big), \end{aligned} \tag{2.67}$$

$$\begin{aligned} \widetilde{G}_{i-1/2+}^{n+1,k} &= \xi\mathbb{1}_{\xi<0}H(M_{i-1/2+}^{n+1,k}, z_{i-1/2}) + \xi\mathbb{1}_{\xi>0}H(M_{i-1/2-}^{n+1,k}, z_{i-1/2}) + \xi H(M_i^{n+1,k}, z_i) \\ &\quad - \xi H(M_{i-1/2+}^{n+1,k}, z_{i-1/2}) + \Big(\eta'(U_i^{n+1,k})\cdot\begin{pmatrix}1\\ \xi\end{pmatrix} + gz_i\Big)\Big(\xi M_{i-1/2+}^{n+1,k} - \xi M_i^{n+1,k} \\ &\quad + (\xi - u_i^{n+1,k})(M_i^{n+1,k} - M_{i-1/2+}^{n+1,k})\Big). \end{aligned} \tag{2.68}$$

We recall that Ψ defined in (2.22) is positive and that the entropy is $\eta(U) = hu^2/2 + gh^2/2$.

Before giving the proof we make the following remark.

Remark 5. *In inequality (2.66) the difference $\widetilde{G}_{i+1/2-}^{n+1,k} - \widetilde{G}_{i-1/2+}^{n+1,k}$ is non conservative at the kinetic level, but becomes conservative when it is integrated over $\xi \in \mathbb{R}$. This is due to the fact that the last two lines of (2.67) and (2.68) are macroscopically zero, see [5] Proposition 3.1. Furthermore, we*

reiterate the comments made in Remark 4 which are to say that in (2.66) the term

$$(1+\alpha)\Delta t^k\Big(\eta'(U_i^{n+1,k+1})\cdot\begin{pmatrix}1\\ \xi\end{pmatrix}+gz_i\Big)Q_i^{n+1,k+1}$$

since $Q_i^{n+1,k+1}$ is a collision term that satisfies the conservation constraints (2.5). Besides, assuming the method converges as $k\to\infty$, the quantity

$$\alpha\Big(H(M_i^{n+1,k},z_i)-H(M_i^{n+1,k+1},z_i)\Big)+(1+\alpha)\Big(\Psi(M_i^{n+1,k},f_i^{n+1,k+1})-\Psi(M_i^{n+1,k+1},f_i^{n+1,k+1})\Big)$$

will eventually become negligible compared to $-\Psi(M_i^{n+1,k},M_i)<0$ from some rank k, similarly to the argument from Remark 4. Integrating inequality (2.66) over $\xi\in\mathbb{R}$, this implies that there exists $K\in\mathbb{N}$ such that for any $k\geq K$ the fully discrete entropy inequality

$$\eta(U_i^{n+1,k+1})\leq\eta(U_i^n)-\sigma^k\bigg(\int_{\mathbb{R}}\xi\,H_{i+1/2}^{n+1,k}(\xi)\,\mathrm{d}\xi-\int_{\mathbb{R}}\xi\,H_{i-1/2}^{n+1,k}(\xi)\,\mathrm{d}\xi\bigg)\tag{2.69}$$

is satisfied at the macroscopic level. Summing inequality (2.69) over every cell $1\leq i\leq P$ we obtain the dissipation of the total energy up to boundary fluxes

$$\frac{1}{\Delta t^k}\sum_{i=1}^{P}\Big(\eta(U_i^{n+1,k+1})-\eta(U_i^n)\Big)+\frac{1}{\Delta x}\bigg(\int_{\mathbb{R}}\xi\,H_{P+1/2}^{n+1,k}(\xi)\,\mathrm{d}\xi-\int_{\mathbb{R}}\xi\,H_{1/2}^{n+1,k}(\xi)\,\mathrm{d}\xi\bigg)\leq 0.\tag{2.70}$$

In addition to the usual tolerance criterion where the iterations are stopped whenever two successive iterates are sufficiently close to each other, we can use (2.70) as a complementary condition to ensure the dissipation of total energy.

Proof of Proposition 12. The proof makes use of the kinetic writing (2.64) of the scheme (2.65).

(i) Remarking that the quantity $\sigma^k(\xi-u_i^{n+1,k})(M_{i+1/2-}^{n+1,k}-M_{i-1/2+}^{n+1,k})$ appearing in the last line of (2.64) defines an odd function of $\xi-u_i^{n+1,k}$, its integral over $\xi\in\mathbb{R}$ vanishes and we have at the macroscopic level

$$(1+\alpha)h_i^{n+1,k+1}=\int_{\mathbb{R}}\Big(M_i+\alpha M_i^{n+1,k}-\sigma^k\xi(M_{i+1/2}^{n+1,k}-M_{i-1/2}^{n+1,k})\Big)\,\mathrm{d}\xi\ .$$

Thus it is enough to prove the positivity of the integrand, whose developed form is

$$M_i+\alpha M_i^{n+1,k}-\sigma^k\xi\Big(\mathbb{1}_{\xi>0}M_{i+1/2-}^{n+1,k}-\mathbb{1}_{\xi<0}M_{i-1/2+}^{n+1,k}\Big)+\sigma^k\xi\Big(\mathbb{1}_{\xi>0}M_{i-1/2-}^{n+1,k}-\mathbb{1}_{\xi<0}M_{i+1/2+}^{n+1,k}\Big).$$

By definition of the water height reconstruction (2.62), we have the inequalities $h_{i+1/2-}^{n+1,k}\leq h_i^{n+1,k}$ and $h_{i-1/2+}^{n+1,k}\leq h_i^{n+1,k}$. As a consequence $M_{i+1/2-}^{n+1,k}\leq M_i^{n+1,k}$ and $M_{i-1/2+}^{n+1,k}\leq M_i^{n+1,k}$, which allows us to bound the integrand from below by

$$M_i+\alpha M_i^{n+1,k}-\sigma^k|\xi|M_i^{n+1,k}\ .$$

If ξ does not belong to $\operatorname{supp}M^{n+1,k}$ this quantity equals M_i which is positive. Otherwise, it is made positive under the condition $\sigma^k|\xi|\leq\alpha+M_i^0/M_i^{n+1,k}$ which gives the desired result.

(ii) We start to rewrite (2.64) as

$$(1+\alpha)(f_i^{n+1,k+1}-M_i^{n+1,k}) = (M_i-M_i^{n+1,k}) - \sigma^k\xi(M_{i+1/2}^{n+1,k}-M_{i-1/2}^{n+1,k}) \\ + \sigma^k(\xi-u_i^{n+1,k})(M_{i+1/2-}^{n+1,k}-M_{i-1/2+}^{n+1,k}). \quad (2.71)$$

The strategy is to multiply (2.71) by $\partial_f H(M_i^{n+1,k},z_i)$ and to write

$$\partial_f H(M_i^{n+1,k},z_i)\Big[(1+\alpha)(f_i^{n+1,k+1}-M_i^{n+1,k}) - (M_i-M_i^{n+1,k})\Big] = \\ -\sigma^k\partial_f H(M_i^{n+1,k},z_i)\Big[\xi(M_{i+1/2}^{n+1,k}-M_{i-1/2}^{n+1,k}) + \delta M_{i+1/2-}^{n+1,k} - \delta M_{i-1/2+}^{n+1,k}\Big], \quad (2.72)$$

where we defined

$$\delta M_{i+1/2-}^{n+1,k} = (\xi-u_i^{n+1,k})(M_i^{n+1,k}-M_{i+1/2-}^{n+1,k}) \\ \delta M_{i-1/2+}^{n+1,k} = (\xi-u_i^{n+1,k})(M_i^{n+1,k}-M_{i-1/2+}^{n+1,k}).$$

We apply Lemma 5 to the left hand side of (2.72) to get

$$\partial_f H(M_i^{n+1,k},z_i)\Big[(1+\alpha)(f_i^{n+1,k+1}-M_i^{n+1,k}) - (M_i-M_i^{n+1,k})\Big] = \\ (1+\alpha)\Big(H(f_i^{n+1,k+1},z_i) - H(M_i^{n+1,k},z_i) - \Psi(M_i^{n+1,k},f_i^{n+1,k+1})\Big) \\ - \Big(H(M_i,z_i) - H(M_i^{n+1,k},z_i) - \Psi(M_i^{n+1,k},M_i)\Big). \quad (2.73)$$

Furthermore, an upper bound on the right hand side of (2.72) is obtained by applying Proposition 3.1 from [5] which directly yields

$$-\partial_f H(M_i^{n+1,k},z_i)\Big[\xi(M_{i+1/2}^{n+1,k}-M_{i-1/2}^{n+1,k}) + \delta M_{i+1/2-}^{n+1,k} - \delta M_{i-1/2+}^{n+1,k}\Big] \le \widetilde{G}_{i-1/2+}^{n+1,k} - \widetilde{G}_{i+1/2-}^{n+1,k}, \quad (2.74)$$

with $\widetilde{G}_{i+1/2-}^{n+1,k}$ and $\widetilde{G}_{i-1/2+}^{n+1,k}$ defined by (2.67) and (2.68). Injecting equality (2.73) and inequality (2.74) into (2.72) we obtain

$$(1+\alpha)H(f_i^{n+1,k+1},z_i) \le H(M_i,z_i) - \sigma^k(\widetilde{G}_{i+1/2-}^{n+1,k} - \widetilde{G}_{i-1/2+}^{n+1,k}) + \alpha H(M_i^{n+1,k},z_i) \\ + (1+\alpha)\Psi(M_i^{n+1,k},f_i^{n+1,k+1}) - \Psi(M_i^{n+1,k},M_i) \quad (2.75)$$

Using again Lemma 5 we can also write

$$\begin{aligned} H(f_i^{n+1,k+1},z_i) &= H(M_i^{n+1,k+1},z_i) + \partial_f H(M_i^{n+1,k+1},z_i)(f_i^{n+1,k+1}-M_i^{n+1,k+1}) \\ &\quad + \Psi(M_i^{n+1,k+1},f_i^{n+1,k+1}) \\ &\ge H(M_i^{n+1,k+1},z_i) + \Big(\eta'(U_i^{n+1,k+1})\cdot\begin{pmatrix}1\\ \xi\end{pmatrix} + gz_i\Big)(f_i^{n+1,k+1}-M_i^{n+1,k+1}) \\ &\quad + \Psi(M_i^{n+1,k+1},f_i^{n+1,k+1}), \end{aligned} \quad (2.76)$$

where we used $\partial_f H = \partial_f H_0 + gz$ and (2.56) to get (2.76). Combining this inequality with (2.75)

we finally get

$$
\begin{aligned}
(1+\alpha)H(M_i^{n+1,k+1}, z_i) \leq H(M_i, z_i) - \sigma^k(\widetilde{G}_{i+1/2-}^{n+1,k} - \widetilde{G}_{i-1/2+}^{n+1,k}) + \alpha H(M_i^{n+1,k}, z_i) \\
+ (1+\alpha)\Psi(M_i^{n+1,k}, f_i^{n+1,k+1}) - \Psi(M_i^{n+1,k}, M_i) \\
- (1+\alpha)\Big(\eta'(U_i^{n+1,k+1}) \cdot \begin{pmatrix} 1 \\ \xi \end{pmatrix} + g z_i\Big)(f_i^{n+1,k+1} - M_i^{n+1,k+1}) \\
- (1+\alpha)\Psi(M_i^{n+1,k+1}, f_i^{n+1,k+1}).
\end{aligned}
$$

After rearranging the terms and using $Q_i^{n+1,k+1} = -(f_i^{n+1,k+1} - M_i^{n+1,k+1})/\Delta t^k$ we obtain the desired kinetic entropy inequality (2.66).

□

2.6 Numerical examples

2.6.1 The one dimensional case

We start by evaluating the qualitative properties and the efficiency related to the fully implicit and iterative kinetic schemes in the one dimensional case.

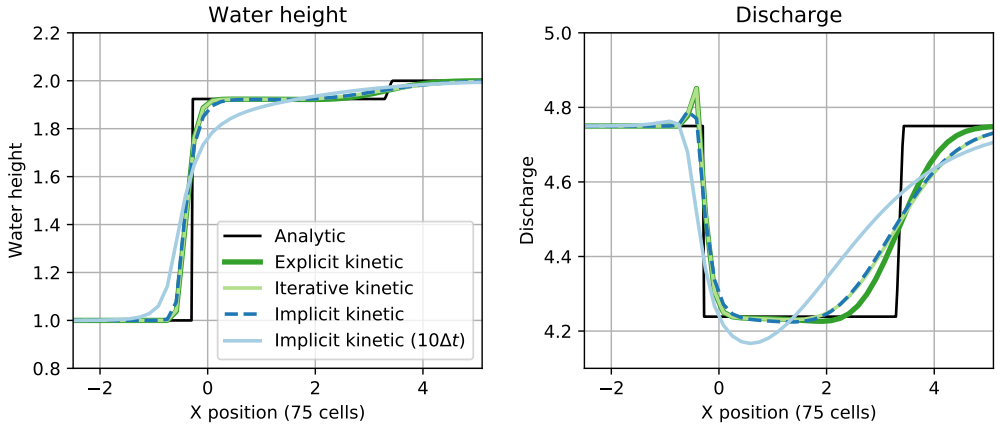

Figure 2.2: Slow moving shock approximated by various kinetic schemes, including explicit, implicit and iterative strategies. The initial condition is given by a Riemann data with discontinuity at position $x = 0$.

Slow moving shock. To assess the efficiency and interest of the implicit scheme (2.18), we perform a numerical test involving a Riemann problem with a slowly moving shock over a flat bottom. This configuration is achieved for a nearly transcritical flow where the material velocity u is positive and satisfies $u - \sqrt{gh} \approx 0$ and $u + \sqrt{gh} \gg 1$. Hence the maximum eigenvalue severely constrains the time step, however a small time step might not be necessary to accurately resolve the slow shock. In Fig. 2.2 we compare several schemes with an explicit time step Δt_{exp} given by the usual CFL condition, as well as the implicit kinetic scheme using a time step $\Delta t_{\text{imp}} = 10\Delta t_{\text{exp}}$. We set $\alpha = 1$ for the iterative scheme (2.50). We notice that in the discharge profile, an oscillation appears downwind of the shock, which is quite pronounced for the explicit and iterative kinetic schemes, and less so

for the fully implicit ones. As expected the implicit scheme using Δt_{imp} strongly diffuses the fast traveling rarefaction. On the other hand the slow shock seems to be slightly less impacted by the large time steps, however it is still less diffused when using Δt_{exp}. Despite requiring ten times less iterations to reach the final time, the use of large time increments for the implicit kinetic scheme only results in around two percents faster computations compared to the explicit strategy which is due to the high quadratic cost of the implicit method. We believe that it is not possible to lower this cost when it comes to unconditionally stable methods, because the associated stencil has to cover the entire computational domain.

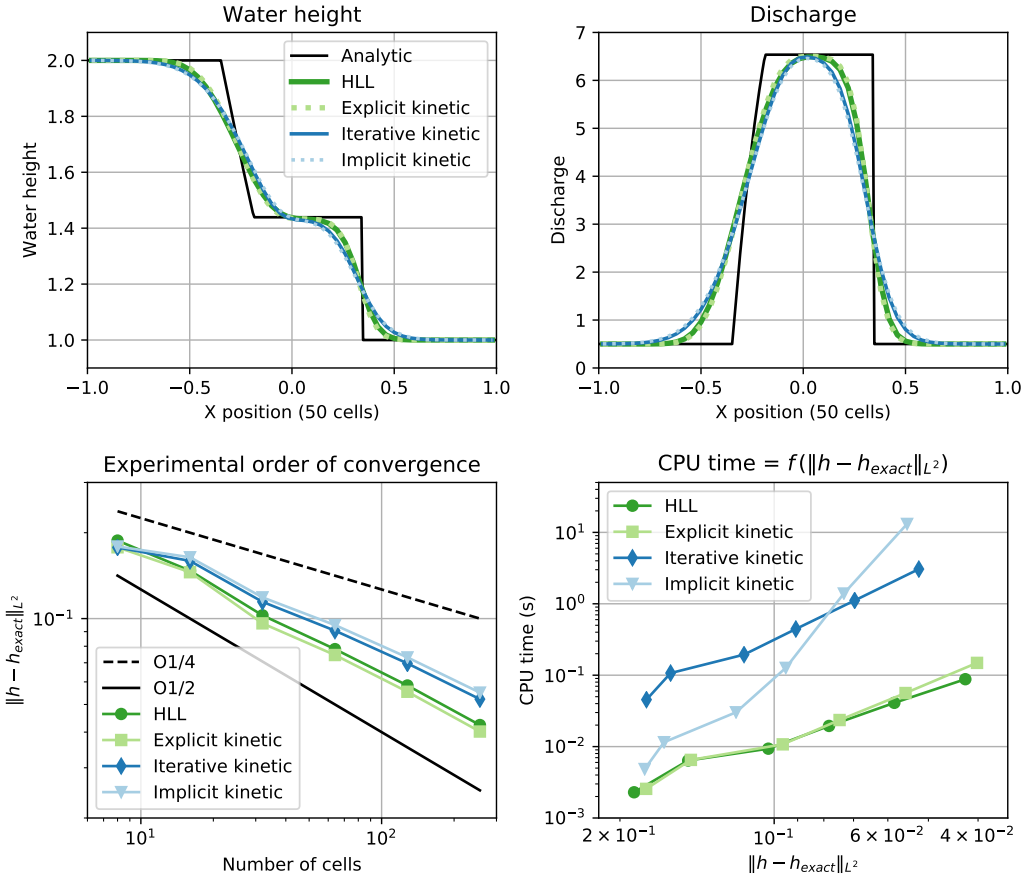

Figure 2.3: Comparing implicit, iterative and explicit kinetic solvers on a Riemann problem.

Riemann problem. We compare the fully implicit kinetic scheme and iterative kinetic scheme to explicit methods. The test case is given by the Riemann problem with initial data $U^0(x) = \mathbb{1}_{x<0} U_L + \mathbb{1}_{x>0} U_R$ where we define

$$U_L = \begin{pmatrix} 2 \\ 1/2 \end{pmatrix}, \quad U_R = \begin{pmatrix} 1 \\ 1/2 \end{pmatrix}.$$

The solution consists in a 1-rarefaction and a 2-shock. The iterative kinetic scheme uses the half-disk Maxwellian, and we choose the parameters $\alpha = 1$ and $\varepsilon_{\text{tol}} = 10^{-9}$ for the stopping criterion. All the

schemes use an explicit time step, and the results are given in Fig. 2.3. Three aspects have to be considered, namely the accuracy, the computational cost and the stability. In the plotted curves, we see that in terms of efficiency both iterative and implicit kinetic schemes are at their disadvantage. Especially, the quadratic complexity of the fully implicit version results in a steeper slope of the efficiency curve. However this is only one part of the picture, and we know from Proposition 11 and Remark 4 that the iterative kinetic scheme (2.50) satisfies a discrete entropy inequality without restriction on the time step, assuming enough iterations are performed. Concretely the greater stability comes with a higher level of diffusion which is noticeable in the first two plots of Fig. 2.3. This increased diffusion remains within acceptable margin, and is the price to pay to have better stability properties.

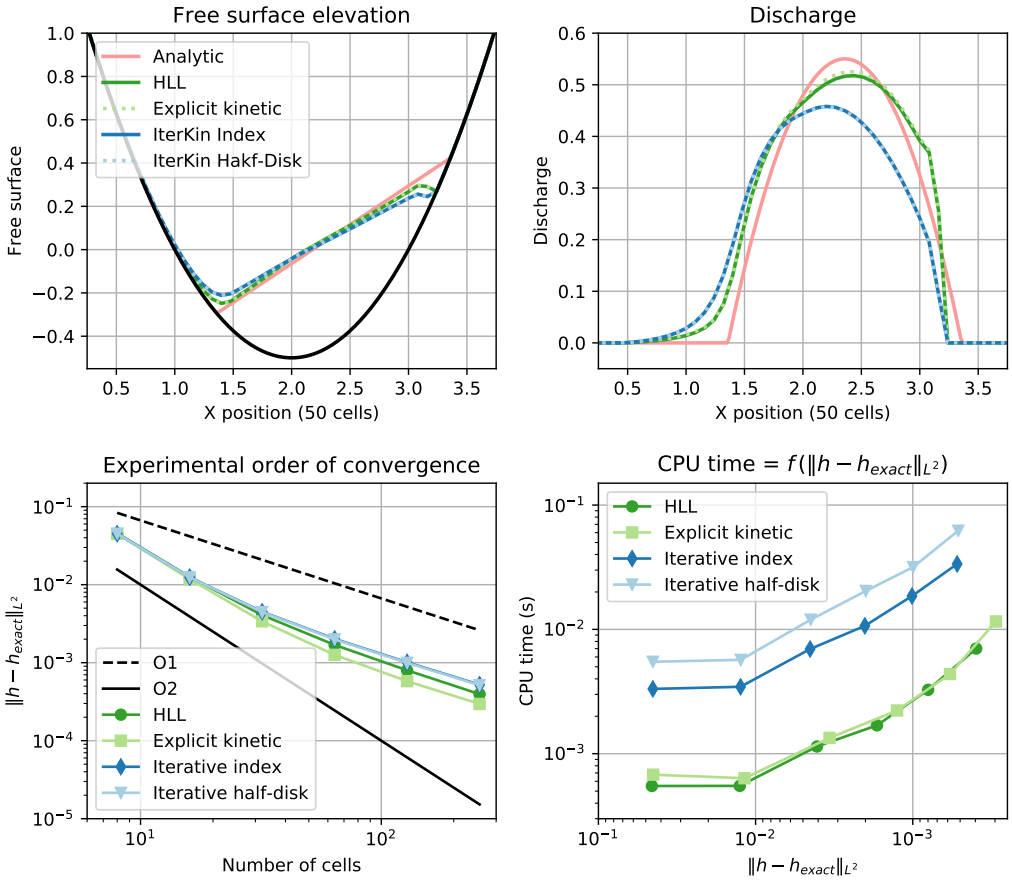

Figure 2.4: Parabolic bowl approximated by explicit and iterative kinetic schemes. First row: elevation and discharge at time 0.75, second row: convergence and efficiency curves. The stopping criteria used in the two kinetic iterative schemes combines the standard tolerance condition with tolerance $\varepsilon = 10^{-9}$ and the entropy condition (2.70).

Parabolic bowl. Next we consider Thacker's testcase, also known as the parabolic bowl testcase, taken from [37]. We plot the numerical solution at time 0.75 in Fig. 2.4. This testcase is relevant as it provides us with a non trivial analytical solution enabling to plot convergence curves, and it is known to be challenging numerically, as it presents a varying bottom together with an evolving wet/dry front and a discontinuous velocity profile. It is interesting to note that the different choice

of Maxwellian used in the two iterative kinetic schemes has very little impact on the approximation. In both cases we obtain a convergence with first order accuracy, and unsurprisingly the numerical cost is higher than for fully explicit methods due to the number of subiterations required to update the solution. One should also note that the use of the half-disk Maxwellian is slightly more expensive than the simpler index Maxwellian. Besides, in this testcase the iterative kinetic scheme with index Maxwellian was always able to fulfill the entropy condition (2.70) after some iterations, which we only proved rigorously for the half-disk Maxwellian. Hence despite using the wrong Maxwellian, it seems that the iterative kinetic scheme in question still has better stability properties than fully explicit methods. This will be further corroborated with the next testcase.

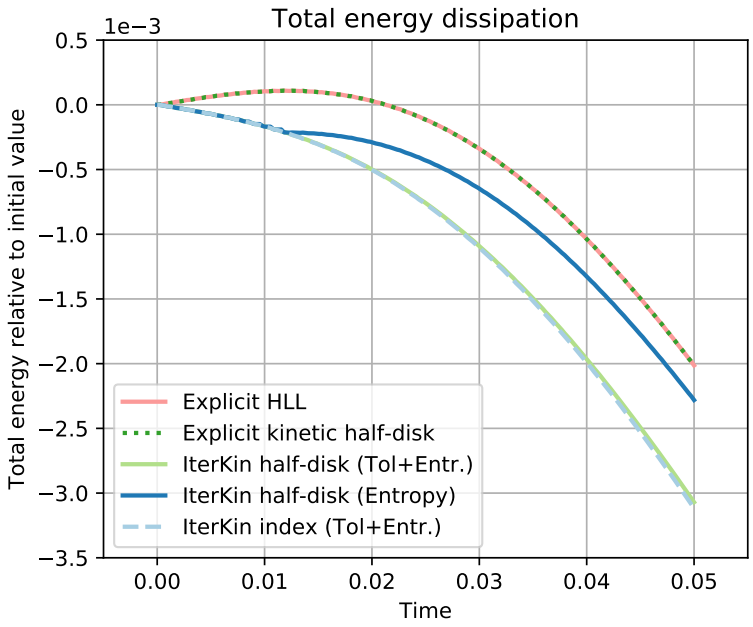

Figure 2.5: Evolution of the relative total energy obtained for various explicit and iterative kinetic schemes.

Total energy dissipation. Given the efficiency curves shown in Fig. 2.3 and Fig. 2.4, the explicit strategy seems preferable in terms of the computational cost at a prescribed accuracy. However we have to stress that among all the considered methods, the iterative kinetic scheme with half-disk Maxwellian is the only one for which we can prove existence of a fully discrete entropy inequality for a large enough but finite number of iterations. We remind that on the opposite, the explicit kinetic scheme with hydrostatic reconstruction does not satisfy a discrete entropy inequality without quadratic error term, however restrictive the CFL condition is, which is the result from Proposition 3.8 in [5]. Therefore the iterative scheme can be considered an improvement over this aspect, and we illustrate this through a numerical test where the explicit strategy increases the total energy, unlike the iterative method.

More precisely we measure the variation of total energy in a configuration with a varying bottom, and where the initial condition is given by a flat free surface and a constant nonzero velocity. Periodic boundary conditions are used, and the results can be seen in Fig. 2.5. Interestingly all the iterative methods manage to dissipate the total energy, even the scheme using the index Maxwellian, for

which there is no proof of discrete entropy inequality. On the contrary, the explicit kinetic scheme with half-disk Maxwellian increases the energy in the first few time steps, after what it decreases. The same goes for the explicit HLL scheme, and as a result these two explicit methods might not converge to the entropy solution. For comparison we also added in dark blue the iterative kinetic scheme with $\alpha = 0$ and whose subiterations stop as soon as the entropy condition (2.70) is verified. We can see that after some time this scheme becomes less dissipative than iterative kinetic methods using the standard tolerance condition.

2.6.2 The two dimensional case

The iterative kinetic scheme (2.49) and its version with hydrostatic reconstruction (2.64) can be easily extended to the two dimensional case. We believe that the results obtained in Section 2.5 the 1D setting carry to the higher dimension. We leave this study for later work, and perform a numerical experiment consisting of the 2D parabolic bowl [37] with a cartesian mesh. The results are obtained with the 2D version of the iterative kinetic scheme (2.64) with the Maxwellian defined by (2.42)(2.43) and are displayed in Fig. 2.7 and Fig. 2.6. We see that when increasing the tolerance value to $\varepsilon_{\text{tol}} = 10^{-5}$, the experimental order of convergence of the iterative scheme decreases, which illustrates that a large tolerance error prevents the sub-iterations to converge to the implicit update. On the other hand, the smaller ε_{tol} is, the more iterations are needed to reach the stopping criteria which translates to an increase in computational time. We also note that we needed to decrease the CFL constant form $1/2$ to $1/10$ to converge with a tolerance of $\varepsilon_{\text{tol}} = 10^{-9}$.

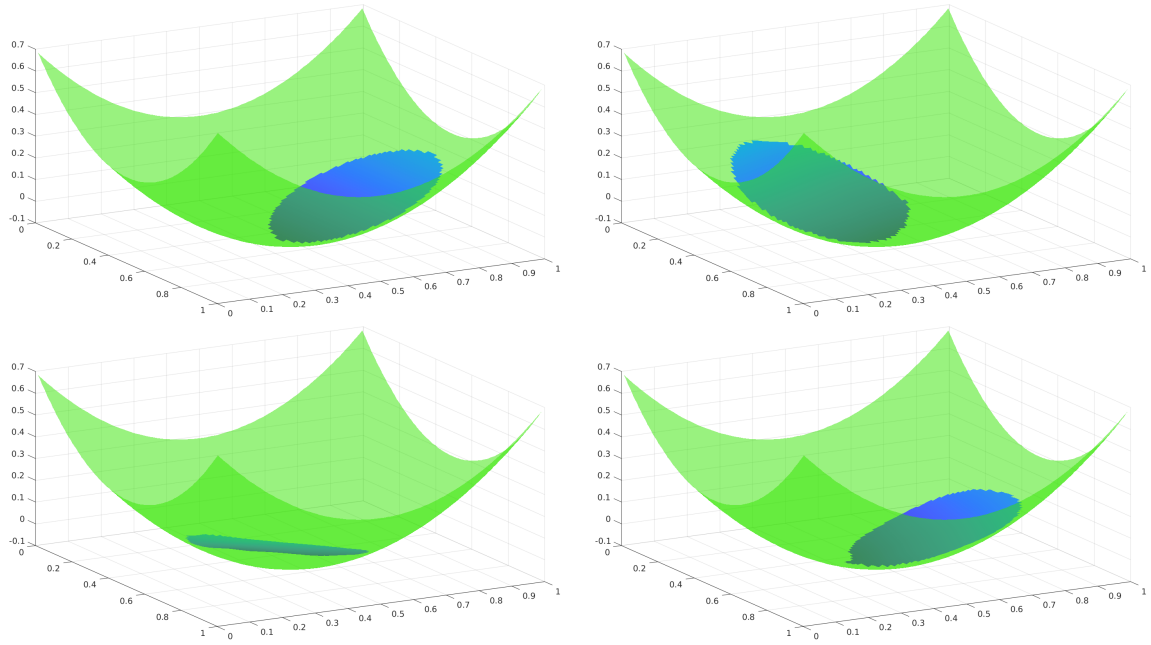

Figure 2.6: Numerical approximation of the 2D parabolic bowl using the iterative kinetic scheme with $\varepsilon = 10^{-7}$ over a 100×100 mesh. From left to right and top to bottom: initial condition, approximation at time $t = 1.1708$, $t = 2.3416$ and $t = 3.5124$.

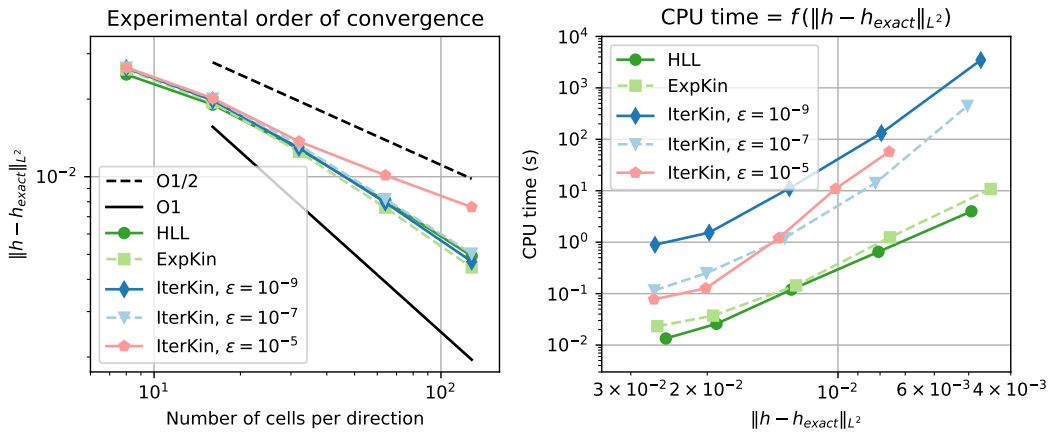

Figure 2.7: Convergence and efficiency curves obtained with the 2D parabolic bowl testcase. Different tolerances ε are compared for the iterative kinetic scheme. A CFL constant of $1/2$ was used, except for the case $\varepsilon = 10^{-9}$ where we set it to $1/10$.

2.A Expression of the numerical fluxes

The optimal choice for the Maxwellian is given by (2.2). Unfortunately the explicit expression for the numerical fluxes appearing in (2.32) is hardly possible with the choice (2.2) and the use of approximate quadrature formula for the integrals in (2.2) will degrade the accuracy of the scheme and increase the computational costs. Hence, we we choose M defined by the first expression in (2.30) and relation (2.32) becomes

$$U_i^{\text{int}} = \frac{1}{2\sqrt{3}}\Big(\int_{\min(0,a_j^n)}^{\min(0,b_j^n)} \sum_{j=i}^{P} \sqrt{\frac{2h_j^n}{g}} \begin{pmatrix} 1 \\ \xi \end{pmatrix} \frac{(-\sigma\xi)^{j-i}}{(1-\sigma\xi)^{j-i+1}} d\xi + \int_{\max(0,a_j^n)}^{\max(0,b_j^n)} \sum_{j=1}^{i} \sqrt{\frac{2h_j^n}{g}} \begin{pmatrix} 1 \\ \xi \end{pmatrix} \frac{(\sigma\xi)^{i-j}}{(1+\sigma\xi)^{i-j+1}} d\xi \Big),$$

with $a_j^n = u_j^n - \sqrt{3}\, c_j^n$ and $b_j^n = u_j^n + \sqrt{3}\, c_j^n$. The expressions of h_i^{int} and $(hu)_i^{\text{int}}$ are given by

$$h_i^{\text{int}} = \frac{1}{2\sqrt{3}}\Big(\sum_{j=i}^{P} \sqrt{\frac{2h_j^n}{g}} \underbrace{\int_{\min(0,a_j^n)}^{\min(0,b_j^n)} \frac{(-\sigma\xi)^{j-i}}{(1-\sigma\xi)^{j-i+1}} d\xi}_{(Ah)_{i,j}} + \sum_{j=1}^{i} \sqrt{\frac{2h_j^n}{g}} \underbrace{\int_{\max(0,a_j^n)}^{\max(0,b_j^n)} \frac{(\sigma\xi)^{i-j}}{(1+\sigma\xi)^{i-j+1}} d\xi}_{(Bh)_{i,j}} \Big)$$

$$(hu)_i^{\text{int}} = \frac{-\sigma^{-1}}{2\sqrt{3}}\Big(\sum_{j=i}^{P} \sqrt{\frac{2h_j^n}{g}} \underbrace{\int_{\min(0,a_j^n)}^{\min(0,b_j^n)} \frac{(-\sigma\xi)^{j-i+1}}{(1-\sigma\xi)^{j-i+1}} d\xi}_{(Ahu)_{i,j}} - \sum_{j=1}^{i} \sqrt{\frac{2h_j^n}{g}} \underbrace{\int_{\max(0,a_j^n)}^{\max(0,b_j^n)} \frac{(\sigma\xi)^{i-j+1}}{(1+\sigma\xi)^{i-j+1}} d\xi}_{(Bhu)_{i,j}} \Big)$$

Now we need to compute analytically the integrals of both expressions using the following lemmas.

Lemma 7. *If we denote* $y = 1 - \dfrac{1}{1+x}$ *for all* $x \in \mathbb{R}\backslash\{-1\}$ *and* $C \in \mathbb{R}$ *we have the following*

primitive:

$$\int \frac{x^k}{(1+x)^{k+1}}dx = \ln(|1+x|) - \sum_{l=1}^{k} \frac{y^l}{l} + C.$$

Lemma 8. *Using the same notation as in the previous lemma, we have*

$$\int \frac{x^k}{(1+x)^{k}}dx = -k\ln(|1+x|) + x + \sum_{l=1}^{k-1} l\frac{y^{k-l}}{k-l} + C'.$$

Proof of Lemma 7. We have

$$I = \int \frac{x^k}{(1+x)^{k+1}}dx = \int \frac{x^k}{(1+x)^{k}}\frac{1}{1+x}dx = \int (1-\frac{x}{1+x})^k \frac{1}{1+x}.$$

We pose $y = 1 - \dfrac{1}{1+x}$, then

$$I = \int y^k(1-y)\frac{dy}{(1-y)^2} = \int \frac{y^k-1}{1-y} + \frac{1}{1-y}dy.$$

Now we use the formula $y^k - 1 = (y-1)(y^{k-1} + y^{k-2} + \ldots + y + 1)$. And we obtain

$$\begin{aligned} I &= -\int \sum_{l=0}^{k-1} y^l dy - \ln(|1-y|) + C, && C \in \mathbb{R} \\ &= \ln(|1+x|) - \sum_{l=1}^{k} \frac{y^l}{l} + C', && C' \in \mathbb{R}. \end{aligned}$$

□

Proof of Lemma 8. We already have denoted $y = \dfrac{x}{1+x} = 1 - \dfrac{1}{1+x}$

$$I = \int (\frac{x}{1+x})^k dx = \int \frac{y^k dy}{(1-y)^2} = \int \Big(\frac{y^k-1}{(1-y)^2} + \frac{1}{(1-y)^2}\Big)dy$$

where the formula $y^k - 1 = (y-1)(y^{k-1} + y^{k-2} + \ldots + y + 1)$ has been used. Hence

$$\begin{aligned} I &= -\int \sum_{l=0}^{k-1} \frac{y^l}{1-y}dy + x + C = -\int \sum_{l=0}^{k-1} \frac{y^l-1}{1-y}dy - \int \frac{1}{1-y}\sum_{l=0}^{k-1} dy + x + C \\ &= \int \sum_{l=1}^{k-1} \frac{y^l-1}{y-1}dy + k\ln(|1-y|) + x + C' = \int \sum_{l=1}^{k-1}\sum_{p=0}^{l-1} y^p dy - k\ln(|1+x|) + x + C' \\ &= \sum_{l=1}^{k-1} l\int y^{k-1-l}dy - k\ln(|1+x|) + x + C' = \sum_{l=1}^{k-1} l\frac{y^{k-l}}{k-l} - k\ln(|1+x|) + x + C'', \end{aligned}$$

with $(C, C', C'') \in \mathbb{R}^3$. □

We are now able to compute the quantities $Ah_{i,j}$ and $Bh_{i,j}$

$$\begin{aligned} Ah_{i,j} &= \int_{\min(0,a_j^n)}^{\min(0,b_j^n)} \frac{(-\sigma\xi)^{j-i}}{(1-\sigma\xi)^{j-i+1}} d\xi = -\frac{1}{\sigma} \int_{-\min(0,a_j^n)\sigma}^{-\min(0,b_j^n)\sigma} \frac{(x)^{j-i}}{(1+x)^{j-i+1}} dx \\ &= \frac{1}{\sigma} \Big[\ln(|1+x|) - \sum_{l=1}^{j-i} \frac{y^l}{l} \Big]_{-\min(0,b_j^n)\sigma}^{-\min(0,a_j^n)\sigma}, \end{aligned} \tag{2.77}$$

$$\begin{aligned} Bh_{i,j} &= \int_{\max(0,a_j^n)}^{\max(0,b_j^n)} \frac{(\sigma\xi)^{i-j}}{(1+\sigma\xi)^{i-j+1}} d\xi = \frac{1}{\sigma} \int_{\max(0,a_j^n)\sigma}^{\max(0,b_j^n)\sigma} \frac{(x)^{i-j}}{(1+x)^{i-j+1}} dx \\ &= \frac{1}{\sigma} \Big[\ln(|1+x|) - \sum_{l=1}^{i-j} \frac{y^l}{l} \Big]_{\max(0,a_j^n)\sigma}^{\max(0,b_j^n)\sigma}. \end{aligned} \tag{2.78}$$

Similarly we obtain the formulas for $Ahu_{i,j}$ and $Bhu_{i,j}$

$$Ahu_{i,j} = \frac{1}{\sigma} \Big[-(j-i+1)\ln(|1+x|) + x + \sum_{l=1}^{j-i} l \frac{y^{j-i+1-l}}{j-i+1-l} \Big]_{-\min(0,b_j^n)\sigma}^{-\min(0,a_j^n)\sigma}, \tag{2.79}$$

$$Bhu_{i,j} = \frac{1}{\sigma} \Big[-(i-j+1)\ln(|1+x|) + x + \sum_{l=1}^{i-j} l \frac{y^{i-j+1-l}}{i-j+1-l} \Big]_{\max(0,a_j^n)\sigma}^{\max(0,b_j^n)\sigma}. \tag{2.80}$$

To conclude this paragraph, we give the final expression of U_i^{int}

$$\begin{aligned} h_i^{\text{int}} = \frac{1}{2\sigma\sqrt{3}} \Big(& \sum_{j=i}^{P} \sqrt{\frac{2h_j^n}{g}} \Big[\ln(|1+x|) - \sum_{l=1}^{j-i} \frac{y^l}{l} \Big]_{-\min(0,b_j^n)\sigma}^{-\min(0,a_j^n)\sigma} \\ & + \sum_{j=1}^{i} \sqrt{\frac{2h_j^n}{g}} \Big[\ln(|1+x|) - \sum_{l=1}^{i-j} \frac{y^l}{l} \Big]_{\max(0,a_j^n)\sigma}^{\max(0,b_j^n)\sigma} \Big) \end{aligned}$$

$$\begin{aligned} (hu)_i^{\text{int}} = \frac{1}{2\sigma^2\sqrt{3}} \Big(& -\sum_{j=i}^{P} \sqrt{\frac{2h_j^n}{g}} \Big[-(j-i+1)\ln(|1+x|) + x + \sum_{k=1}^{j-i} (j-i+1-k)\frac{y^k}{k} \Big]_{-\min(0,b_j^n)\sigma}^{-\min(0,a_j^n)\sigma} \\ & + \sum_{j=1}^{i} \sqrt{\frac{2h_j^n}{g}} \Big[-(i-j+1)\ln(|1+x|) + x + \sum_{k=1}^{i-j} (i-j+1-k)\frac{y^k}{k} \Big]_{\max(0,a_j^n)\sigma}^{\max(0,b_j^n)\sigma} \Big). \end{aligned}$$

2.B Computations of the fluxes involving the boundary conditions

We assume the ghost quantities U_0^{n+1} and U_{P+1}^{n+1} at time t^{n+1} to be known. The exterior contribution given in (2.31) also writes

$$U_i^{\text{ext}} = \int_{\mathbb{R}^-} \begin{pmatrix} 1 \\ \xi \end{pmatrix} \frac{(-\sigma\xi)^{P-i+1}}{(1-\sigma\xi)^{P-i+1}} M_{P+1}^{n+1} d\xi + \int_{\mathbb{R}^+} \begin{pmatrix} 1 \\ \xi \end{pmatrix} \frac{(\sigma\xi)^i}{(1+\sigma\xi)^i} M_0^{n+1} d\xi.$$

Using computations similar to what has been proposed in Appendix 2.A, we get

$$U_i^{\text{ext}} = \frac{1}{2\sqrt{3}}\left[\sqrt{\frac{2h_{P+1}^{n+1}}{g}}\int_{\min(0,a_{P+1}^{n+1})}^{\min(0,b_{P+1}^{n+1})}\binom{1}{\xi}\frac{(-\sigma\xi)^{P-i+1}}{(1-\sigma\xi)^{P-i+1}}d\xi \right. \\ \left. + \sqrt{\frac{2h_0^{n+1}}{g}}\int_{\max(0,a_0^{n+1})}^{\max(0,b_0^{n+1})}\binom{1}{\xi}\frac{(\sigma\xi)^i}{(1+\sigma\xi)^i}d\xi\right],$$

or equivalently

$$h_i^{\text{ext}} = \frac{1}{2\sqrt{3}}\left[\sqrt{\frac{2h_{P+1}^{n+1}}{g}}\int_{min(0,a_{P+1}^{n+1})}^{min(0,b_{P+1}^{n+1})}\frac{(-\sigma\xi)^{P-i+1}}{(1-\sigma\xi)^{P-i+1}}d\xi + \sqrt{\frac{2h_0^{n+1}}{g}}\int_{max(0,a_0^{n+1})}^{max(0,b_0^{n+1})}\frac{(\sigma\xi)^i}{(1+\sigma\xi)^i}d\xi\right],$$

$$(hu)_i^{\text{ext}} = \frac{1}{2\sqrt{3}}\left[\sqrt{\frac{2h_{P+1}^{n+1}}{g}}\int_{min(0,a_{P+1}^{n+1})}^{min(0,b_{P+1}^{n+1})}\xi\frac{(-\sigma\xi)^{P-i+1}}{(1-\sigma\xi)^{P-i+1}}d\xi + \sqrt{\frac{2h_0^{n+1}}{g}}\int_{max(0,a_0^{n+1})}^{max(0,b_0^{n+1})}\xi\frac{(\sigma\xi)^i}{(1+\sigma\xi)^i}d\xi\right].$$

As explained in Section 2.3.4, in practice we will replace the unknown values of U_0^{n+1}, U_{P+1}^{n+1} with that of U_0^n, U_{P+1}^n. The expression of h_i^{ext} can then be established by the means of Lemma 8. We now have to find an analytic expression for the quantity $\int \frac{x^{k+1}}{(1+x)^k}dx$ in order to obtain the final expression of $(hu)_i^{\text{ext}}$. The following lemma holds.

Lemma 9. *Let $k \in \mathbb{N}^*$. If we denote $y = 1 - \frac{1}{1+x}$ for all $x \in \mathbb{R}\backslash\{-1\}$ and $C \in \mathbb{R}$ we have the following expression*

$$\int \frac{x^{k+1}}{(1+x)^k}dx = \left(-\sum_{r=1}^{k-2}(k-r-1)\frac{k-r}{2}\frac{y^r}{r}\right)\mathbb{1}_{k\geq 3} - \left(\frac{k(k-1)}{2}\ln|1-y|\right)\mathbb{1}_{k\geq 2} \\ - \frac{k+1}{(1-y)} + \frac{1}{2(1-y)^2} - \left(\sum_{q=1}^{k-1}(k-q)\frac{y^q}{q}\right)\mathbb{1}_{k\geq 2} - k\ln|1-y| + C.$$

Proof. We begin by performing the change of variable $y = 1 - \frac{1}{1+x}$,

$$\int \frac{x^{k+1}}{(1+x)^k}dx = \int y^k\Big(\frac{1}{1-y}-1\Big)\frac{\mathrm{d}y}{(1-y)^2} = \int \frac{y^k}{(1-y)^3}\,\mathrm{d}y - \int \frac{y^k}{(1-y)^2}\,\mathrm{d}y\ .$$

Making use of $y^k - 1 = (y-1)(y^{k-1} + y^{k-2} + \cdots + 1)$ as before, we remark the following relation for $k \geq 1$

$$\frac{y^k}{1-y} = \frac{y^k-1}{1-y} + \frac{1}{1-y} = -\sum_{p=0}^{k-1} y^p + \frac{1}{1-y}\ .$$

Dividing this by $1-y$ leads to

$$\begin{aligned}\frac{y^k}{(1-y)^2} &= -\sum_{p=0}^{k-1}\frac{y^p}{1-y}+\frac{1}{(1-y)^2} = -\sum_{p=0}^{k-1}\Big(\frac{y^p-1}{1-y}+\frac{1}{1-y}\Big)+\frac{1}{(1-y)^2}\\ &= \Big(\sum_{p=1}^{k-1}\sum_{q=0}^{p-1} y^q\Big)\mathbb{1}_{k\geq 2}-\frac{k}{1-y}+\frac{1}{(1-y)^2}.\end{aligned}$$

Iterating this one more time we find

$$\begin{aligned}\frac{y^k}{(1-y)^3} &= \Big(\sum_{p=1}^{k-1}\sum_{q=0}^{p-1}\frac{y^q}{1-y}\Big)\mathbb{1}_{k\geq 2}-\frac{k}{(1-y)^2}+\frac{1}{(1-y)^3}\\ &= \Big(\sum_{p=1}^{k-1}\sum_{q=0}^{p-1}\frac{y^q-1}{1-y}+\frac{1}{1-y}\Big)\mathbb{1}_{k\geq 2}-\frac{k}{(1-y)^2}+\frac{1}{(1-y)^3}\\ &= \Big(-\sum_{p=2}^{k-1}\sum_{q=1}^{p-1}\sum_{r=0}^{q-1} y^r\Big)\mathbb{1}_{k\geq 3}+\frac{k(k-1)}{2(1-y)}\mathbb{1}_{k\geq 2}-\frac{k}{(1-y)^2}+\frac{1}{(1-y)^3}.\end{aligned}$$

As a consequence we get the following primitives up to a constant

$$\begin{aligned}\int\frac{y^k}{(1-y)^2}\,\mathrm{d}y &= \Big(\sum_{p=1}^{k-1}\sum_{q=1}^{p}\frac{y^q}{q}\Big)\mathbb{1}_{k\geq 2}+k\ln|1-y|+\frac{1}{(1-y)},\\ \int\frac{y^k}{(1-y)^3}\,\mathrm{d}y &= \Big(-\sum_{p=2}^{k-1}\sum_{q=1}^{p-1}\sum_{r=1}^{q}\frac{y^r}{r}\Big)\mathbb{1}_{k\geq 3}-\frac{k(k-1)}{2}\ln|1-y|\mathbb{1}_{k\geq 2}-\frac{k}{(1-y)}+\frac{1}{2(1-y)^2}.\end{aligned}$$

Finally, we simplify the double and triple sums as follows:

$$\sum_{p=1}^{k-1}\sum_{q=1}^{p}\frac{y^q}{q} = \sum_{q=1}^{k-1}\sum_{p=q}^{k-1}\frac{y^q}{q} = \sum_{q=1}^{k-1}(k-q)\frac{y^q}{q},$$

and from the above we deduce that

$$\begin{aligned}\sum_{p=2}^{k-1}\sum_{q=1}^{p-1}\sum_{r=1}^{q}\frac{y^r}{r} &= \sum_{p=2}^{k-1}\sum_{r=1}^{p-1}(p-r)\frac{y^r}{r}\\ &= \sum_{p=1}^{k-2}\sum_{r=1}^{p}(p-r+1)\frac{y^r}{r} = \sum_{r=1}^{k-2}\sum_{p=r}^{k-2}(p-r+1)\frac{y^r}{r}\\ &= \sum_{r=1}^{k-2}\Big(\frac{(k-r-1)(k+r-2)}{2}+(k-r-1)(1-r)\Big)\frac{y^r}{r}\\ &= \sum_{r=1}^{k-2}(k-r-1)\frac{k-r}{2}\frac{y^r}{r}.\end{aligned}$$

As a conclusion we have the expression

$$\int \frac{x^{k+1}}{(1+x)^k}dx = \left(-\sum_{r=1}^{k-2}(k-r-1)\frac{k-r}{2}\frac{y^r}{r}\right)\mathbb{1}_{k\geq 3} - \left(\frac{k(k-1)}{2}\ln|1-y|\right)\mathbb{1}_{k\geq 2} \\ - \frac{k+1}{(1-y)} + \frac{1}{2(1-y)^2} - \left(\sum_{q=1}^{k-1}(k-q)\frac{y^q}{q}\right)\mathbb{1}_{k\geq 2} - k\ln|1-y| + C,$$

where $C \in \mathbb{R}$ and with $y = x/(x+1)$. □

Chapter 3

Hydrostatic Euler equations

Outline of the current chapter

This chapter is written in collaboration with *Bernard Di Martino*, *Edwige Godlewski*, *Jacques Sainte-Marie*, and *Julien Guillod*. The pre-print [38] can be found on *HAL* (hal-04190892) or *arxiv* (arXiv:2308.15083).

By a semi-Lagrangian change of coordinates, the hydrostatic Euler equations describing free-surface sheared flows is rewritten as a system of quasilinear equations, where stability conditions can be determined by the analysis of its hyperbolic structure. This new system can be written as a quasi linear system in time and horizontal variables and involves no more vertical derivatives. However, the coefficients in front of the horizontal derivatives include an integral operator acting on the new vertical variable. The spectrum of these operators is studied in detail, in particular it

includes a continuous part. Riemann invariants are then determined as conserved quantities along the characteristic curves. Examples of solutions are provided, in particular stationary solutions and solutions blowing-up in finite time. Eventually, we propose an exact multi-layer $\mathbb{P}_0$-discretization, which could be used to solve numerically this semi-Lagrangian system, and analyze the eigenvalues of the corresponding discretized operator to investigate the hyperbolic nature of the approximated system.

3.1 Introduction

The classical shallow-water equations are widely used to describe irrotational flows of incompressible fluids in large-scale domains and for which the pressure is assumed to be hydrostatic [10, 48, 73]. The range of applications of the shallow-water models often includes flows in coastal regions, rivers, and channels as well as atmospheric flows or debris flows [19, 47, 69, 96]. However the shallow-water equations, due to their depth-averaged structure, describe only the horizontal profile of the velocity and information on the vertical component of the velocity is lost. Through the averaging process, the modeling of a $(d+1)$-dimensional flow (where d represents the number of horizontal coordinates) is reduced to a set of equations in dimension d. For a better approximation of some truly $(d+1)$-dimensional physical phenomena such as shear flows, the vertical contribution of the horizontal velocity must be taken into consideration.

In this paper we shall consider the motion of an incompressible ideal fluid described by the Euler system under the assumption that the pressure is hydrostatic. This system provides full access to the vertical velocity profile under the expense of losing the hyperbolic structure of the system. Indeed unlike the irrotational setting of the shallow-water flow which can be described by a system of hyperbolic partial differential equations, the rotational flow described by the hydrostatic Euler equations do not seem to fall under any classical classification. However, the hydrostatic Euler system with free surface evolution that we consider can be rewritten using a specific change of coordinates referred to as the semi-Lagrangian formulation. This formulation was first introduced by Zakharov [95] to derive an infinite system of conservation laws for shear flows in shallow water. The transformation maps the free-surface flow onto a fixed domain with flat boundaries. Note that it differs from the classical so-called sigma-transformation [36] and it also differs from the Lagrangian description of the flow. The advantage of the change of variables is that it allows us to rewrite the hydrostatic Euler system in a generalized quasi-linear form. More precisely, the rewriting gives a quasi-linear system of equations where the differential operator involves only time and horizontal derivatives but the coefficient in front of the horizontal derivatives has an integral operator in the new vertical variable. Therefore the classical hyperbolicity condition on a matrix is replaced by an analysis of the spectrum of an operator including both multiplicative and integral terms. Unfortunately, in general, explicit expressions of the eigenvalues are difficult to obtain, and the attempt to diagonalize the hydrostatic Euler system is only partially successful.

The approach we follow was presented in [91] by Teshukov who introduced a generalized notion of hyperbolicity and extended the theory of characteristics. With this generalized notion of hyperbolicity, sufficient criteria for the stability of flows were established in the case of a monotonic velocity profile [86] or for non-monotonic profiles satisfying further assumptions [89]. It is shown in [30] that all shear flows with a monotone and convex velocity profile are stable, and moreover sufficient conditions for the stability of piecewise linear continuous and discontinuous velocity profiles are determined.

Local well-posedness of the hydrostatic Euler system with free surface seems to be open in general, however results without free-surface are available. In [23], Brenier shows that the hydrostatic Euler equations without free-surface are locally solvable for smooth solutions with a strictly convex

velocity profile (local Rayleigh condition) using a similar semi-Lagrangian reformulation. We note that the assumptions of constant slopes at the boundaries of the domain required in [23] was removed later on in [74], without using a semi-Lagrangian change of variable. However, the local convexity assumption on the velocity field is required: [81] shows that the linearization around some velocity profile is ill-posed. For analytical initial data, local well-posedness was obtained in [65] using a Cauchy–Kovalevskaya type argument.

The paper presents several improvements and a few novelties. First we provide rigorous results about the existence of the semi-Lagrangian change of variable and thus provide that the two formulations are equivalent, at least locally in time. Secondly, we extend the results given by Teshukov in [91] on the spectrum of the given integral and multiplicative operator, by rigorously analyzing its spectrum and determining its different parts (point and discrete spectrum as well as continuous, residual and essential spectrum). We localize the spectrum more finely and determine limiting cases for which the existence of the two real eigenvalues given in [30] is guaranteed or not. Eventually a multi-layer formulation of the transformed system is proposed and analyzed by localizing the eigenvalues of the resulting discrete system and providing conditions for having real eigenvalues.

The assumptions made to establish that the spectrum is real (hence some sort of generalized hyperbolicity) together with some numerical experiments suggest that complex eigenvalues may exist for some physical velocity profiles. It would be interesting to exhibit precisely and study these situations, in both original and semi-Lagrangian frameworks.

Because when the number of layers increases, our discrete results are consistent with those obtained at the continuous level, the multi-layer formulation [6, 7, 25, 46] is interesting in practice for the numerical approximation of the transformed system and of the hydrostatic Euler system with free-surface. The simulations of geophysical flows take place over large domains in space and time and thus a compromise between stability and accuracy has to be found for their numerical approximation [5]. Compared to existing techniques for the approximation of hydrostatic models (see [3]), a discretization of the multi-layer version of the transformed system has several advantages:

- only derivatives along the horizontal axis are present;
- no exchange term between the layers appears in the formulation;
- estimates of the eigenvalues are available, which eventually could provide stability conditions for the time discretization.

The price to pay is the requirement to reinterpolate the variables when the change of variable becomes singular and thus hardly invertible. Numerical analysis of the discretized system as well as numerical results and comparison with traditional multi-layer formulations will be the subject of a future study.

The paper is organized as follows. In Section 3.2, we introduce the semi-Lagrangian formulation of the hydrostatic Euler system and prove that the change of variable is locally well-defined (Theorem 2) and that the two formulations are equivalent, at least locally in time (Theorems 3 and 4).

In Section 3.3, we discuss some examples of more or less explicit solutions in both formulations. In particular, we emphasize that the change of variable can become ill-defined in finite-time and we underline the degree of freedom in the definition of the change of variable (Section 3.3.1). We analyze the stationary solutions in both formulations (Section 3.3.2). The shallow water limit and an example of a flow with non-trivial vorticity are given respectively in Sections 3.3.3 and 3.3.4.

In Section 3.4, the spectrum of the integral and multiplicative operator appearing in front of the horizontal derivatives is analyzed and the associated Riemann invariants are discussed. In Section 3.4.1, we characterize rigorously the different parts of the spectrum, namely, the point, discrete, continuous, residual and essential spectrum (Theorem 6 and Corollary 2). Moreover we localize the spectrum with precise explicit bounds (Proposition 14). In Section 3.4.2 we provide

conditions under which two real eigenvalues exist (Propositions 15 and 16) and a counterexample showing our result in nearly optimal (Remark 13 and Proposition 17). In Section 3.4.3, the Riemann invariants associated to the two real eigenvalues and to the continuous part of the spectrum are discussed. All this section is done under the assumption that the topography is flat. However, this can be generalized to variable topography as explained in Section 3.4.4.

Finally, Section 3.5 is dedicated to a multi-layer approximation in the vertical variable of the system in the semi-Lagrangian formulation. Surprisingly, the multi-layer discretization of this system is an exact solution of the continuous system since the dependence on the vertical variable in only through an integral (Corollary 4). The eigenvalue of the corresponding matrix are determined (Proposition 25) and analyzed (Propositions 24 and 25). In the end, we analyze the convergence of the spectrum of the approximation towards the spectrum of the continuous case under the assumption that the velocity profile is monotonic and convex (Proposition 26).

Let us note that for simplicity, most of the results concern the one-dimensional case, *i.e.*, one horizontal variable ($d = 1$) and one vertical variable, however some of them also concern the case of two horizontal variables ($d = 2$).

3.2 Semi-Lagrangian formulation of the hydrostatic Euler system

The aim of this section is to rewrite the hydrostatic Euler system in a generalized quasi-linear form using a semi-Lagrangian formulation, and prove that both formulations are in some sense equivalent at least for short time.

3.2.1 Free-surface hydrostatic Euler system

We denote respectively by $\boldsymbol{x} \in \mathbb{R}^d$ for $d = 1, 2$ and z the coordinates in the horizontal and vertical directions (see Fig. 3.1). The surface elevation is described by a function $\eta(t, \boldsymbol{x})$ and the bottom is given by a function $z_b(\boldsymbol{x})$ independent of time; we denote by $h(t, \boldsymbol{x}) = \eta(t, \boldsymbol{x}) - z_b(\boldsymbol{x})$ the water depth and by Ω_t the domain occupied by the fluid at time t:

$$\Omega_t = \big\{(\boldsymbol{x}, z) \in \mathbb{R}^d \times \mathbb{R} \,:\, z_b(\boldsymbol{x}) < z < \eta(t, \boldsymbol{x})\big\}. \tag{3.1}$$

Note that the variable $\boldsymbol{x}$ is assumed to lie in an infinite domain, which is counter intuitive at first glance. However, this choice circumvents the difficulties arising if boundary conditions were to be imposed on the horizontal directions which is not the interest of this paper. The velocity field is given by $(\boldsymbol{u}(t, \boldsymbol{x}, z), w(t, \boldsymbol{x}, z))$ where $\boldsymbol{u}(t, \boldsymbol{x}, z)$ and $w(t, \boldsymbol{x}, z)$ represent the horizontal and vertical components, respectively.

The hydrostatic Euler system is given by:

$$\frac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u} \cdot \nabla_{\boldsymbol{x}} \boldsymbol{u} + w \partial_z \boldsymbol{u} + g \nabla_{\boldsymbol{x}} \eta = -\nabla_{\boldsymbol{x}} p^a, \tag{3.2}$$

$$\nabla_{\boldsymbol{x}} \cdot \boldsymbol{u} + \partial_z w = 0, \tag{3.3}$$

where $p^a = p^a(t, \boldsymbol{x})$ is a given function corresponding to the atmospheric pressure. This system has to be completed with boundary conditions. At the free surface we have the kinematic boundary condition

$$\frac{\partial \eta}{\partial t} + \boldsymbol{u}_s \cdot \nabla_{\boldsymbol{x}} \eta - w_s = 0, \tag{3.4}$$

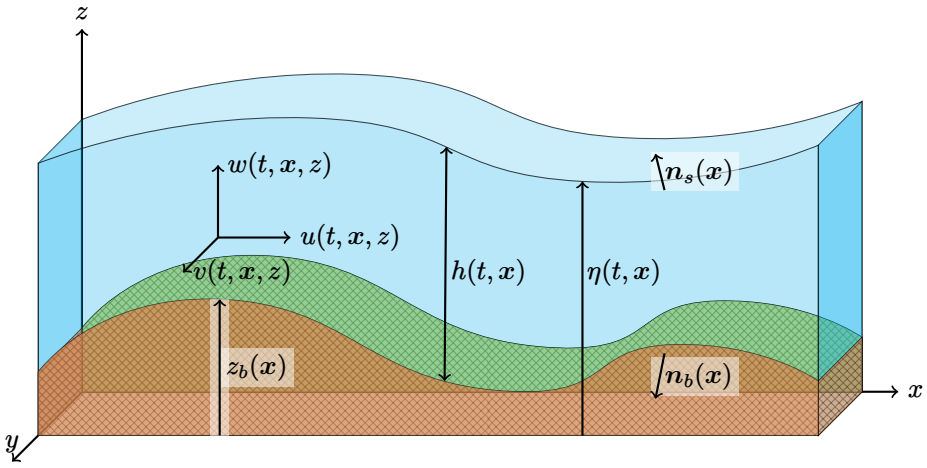

Figure 3.1: The three dimensional set-up for the hydrostatic Euler system with free-surface, where $\boldsymbol{x} = (x, y)$ and $\boldsymbol{u} = (u, v)$.

where the subscript s indicates the value of the considered quantity at the free surface, for example $w_s(t, \boldsymbol{x}) = w(t, \boldsymbol{x}, \eta(t, \boldsymbol{x}))$. The kinematic boundary condition at the bottom consists in a classical no-penetration condition,

$$\boldsymbol{u}_b \cdot \nabla_{\boldsymbol{x}} z_b - w_b = 0, \tag{3.5}$$

where the subscript b indicates the value of the considered quantity at bottom, for example $w_b(t, \boldsymbol{x}) = w(t, \boldsymbol{x}, z_b(\boldsymbol{x}))$.

Together these equations define the following Cauchy problem for $\eta, \boldsymbol{u}, w$:

$$\begin{cases} \dfrac{\partial \eta}{\partial t} + \boldsymbol{u}_s \cdot \nabla_{\boldsymbol{x}} \eta - w_s = 0 & \text{in } (0,T) \times \mathbb{R}^d, \\ \boldsymbol{u}_b \cdot \nabla_{\boldsymbol{x}} z_b - w_b = 0 & \text{in } (0,T) \times \mathbb{R}^d, \\ \dfrac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u} \cdot \nabla_{\boldsymbol{x}} \boldsymbol{u} + w \partial_z \boldsymbol{u} + g \nabla_{\boldsymbol{x}} \eta = -\nabla_{\boldsymbol{x}} p^a & \text{in } (0,T) \times \Omega_t, \\ \nabla_{\boldsymbol{x}} \cdot \boldsymbol{u} + \partial_z w = 0 & \text{in } (0,T) \times \Omega_t, \\ \boldsymbol{u}|_{t=0} = \boldsymbol{u}_0 & \text{in } \Omega_0, \\ \eta|_{t=0} = \eta_0 & \text{in } \mathbb{R}^d, \end{cases} \tag{3.6}$$

where $\boldsymbol{u}_0, \eta_0$ are given, and $(0,T) \times \Omega_t$ is a sloppy notation for the space-time domain

$$\begin{aligned} (0,T) \times \Omega_t &= \big\{ (t, \boldsymbol{x}, z) \in (0,T) \times \mathbb{R}^d \times \mathbb{R} \,:\, (\boldsymbol{x}, z) \in \Omega_t \big\} \\ &= \big\{ (t, \boldsymbol{x}, z) \in (0,T) \times \mathbb{R}^d \times \mathbb{R} \,:\, z_b(\boldsymbol{x}) < z < \eta(t, \boldsymbol{x}) \big\}. \end{aligned}$$

Remark 6. *The variable w can be eliminated by integrating the divergence-free condition $(3.6)_4$:*

$$w = w_b - \int_{z_b(\boldsymbol{x})}^{z} (\nabla_{\boldsymbol{x}} \cdot \boldsymbol{u}) \, \mathrm{d}Z = \boldsymbol{u}_b \cdot \nabla_{\boldsymbol{x}} z_b - \int_{z_b(\boldsymbol{x})}^{z} (\nabla_{\boldsymbol{x}} \cdot \boldsymbol{u}) \, \mathrm{d}Z. \tag{3.7}$$

In particular, we have

$$w_s = w_b - \int_{z_b(\boldsymbol{x})}^{\eta(t,\boldsymbol{x})} (\nabla_{\boldsymbol{x}} \cdot \boldsymbol{u}) \, \mathrm{d}Z \,, \tag{3.8}$$

so after some calculations one gets the following conservative form

$$\frac{\partial \eta}{\partial t} + \nabla_{\boldsymbol{x}} \cdot \left(\int_{z_b(\boldsymbol{x})}^{\eta(t,\boldsymbol{x})} \boldsymbol{u}(t,\boldsymbol{x},Z) \, \mathrm{d}Z \right) = 0, \tag{3.9}$$

$$\frac{\partial \boldsymbol{u}}{\partial t} + \nabla_{\boldsymbol{x}} \cdot (\boldsymbol{u} \otimes \boldsymbol{u}) + \partial_z (w\boldsymbol{u}) + g \nabla_{\boldsymbol{x}} \eta = -\nabla_{\boldsymbol{x}} p^a \,. \tag{3.10}$$

From now on, we assume that p^a is a constant unless otherwise stated (as in Section 3.3.4).

3.2.2 Derivation of the semi-Lagrangian formulation

The aim is now to reformulate (3.6) using semi-Lagrangian coordinates. Following [89], let us introduce a new variable $\lambda \in I = (0,1)$ and the function $\phi = \phi(t,\boldsymbol{x},\lambda)$ solution of the following Cauchy problem in $(0,T) \times \tilde{\Omega}$ where $\tilde{\Omega} = \mathbb{R}^d \times I$:

$$\begin{cases} \dfrac{\partial \phi}{\partial t} + \boldsymbol{u}(t,\boldsymbol{x},\phi) \cdot \nabla_{\boldsymbol{x}} \phi = w(t,\boldsymbol{x},\phi) & \text{in } (0,T) \times \tilde{\Omega}, \\ \phi(0,\boldsymbol{x},\lambda) = \phi_0(\boldsymbol{x},\lambda) & \text{in } \tilde{\Omega}. \end{cases} \tag{3.11}$$

Integrating the divergence-free condition $(3.6)_4$ from $z_b(\boldsymbol{x})$ to $\phi(t,\boldsymbol{x},\lambda)$ and using $(3.6)_2$, the first equation in (3.11) can be rewritten in a conservative form similar to (3.9)

$$\frac{\partial \phi}{\partial t} + \nabla_{\boldsymbol{x}} \cdot \left(\int_{z_b(\boldsymbol{x})}^{\phi(t,\boldsymbol{x},\lambda)} \boldsymbol{u}(t,\boldsymbol{x},z) \, \mathrm{d}z \right) = 0. \tag{3.12}$$

The initial condition in (3.11) is chosen to verify:

$$\phi_0(\boldsymbol{x},0) = z_b(\boldsymbol{x}), \qquad \phi_0(\boldsymbol{x},1) = \eta_0(\boldsymbol{x}), \qquad \partial_\lambda \phi_0(\boldsymbol{x},\lambda) > 0. \tag{3.13}$$

We note that there is not a unique choice of ϕ_0; a canonical choice is

$$\phi_0(\boldsymbol{x},\lambda) = (1-\lambda) z_b(\boldsymbol{x}) + \lambda \eta_0(\boldsymbol{x}). \tag{3.14}$$

Since the equation (3.11) is quasilinear, its local in time well-posedness stated in the following theorem is rather standard and its proof will be postponed to the end of the section.

Theorem 2. *Let $s \geq 1$, $T > 0$, $p^a = 0$, and $z_b \in C_b^s(\mathbb{R}^d)$. If $(\eta, \boldsymbol{u}, w)$ is a solution of (3.6) such that $\eta \in C_b^s((0,T) \times \mathbb{R}^d)$, $\boldsymbol{u} \in C_b^s((0,T) \times \Omega_t)^d$, and $w \in C_b^s((0,T) \times \Omega_t)$, then for $\phi_0 \in C_b^s(\tilde{\Omega})$, there exists $T^* \in (0,T]$ such that (3.11) admits a unique solution $\phi \in C^s((0,T^*) \times \tilde{\Omega})$.*

Moreover if $\inf_{(\boldsymbol{x},\lambda) \in \mathbb{R}^d \times I} \partial_\lambda \phi_0(\boldsymbol{x},\lambda) > 0$ then for $t \in (0,T^*)$ we have $\partial_\lambda \phi(t,\boldsymbol{x},\lambda) > 0$, so in particular

$$\begin{aligned} (\mathrm{Id}, \phi(t)) : \tilde{\Omega} &\to \Omega_t \\ (\boldsymbol{x},\lambda) &\mapsto (\boldsymbol{x}, \phi(t,\boldsymbol{x},\lambda)) \end{aligned}$$

is an orientation preserving C^s-diffeomorphism, provided it is a diffeomorphism at $t = 0$, i.e. *provided that $\phi_0(\boldsymbol{x},0) = z_b(\boldsymbol{x})$ and $\phi_0(\boldsymbol{x},1) = \eta_0(\boldsymbol{x})$.*

Using this result, we can perform a semi-Lagrangian change of variable and obtain the following reformulation of the hydrostatic Euler system.

Theorem 3. *Assuming that the hypotheses of the previous theorem are satisfied with $s \geq 2$, then the functions defined by*

$$H(t,x,\lambda) = \partial_\lambda \phi(t,x,\lambda)\,, \qquad \tilde{u}(t,x,\lambda) = u(t,x,\phi(t,x,\lambda)) \tag{3.15}$$

have regularities $H \in C^{s-1}((0,T^*)\times\tilde{\Omega})$ and $\tilde{u} \in C^s((0,T^*)\times\tilde{\Omega})^d$ and are solutions of

$$\begin{cases} \dfrac{\partial H}{\partial t} + \nabla_x (H\tilde{u}) = 0 & \text{in } (0,T)\times\tilde{\Omega}\,, \\ \dfrac{\partial \tilde{u}}{\partial t} + \tilde{u}\cdot\nabla_x \tilde{u} + g\nabla_x \displaystyle\int_0^1 H\,\mathrm{d}\lambda = -g\nabla_x z_b & \text{in } (0,T)\times\tilde{\Omega}\,, \\ H(0,x,\lambda) = H_0(x,\lambda) & \text{in } \tilde{\Omega}\,, \\ \tilde{u}(0,x,\lambda) = \tilde{u}_0(x,\lambda) & \text{in } \tilde{\Omega}\,, \end{cases} \tag{3.16}$$

where the initial data are given by

$$H_0(x,\lambda) = \partial_\lambda \phi_0(x,\lambda), \qquad \tilde{u}_0(x,\lambda) = u(0,x,\phi_0(x,\lambda)). \tag{3.17}$$

We can also go back from the new formulation to the original system.

Theorem 4. *Let $s \geq 1$, $T > 0$, and $z_b \in C_b^s(\mathbb{R}^d)$. If $H \in C_b^s((0,T)\times\tilde{\Omega})$ and $\tilde{u} \in C_b^{s+1}((0,T)\times\tilde{\Omega})^d$ are solutions of (3.16) with $\inf_{(x,\lambda)\in\mathbb{R}^d\times I} H_0(x,\lambda) > 0$, then there exists $T^* \in (0,T]$ such that*

$$\begin{aligned} (\mathrm{Id},\phi(t)) : \tilde{\Omega} &\to \Omega_t \\ (x,\lambda) &\mapsto (x,\phi(t,x,\lambda)) \end{aligned}$$

is an orientation preserving C^s-diffeomorphism for $t \in (0,T^*)$ where $\phi \in C_b^s((0,T)\times\tilde{\Omega})$ and Ω_t are defined by

$$\begin{aligned} \phi(t,x,\lambda) &= z_b(x) + \int_0^\lambda H(t,x,\lambda')\,\mathrm{d}\lambda'\,, \\ \Omega_t &= \{(x,z)\in\mathbb{R}^d\times\mathbb{R} \,:\, z_b(x) < z < \phi(t,x,1)\}. \end{aligned}$$

Moreover, if $(\mathrm{Id},\phi(t)^{-1})$ denotes the inverse of $(\mathrm{Id},\phi(t))$, then

$$\begin{aligned} \eta(t,x) &= \phi(t,x,1)\,, \\ u(t,x,z) &= \tilde{u}(t,x,\phi^{-1}(t,x,z))\,, \\ w(t,x,z) &= \tilde{u}(t,x,0)\cdot\nabla_x z_b - \int_{z_b(x)}^z (\nabla_x\cdot u)\,\mathrm{d}Z\,, \end{aligned}$$

have regularities $\eta \in C_b^s((0,T)\times\mathbb{R}^d)$, $u \in C^s((0,T^*)\times\tilde{\Omega})^d$ and $w \in C^s((0,T^*)\times\tilde{\Omega})$ and are solutions of (3.6) with

$$\begin{aligned} \eta_0(x) &= z_b(x) + \phi(0,x,1) \\ u_0(x,z) &= \tilde{u}_0(x,\phi^{-1}(0,x,z)) \end{aligned}$$

Remark 7. *We note that to go from the original system (3.6) to the semi-Lagrangian formulation*

(3.16) an initial data ϕ_0 satisfying (3.13) has to be chosen. In the opposite direction, going from the new formulation (3.16) to the original Euler system (3.6), one does not have this degree of freedom.

Note also the relation between $h = \eta - z_b$ and H:

$$h(x,t) = \int_0^1 H(t,\boldsymbol{x},\lambda)\,\mathrm{d}\lambda.$$

Remark 8. *The previous results require the existence of a solution to (3.6), which is currently an open problem. Even without free-surface, the hydrostatic Euler equations are well-posed only for analytical data [65, 81], and it is conceivable to expect similar results in the case of a free surface.*

Remark 9. *We note that as for the hydrostatic Euler system (3.6), no local energy balance holds for (3.16). Indeed, after some lengthy but straightforward calculations, one gets*

$$\frac{\partial}{\partial t}\left(\frac{H|\tilde{\boldsymbol{u}}|^2}{2} + gz_bH\right) + g\int_0^1 H\mathrm{d}\lambda\frac{\partial H}{\partial t} + \nabla_{\boldsymbol{x}}\cdot\left(H\tilde{\boldsymbol{u}}\left(\frac{|\tilde{\boldsymbol{u}}|^2}{2} + gz_b + g\int_0^1 H\mathrm{d}\lambda\right)\right) = 0,$$

which is not a local energy equality. However by integrating this equation in λ we get the following energy balance:

$$\frac{\partial}{\partial t}\left(\int_0^1 H\left(\frac{|\tilde{\boldsymbol{u}}|^2}{2} + gz_b + \frac{g}{2}\int_0^1 H\mathrm{d}\lambda\right)\mathrm{d}\lambda\right) + \nabla_{\boldsymbol{x}}\cdot\left(\int_0^1 H\tilde{\boldsymbol{u}}\left(\frac{|\tilde{\boldsymbol{u}}|^2}{2} + gz_b + g\int_0^1 H\mathrm{d}\lambda\right)\mathrm{d}\lambda\right) = 0.$$

3.2.3 Proofs of the equivalence of the two formulations

We begin by the proof of the local well-posedness for (3.11).

Proof of Theorem 2. The equation being quasilinear, we use the method of characteristics. One difficulty is that $\eta(t)$, $\boldsymbol{u}(t)$ and $w(t)$ are only defined on Ω_t and not on $\mathbb{R}^{d+1}$. Hence, we extend the fields $w(t)$ and $\boldsymbol{u}(t)$ defined on the moving domain Ω_t to $\mathbb{R}^{d+1}$. By using [45, Theorem 1], it is possible to define C^s extensions of $\boldsymbol{u}$ and w on $(0,T)\times\mathbb{R}^{d+1}$ in a way that their norms remain bounded by a constant depending only on s and d. Therefore from now on we assume that $\boldsymbol{u} \in C^s_b((0,T)\times\mathbb{R}^{d+1})^d$, and $w \in C^s_b((0,T)\times\mathbb{R}^{d+1})$.

Let us introduce the characteristics $\boldsymbol{X}(t,\boldsymbol{y},v)$ and $\Phi(t,\boldsymbol{y},v)$ defined respectively as the solutions of

$$\begin{aligned} &\frac{d\boldsymbol{X}(t,\boldsymbol{y},v)}{dt} = \boldsymbol{u}(t,\boldsymbol{X}(t,\boldsymbol{y},v),\Phi(t,\boldsymbol{y},v))\,, && \frac{d\Phi(t,\boldsymbol{y},v)}{dt} = w(t,\boldsymbol{X}(t,\boldsymbol{y},v),\Phi(t,\boldsymbol{y},v))\,,\\ &\boldsymbol{X}(0,\boldsymbol{y},v) = \boldsymbol{y}\,, && \Phi(0,\boldsymbol{y},v) = v\,, \end{aligned}$$

and we know by the Cauchy-Lipschitz theorem that $\boldsymbol{X} \in C^s((0,T)\times\mathbb{R}^{d+1})^d$ and $\Phi \in C^s((0,T)\times\mathbb{R}^{d+1})$.
If ϕ is a solution of (3.11), then

$$\phi(t,\boldsymbol{X}(t,\boldsymbol{y},v),\lambda) = \Phi(t,\boldsymbol{y},v)\,, \qquad \text{with} \qquad v = \phi_0(\boldsymbol{y},\lambda)\,.$$

Since $\boldsymbol{X}(0,\boldsymbol{y},\phi_0(\boldsymbol{y},\lambda)) = \boldsymbol{y}$, the map $\boldsymbol{y}\mapsto\boldsymbol{X}(t,\boldsymbol{y},\phi_0(\boldsymbol{y},\lambda))$ is invertible at least on some small time interval $(0,T^*)$. Therefore, denoting its inverse by $\boldsymbol{Y}(t,\boldsymbol{x},\lambda)$, the solution of (3.11) is given by

$$\phi(t,\boldsymbol{x},\lambda) = \Phi(t,\boldsymbol{Y}(t,\boldsymbol{x},\lambda),\phi_0(\boldsymbol{Y}(t,\boldsymbol{x},\lambda),\lambda))\,,$$

and by composition of C^s functions, we get $\phi \in C^s((0,T^*)\times\tilde{\Omega})$.

If $\inf_{(x,\lambda)\in\mathbb{R}^d\times I}\partial_\lambda\phi_0(x,\lambda)>0$, there exists some $\varepsilon>0$ such that $\partial_\lambda\phi_0(x,\lambda)\geq\varepsilon$ for all $x\in\mathbb{R}^d$ and $\lambda\in I$. Then the regularity on ϕ ensures that there exists some $T^*\in(0,T]$ such that for $t\in(0,T^*)$, $\partial_\lambda\phi(t,x,\lambda)\geq\frac{\varepsilon}{2}$. The gradient of the function $(\mathrm{Id},\phi(t))$ is

$$\nabla_{x,\lambda}\begin{pmatrix}x\\ \phi(t,x,\lambda)\end{pmatrix}=\begin{pmatrix}\mathbf{1} & \mathbf{0}\\ \nabla_x\phi(t,x,\lambda) & \partial_\lambda\phi(t,x,\lambda)\end{pmatrix},$$

which is invertible for $t\in(0,T^*)$, so $(\mathrm{Id},\phi(t))$ is an orientation preserving diffeomorphism from $\tilde{\Omega}$ to Ω_t defined by

$$\Omega_t=\{(x,z)\in\mathbb{R}^d\times\mathbb{R}\,:\,\phi(t,x,0)<z<\phi(t,x,1)\}.$$

It remains to show that this definition of the domain Ω_t is equivalent to the one given by (3.1). Since $u(t,x,z_b(x))=u_b(t,x)$, and $w(t,x,z_b(x))=w_b(t,x)$ satisfies $(3.6)_2$, $z_b(x)$ is a solution of (3.11) for $\lambda=0$, hence by uniqueness $\phi(t,x,0)=z_b(x)$. Similarly, since $\eta(t,x)$ satisfies $(3.6)_1$, $\eta(t,x)$ is a solution of (3.11) for $\lambda=1$, hence by uniqueness $\phi(t,x,1)=\eta(t,x)$. □

Then the transformation of (3.6) to the new system (3.16) can be easily deduced.

Proof of Theorem 3. The claimed regularities on H and $\tilde{u}$ follow from the regularity on ϕ and u.

We have

$$\begin{aligned}\partial_\lambda\left(\partial_t\phi+\tilde{u}\cdot\nabla_x\phi-\tilde{w}\right)&=\partial_tH+\tilde{u}\cdot\nabla_xH+H\partial_zu\cdot\nabla_x\phi-H\partial_zw\\&=\partial_tH+\nabla_x(H\tilde{u})-H(\nabla_x\cdot u+\partial_zw),\end{aligned}\tag{3.18}$$

since

$$\nabla_x(H\tilde{u})=\tilde{u}\cdot\nabla_xH+H\nabla_x\cdot u+H\partial_zu\cdot\nabla_x\phi.$$

Therefore, using $(3.11)_1$ and the divergence-free condition $(3.6)_4$, we deduce the first equation $(3.16)_1$.

It follows from (3.15) that

$$\begin{aligned}\partial_t\tilde{u}=\partial_tu+\partial_t\phi\,\partial_zu=\partial_tu+w\partial_zu-u\cdot\nabla_x\phi\,\partial_zu\\ \tilde{u}\cdot\nabla_x\tilde{u}=u\cdot\nabla_xu+u\cdot\nabla_x\phi\,\partial_zu\end{aligned}$$

and therefore

$$\partial_t\tilde{u}+\tilde{u}\cdot\nabla_x\tilde{u}=\partial_tu+u\cdot\nabla_xu+w\partial_zu\,.\tag{3.19}$$

Finally, since $\eta(t,x)=\phi(t,x,1)$ and $z_b(x)=\phi(t,x,0)$ as seen in Theorem 2, we have

$$\eta(t,x)=\phi(t,x,1)=\phi(t,x,0)+\int_0^1\partial_\lambda\phi(t,x,\lambda)\,\mathrm{d}\lambda=z_b(x)+\int_0^1H(t,x,\lambda)\,\mathrm{d}\lambda\,,$$

so

$$\nabla_x\eta=\nabla_xz_b+\nabla_x\int_0^1H(t,x,\lambda)\,\mathrm{d}\lambda\,,\tag{3.20}$$

and the second equation $(3.16)_2$ is deduced using $(3.6)_3$. □

Finally, the proof of converse is in the same spirit.

Proof of Theorem 4. Since $(3.16)_1$ can be viewed as a linear transport equation for H with $\tilde{u}$ known, one can deduce that there exists $T^*\in(0,T)$ such that $H(t,x,\lambda)>0$. Therefore the map $(\mathrm{Id},\phi(t))$ is an orientation preserving C^s-diffeomorphism for $t\in(0,T^*)$. Then, we can define the inverse $\phi(t)^{-1}$ such that $\phi(t)^{-1}(x,\phi(t,x,\lambda))=\lambda$ and $\phi(t,x,\phi^{-1}(t,x,z))=z$. The claimed regularity of η, u and

w follows by standard arguments. It remains to prove that η, $\boldsymbol{u}$ and w are solutions of (3.6). One deduces directly that equations $(3.6)_{2,4}$ are satisfied. Since $H = \partial_\lambda \phi$, using (3.18) we have

$$\partial_\lambda \left(\partial_t \phi + \tilde{\boldsymbol{u}} \cdot \nabla_{\boldsymbol{x}} \phi - \tilde{w} \right) = \partial_t H + \nabla_{\boldsymbol{x}} (H \tilde{\boldsymbol{u}}) = 0 \,,$$

and therefore

$$\partial_t \eta + \boldsymbol{u}_s \cdot \nabla_{\boldsymbol{x}} \eta - w_s = \partial_t \phi + \tilde{\boldsymbol{u}} \cdot \nabla_{\boldsymbol{x}} \phi - \tilde{w}|_{\lambda=1} = \partial_t \phi + \tilde{\boldsymbol{u}} \cdot \nabla_{\boldsymbol{x}} \phi - \tilde{w}|_{\lambda=0} = 0 \,,$$

which gives $(3.6)_1$. In view of (3.19)-(3.20) we deduce $(3.6)_3$ from $(3.16)_2$. □

3.3 Particular solutions

In general, solutions of non-linear hyperbolic equations may become discontinuous after a finite critical time at which the space derivative of the transported quantity blows up, see [68]. This may occur for the transformation $\phi(t, \boldsymbol{x}, \lambda)$ solution of (3.11), in which case it will eventually evolve towards a discontinuous solution, as we will see just below. Indeed, in this section, we discuss some properties of particular solutions and the link between the two formulations.

More precisely, we will consider first stationary solutions of (3.6) depending only on the vertical variable z, then we characterize the stationary solutions of (3.16) when $d = 1$. We also consider the shallow water regime and finally exhibit an explicit solution with vorticity. These particular solutions allow us to highlight some properties of the two formulations and the link between them.

3.3.1 Stationary solutions of the hydrostatic Euler system depending only on z

In this subsection, we assume z_b to be constant (null for simplicity). If $(\eta, \boldsymbol{u}, w)$ is a stationary solution of (3.6) such that $\boldsymbol{u}$ does not depend on $\boldsymbol{x}$, then there exist a constant $\eta_0 > 0$ and a function $\boldsymbol{f} : \mathbb{R} \to \mathbb{R}^d$ such that

$$\eta(t, \boldsymbol{x}) = \eta_0, \qquad \boldsymbol{u}(t, \boldsymbol{x}, z) = \boldsymbol{f}(z), \qquad w(t, \boldsymbol{x}, z) = 0.$$

In this case, (3.11) defining the evolution of ϕ reduces to

$$\begin{cases} \dfrac{\partial \phi}{\partial t} + \boldsymbol{f}(\phi) \cdot \nabla_{\boldsymbol{x}} \phi = 0 \\ \qquad\qquad \phi(0, \boldsymbol{x}, \lambda) = \phi_0(\boldsymbol{x}, \lambda). \end{cases} \tag{3.21}$$

We assume that $\boldsymbol{f}$ is regular enough. However, several choices for the initial condition ϕ_0 are possible as will be discussed below:

- Taking the initial condition
$$\phi_0(\boldsymbol{x}, \lambda) = \varphi_0(\lambda),$$
with $\varphi_0(0) = 0$, $\varphi_0(1) = \eta_0$ and $\varphi_0' > 0$, the solution of (3.21) remains constant in $t, \boldsymbol{x}$ and is given by
$$\phi(t, \boldsymbol{x}, \lambda) = \varphi_0(\lambda).$$

- For a more general example in the one-dimensional case $d = 1$, we can consider
$$\phi_0(x, \lambda) = \lambda(\eta_0 + (1 - \lambda) a(x)),$$

for some arbitrary function $a \in C_b^1(\mathbb{R})$. The condition $\partial_\lambda \phi_0 > 0$ holds provided $\|a\|_{L^\infty} < \eta_0$. In particular taking $\boldsymbol{f}(\phi) = \phi$ leads to a Burgers equation. The time T for which a smooth solution ϕ exists is given by the method of characteristics, see [51, Chapter I]. More precisely, since

$$\partial_x \phi_0(x, \lambda) = \lambda(1-\lambda) a'(x),$$

one has

$$\frac{1}{T} = -\lambda(1-\lambda) \inf\big(\{0\} \cup a'(\mathbb{R})\big),$$

which means that for each λ in I there is a critical time T after which ϕ is no longer well-defined.

This example enlightens the dependence of the change of variable ϕ with respect to the choice of the initial data ϕ_0. Besides, this proves that the local well-posedness established in Theorem 2 cannot be global in general. It also illustrates the fact that a stationary solution of (3.6) in the domain Ω_t does not necessarily correspond to a stationary solution of (3.16) in the domain $\tilde{\Omega}$.

3.3.2 Stationary solutions of the semi-Lagrangian formulation

It is important to note that the original Euler system (3.6) and the semi-Lagrangien formulation (3.16) do not share the same stationary solutions. While stationary solutions of (3.16) are stationary solutions of (3.6), the converse is not true in general, as enlightened in Section 3.3.1. As seen before, if $(\eta, \boldsymbol{u}, w)$ is a stationary solution of (3.6), then ϕ might depend on time according to the choice of the initial data ϕ_0. However, if the equation

$$\begin{cases} \boldsymbol{u}(\boldsymbol{x}, \phi(\boldsymbol{x}, \lambda)) \cdot \nabla_{\boldsymbol{x}} \phi = w(\boldsymbol{x}, \phi(\boldsymbol{x}, \lambda)) \\ \phi(\boldsymbol{x}, 0) = z_b(\boldsymbol{x}) \\ \phi(\boldsymbol{x}, 1) = \eta(\boldsymbol{x}), \end{cases}$$

has a solution, then (3.11) admits a steady state, which leads to a stationary solution $(H, \tilde{\boldsymbol{u}})$ of (3.16). We note that the first equation above is satisfied for $\lambda \in (0,1)$. This equation being quasilinear, in theory it is possible to solve it for example by starting with some initial data on a manifold of co-dimension one. More precisely, the strategy is to take $\phi(\boldsymbol{x}, \lambda) = (1-\lambda) z_b(\boldsymbol{x}) + \lambda \eta(\boldsymbol{x})$ as initial data on a smooth manifold $\mathcal{M} \subset \mathbb{R}^d$ of co-dimension one such that $\boldsymbol{u}(\boldsymbol{x}, \phi(\boldsymbol{x}, \lambda)) \cdot \boldsymbol{n} \neq 0$ for $(\boldsymbol{x}, \lambda) \in \partial\mathcal{M} \times I$ where $\boldsymbol{n}$ denotes the normal to $\mathcal{M}$. Then by the method of characteristics, a solution exists provided the characteristics are defined globally.

In the remaining of this section we characterize the stationary solutions of the new system (3.16) in one dimension.

Proposition 13. *For $d = 1$, the following properties are equivalent:*

- *$(H, \tilde{u})$ is a stationary solution of (3.16) with $H > 0$ and $\tilde{u} > 0$;*
- *there exists three functions $F, G, Q : I \to \mathbb{R}$, with $F > G$, such that:*

$$\begin{aligned} \tilde{u}(t, x, \lambda) &= \sqrt{F(\lambda) - G(x)} > 0\,, \\ H(t, x, \lambda) &= \frac{Q(\lambda)}{\sqrt{F(\lambda) - G(x)}} > 0\,, \\ z_b(x) &= \frac{G(x)}{2g} - \int_0^1 \frac{Q(\lambda)}{\sqrt{F(\lambda) - G(x)}} \mathrm{d}\lambda\,. \end{aligned}$$

Proof. For $d = 1$, stationary solutions are characterized by

$$\partial_x(H\tilde{u}) = 0\,, \qquad\qquad \partial_x\left(\frac{\tilde{u}^2}{2} + g\int_0^1 H\mathrm{d}\lambda + gz_b\right) = 0,$$

which is equivalent to the existence of three functions $F, G, Q : I \to \mathbb{R}$ such that:

$$H\tilde{u} = Q(\lambda)\,, \qquad\qquad \tilde{u}^2 + G(x) = F(\lambda)\,,$$

where

$$G(x) = 2g\int_0^1 H\mathrm{d}\lambda + 2gz_b\,.$$

The equivalence then follows by using the positivity assumptions. □

The previous result is not very explicit as the topography is given in terms of the functions F, G, and Q. However, when $Q = F'$ this simplifies to a more explicit form:

Corollary 1. *Let $F : I \to \mathbb{R}$ and $G : \mathbb{R} \to \mathbb{R}$ be such that $F(\lambda) - G(x) > 0$ and $F'(\lambda) > 0$ for all $(x, \lambda) \in \tilde{\Omega}$. Then the functions H and $\tilde{u}$ defined by:*

$$\begin{aligned} H(t, x, \lambda) &= \frac{F'(\lambda)}{\sqrt{F(\lambda) - G(x)}}, \\ \tilde{u}(t, x, \lambda) &= 2\sqrt{F(\lambda) - G(x)}, \end{aligned}$$

are stationary solutions of (3.16), provided the topography satisfies

$$z_b(x) = \frac{G(x)}{2g} - 2\sqrt{F(1) - G(x)} + 2\sqrt{F(0) - G(x)}.$$

Moreover the quantities h, η, ϕ can be computed explicitly:

$$\begin{aligned} h &= \int_0^1 H\mathrm{d}\lambda = 2\sqrt{F(1) - G(x)} - 2\sqrt{F(0) - G(x)}, \\ \eta &= z_b + h = \frac{G(x)}{2g}, \\ \phi &= z_b + \int_0^\lambda H\mathrm{d}\lambda = \frac{G(x)}{2g} + 2\sqrt{F(\lambda) - G(x)} - 2\sqrt{F(1) - G(x)}. \end{aligned}$$

Proof. By taking $Q = F'$ in Proposition 13, the results follow from explicit calculations since

$$2\partial_\lambda\left(\sqrt{F(\lambda) - G(x)}\right) = \frac{F'(\lambda)}{\sqrt{F(\lambda) - G(x)}}.$$

□

3.3.3 Shallow water flows

Shallow water flows are characterized by the fact that the horizontal velocity $\boldsymbol{u}$ does not depend on the vertical variable z, *i.e.*,

$$\boldsymbol{u}(t, \boldsymbol{x}, z) = \boldsymbol{u}(t, \boldsymbol{x}).$$

In this case, (3.6) reduces exactly to the shallow water system. In such a situation, $(3.11)_1$ becomes

$$\frac{\partial \phi}{\partial t} + \nabla_{\boldsymbol{x}} \cdot \big((\phi - z_b)\boldsymbol{u}\big) = 0,$$

as results from the conservative form (3.12). Considering the canonical initial condition (3.14)

$$\phi_0(\boldsymbol{x}, \lambda) = \lambda \eta_0(\boldsymbol{x}) + (1-\lambda) z_b(\boldsymbol{x}),$$

one can check explicitly using (3.8) that the solution of (3.11) is given by

$$\phi(t, \boldsymbol{x}, \lambda) = \lambda \eta(t, \boldsymbol{x}) + (1-\lambda) z_b(\boldsymbol{x}).$$

In view of (3.15), H and $\tilde{\boldsymbol{u}}$ do not depend on λ:

$$H(t, \boldsymbol{x}, \lambda) = \eta(t, \boldsymbol{x}) - z_b(\boldsymbol{x}), \qquad \tilde{\boldsymbol{u}}(t, \boldsymbol{x}, \lambda) = \boldsymbol{u}(t, \boldsymbol{x}),$$

and (3.16) reduces to the shallow water equations. Hence with a suitable initial condition for ϕ, (3.16) and the classical shallow water system share the same solutions.

3.3.4 A flow with an horizontal velocity depending on the vertical coordinate

In this subsection, p^a is not supposed constant. The following paragraph exhibits a situation where the solution ϕ of (3.11) is global but converges asymptotically for large time to a discontinuous function. This shows a situation where the change of variable can hardly be used for numerical implementation.

For $\eta_0 \in \mathbb{R}$, the functions η, $\boldsymbol{u}$, w defined by

$$\eta(t, \boldsymbol{x}) = \eta_0, \qquad \boldsymbol{u}(t, \boldsymbol{x}, z) = \boldsymbol{x}\left(z - \frac{\eta_0}{2}\right), \qquad w(t, \boldsymbol{x}, z) = z(\eta_0 - z),$$

are solutions of (3.6) for $z_b = 0$ and the atmospheric pressure given by

$$p^a = -|\boldsymbol{x}|^2 \frac{\eta_0^2}{8}.$$

See Fig. 3.2 for a graphical representation of the solution. With the initial condition

$$\phi_0(\boldsymbol{x}, \lambda) = \lambda \eta_0,$$

the solution of (3.11) is given by

$$\phi(t, \boldsymbol{x}, \lambda) = \frac{\lambda \eta_0}{(1-\lambda)e^{-t\eta_0} + \lambda}. \tag{3.22}$$

Hence when t becomes large, $H = \partial_\lambda \phi$ goes to zero when $\lambda \neq 0$, ϕ evolves towards a discontinuous function asymptotically in time and the change of variable (3.15) becomes difficult to implement. This difficulty is presented in Fig. 3.3, which shows the evolution of ϕ for $t \in (0, 6)$ and $\eta_0 = 1$. One can clearly see that ϕ evolves towards a discontinuous function in infinite time.

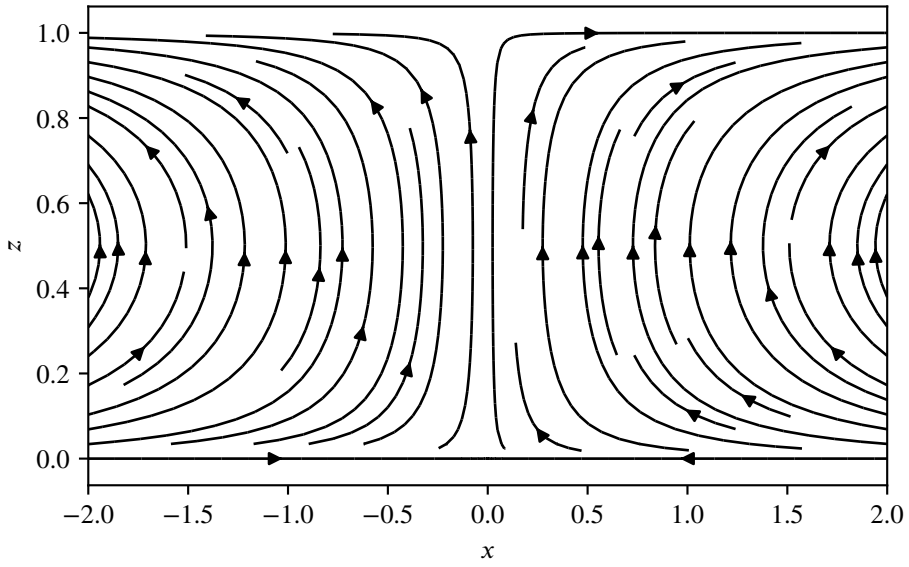

Figure 3.2: Explicit flow with vorticity proposed in Section 3.3.4 with $\eta_0 = 1$.

3.4 Spectrum and Riemann invariants

The semi-Lagrangian formulation (3.16) has a chance for having an hyperbolic structure so it is quite natural to try to determine the eigenvalues and Riemann invariants of the associated operator. Due to the presence of the integral in λ this is not a standard hyperbolic system, the matrix is replaced by an operator acting on functions of λ.

For simplicity we first consider the case $z_b = 0$. The case of non-zero topography will be treated in Section 3.4.4. For $d = 2$, the system (3.16) can be rewritten as

$$\begin{cases} \dfrac{\partial \boldsymbol{U}}{\partial t} + A_1(\boldsymbol{U})\dfrac{\partial \boldsymbol{U}}{\partial x} + A_2(\boldsymbol{U})\dfrac{\partial \boldsymbol{U}}{\partial y} = 0 \\ \boldsymbol{U}(0) = (H_0, \tilde{u}_0, \tilde{v}_0)^T \end{cases} \tag{3.23}$$

where $\boldsymbol{U} = (H, \tilde{\boldsymbol{u}})^T = (H, \tilde{u}, \tilde{v})^T$ and

$$A_1(\boldsymbol{U}) = \begin{pmatrix} \tilde{u} & H & 0 \\ g\int_0^1 \cdot \,\mathrm{d}\lambda & \tilde{u} & 0 \\ 0 & 0 & \tilde{u} \end{pmatrix}, \qquad A_2(\boldsymbol{U}) = \begin{pmatrix} \tilde{v} & 0 & H \\ 0 & \tilde{v} & 0 \\ g\int_0^1 \cdot \,\mathrm{d}\lambda & 0 & \tilde{v} \end{pmatrix}, \tag{3.24}$$

are non-local operators acting on functions in the variable λ. The first aim is to study the hyperbolicity of (3.23). Hence, we may consider that the variables $(t, \boldsymbol{x})$ are non-essential for the study of hyperbolicity of the operators $A_1(\boldsymbol{U})$ and $A_2(\boldsymbol{U})$ so we will often not write the dependence on $(t, \boldsymbol{x})$ explicitly for the sake of simplicity. Let $\boldsymbol{\xi} = (\xi_1, \xi_2) \in \mathbb{R}^2$ such that $\|\boldsymbol{\xi}\| = 1$ and consider a linear combination of the operators $A_1(\boldsymbol{U})$ and $A_2(\boldsymbol{U})$ given by (3.24):

$$A_{\boldsymbol{\xi}}(\boldsymbol{U}) := \xi_1 A_1(\boldsymbol{U}) + \xi_2 A_2(\boldsymbol{U}).$$

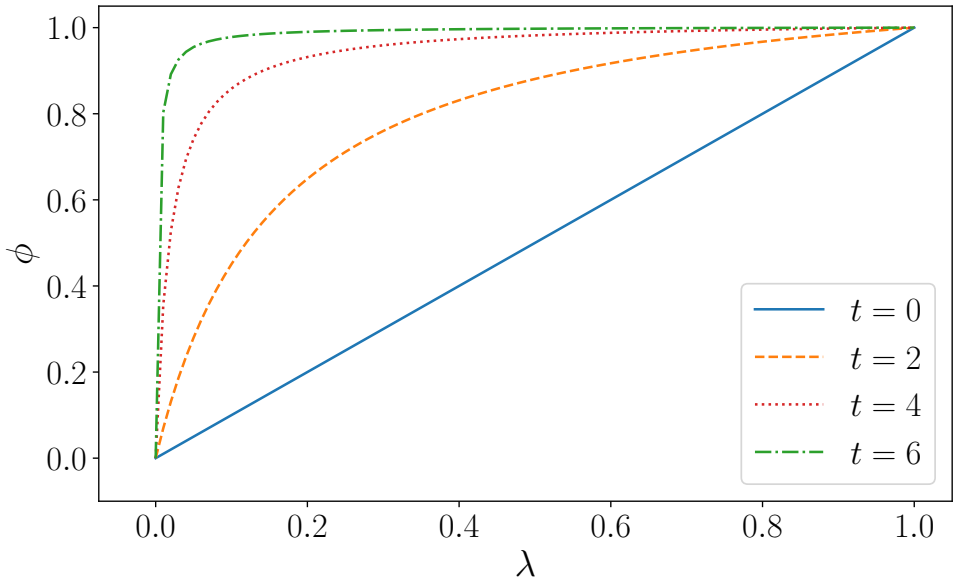

Figure 3.3: Graph of ϕ defined by (3.22) for $\eta_0 = 1$ at times $t = 0, 2, 4, 6$.

System (3.23) is invariant by the rotation along the vertical axis, see [51, Chapter 5]. By defining the following rotation matrix

$$R = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{pmatrix},$$

we have

$$R^{-1} A_\xi(U) R = \begin{pmatrix} \xi \cdot \tilde{u} & H & 0 \\ g \int_0^1 \cdot d\lambda & \xi \cdot \tilde{u} & 0 \\ 0 & 0 & \xi \cdot \tilde{u} \end{pmatrix},$$

with $\xi_1 = \cos\theta, \xi_2 = \sin\theta$. Hence the operators $R^{-1} A_\xi(U) R$ and $A_\xi(U)$ share the same spectrum. Note that $R^{-1} A_\xi(U) R = A_1(U_\xi)$ where $U_\xi = (H, \xi \cdot \tilde{u})^T$. Moreover in view of the structure of $R^{-1} A_\xi(U) R$, the third variable $\tilde{v}$ is decoupled from the first two variables, which motivates the study of the spectrum of the 2×2 matrix $A_0(\tilde{U})$ corresponding to the case $d = 1$ given by

$$A_0(\tilde{U}) = \begin{pmatrix} \tilde{u} & H \\ g \int_0^1 \cdot \, \mathrm{d}\lambda & \tilde{u} \end{pmatrix}, \quad \text{where} \quad \tilde{U} = \begin{pmatrix} H \\ \tilde{u} \end{pmatrix}.$$

Provided that $\tilde{u} \in L^\infty(I)$ and $H \in L^\infty(I)$, $A_0(\tilde{U}) : L^2(I)^2 \to L^2(I)^2$ is a bounded operator. We can rewrite (3.16) when $d = 1$ in the following form

$$\frac{\partial \tilde{U}}{\partial t} + A_0(\tilde{U}) \frac{\partial \tilde{U}}{\partial x} = 0, \tag{3.25}$$

and the aim of the next section is to study the spectrum of the operator A_0.

3.4.1 Characterization of the spectrum

For a linear operator on an infinite dimensional space, say $A_0(\tilde{U}) : L^2(I)^2 \to L^2(I)^2$, the notion of eigenvalues becomes more inclusive and an analysis of the spectrum is essential. To simplify the notation, and as $\tilde{U}$ is fixed, we omit the argument $\tilde{U}$ in $A_0(\tilde{U})$ and simply use the notation A_0, except in certain situations where there is ambiguity. The spectrum of A_0 is defined as the set $\sigma(A_0)$ of all $c \in \mathbb{C}$ for which the operator $A_0 - c\mathbf{I}$ is not invertible. We will use some standard definitions related to the decomposition of the spectrum (for the reader's convenience, these definitions are recalled in Appendix 3.A).

We denote by $C^{0,\alpha}(I)$ the space of Hölder continuous functions of index α over I. The authors in [91] and [89] derived a formula satisfied by the eigenvalues of A_0 given by equation (3.32) below. However, no regularity assumptions had been imposed and the computation of eigenelements was performed without precision on the regularity of the variables H and $\tilde{u}$. The following theorem is an extension of the results given in [91] into a complete characterization of the spectrum of the operator A_0 precisely when the velocity $\tilde{u}$ is $\frac{1}{4}$-Hölder continuous over I. This regularity for $\tilde{u}$ turns out to be the limiting case for the existence of at least two real and distinct eigenvalues outside the range of values of the velocity $\tilde{u}$ as presented in Proposition 15.

Let us state our main result on the characterization of the spectrum of A_0.

Theorem 5. *If $\tilde{u} \in C^{0,1/4}(I)$, $H \in C^0(I)$ and $H > 0$, then the spectrum of $A_0 : L^2(I)^2 \to L^2(I)^2$ is characterized by*

$$\begin{aligned}
\sigma_p(A_0) &= \left\{ c \in \mathbb{C} \setminus \tilde{u}(I) \,:\, \int_0^1 \frac{gH}{(c-\tilde{u})^2} \mathrm{d}\lambda = 1 \right\} \cup \left\{ c \in \tilde{u}(I) \,:\, \text{meas}\left(\tilde{u}^{-1}(c)\right) > 0 \right\}, \\
\sigma_c(A_0) &= \left\{ c \in \tilde{u}(I) \,:\, \text{meas}\left(\tilde{u}^{-1}(c)\right) = 0 \right\}, \\
\sigma_r(A_0) &= \emptyset,
\end{aligned}$$

where the definitions of the point spectrum $\sigma_p(A_0)$, continuous spectrum $\sigma_c(A_0)$, and residual spectrum $\sigma_r(A_0)$ are given in Appendix 3.A.

Proof. The proof is divided into three steps: injectivity, surjectivity, and the characterization of the range.

1. *Injectivity:* The first step is to determine when $A_0 - c\mathbf{I}$ is not injective, *i.e.* we determine the values c such that there exists a non-zero solution $\mathbf{\Phi} = (\phi_1, \phi_2) \in L^2(I)^2$ of

$$(A_0 - c\mathbf{I})\,\mathbf{\Phi} = 0\,,$$

or explicitly

$$H\phi_2 = (c - \tilde{u})\phi_1\,, \tag{3.26}$$

$$g\int_0^1 \phi_1 \mathrm{d}\lambda = (c-\tilde{u})\phi_2\,, \tag{3.27}$$

which yields in particular

$$gH\int_0^1 \phi_1 \mathrm{d}\lambda = (c-\tilde{u})H\phi_2 = (c-\tilde{u})^2\phi_1\,.$$

One has to distinguish the cases where $\tilde{u}^{-1}(c)$ has zero or nonzero measure.

(a) If $\tilde{u}^{-1}(c)$ has nonzero measure, then there exists a nontrivial function $\phi_1 \in L^2(I)$ with $\phi_1 = 0$ on $I \setminus \tilde{u}^{-1}(c)$ and such that

$$\int_0^1 \phi_1 \mathrm{d}\lambda = \int_{\tilde{u}^{-1}(c)} \phi_1 \mathrm{d}\lambda = 0\,.$$

One can check easily that such a ϕ_1 is a solution of (3.26) and (3.27) together with $\phi_2 = 0$. Thus if $\tilde{u}^{-1}(c)$ has nonzero measure, then $A_0 - c\mathbf{I}$ is not injective, and these values c are in the point spectrum $\sigma_p(A_0)$.

(b) If $\tilde{u}^{-1}(c)$ has zero measure, the solution of (3.26)-(3.27) is given by

$$\phi_1 = \frac{gH}{(c-\tilde{u})^2} \int_0^1 \phi_1 \mathrm{d}\lambda\,, \tag{3.28}$$

$$\phi_2 = \frac{g}{c-\tilde{u}} \int_0^1 \phi_1 \mathrm{d}\lambda\,, \tag{3.29}$$

and it remains to check whether $\mathbf{\Phi} = (\phi_1, \phi_2)$ is in $L^2(I)^2$.
Obviously $\mathbf{\Phi}$ is trivial if and only if $\int_0^1 \phi_1 \mathrm{d}\lambda = 0$, so without lost of generality ϕ_1 can be normalized such that $\int_0^1 \phi_1 \mathrm{d}\lambda = 1$. Then equations (3.28), (3.29) reduce to

$$\phi_1 = \frac{gH}{(c-\tilde{u})^2}\,, \tag{3.30}$$

$$\phi_2 = \frac{g}{(c-\tilde{u})}\,. \tag{3.31}$$

We now have to distinguish the cases $c \in \tilde{u}(I)$ and $c \notin \tilde{u}(I)$:

- If $c \in \tilde{u}(I)$, there exists $\lambda_0 \in I$ such that $\tilde{u}(\lambda_0) = c$ and since $\tilde{u}$ is $\frac{1}{4}$- Hölder continuous, there exists a constant $K > 0$ such that for all $\lambda \in I$,

$$|c - \tilde{u}(\lambda)| = |\tilde{u}(\lambda) - \tilde{u}(\lambda_0)| \leq K\,|\lambda - \lambda_0|^{1/4}\,.$$

Therefore

$$\frac{1}{|c-\tilde{u}(\lambda)|} \geq \frac{1}{K\,|\lambda-\lambda_0|^{1/4}}\,,$$

and $\dfrac{1}{(c-\tilde{u})^2} \notin L^2(I)$, so there is no non-trivial solution for $\mathbf{\Phi}$. Thus if $c \in \tilde{u}(I)$ and $\tilde{u}^{-1}(c)$ has zero measure, then $A_0 - c\mathbf{I}$ is injective, so such values c are not in the point spectrum $\sigma_p(A_0)$.

- On the other hand, if $c \notin \tilde{u}(I)$, then by continuity of $\tilde{u}$ there exists $\alpha > 0$ such that $|c - \tilde{u}| > \alpha$, and therefore $\dfrac{1}{c-\tilde{u}} \in C^0(I)$ and $\dfrac{1}{(c-\tilde{u})^2} \in C^0(I)$. Hence, equations (3.26) and (3.27) admit a solution $\mathbf{\Phi} \in L^2(I)^2$. As explained above the solution $\mathbf{\Phi}$ is not trivial as long as $\int_0^1 \phi_1 \mathrm{d}\lambda \neq 0$, so integrating (3.28) on λ implies the following characterization for having a non-trivial solution $\mathbf{\Phi}$:

$$\int_0^1 \frac{gH}{(c-\tilde{u})^2} \mathrm{d}\lambda = 1\,. \tag{3.32}$$

This terminates the characterization of the injectivity of $A_0 - c\mathbf{I}$, hence the claimed charac-

terization of the point-spectrum.

2. *Surjectivity:* One has to determine the range of $A_0 - c\mathbf{I}$, *i.e.* determining for which $\boldsymbol{b} = (b_1, b_2)$ the system

$$\begin{aligned} H\phi_2 &= (c-\tilde{u})\phi_1 - b_1\,, \\ g\int_0^1 \phi_1 \mathrm{d}\lambda &= (c-\tilde{u})\phi_2 - b_2\,, \end{aligned}$$

has a solution $\boldsymbol{\Phi} = (\phi_1, \phi_2) \in L^2(I)^2$. Obviously one has to consider only the case where $\tilde{u}^{-1}(c)$ has zero measure and $c \in \tilde{u}(I)$. One has

$$gH\int_0^1 \phi_1 \mathrm{d}\lambda = (c-\tilde{u})H\phi_2 - Hb_2 = (c-\tilde{u})^2\phi_1 - (c-\tilde{u})b_1 - Hb_2\,,$$

therefore

$$\begin{aligned} \phi_1 &= \frac{H}{(c-\tilde{u})^2}\left(g\int_0^1 \phi_1 \mathrm{d}\lambda + b_2\right) + \frac{b_1}{c-\tilde{u}}\,, \\ \phi_2 &= \frac{1}{c-\tilde{u}}\left(g\int_0^1 \phi_1 \mathrm{d}\lambda + b_2\right)\,. \end{aligned}$$

Since $c \in \tilde{u}(I)$, then as before $\dfrac{1}{(c-\tilde{u})^2} \notin L^2(I)$ and $A_0 - c\mathbf{I}$ is not surjective so $\sigma_c(A_0) \cup \sigma_r(A_0) = \left\{c \in \tilde{u}(I) \,:\, \mathrm{meas}\left(\tilde{u}^{-1}(c)\right) = 0\right\}$.

3. *Density of the range:* It remains to characterize the range of $A_0 - c\mathbf{I}$ when $c \in \tilde{u}(I)$ and $\tilde{u}^{-1}(c)$ has zero measure. One first proves that the range is dense in $L^2(I)^2$, so that $\sigma_r(A_0) = \emptyset$ and consequently $\sigma_c(A_0) = \{c \in \tilde{u}(I) \,:\, \mathrm{meas}\left(\tilde{u}^{-1}(c)\right) = 0\}$. Let us define the following characteristic function

$$\chi_n(\lambda) = \begin{cases} 1 & \text{if } |\tilde{u}(c) - \lambda| > \frac{1}{n}\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

Let $\boldsymbol{b} = (b_1, b_2) \in L^2(I)^2$ and consider $\boldsymbol{\Phi}_n = (\phi_{1n}, \phi_{2n}) \in L^2(I)^2$ defined by

$$\begin{aligned} \phi_{1n} &= \frac{\chi_n H}{(c-\tilde{u})^2}\left[g\alpha_n + b_2\right] + \frac{\chi_n b_1}{c-\tilde{u}}\,, \\ \phi_{2n} &= \frac{\chi_n}{(c-\tilde{u})}\left[g\alpha_n + b_2\right]\,, \end{aligned}$$

where

$$\alpha_n = \kappa^{-1}\int_0^1 \chi_n\left(\frac{b_1}{c-\tilde{u}} + \frac{Hb_2}{(c-\tilde{u})^2}\right)\mathrm{d}\lambda\,,$$

and

$$\kappa = 1 - \int_0^1 \frac{gH}{(c-\tilde{u})^2}\mathrm{d}\lambda\,.$$

Therefore $\boldsymbol{b}_n = (A_0 - c\mathbf{I})\boldsymbol{\Phi}_n$ is explicitly given by

$$\begin{aligned} b_{n1} &= \chi_n b_1\,, \\ b_{n2} &= \chi_n b_2 + (\chi_n - 1) g\alpha_n - g\alpha_n \int_0^1 \frac{(\chi_n - 1)gH}{(c-\tilde{u})^2} \mathrm{d}\lambda\,. \end{aligned}$$

Since $\boldsymbol{b}_n$ converges pointwise to $\boldsymbol{b}$ and $\boldsymbol{b}_n$ is uniformly bounded in $L^2(I)^2$, the Lebesgue dominated convergence theorem implies that $\boldsymbol{b}_n$ converges to $\boldsymbol{b}$ in $L^2(I)^2$, proving the density of the range of $A_0 - c\mathbf{I}$ in $L^2(I)^2$ when $c \in \sigma_c(A_0) \cup \sigma_r(A_0)$. Hence $\sigma_r(A_0) = \emptyset$ and therefore $\sigma_c(A_0) = \{c \in \tilde{u}(I) \,:\, \text{meas}\left(\tilde{u}^{-1}(c)\right) = 0\}$.

□

From this result, we can easily deduce the following alternative characterization of the spectrum which is useful to distinguish between two types of elements in the spectrum: the elements that satisfy the integral condition and the range of values of the velocity $\tilde{u}$ on the interval I.

Corollary 2. *Under the hypotheses of Theorem 5, the spectrum of $A_0 : L^2(I)^2 \to L^2(I)^2$ is characterized by*

$$\begin{aligned} \sigma_d(A_0) &= \left\{ c \in \mathbb{C} \setminus \tilde{u}(I) \,:\, \int_0^1 \frac{gH}{(c-\tilde{u})^2} \mathrm{d}\lambda = 1 \right\}, \\ \sigma_e(A_0) &= \tilde{u}(I)\,, \end{aligned}$$

where $\sigma_d(A_0)$ is the discrete spectrum and $\sigma_e(A_0)$ is the essential spectrum defined in Appendix 3.A.

Proof. The operator A_0 is a compact perturbation of the operator A_0 with $g = 0$. More precisely, we have $A_0 = D_0 + K_0$, where

$$D_0 = \begin{pmatrix} \tilde{u} & H \\ 0 & \tilde{u} \end{pmatrix}, \qquad\qquad K_0 = \begin{pmatrix} 0 & 0 \\ g\int_0^1 \cdot \,\mathrm{d}\lambda & 0 \end{pmatrix},$$

and K_0 is compact (even finite-rank). Using the fact that the essential spectrum is invariant under compact perturbations, we directly deduce the corollary using Theorem 5. □

Remark 10. *The above results can be easily generalized for $A_0 : L^p(I)^2 \mapsto L^p(I)^2$ and $H \in C(I)$, $\tilde{u} \in C^{0,\alpha}(I)$ such that $\alpha p \geq 1/2$.*

Remark 11. *The integral condition (3.32) can be rewritten as follows in the original variables:*

$$1 = \int_0^1 \frac{gH}{(c-\tilde{u})^2} \mathrm{d}\lambda = \int_{z_b}^{\eta} \frac{g}{(c-u)^2} \mathrm{d}z.$$

Remark 12. *In the shallow water regime (see Section 3.3.3),* i.e., *when $\tilde{u}$ is independent of λ, then the spectrum reduces to*

$$\sigma(A_0) = \left\{ \tilde{u} + \sqrt{gh}, \tilde{u} - \sqrt{gh}, \tilde{u} \right\}, \qquad \textit{where} \qquad h = \int_0^1 H \mathrm{d}\lambda = \eta - z_b.$$

The first two eigenvalues $\tilde{u} \pm \sqrt{gh}$ are the usual eigenvalues of the Saint-Venant system, whereas $\tilde{u}$ is a spurious eigenvalues coming from the fact that adding a dependency in λ in Saint-Venant, which is expressed through a transport equation with velocity u for ϕ, is artificial.

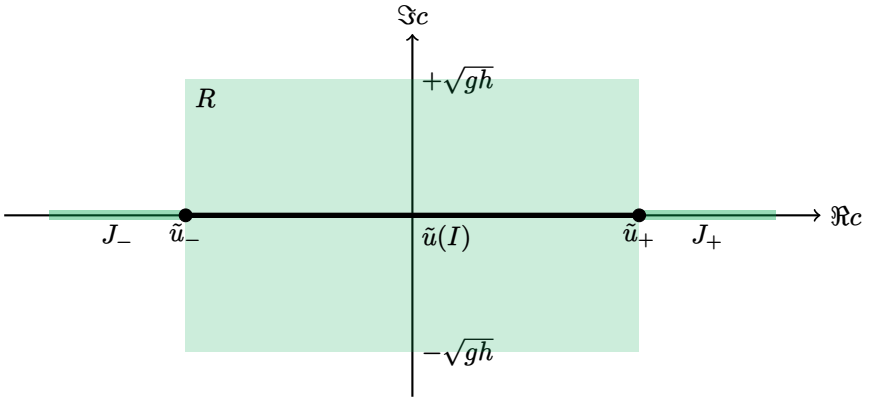

Figure 3.4: The spectrum $\sigma(A)$ is included in the rectangle R and intervals $J_\pm$ defined in Proposition 14. The line $\tilde{u}(I)$ between $\tilde{u}_-$ and $\tilde{u}_+$ is the continuous part of the spectrum.

We have the following localization result of the spectrum.

Proposition 14. *Under the assumptions of Theorem 5, $\sigma(A_0)$ is contained in the union of the following three sets represented on Fig. 3.4:*

$$J_- = \left[\tilde{u}_- - \sqrt{gh}, \tilde{u}_-\right], \qquad J_+ = \left[\tilde{u}_+, \tilde{u}_+ + \sqrt{gh}\right]. \tag{3.33}$$

$$R = \left\{ z \in \mathbb{C} \,:\, \tilde{u}_- \leq \Re z \leq \tilde{u}_+ \text{ and } |\Im z| \leq \sqrt{gh} \right\}, \tag{3.34}$$

where

$$\tilde{u}_- = \inf_I \tilde{u}, \qquad \tilde{u}_+ = \sup_I \tilde{u}, \qquad h = \int_0^1 H \mathrm{d}\lambda.$$

Proof. It suffices to prove that $\sigma_d(A_0)$ is included in the claimed set, so to consider the values c satisfying the integral condition (3.32). Therefore this reduces to study the function F defined on $\mathbb{C} \setminus \tilde{u}(I)$ by:

$$F(c) = \int_0^1 \frac{gH}{(c-\tilde{u})^2} \mathrm{d}\lambda \,. \tag{3.35}$$

On the one hand, since

$$\Im F(c) = -2 \int_0^1 \frac{gH(\Re c - \tilde{u})\Im c}{|c-\tilde{u}|^4} \mathrm{d}\lambda \,,$$

if $\Re c > \sup_I \tilde{u}$ or $\Re c < \inf_I \tilde{u}$, then $(\Re c - \tilde{u})$ is either strictly positive or strictly negative on I, so the only way to make the integral $\Im F(c)$ zero, is that $\Im c = 0$. This proves that

$$\sigma(A_0) \subset \{ z \in \mathbb{C} \,:\, \tilde{u}_- \leq \Re z \leq \tilde{u}_+ \} \cup \mathbb{R} \,.$$

On the other hand, if $\mathrm{dist}(c, \tilde{u}(I)) > \sqrt{gh}$, then for any $\lambda \in I$, $|c - \tilde{u}| > \sqrt{gh}$ and therefore

$$|F(c)| = \int_0^1 \frac{gH}{|c-\tilde{u}|^2} \mathrm{d}\lambda < \frac{1}{h} \int_0^1 H \mathrm{d}\lambda = 1 \,,$$

so c cannot be an eigenvalue. Therefore, we have shown that

$$\sigma(A_0) \subset \left\{ z \in \mathbb{C} \,:\, \operatorname{dist}(z, \tilde{u}(U)) \leq \sqrt{gh} \right\} ,$$

which ends the proof. □

3.4.2 Limiting cases

Finding whether the values $c \in \sigma_d(A_0)$ are real or complex is essential to determine the hyperbolicity conditions for the associated system (3.25). As noted in [30,31,87,89,90] for regular velocity profiles there exists at least two real eigenvalues of the operator A_0 outside the range of values of the velocity $\tilde{u}$. In [30], generalized hyperbolicity conditions for system (3.25) with a monotonic velocity profile are formulated using the limit values of (3.32) in the upper and lower complex half-planes. In the general case, as shown in this subsection, these two real solutions can only exist under certain regularity assumptions on the velocity $\tilde{u}(\lambda)$.

We have the following result on the minimum regularity criterion for the existence of these two real eigenvalues.

Proposition 15. *Under the assumptions of Theorem 5 , if either:*

1. $\tilde{u} \in C^{0,1/2}(I)$, *or*
2. $\tilde{u} \in C^{0,1/4}(I)$ *with Hölder constant* $K > 0$ *satisfying* $K \leq \sqrt{g \inf_I H}$,

there exists exactly two real eigenvalues $c_\pm \in J_\pm$, *where the intervals* $J_\pm$ *are defined in (3.33), such that*

$$\sigma(A_0) \cap \mathbb{R} = \{c_-, c_+\} \cup \tilde{u}(I), \quad \text{or} \quad \sigma_d(A_0) \cap \mathbb{R} = \{c_-, c_+\} \, .$$

Proof. In view of (3.32), this reduces to study the function F defined on $J_+ \cup J_-$ by (3.35). One can check that the function F is continuous on $J_\pm$. The first step is to show that

$$\lim_{c \to \tilde{u}_-, c < \tilde{u}_-} F(c) \geq 1 \quad \text{and} \quad \lim_{c \to \tilde{u}_+, c > \tilde{u}_+} F(c) \geq 1.$$

Since $\tilde{u}$ is continuous, there exists $\lambda_\pm \in I$ such that $\tilde{u}(\lambda_\pm) = \tilde{u}_\pm$.

1. If $\tilde{u}$ is $\frac{1}{2}$-Hölder continuous, there exists $K > 0$ such that for all $\lambda \in I$,

   $$|\tilde{u}(\lambda) - \tilde{u}_\pm| \leq K \, |\lambda - \lambda_\pm|^{1/2} \, ,$$

   so for $c = \tilde{u}_\pm \pm \delta$ with $\delta > 0$, we have

   $$|\tilde{u}(\lambda) - c| \leq |\tilde{u}(\lambda) - \tilde{u}_\pm| + \delta \leq K \, |\lambda - \lambda_\pm|^{1/2} + \delta \, .$$

   Therefore, there exists a constant $L > 0$ such that for all $\delta \in (0,1)$,

   $$F(c) = F(\tilde{u}_\pm \pm \delta) \geq \frac{g}{2} \inf_I H \int_0^1 \frac{1}{K^2 \, |\lambda - \lambda_\pm| + \delta^2} \mathrm{d}\lambda \geq L |\log \delta^2| \, ,$$

   which proves

   $$\lim_{c \to \tilde{u}_-, c < \tilde{u}_-} F(c) = \infty \quad \text{and} \quad \lim_{c \to \tilde{u}_+, c > \tilde{u}_+} F(c) = \infty.$$

2. If $\tilde{u}$ is $\frac{1}{4}$-Hölder continuous with constant $K > 0$, then for all $\lambda \in I$,

$$|\tilde{u}(\lambda) - \tilde{u}_\pm| \leq K \left|\lambda - \lambda_\pm\right|^{1/4} ,$$

therefore

$$\begin{aligned} F(\tilde{u}_\pm) &\geq \frac{g}{2} \inf_I H \int_0^1 \frac{1}{K^2 \left|\lambda - \lambda_\pm\right|^{1/2}} \mathrm{d}\lambda = \frac{g}{K^2} \inf_I H \left(\sqrt{1-\lambda_\pm} + \sqrt{\lambda_\pm} \right) \\ &\geq \frac{g}{K^2} \inf_I H \,. \end{aligned}$$

Hence on the one hand if the constant $0 < K \leq 1$ satisfies $K \leq \sqrt{g \inf_I H}$, then

$$\lim_{c \to \tilde{u}_-, c < \tilde{u}_-} F(c) \geq 1 \quad \text{and} \quad \lim_{c \to \tilde{u}_+, c > \tilde{u}_+} F(c) \geq 1.$$

On the other hand, for all $\lambda \in I$ one has $|\tilde{u}_\pm \pm \sqrt{gh} - \tilde{u}| \geq \sqrt{gh}$ so

$$F(\tilde{u}_\pm \pm \sqrt{gh}) \leq 1 \,.$$

Taking the derivative, one has

$$F'(c) = -\int_0^1 \frac{2gH}{(c-\tilde{u})^3} \mathrm{d}\lambda \,,$$

which shows that F is strictly decreasing on J_+ and strictly increasing on J_-. Therefore, there exists exactly one solution of $F(c) = 1$ in J_- and exactly one in J_+. □

Remark 13. *The previous result is almost optimal in two aspects. The first is that no smallness assumption is required in $C^{0,1/2}(I)$ whereas a smallness assumption is required in $C^{0,1/4}(I)$. In fact one can show that $C^{0,1/2}(I)$ is the critical space, as a (more complicated) smallness assumption is also required in $C^{0,\alpha}(I)$ for $\frac{1}{4} < \alpha < \frac{1}{2}$. The second aspect is that this result is also optimal despite the smallness assumption required in $C^{0,1/4}(I)$, in view of the following example. Consider $H(\lambda) = 1$ and $\tilde{u}(\lambda) = K\lambda^{1/4}$ for some $K > 0$, then $\tilde{U} = (H, \tilde{u})$ is a stationary solution of system (3.25). An explicit computation shows that $\tilde{u}_- = 0$, $\tilde{u}_+ = K$, and that F is strictly decreasing on J_+ and strictly increasing on J_-. Moreover, we have the following limits:*

$$\lim_{c \to \tilde{u}_-} F(c) = \frac{g}{K^2} \,, \qquad\qquad \lim_{c \to \tilde{u}_+} F(c) = \infty \,.$$

This proves that there are exactly two real eigenvalues when $K \leq \sqrt{g}$ and exactly one (which is in J_+) when $K > \sqrt{g}$, see Fig. 3.5.

Finally, under additional hypotheses, we can ensure that the spectrum is fully real.

Proposition 16. *If $H \in C^1(I)$ with $H > 0$ and $\tilde{u} \in C^2(I)$ is strictly monotonic in λ and $\partial_\lambda \big(H/\partial_\lambda u\big) \neq 0$ for all $\lambda \in I$ then*

$$\sigma_p(A_0) = \sigma_d(A_0) = \{c_-, c_+\} \,, \qquad \sigma_c(A_0) = \sigma_e(A_0) = \tilde{u}(I), \qquad \sigma_r(A_0) = \emptyset,$$

where $c_\pm$ are defined in Proposition 15.

Remark 14. *The condition $\partial_\lambda \big(H/\partial_\lambda \tilde{u}\big) \neq 0$ is equivalent to the condition $\partial_{zz} u \neq 0$ in the original domain Ω_t. We note that this is precisely the condition proposed in [30, Lemma 2.2)] to study the*

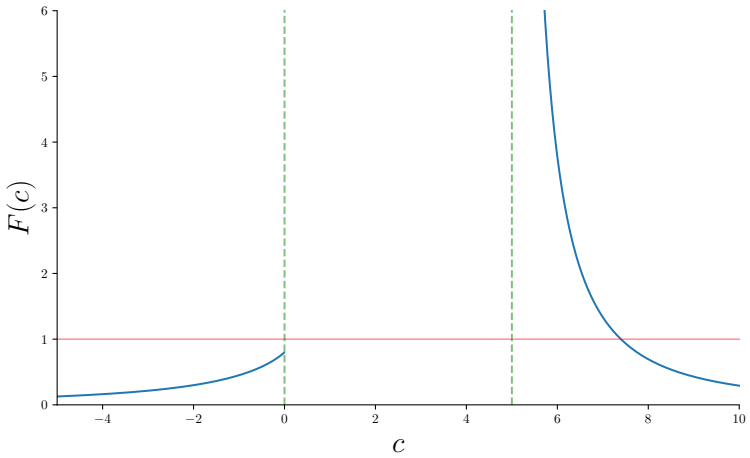

Figure 3.5: The limiting case for $\tilde{u}(\lambda) = K\lambda^{1/4}$, $H(\lambda) = 1$, and $K > \sqrt{g}$ for $g = 10$.

stability of the flow using the Sokhotski–Plemelj formulae.

Proof. Since $\tilde{u}$ is strictly monotonic, we have that $\operatorname{meas}\left(\tilde{u}^{-1}(c)\right) = 0$ for $c \in \tilde{u}(I)$, hence $\sigma_p(A_0) = \sigma_d(A_0)$ and $\sigma_c(A_0) = \sigma_r(A_0) = \tilde{u}(I)$. Therefore, it remains to study the function F defined by (3.35) on $\mathbb{C} \setminus \tilde{u}(I)$. In view of Propositions 14 and 15 it suffices to prove that $F(c) \neq 1$ for all $c \in \mathbb{C}$ such that $\Im c \neq 0$ and $\tilde{u}_- \leq \Re c \leq \tilde{u}_+$. In the following we assume that c satisfies these conditions.

Without loss of generality, we can assume that $\partial_\lambda \tilde{u} > 0$, and therefore, we have

$$F(c) = \int_0^1 \frac{gH}{(c-\tilde{u})^2} \mathrm{d}\lambda = \int_0^1 \frac{gH}{\partial_\lambda \tilde{u}} \partial_\lambda \left(\frac{1}{c-\tilde{u}} \right) \mathrm{d}\lambda = \left[\frac{gH}{\partial_\lambda \tilde{u}} \frac{1}{c-\tilde{u}} \right]_0^1 - G(c)\,,$$

where

$$G(c) = \int_0^1 \partial_\lambda \left(\frac{gH}{\partial_\lambda \tilde{u}} \right) \frac{1}{c-\tilde{u}} \mathrm{d}\lambda = \int_0^1 \partial_\lambda \left(\frac{gH}{\partial_\lambda \tilde{u}} \right) \frac{\bar{c}-\tilde{u}}{|c-\tilde{u}|^2} \mathrm{d}\lambda\,.$$

Since $\partial_\lambda (H/\partial_\lambda \tilde{u}) \neq 0$, by continuity $\partial_\lambda (H/\partial_\lambda \tilde{u})$ does not change sign on I so by using the mean value theorem on the real part of $G(c)$, there exists $\lambda_c \in I$ such that

$$G(c) = (\bar{c} - \tilde{u}(\lambda_c)) \int_0^1 \partial_\lambda \left(\frac{gH}{\partial_\lambda \tilde{u}} \right) \frac{1}{|c-\tilde{u}|^2} \mathrm{d}\lambda\,,$$

and therefore

$$F(c) = \left[\frac{gH}{\partial_\lambda \tilde{u}} \frac{\bar{c}-\tilde{u}}{|c-\tilde{u}|^2} \right]_0^1 - (\bar{c} - \tilde{u}(\lambda_c)) \int_0^1 \partial_\lambda \left(\frac{gH}{\partial_\lambda \tilde{u}} \right) \frac{1}{|c-\tilde{u}|^2} \mathrm{d}\lambda\,.$$

Since $\Im c \neq 0$, by explicitly computing the real and imaginary parts of the previous equation, one

has

$$\Re F(c) + \frac{\Re c - \tilde{u}(\lambda_c)}{\Im c}\Im F(c) = \left[\frac{gH}{\partial_\lambda \tilde{u}}\frac{\Re c - \tilde{u}}{|c-\tilde{u}|^2}\right]_0^1 - (\Re c - \tilde{u}(\lambda_c))\left[\frac{gH}{\partial_\lambda \tilde{u}}\frac{1}{|c-\tilde{u}|^2}\right]_0^1$$
$$= \left[\frac{gH}{\partial_\lambda \tilde{u}}\frac{1}{|c-\tilde{u}|^2}\left(\tilde{u}(\lambda_c) - \tilde{u}\right)\right]_0^1.$$

Since $\tilde{u}(0) \leq \tilde{u}(\lambda_c) \leq \tilde{u}(1)$, we deduce that

$$\Re F(c) + \frac{\Re c - \tilde{u}(\lambda_c)}{\Im c}\Im F(c) \leq 0\,,$$

so

$$\Re F(c) \leq \frac{|\Re c - \tilde{u}(\lambda_c)|}{|\Im c|}|\Im F(c)| \tag{3.36}$$

which is incompatible with having a solution of $F(c) = 1$. □

The convexity assumption $\partial_\lambda\left(H/\partial_\lambda u\right) \neq 0$ cannot be omitted regardless of the sign of the derivative as enlightened by the following counter-example showing the existence of complex eigenvalues hence possible instability.

Proposition 17. *For $a \neq 0$ and $b > 0$ such that*

$$b\tanh b > 1, \qquad |a| < \sqrt{1 - (b\tanh b)^{-1}},$$

the spectrum of A_0 for

$$H(\lambda) = g^{-1}, \qquad \tilde{u}(\lambda) = a\tanh(b(2\lambda - 1))$$

has at least two complex conjugate eigenvalues with $\Re c = 0$ and $\Im c \neq 0$.

Proof. The aim is to study the function $F(c)$ defined by (3.35) for $c \in i\mathbb{R}^*$, *i.e.* $\Re c = 0$ and $\Im c \neq 0$. For simplicity, we write $\nu = \Im c$, so

$$\Re F(i\nu) = \int_0^1 \frac{\tilde{u}^2 - \nu^2}{(\tilde{u}^2 + \nu^2)^2}\mathrm{d}\lambda\,, \qquad \Im F(i\nu) = 2\int_0^1 \frac{\tilde{u}\nu}{(\tilde{u}^2 + \nu^2)^2}\mathrm{d}\lambda\,,$$

The function $\tilde{u}$ being odd with respect to $\lambda = \frac{1}{2}$, we deduce directly that $\Im F(i\nu) = 0$. We can compute $\Re F(i\nu)$ explicitly and in particular we can show that:

$$\lim_{\nu\to 0}\Re F(i\nu) = \frac{1 - (b\tanh b)^{-1}}{a^2}, \qquad \lim_{\nu\to\pm\infty}\Re F(i\nu) = 0.$$

Therefore, under the above hypotheses, $\lim_{\nu\to 0}\Re F(i\nu) > 1$ and by continuity, there exists some $\nu \in (0,\infty)$ such that $\Re F(i\nu) = 1$. □

3.4.3 Generalized Riemann invariants

In this section, we attempt to define a generalized notion of Riemann invariant associated to the eigenvalues c, by analogy to the classical theory of hyperbolic systems. For simplicity only the case $d = 1$ is considered. Given a solution $\tilde{U}$ to (3.25), which we assume regular enough, an associated

eigenvalue $c \in \sigma_p(A_0)$ depends implicitly on t and x (through $\tilde{U}$), but not on λ. We shall call a generalized Riemann invariant associated to $c(t,x)$, any function $R(t,x)$ satisfying:

$$\frac{\partial R}{\partial t} + c\frac{\partial R}{\partial x} = 0.$$

This new definition, introduced in analogy with the Saint-Venant system, allows us to view the generalized Riemann invariants as quantities transported by the velocity $c(t,x)$ for all $\lambda \in I$, see [90].

Proposition 18. *Let $c = c(t,x) \in \sigma_p(A_0)$ be a solution of (3.32), then the relation*

$$\frac{\partial}{\partial t}\left(c - g\int_0^1 \frac{H}{\tilde{u}-c}\mathrm{d}\lambda\right) + c\frac{\partial}{\partial x}\left(c - g\int_0^1 \frac{H}{\tilde{u}-c}\mathrm{d}\lambda\right) = 0, \tag{3.37}$$

holds and the quantity $\tilde{r} = \tilde{r}(t,x)$ defined by

$$\tilde{r} = c - g\int_0^1 \frac{H}{\tilde{u}-c}\mathrm{d}\lambda,$$

can be seen as a generalized Riemann invariant associated to the eigenvalue $c = c(t,x)$.

Proof. Let us consider $c = c(t,x) \in \sigma_p(A_0)$. Computing the derivatives in (3.37) and replacing the time derivatives of H and $\tilde{u}$ using (3.16) leads to:

$$\partial_t\tilde{r} + c\partial_x\tilde{r} = \left(\partial_t c + c\partial_x c + g\int_0^1 \partial_x H\mathrm{d}\lambda\right)\left[1 - \int_0^1 \frac{gH}{(c-\tilde{u})^2}\mathrm{d}\lambda\right].$$

The result follows using the fact that c satisfies (3.32).

As for the classical hyperbolic Saint-Venant system, the result can also be obtained multiplying the system (3.25) on the left by φ, where φ is a vector orthogonal to an eigenvector $\mathbf{\Phi}$ of A_0, see (3.30)-(3.31). Then after simple computations, an integration in λ on $I = (0,1)$ gives the result. □

For the hydrostatic Euler system (3.2)-(3.3) with $d = 1$, the vorticity is defined by $\omega = \partial_z u$ and satisfies the equation

$$\frac{\partial \omega}{\partial t} + u\partial_x\omega + w\partial_z\omega = 0.$$

The following proposition gives a definition of the vorticity for the system (3.16).

Proposition 19. *We define the vorticity of the flow in the domain $\tilde{\Omega}$ as follows:*

$$\tilde{\omega}(t,x,\lambda) = \omega(t,x,\phi(t,x,\lambda)) = \frac{\partial_\lambda \tilde{u}}{H}.$$

The vorticity satisfies the following transport equation:

$$\frac{\partial \tilde{\omega}}{\partial t} + \tilde{u}\frac{\partial \tilde{\omega}}{\partial x} = 0.$$

Therefore, assuming that $\tilde{\omega}$ is smooth enough, for a fixed value of $\lambda \in I$ one can consider $\tilde{\omega}_\lambda(t,x) = \tilde{\omega}(t,x,\lambda)$ as a generalized Riemann invariant associated to the velocity $\tilde{u}_\lambda(t,x) = \tilde{u}(t,x,\lambda)$ in $\sigma_e(A_0)$.

Proof. An explicit calculation leads to

$$\partial_t\tilde{\omega} + \tilde{u}\partial_x\tilde{\omega} = \frac{1}{H}\partial_\lambda\left(\partial_t\tilde{u} + \tilde{u}\partial_x\tilde{u} + g\partial_x\int_0^1 H\mathrm{d}\lambda\right) - \frac{\partial_x\tilde{u}}{H^2}\Big(\partial_t H + \partial_x(H\tilde{u})\Big).$$

Since system (3.25), coming from (3.16) for $d = 1$, writes

$$\frac{\partial H}{\partial t} + \partial_x(H\tilde{u}) = 0, \tag{3.38}$$

$$\frac{\partial \tilde{u}}{\partial t} + \tilde{u}\partial_x\tilde{u} + g\partial_x \int_0^1 H \mathrm{d}\lambda = -g\partial_x z_b. \tag{3.39}$$

and z_b is constant, the proof is straightforward. □

We can now collect theses results with the existence of two real eigenvalues ensured by Proposition 15.

Corollary 3. *If $\tilde{U}$ is an enough regular solution to (3.25), let $c_\pm = c_\pm(t,x)$ denote the two real eigenvalues given by Proposition 15. We have the following list of Riemann invariants for any $\lambda \in I$:*

$$\frac{\partial r_\pm}{\partial t} + c_\pm \frac{\partial r_\pm}{\partial x} = 0, \qquad\qquad \frac{\partial \tilde{\omega}_\lambda}{\partial t} + \tilde{u}_\lambda \frac{\partial \tilde{\omega}_\lambda}{\partial x} = 0,$$

where

$$r_\pm = c_\pm - g\int_0^1 \frac{H}{\tilde{u} - c_\pm}\mathrm{d}\lambda, \qquad \tilde{\omega}_\lambda(t,x) = \tilde{\omega}(t,x,\lambda) = \frac{\partial_\lambda \tilde{u}}{H}, \qquad \tilde{u}_\lambda(t,x) = \tilde{u}(t,x,\lambda).$$

The previous result is an attempt to somehow diagonalize (3.25) and is consistent with the decomposition of the spectrum given by Corollary 2: the two real eigenvalues $c_\pm$ in $\sigma_d(A_0)$ lead to two equations, and the continuous spectrum $\tilde{u}(t,x,I)$ to an infinite number of equations parameterized by λ. However, in the case where complex eigenvalues exist in $\sigma_d(A_0)$, it seems quite complicated to understand and justify how the previous list of Riemann invariant equations are equivalent to the original system (3.25).

Remark 15. *Riemann invariants are often used to build particular solutions of the considered system* e.g. *self-similar solutions. This has not been investigated in this paper but this could help to understand the behavior of the system (3.38)-(3.39) in valuable situations.*

3.4.4 Case with variable topography

In the classical Saint-Venant system, one way to deal with non trivial topography allowing to obtain an hyperbolic system is to add z_b as a variable together with the equation $\partial_t z_b = 0$, see for example [51,53,56]. For $d = 2$ and by following the same idea, we can rewrite (3.23) coupled with topography $z_b(t,x)$ satisfying $\partial_t z_b = 0$:

$$\frac{\partial \hat{U}}{\partial t} + B_1(\hat{U})\frac{\partial \hat{U}}{\partial x} + B_2(\hat{U})\frac{\partial \hat{U}}{\partial y} = 0, \tag{3.40}$$

where $\hat{U} = (H, \tilde{u}, z_b)^T$ and

$$B_1(\hat{U}) = \begin{pmatrix} \tilde{u} & H & 0 & 0 \\ g\int_0^1 \cdot \mathrm{d}\lambda & \tilde{u} & 0 & g \\ 0 & 0 & \tilde{u} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad B_2(\hat{U}) = \begin{pmatrix} \tilde{v} & 0 & H & 0 \\ 0 & \tilde{v} & 0 & 0 \\ g\int_0^1 \cdot \mathrm{d}\lambda & 0 & \tilde{v} & g \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

are non-local operators acting on functions in the variable λ. Hence, we will again write $\tilde{\boldsymbol{u}}(\lambda)$ for the sake of simplicity. As in the case of constant topography presented above, system (3.40) is invariant

by rotation along the vertical axis and by a similar argument, this motivates the analysis of the 3×3 matrix,

$$B_0(\bar{U}) = \begin{pmatrix} \tilde{u} & H & 0 \\ g\int_0^1 \cdot \,\mathrm{d}\lambda & \tilde{u} & g \\ 0 & 0 & 0 \end{pmatrix}.$$

Provided that $\tilde{u} \in L^\infty(I)$ and $H \in L^\infty(I)$, $B_0(\bar{U}) : L^2(I)^3 \to L^2(I)^3$ is a bounded operator. We shall consider the following system corresponding to the case $d = 1$:

$$\frac{\partial \bar{U}}{\partial t} + B_0(\bar{U})\frac{\partial \bar{U}}{\partial x} = 0, \quad \text{where} \quad \bar{U} = \begin{pmatrix} H \\ \tilde{u} \\ z_b \end{pmatrix}.$$

In this case the spectrum decomposition is not changed much, except that zero is always in the point spectrum, similarly to the fact that zero is an eigenvalue for the Saint-Venant system with topography.

Theorem 6. *If $\tilde{u} \in C^{0,1/4}(I)$, $H \in C^0(I)$ and $H > 0$, then the spectrum of $B_0(\bar{U}) : L^2(I)^3 \to L^2(I)^3$ is characterized by*

$$\begin{aligned}
&\sigma_p(B_0) = \{0\} \cup \left\{ c \in \mathbb{C} \setminus \tilde{u}(I) \,:\, \int_0^1 \frac{gH}{(c-\tilde{u})^2}\mathrm{d}\lambda = 1 \right\} \cup \left\{ c \in \tilde{u}(I) \setminus \{0\} \,:\, \operatorname{meas}\left(\tilde{u}^{-1}(c)\right) > 0 \right\}, \\
&\sigma_c(B_0) = \left\{ c \in \tilde{u}(I) \setminus \{0\} \,:\, \operatorname{meas}\left(\tilde{u}^{-1}(c)\right) = 0 \right\}, \\
&\sigma_r(B_0) = \emptyset,
\end{aligned}$$

where the definitions of the point spectrum $\sigma_p(B_0)$, continuous spectrum $\sigma_c(B_0)$, and residual spectrum $\sigma_r(B_0)$ are given in Appendix 3.A.

The proof is very similar to the proof of Theorem 5 and is omitted. The only change is that we have to distinguish the cases $c = 0$ and $c \neq 0$. In addition Propositions 14 to 16 also hold with completely similar proofs.

In the shallow water regime, the spectrum reduces as explained in Remark 12 however with zero as an addition eigenvalue, like for the classical Saint-Venant system with topography.

The Riemann invariants corresponding to the eigenvalues $c \in \sigma_p(B_0)$ are slightly modified to take into account the topography.

Proposition 20. *Let $c = c(t,x) \in \sigma_p(B_0)$ be a solution of (3.32), then the relation*

$$\frac{\partial}{\partial t}\left(c - g\int_0^1 \frac{H}{\tilde{u}-c}\mathrm{d}\lambda - gz_b \right) + c\frac{\partial}{\partial x}\left(c - g\int_0^1 \frac{H}{\tilde{u}-c}\mathrm{d}\lambda - gz_b \right) = 0,$$

holds and the quantity $\bar{r} = \bar{r}(t,x)$ with

$$\bar{r} = c - g\int_0^1 \frac{H}{\tilde{u}-c}\mathrm{d}\lambda - gz_b,$$

can be seen as a generalized Riemann invariant associated to the eigenvalue $c = c(t,x)$.

Remark 16. *For $c = 0$, the associated Riemann invariant is trivially z_b as $\partial_t z_b = 0$. However in case $c = 0$ is solution of (3.32), another Riemann invariant given by the previous proposition might exist.*

3.5 A multi-layer approach

In order to better understand the behavior of the eigenvalues in the interval $\tilde{u}(I)$, and thinking about a possible numerical approximation, we will use a multi-layer discretization of the vertical domain. This approach was introduced by Audusse et al. (see [7]) in order to access the vertical variation of the horizontal velocity profile of the Saint-Venant system. Following this approach we consider a piecewise constant approximation of the velocity $\tilde{u}$ in the variable $\lambda \in I$ and we show in Proposition 23 that for a monotone velocity profile and for a large number of horizontal layers we recover only two real eigenvalues $c_\pm$ corresponding to system (3.25) under the assumptions of Proposition 15.

The vertical coordinate plays a particular role in geophysical flow models. For the numerical approximation of the Navier–Stokes or Euler equations with free surface, multilayer models have been proposed [4,6,25,46] consisting in a piecewise constant approximation of the horizontal velocity field along the vertical axis. Efficient and robust numerical schemes endowed with strong stability properties (well-balancing, discrete maximum principle, discrete entropy inequality) can be obtained to approximate these multilayer models [3]. In this section we propose and study a multilayer version of the model (3.16). In contrast to other multilayer schemes for other problems, in the present case, the piecewise constant approximation is exact, *i.e.*, the truncation error is zero. The objective is to examine the eigenvalues of this multilayer version of the model (3.16).

In this section, we only consider $d = 1$. As described by Fig. 3.6, we consider a discretization of the fluid domain $\tilde{\Omega} = \mathbb{R} \times I$ into N different layers. The layer with index α contains the points of coordinates (x, λ) with $\lambda \in L_\alpha = [\lambda_{\alpha-1/2}, \lambda_{\alpha+1/2}]$ with

$$0 = \lambda_{1/2} < \lambda_{3/2} < \cdots < \lambda_{N+1} = 1.$$

For $\alpha = 1, \ldots, N$, the width of the layer α is given by

$$\gamma_\alpha = \lambda_{\alpha+1/2} - \lambda_{\alpha-1/2},$$

with $\sum_{\beta=1}^{N} \gamma_\beta = 1$. Note that each $\lambda_{\alpha+1/2}$ is a constant corresponding to the interface between the two layers α and $\alpha + 1$ with

$$\lambda_{\alpha+1/2} = \sum_{\beta=1}^{\alpha} \gamma_\beta.$$

For a given $X : \tilde{\Omega} \to \mathbb{R}$, we consider its $\mathbb{P}_0-$approximation in λ having the form:

$$X^N(t, x, \lambda) := \sum_{\alpha=1}^{N} \mathbb{1}_{\lambda \in L_\alpha}(\lambda) X_\alpha(t, x),$$

where $X_\alpha(t, x)$ defined by

$$X_\alpha(t, x) = \frac{1}{\gamma_\alpha} \int_{\lambda_{\alpha-1/2}}^{\lambda_{\alpha+1/2}} X(t, x, \lambda)\, \mathrm{d}\lambda, \tag{3.41}$$

is the average of $X(t, x, \lambda)$ over the layer L_α. We will denote by $\boldsymbol{X}_N$ the vector $(X_1, \ldots, X_N)^T$. Then the following proposition holds.

Proposition 21. *Let $(H, \tilde{u})$ be a solution of the system (3.16) completed with initial conditions*

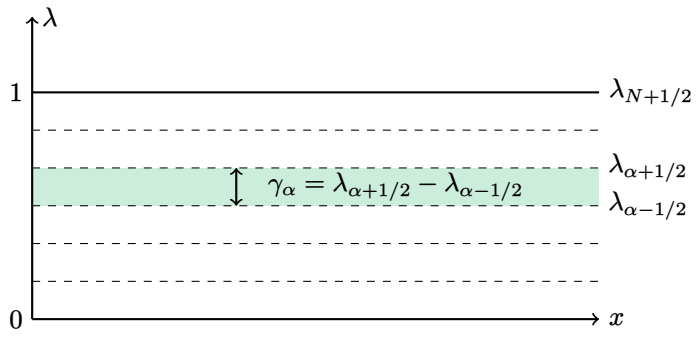

Figure 3.6: Multi-layer discretization of $\tilde{\Omega}$

(3.17). For $\alpha = 1, ..., N$, the multi-layer formulation given by

$$\begin{cases} \dfrac{\partial H_\alpha}{\partial t} + \partial_x(H_\alpha \tilde{u}_\alpha) = 0, \\ \dfrac{\partial \tilde{u}_\alpha}{\partial t} + \tilde{u}_\alpha \partial_x \tilde{u}_\alpha + g\partial_x \displaystyle\sum_{\beta=1}^{N} \gamma_\beta H_\beta = -g\partial_x z_b, \end{cases} \tag{3.42}$$

with initial conditions:

$$\begin{cases} H_\alpha(0,x) = \dfrac{1}{\gamma_\alpha} \displaystyle\int_{\lambda_{\alpha-1/2}}^{\lambda_{\alpha+1/2}} H(0,x,\lambda)\,\mathrm{d}\lambda, \\ \tilde{u}_\alpha(0,x) = \dfrac{1}{\gamma_\alpha} \displaystyle\int_{\lambda_{\alpha-1/2}}^{\lambda_{\alpha+1/2}} \tilde{u}(0,x,\lambda)\,\mathrm{d}\lambda, \end{cases} \tag{3.43}$$

is the $\mathbb{P}_0$ Galerkin approximation of (3.16).

Proof. The one-dimensional form of equations (3.16) reads as written in (3.38)-(3.39),

$$\frac{\partial H}{\partial t} + \partial_x(H\tilde{u}) = 0, \qquad \frac{\partial \tilde{u}}{\partial t} + \tilde{u}\partial_x\tilde{u} + g\partial_x \int_0^1 H\,\mathrm{d}\lambda = -g\partial_x z_b.$$

Multiplying these equations by $\mathbf{1}_\alpha = \mathbf{1}_{\lambda \in L_\alpha}$ and integrating over $\lambda \in I$, we get

$$\begin{aligned} &\frac{\partial}{\partial t}\int_0^1 H\mathbf{1}_\alpha \mathrm{d}\lambda + \partial_x \int_0^1 H\tilde{u}\mathbf{1}_\alpha \mathrm{d}\lambda = 0, \\ &\frac{\partial}{\partial t}\int_0^1 \tilde{u}\mathbf{1}_\alpha \mathrm{d}\lambda + \partial_x \int_0^1 \tilde{u}^2\mathbf{1}_\alpha \mathrm{d}\lambda + g\partial_x \int_0^1 H\mathrm{d}\lambda = -g\partial_x \tilde{z}_b. \end{aligned}$$

Replacing H and $\tilde{u}$ by their approximations defined by (3.41), we recover the multi-layer formulation under the form (3.42) for $\alpha = 1, \ldots, N$. □

The following corollary emphasizes the interest of the multilayer model for a class of piecewise constant functions.

Corollary 4. *Let us consider $(H, \tilde{u})$ defined by*

$$H = \sum_{\alpha=1}^{N} \mathbf{1}_{\lambda \in L_\alpha}(\lambda) H_\alpha(t,x), \qquad \tilde{u} = \sum_{\alpha=1}^{N} \mathbf{1}_{\lambda \in L_\alpha}(\lambda) \tilde{u}_\alpha(t,x).$$

Then $(H, \tilde{u})$ are solutions of system (3.16) if and only if $(\boldsymbol{H}_N, \tilde{\boldsymbol{U}}_N)$ defined by $\boldsymbol{H}_N = (H_1, \dots, H_N)^T$ and $\tilde{\boldsymbol{U}}_N = (\tilde{u}_1, \dots, \tilde{u}_N)^T$ are solutions of system (3.42) with initial data (3.43).

Proof. The corollary is deduced directly from the previous proof of Proposition 21, the approximation commuting with the nonlinearities and integration. □

We remark that system (3.42) can be viewed as a Saint-Venant equation in each layer with some coupling through the summation. In particular the interaction is between all the layers and not only adjacent ones; in fact there is no exchange term between the layer and the coupling is exactly the same for all the layers.

Remark 17. *In terms of numerical approximation, the system (3.42) can be an interesting alternative to existing multi-layer models simulating hydrostatic free-surface flows, see [3] and references therein. Indeed the system (3.42) can also be written under the form*

$$\begin{cases} \dfrac{\partial H_\alpha}{\partial t} + \partial_x(H_\alpha \tilde{u}_\alpha) = 0, \\ \dfrac{\partial (H_\alpha \tilde{u}_\alpha)}{\partial t} + \partial_x\Big(H_\alpha \tilde{u}_\alpha^2 + g H_\alpha \displaystyle\sum_{\beta=1}^{N} \gamma_\beta H_\beta\Big) = -g H_\alpha \partial_x z_b + g(\partial_x H_\alpha) \displaystyle\sum_{\beta=1}^{N} \gamma_\beta H_\beta, \end{cases} \tag{3.44}$$

and the numerical scheme proposed in [3] can be adapted. Notice that the multi-layer formulation has also to deal with the change of variable that can become singular when time evolves and thus hardly invertible. Therefore a reinterpolation of the variables $(H_\alpha, \tilde{u}_\alpha)$ has to be done when quantities vary greatly in the domain under consideration. The total mass, the total momentum and the total energy of a column being conserved for smooth solutions, the reinterpolations should also conserve the corresponding quantities. A comparison of performance (computational cost versus accuracy) with classical approximation methods (see [3] in the multi-layer context) would be useful.

3.5.1 Characterization of the spectrum in the discrete case

It is important to study the hyperbolic nature of the multi-layer system (3.42), in particular in view of analyzing the stability of a numerical scheme. To this end, the multi-layer model (3.42) can be rewritten abstractly as the following quasi-linear system.

Proposition 22. *System (3.42) can be written in a quasi-linear form*

$$\frac{\partial \tilde{\boldsymbol{\mathcal{U}}}}{\partial t} + A_N(\tilde{\boldsymbol{\mathcal{U}}}) \frac{\partial \tilde{\boldsymbol{\mathcal{U}}}}{\partial x} = \tilde{\boldsymbol{S}}, \tag{3.45}$$

with $\tilde{\boldsymbol{\mathcal{U}}} = (\boldsymbol{H}_N, \tilde{\boldsymbol{U}}_N)^T$ and $\tilde{\boldsymbol{S}} = (\boldsymbol{0}_N, -g\partial_x z_b \mathbf{1}_N)$ being block vectors and $A_N(\tilde{\boldsymbol{\mathcal{U}}})$ being the 2×2 block matrix defined by

$$A_N(\tilde{\boldsymbol{\mathcal{U}}}) = \begin{pmatrix} \mathrm{diag}(\tilde{\boldsymbol{U}}_N) & \mathrm{diag}(\boldsymbol{H}_N) \\ g \mathbf{1}_N \otimes \boldsymbol{\gamma}_N & \mathrm{diag}(\tilde{\boldsymbol{U}}_N) \end{pmatrix},$$

where $\boldsymbol{\gamma}_N = (\gamma_1, \dots, \gamma_N)^T$.

Proof. The result follows easily by explicitly computing the matrix product. □

The following result corresponds to a discrete version of the definition of the point spectrum given in Theorem 5 in the case of flat topography.

Proposition 23. *Let $\tilde{\boldsymbol{\mathcal{U}}} = (\boldsymbol{H}_N, \tilde{U}_N)^T$ be a solution of (3.45) with $z_b = 0$. If $\boldsymbol{H}_N > 0$ and $g > 0$ then the matrix $A_N(\tilde{\boldsymbol{\mathcal{U}}})$ admits at most $2N$ eigenvalues given by*

$$\sigma(A_N) = \left\{ c \in \mathbb{C} \setminus \tilde{U}_N \,:\, \sum_{i=1}^{N} \frac{g\gamma_i H_i}{(\tilde{u}_i - c)^2} = 1 \right\} \cup \left\{ c \in \tilde{U}_N \,:\, \operatorname{card}\big(\tilde{U}_N^{-1}(c)\}\big) > 1 \right\}, \tag{3.46}$$

where $\tilde{U}_N^{-1}(c) = \big\{ j \in \{1, \dots, N\} \,:\, \tilde{u}_j = c \big\}$ is the set of indices for which $\tilde{u}_j = c$. In particular if all the $\tilde{u}_i$ are distinct for $i \in \{1, \dots, N\}$, then the second part in the union is empty.

In the shallow water regime, *i.e.* when $\tilde{u}_i = \tilde{u}_1$ for all i, the spectrum reduces to three elements as explained for the continuous case in Remark 12. We note that in view of Corollary 4, Proposition 23 can be viewed as a direct corollary of Theorem 5 by considering the finite-dimensional subspace of $L^2(I)$ of $\mathbb{P}_0$-approximations and modifying the Lebesgue measure appropriately on this space. However we provide here a more straightforward proof.

Proof. In view of the structure of the matrix $A_N(\tilde{\boldsymbol{\mathcal{U}}})$, its characteristic polynomial is:

$$\det\big(A_N(\tilde{\boldsymbol{\mathcal{U}}}) - cI_{2N}\big) = \det\big(A_N(\tilde{\boldsymbol{\mathcal{U}}} - c(\mathbf{0}_N, \mathbf{1}_N)\big)$$

so by shifting $\tilde{U}_N \mapsto \tilde{U}_N - c\mathbf{1}_N$ we can compute the characteristic polynomial for $c = 0$ without loss of generality. Since $\operatorname{diag}(\tilde{U}_N)$ and $\operatorname{diag}(\boldsymbol{H}_N)$ commute, using [85, equation (16)] we obtain

$$\begin{aligned} \det(A_N\tilde{\boldsymbol{\mathcal{U}}}) &= \det\big(\operatorname{diag}(\tilde{U}_N)\operatorname{diag}(\tilde{U}_N) - g\mathbf{1}_N \otimes \gamma_N \operatorname{diag}(\boldsymbol{H}_N)\big) \\ &= \det\big(\operatorname{diag}(\tilde{U}_N)^2 - \mathbf{1}_N \otimes \boldsymbol{g}_N\big)\,, \end{aligned}$$

where $\boldsymbol{g}_N = g\gamma_N \odot \boldsymbol{H}_N$ denotes the element-wise or Hadamard product of γ_N with $\boldsymbol{H}_N$, *i.e* $g_\alpha = g\gamma_\alpha H_\alpha$. Hence $A_N(\tilde{\boldsymbol{\mathcal{U}}})$ is invertible if and only if $\operatorname{diag}(\tilde{U}_N)^2 - \mathbf{1}_N \otimes \boldsymbol{g}_N$ is invertible. Now one has to distinguish the cases where $\operatorname{diag}(\tilde{U}_N)$ is invertible or not:

1. If $\operatorname{diag}(\tilde{U}_N)$ is invertible, by using the Sherman–Morrison formula [84], this is equivalent to the fact that $1 - \mathbf{1}_N \cdot \operatorname{diag}(\tilde{U}_N)^{-2}\boldsymbol{g}_N \neq 0$. This later condition writes equivalently

$$1 - \mathbf{1}_N \cdot \operatorname{diag}(\tilde{U}_N)^{-2}\boldsymbol{g}_N = 1 - \tilde{U}_N^{\odot -2} \cdot \boldsymbol{g}_N = 1 - \sum_{i=1}^{N} \frac{g\gamma_i H_i}{\tilde{u}_i^2} \neq 0\,,$$

   where $\tilde{U}_N^{\odot -2} = (\tilde{u}_1^{-2}, \dots, \tilde{u}_N^{-2})$ denotes the element-wise or Hadamard power. Performing the shift back in $\tilde{U}$ leads to the first part in the union given by (3.46).

2. If $\operatorname{diag}(\tilde{U}_N)$ is not invertible, then one of its diagonal element is zero, *i.e.* there exists $j \in \{0, \dots, N\}$ such that $\tilde{u}_j = 0$. Therefore the j-th row of the matrix $\operatorname{diag}(\tilde{U}_N)^2 - \mathbf{1}_N \otimes \boldsymbol{g}_N$ is exactly $-\boldsymbol{g}_N^T$ and by subtracting this row to all the others, we obtain that $\operatorname{diag}(\tilde{U}_N)^2 - \mathbf{1}_N \otimes \boldsymbol{g}_N$ is invertible precisely when $\operatorname{diag}(\tilde{U}_N)^2 - \mathbf{e}_j \otimes \boldsymbol{g}_N$ is invertible, where $\boldsymbol{e}_j = (\delta_{1j}, \dots, \delta_{Nj})$ is the j-th basis vector. Therefore we proved that $A_N(\tilde{\boldsymbol{\mathcal{U}}})$ is invertible if and only if:

$$\det\big(\operatorname{diag}(\tilde{U}_N)^2 - \mathbf{e}_j \otimes \boldsymbol{g}_N\big) = -g\gamma_j H_j \prod_{i \neq j} \tilde{u}_j^2 \neq 0.$$

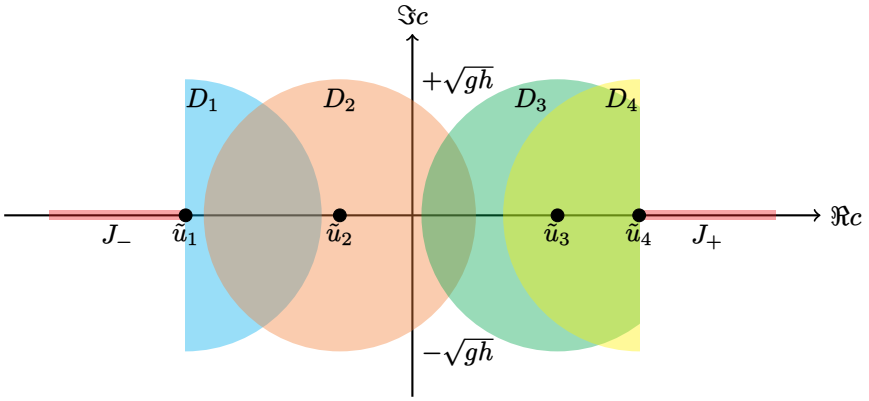

Figure 3.7: Example for $N = 4$: the spectrum $\sigma(A_N)$ is included in the intervals $J_\pm$ and disks D_i defined in Proposition 24. Moreover exactly one eigenvalue $c_\pm$ is in each interval $J_\pm$.

Since by assumption $H_i > 0$, we obtain that $c = 0$ is an eigenvalue if and only if $\tilde{u}_i = \tilde{u}_j = 0$ for some $i \neq j$, which is precisely the condition $\mathrm{card}\big(\tilde{U}_N^{-1}(0)\}\big) > 1$. Performing the shift back in $\tilde{U}$ leads to the second part in the union given by (3.46).

□

In the discrete case, we can slightly strengthen Propositions 14 and 15.

Proposition 24. *For $H_N > 0$, the eigenvalues $\sigma(A_N)$ of A_N are contained in the union of the following $N+2$ sets:*

$$J_- = \big[\tilde{u}_- - \sqrt{gh_N}, \tilde{u}_-\big], \qquad J_+ = \big[\tilde{u}_+, \tilde{u}_+ + \sqrt{gh_N}\big],$$

$$D_i = \Big\{ z \in \mathbb{C} \,:\, |z - \tilde{u}_i| \leq \sqrt{gh_N} \text{ and } \tilde{u}_- \leq \Re z \leq \tilde{u}_+ \Big\},$$

for $1 \leq i \leq N$ (illustrated on Fig. 3.7 for $N = 4$), where

$$\tilde{u}_- = \inf_\alpha \tilde{u}_\alpha, \qquad \tilde{u}_+ = \sup_\alpha \tilde{u}_\alpha, \qquad h_N = \sum_{i=1}^N \gamma_i H_i.$$

Moreover, there exist exactly two eigenvalues $c_\pm$ in the intervals $J_\pm$:

$$\sigma(A_N) \cap J_\pm = \{c_\pm\}.$$

Proof. In view of Proposition 23, the result reduces to the study of the function F_N defined on $\mathbb{C} \setminus \tilde{U}_N$ by:

$$F_N(c) = \sum_{i=1}^N \frac{g\gamma_i H_i}{(\tilde{u}_i - c)^2}. \tag{3.47}$$

We note that $F_N(c)$ is a Riemann sum approximation of $F(c)$ given by (3.35). As in the proof of Proposition 14, the first step is to prove:

$$\sigma(A_N) \subset \{z \in \mathbb{C} \,:\, \tilde{u}_i \leq \Re z \leq \tilde{u}_+\} \cup \mathbb{R}.$$

Since

$$\Im F_N(c) = \sum_{i=1}^{N} \frac{g\gamma_i H_i (\Re c - \tilde{u}_i)\Im c}{|\tilde{u}_i - c|^2},$$

if $\Re c > \tilde{u}_+$ or $\Re c < \tilde{u}_-$, then $(\Re c - \tilde{u})$ is either strictly positive or strictly negative on I, so the only way to make the sum $\Im F(c)$ zero, is that $\Im c = 0$.

The second step is to prove that the eigenvalues are in the union of the disks $\{D_i\}_{i=1}^N$. If c is not in the union of theses disks, then we have $|c - \tilde{u}_i| > \sqrt{gh_n}$ for all $1 \leq i \leq N$, and therefore

$$|F_N(c)| \leq \sum_{i=1}^{N} \frac{g\gamma_i H_i}{|\tilde{u}_i - c|^2} < \sum_{i=1}^{N} \frac{g\gamma_i H_i}{gh_N} = 1,$$

and c cannot satisfy $F_N(c) = 1$. We note that this second step is exactly given by the Gershgorin circle theorem.

Finally, the third step is to show the existence of $c_\pm \in J_\pm$ such that $\sigma(A_N) \cap J_\pm = \{c_\pm\}$. The strategy of proof is completely similar to the one in Proposition 15 so we only state the main steps. We have:

$$\lim_{c \to \tilde{u}_-} F_N(c) = \infty, \qquad \lim_{c \to \tilde{u}_+} F_N(c) = \infty, \qquad F_N(\tilde{u}_\pm \pm \sqrt{gh}) \leq 1\,,$$

and since F is strictly decreasing on J_+ and strictly increasing on J_-, there exists exactly one solution of $F(c) = 1$ in J_- and exactly one in J_+. □

In contrast to the continuous case, the previous result is not characterizing all real eigenvalues, as real solutions of summation condition could be in the interval $[\tilde{u}_-, \tilde{u}_+]$. This can be ruled out, under the following assumptions.

Proposition 25. *If $\boldsymbol{H}_N > 0$, and either:*

1. $\tilde{u}_+ - \tilde{u}_- < \sqrt{gh_N}$, *or*
2. $\max_i(|\tilde{u}_i - \tilde{u}_{i+1}|^2) < 8g \min_i(\gamma_i H_i)$

then:

$$\sigma(A_N) \cap \mathbb{R} = \{c_-, c_+\} \cup \left\{c \in \tilde{\boldsymbol{U}}_N \,:\, \mathrm{card}(\tilde{\boldsymbol{U}}_N^{-1}(c)\}) > 1\right\},$$

and in particular if all the $\tilde{u}_i$ are distinct this reduces to $\sigma(A_N) \cap \mathbb{R} = \{c_-, c_+\}$.

Proof. We treat each case separately:

1. If $(\tilde{u}_+ - \tilde{u}_-)^2 < gh_N$, then we have for all $c \in (\tilde{u}_-, \tilde{u}_+)$

$$F_N(c) \geq \sum_{i=1}^{N} \frac{g\gamma_i H_i}{(\tilde{u}_+ - \tilde{u}_-)^2} \geq \frac{gh_N}{(\tilde{u}_+ - \tilde{u}_-)^2} > 1\,.$$

2. Let $j \in \{1, N-1\}$, $m_j = \min(\tilde{u}_j, \tilde{u}_{j+1})$, $M_j = \max(\tilde{u}_j, \tilde{u}_{j+1})$. For $c \in (m_j, M_j)$, we have

$$F_N(c) = \sum_{i=1}^{N} \frac{g\gamma_i H_i}{(c - \tilde{u}_i)^2} \geq g \min_i(\gamma_i H_i) \sum_{i=1}^{N} \frac{1}{(c - \tilde{u}_i)^2} \geq \min_i(\gamma_i H_i) F_j(c)\,,$$

where

$$F_j(c) = \frac{1}{(c - \tilde{u}_j)^2} + \frac{1}{(c - \tilde{u}_{j+1})^2}\,.$$

On the interval (m_j, M_j), the function F_j has a minimum at $c^* = \frac{\tilde{u}_j + \tilde{u}_{j+1}}{2}$, so

$$F_j(c) \geq F_j(c^*) = \frac{8}{(\tilde{u}_{j+1} - \tilde{u}_j)^2} \geq \frac{8}{\max_i |\tilde{u}_i - \tilde{u}_{i+1}|^2}.$$

Therefore, this proves that for $c \in [m_j, M_j]$, $F(c) > 1$. Since

$$\bigcup_{j=1}^{N-1} (m_j, M_j) \subset (\tilde{u}_-, \tilde{u}_+)$$

we also have that $F(c) > 1$ for all $c \in (\tilde{u}_-, \tilde{u}_+)$, which concludes the proof

□

3.5.2 A convergence result of the spectrum to the continuous case

A natural question is whether the discrete approximation can be linked in some way to the continuous case as $N \to \infty$, for example do we have $\lim_{N\to\infty} \sigma(A_N) = \sigma(A_0)$ in some sense? Since A_0 is not compact in general, it is not trivial to approximate it with finite-rank operators and obtain some convergence of the spectrum. However, we note that for $g = 0$, the spectrum of A_N is $\tilde{\boldsymbol{U}}_N$ and so by a suitable discretization, we have $\lim_{N\to\infty} \sigma(A_N) = \sigma(A_0) = \tilde{u}(I)$, in the following sense:

$$\lim_{N\to\infty} d_H(\sigma(A_N), \sigma(A_0)) = \lim_{N\to\infty} \sup_{\lambda\in I} d(\sigma(A_N), \tilde{u}(\lambda)) = 0,$$

where d_H denotes the Hausdorff distance. However, when $g > 0$ it seems difficult to prove and maybe wrong that an eigenvalue of A_N can become as close as possible to any point of $\sigma_e(A_0) = \tilde{u}(I)$, but also that solutions of $F_N(c) = 1$ converge to solutions of $F(c) = 1$, where $F_N(c)$ and $F(c)$ are respectively defined by (3.35) and (3.47).

In the continuous case, Proposition 16 provides conditions under which the operator has only a real spectrum, hence is hyperbolic in some generalized sense. We propose here to look at what happens when we use the discretized version of the equations on a profile with the same properties. Proposition 25 ensures that only two eigenvalues are real and the others might have a nonzero imaginary part. In the following proposition, we prove that this imaginary part goes to zero when N goes to infinity assuming that the vertical profile of $\tilde{u}$ is strictly monotonic in λ and $\partial_\lambda(H/\partial_\lambda u) \neq 0$.

Proposition 26. *Let $H \in C^1(I)$ and $\tilde{u} \in C^2(I)$. For any $N \geq 1$, let $\{u_i\}_{1\leq i\leq N}$ and $\{H_i\}_{1\leq i\leq N}$ be the $\mathbb{P}_0-$approximation in λ of $\tilde{u}$ and H for $\gamma_\alpha = \frac{1}{N}$. If $H > 0$ and $\tilde{u}$ is strictly monotonic in λ with $\partial_\lambda(H/\partial_\lambda u) \neq 0$ for all $\lambda \in I$, then for N large enough*

$$\sup_{c\in\sigma(A_N)} |\Im c| \leq \left(\frac{6gC^3}{N}\right)^{1/4},$$

where $C > 0$ depends only on $\tilde{u}$ and H:

$$C = \max\left(1, \|H\|_{L^\infty}, \|\partial_\lambda H\|_{L^\infty}, \|\tilde{u}\|_{L^\infty}, \|\partial_\lambda \tilde{u}\|_{L^\infty}\right).$$

Proof. In view of Proposition 24, it suffices to consider $c \in \mathbb{C}$ such that $|\Im c| \neq 0$ and $\tilde{u}_- \leq \Re c \leq \tilde{u}_+$.

Since $F_N(c)$ is a Riemann sum approximation of $F(c)$, the first step is to prove that:

$$|F_N(c) - F(c)| \leq \frac{3gC^2}{|\Im c|^3 N}. \tag{3.48}$$

We have

$$|F_N(c) - F(c)| \leq g \sum_{i=1}^{N} \int_{L_i} \left| \frac{H_i}{(c-\tilde{u}_i)^2} - \frac{H}{(c-\tilde{u})^2} \right| \mathrm{d}\lambda \,.$$

Using many times the mean value theorem, for any $\lambda \in L_i$, we have

$$\begin{aligned} \left| \frac{H_i}{(c-\tilde{u}_i)^2} - \frac{H(\lambda)}{(c-\tilde{u}(\lambda)^2} \right| &\leq \left| \frac{H_i - H(\lambda)}{(c-\tilde{u}_i)^2} \right| + H(\lambda) \left| \frac{1}{(c-\tilde{u}_i)^2} - \frac{1}{(c-\tilde{u}(\lambda)^2} \right| \\ &\leq \frac{|H_i - H(\lambda)|}{|\Im c|^2} + \frac{2H(\lambda)|\tilde{u}_i - \tilde{u}(\lambda)|}{|\Im c|^3} \leq \frac{3C^2}{N|\Im c|^3} \,, \end{aligned}$$

since

$$|H_i - H(\lambda)| \leq CN^{-1} \,, \qquad\qquad |\tilde{u}_i - \tilde{u}(\lambda)| \leq CN^{-1} \,.$$

and therefore (3.48) is proven.

Since $\tilde{u}$ and H satisfy the hypotheses of Proposition 16, we know that (3.36) holds, so

$$\Re F(c) \leq \frac{C}{|\Im c|} |\Im F(c)| \,,$$

and we deduce that

$$\begin{aligned} \Re F_N(c) &\leq \Re F(c) + |F_N(c) - F(c)| \\ &\leq \frac{C}{|\Im c|} |\Im F(c)| + |F_N(c) - F(c)| \\ &\leq \frac{C}{|\Im c|} \left[|\Im F_N(c)| + |F_N(c) - F(c)| \right] + |F_N(c) - F(c)| \\ &\leq \frac{C}{|\Im c|} |\Im F_N(c)| + \left[1 + \frac{C}{|\Im c|} \right] |F_N(c) - F(c)|. \end{aligned}$$

In particular if $|\Im c|^4 > 6gC^3/N$ and for N large enough we get

$$\Re F_N(c) \leq \frac{C}{|\Im c|} |\Im F_N(c)| + \left[1 + \frac{C}{|\Im c|} \right] \frac{3gC^2}{|\Im c|^3 N} < \frac{C}{|\Im c|} |\Im F_N(c)| + 1,$$

which is incompatible with having a solution of $F_N(c) = 1$, hence $c \notin \sigma(A_N)$. □

3.A Spectrum definition

Definition 1. *The spectrum of an operator A is defined as the set $\sigma(A)$ of all $c \in \mathbb{C}$ for which the operator $A - c\mathbf{I}$ is not invertible. The following classification of the spectrum is used:*

- *The point spectrum $\sigma_p(A)$ is defined as the set of all $c \in \mathbb{C}$ for which the operator $A - c\mathbf{I}$ is not injective.*
- *The continuous spectrum $\sigma_c(A)$ is defined as the set of all $c \in \mathbb{C}$ for which the operator $A - c\mathbf{I}$ is injective and its range is dense in $L^2(I)$ but not equal to $L^2(I)$.*
- *The residual spectrum $\sigma_r(A)$ is defined as the set of all $c \in \mathbb{C}$ for which the operator $A - c\mathbf{I}$ is injective and its range is not dense in $L^2(I)$.*

- *The discrete spectrum $\sigma_d(A)$ is defined as the set of all $c \in \mathbb{C}$ for which the operator $A - c\mathbf{I}$ is not invertible but Fredholm.*
- *The essential spectrum $\sigma_e(A)$ is defined as the set of all $c \in \mathbb{C}$ for which the operator $A - c\mathbf{I}$ is not Fredholm.*

Therefore:

$$\sigma(A) = \sigma_p(A) \cup \sigma_c(A) \cup \sigma_r(A) = \sigma_d(A) \cup \sigma_e(A)\,.$$

Chapter 4

Boussinesq Equations

Outline of the current chapter

This chapter is a work in progress in collaboration with *Samer Israwi*, *Henrik Kalisch*, and *Dimitrios Mitsotakis*.

Most of the asymptotically derived Boussinesq systems of water wave theory for long waves of small amplitude fail to satisfy exact mechanical conservation laws for mass, momentum and energy. It is thus only fair to consider approximate conservation laws that hold in the context of these systems. Although such approximate mass, momentum and energy conservation laws can be derived, the question of a rigorous mathematical justification still remains unanswered. The aim of this chapter is to justify the formally derived mechanical balance laws for weakly nonlinear and weakly dispersive water wave Boussinesq systems. In particular, the asymptotic expansion used for the derivation of the Boussinesq system is the same one employed for the derivation and rigorous justification of the balance laws.

4.1 Introduction

The propagation of surface water waves is described by the Euler equations of fluid mechanics accompanied by dynamic boundary conditions at the free surface and at the sea floor, [92]. The solutions of the Euler equations also satisfy certain conservation equations such mass, momentum

and energy. Although the Euler equations is a completely justified model from physical and mathematical aspects, [67], it is still a very difficult problem to solve, theoretically and numerically. This is due to the fact that domain in which the dependent variables such as the velocity field, the pressure are defined is changing with time and is bounded by the unknown free surface. For this reason, several approximate models have been derived aiming to approximate the solutions of the Euler equations. The derivation of such approximate systems is based on simplification assumptions on the characteristics of the waves to be described. These assumptions usually lead to the so-called wave regimes. We will focus on the case of long waves, which means that the waves considered to have large wavelength compared to the depth of the water. Two significant examples of long wave regimes are the small amplitude and large amplitude regimes. In the small amplitude regime the waves are assumed to be of small amplitude compared to the depth while in the large amplitude regime there is no restriction in the amplitude of the waves.

Typically, formal approximations of the Euler equations in both regimes are usually called Boussinesq systems but in the particular case of large amplitude waves the equations are referred to as the Serre-Green-Naghdi equations, [67]. This is because such approximations were first derived by J. Boussinesq [22] with the small amplitude assumption. Later the small amplitude assumption was removed by F. Serre [83] and independently by Green and Naghdi in [55] who derived equations for strongly nonlinear and weakly dispersive water waves. For more information we refer to [67]. These first candidates of approximations of the Euler equations appeared to have several theoretical and sometimes practical limitations such as well-posedness problems. Research was focused on the derivation of Boussinesq systems that have favourable nonlinear and dispersive properties. Extending the theory of Boussinesq, a whole class of Boussinesq-type systems were derived [14, 15] and justified theoretically and numerically, [39, 40, 41, 42, 59, 63, 67]. As these models can be used for applications in the nearshore zone it is essential to be able to estimate their accuracy in the mass, momentum and energy balance laws.

For the Serre-Green-Naghdi equations in one-dimension the respective balance laws have been estimated using asymptotic techniques and they have been found to be satisfied by the solutions of the Serre-Green-Naghdi equations with the same order of accuracy as they approximate the Euler equations, [61]. In the case of Boussinesq systems in the one-dimensional case, similar estimates were obtained in [1]. In this paper we consider a two-dimensional shallow water wave regime modelled by some Boussinesq-type equations first presented in [15]. The two-dimensional Boussinesq equations can be derived from the water waves problem by a simple asymptotic expansion on the potential Φ, or by a finite expansion on the shallowness parameter which is well-justified for shallow water regimes. We will adopt the latter technique throughout this chapter. After reviewing the derivation of the systems using asymptotic techniques, we present an asymptotic analysis of the mass, momentum and energy balance laws. These balance laws are justified theoretically along with the approximations of the velocity field. The justification is based on the error estimation between solutions of the Euler equations and the Boussinesq systems with the same initial data.

This paper is organized as follows. In Section 4.3 we present an asymptotic derivation of the two-dimensional Boussinesq systems of [14,15]. We present in Section 4.4 the necessary preliminary results that justify the accuracy of the asymptotics used in Section 4.3 for the derivation of the Boussinesq system. In Section 4.5, a mathematically rigorous derivation of the balance laws is presented, expanding significantly the preliminary work in [57, 58]. Moreover, an analysis of the errors in the balance laws is furnished. These mechanical balance laws depend on the usual two small parameters α and β that measure the non-linearity and the dispersion of the system respectively.

4.2 Notation

We start by introducing some notation. In what follows we denote by $X \in \mathbb{R}^2$ the horizontal variables $X = (x, y)$. We denote by α and β the non-linearity and shallowness parameters respectively given below by (4.6). We use the following notations for the gradient and Laplacian:

$$\nabla = (\partial_x, \partial_y)^T, \qquad \Delta = \partial_x^2 + \partial_y^2, \qquad \nabla_{X,z} = (\partial_x, \partial_y, \partial_z)^T, \qquad \nabla_{X,z}^{\beta} = (\sqrt{\beta}\nabla, \partial_z)^T.$$

We denote by e_z the upward normal unit vector in the vertical direction, while $\partial_n u$ is the upward co-normal derivative of u. The non-dimensional fluid domain will be defined as

$$\Omega_t = \left\{(X, z) \in \mathbb{R}^3, 0 \leq z \leq 1 + \alpha\eta(t, X)\right\},$$

where η denotes the free-surface. We denote by g the gravitational acceleration and by ρ the density of the fluid (we will later assume that $\rho = 1$).

We denote by $C(\lambda_1, \lambda_2, \cdots)$ a constant depending on the parameters λ_1, $\lambda_2, \cdots$, and *whose dependence on the λ_j is always assumed to be non-decreasing.*

We denote by $(H^{s,k}, |\cdot|_{H^{s,k}})$ the Banach space over $S = (-1, 0) \times \mathbb{R}^2$ defined by

$$H^{s,k} = \bigcap_{j=0}^{k} H^j\left((-1, 0); H^{s-j}(\mathbb{R}^2)\right), \qquad |u|_{H^{s,k}} = \sum_{j=0}^{k} |\Lambda^{s-j}\partial_z^j u|_{L^2},$$

where $\Lambda = (1 - \Delta)^{1/2}$ is the fractional derivative.

We denote by $\dot{H}^{s+1}(\mathbb{R}^2)$ the topological vector space

$$\dot{H}^{s+1}(\mathbb{R}^2) = \left\{f \in L^2_{loc}(\mathbb{R}^2), \quad \nabla f \in H^s(\mathbb{R}^2) \times H^s(\mathbb{R}^2)\right\},$$

equipped with the semi-norm $|f|_{\dot{H}^{s+1}(\mathbb{R}^2)} = |\nabla f|_{H^s(\mathbb{R}^2)}$.

For all $a, b \in \mathbb{R}$ we write $a \vee b = \max\{a, b\}$.

4.3 Formal derivation of the Boussinesq systems

We start with the derivation of the $a - b - c - d$ family of Boussinesq systems of [15]. We consider the domain $\left\{(x, y, z) \in \mathbb{R}^3, -h_0 < z < \eta(x, y, t)\right\}$ where the parameter h_0 represents the undisturbed depth of the fluid and $\eta(x, y, t)$ represents the free surface deviation above its rest position. The Euler equations are given by:

$$\rho_t + \nabla_{X,z} \cdot (\rho U) = 0, \tag{4.1a}$$

$$U_t + U \cdot \nabla_{X,z} U + \frac{1}{\rho}\nabla_{X,z} P = -g e_z, \tag{4.1b}$$

where $U = (u, w)^T$ is the velocity vector such that $\boldsymbol{u}(x, y, z, t) = (u(x, y, z, t), v(x, y, z, t))$ denotes the horizontal components and $w(x, y, z, t)$ the vertical component of the fluid velocity vector field and $P(t, x, y, z)$ denotes the pressure. Assuming an ideal and irrotational flow we can define the velocity potential $\mathring{\Phi}$ as

$$\boldsymbol{u} = \nabla\mathring{\Phi}, \qquad w = \mathring{\Phi}_z\ . \tag{4.2}$$

Then, the Bernoulli equation and the free-surface boundary condition governing the motion of the fluid are formulated in terms of the potential and the free-surface [92]

$$\left.\begin{array}{c} \mathring{\Phi}_t + \frac{1}{2}(\mathring{\Phi}_x^2 + \mathring{\Phi}_y^2 + \mathring{\Phi}_z^2) + g\eta = 0 \\ \eta_t + \mathring{\Phi}_x\eta_x + \mathring{\Phi}_y\eta_y - \mathring{\Phi}_z = 0 \end{array}\right\} \quad \text{on} \quad z = \eta(x,y,t)\ . \tag{4.3}$$

The derivation of asymptotic models relies on appropriate scaling. Here, we use non-dimensional variables:

$$\tilde{x} = \frac{x}{l}, \quad \tilde{y} = \frac{y}{l}, \quad \tilde{z} = \frac{z+h_0}{h_0}, \quad \tilde{t} = \frac{\sqrt{gh_0}t}{l}, \quad \tilde{\eta} = \frac{\eta}{A}, \quad \tilde{\Phi} = \frac{h_0}{Al\sqrt{gh_0}}\mathring{\Phi}, \tag{4.4}$$

where tilde denotes non-dimensional variables in the domain Ω_t (given in Section 4.2), l and A denote a characteristic wavelength and wave amplitude. The governing equations and boundary conditions for the fully dispersive and fully non-linear irrotational water wave problem are given by the Euler equations. These equations can be expressed in terms of the velocity potential and the free-surface elevation as follows:

$$\beta\nabla^2\tilde{\Phi} + \tilde{\Phi}_{zz} = 0, \quad 0 < \tilde{z} < \alpha\tilde{\eta} + 1\ , \tag{4.5a}$$

$$\tilde{\Phi}_{\tilde{z}} = 0, \quad \tilde{z} = 0\ , \tag{4.5b}$$

$$\tilde{\Phi}_{\tilde{t}} + \frac{\alpha}{2}\left(\tilde{\Phi}_{\tilde{x}}^2 + \tilde{\Phi}_{\tilde{y}}^2 + \frac{1}{\beta}\tilde{\Phi}_{\tilde{z}}^2\right) + \tilde{\eta} = 0, \quad \text{on } \tilde{z} = \alpha\tilde{\eta} + 1\ , \tag{4.5c}$$

$$\tilde{\eta}_{\tilde{t}} + \alpha\left[\tilde{\eta}_{\tilde{x}}\tilde{\Phi}_{\tilde{x}} + \tilde{\eta}_{\tilde{y}}\tilde{\Phi}_{\tilde{y}}\right] - \frac{1}{\beta}\tilde{\Phi}_{\tilde{z}} = 0, \quad \text{on } \tilde{z} = \alpha\tilde{\eta} + 1\ , \tag{4.5d}$$

where α and β measure the non-linearity and frequency dispersion defined as

$$\alpha = \frac{A}{h_0}, \qquad \beta = \frac{h_0^2}{l^2}\ . \tag{4.6}$$

Appropriate assumptions on the respective magnitude of the parameters α and β, lead to the derivation of (simpler) asymptotic models from the Euler equations. The Stokes number

$$S = \frac{\alpha}{\beta},$$

is introduced in order to quantify the applicability of the equation to a particular regime of surface water waves. For the Boussinesq regime, the Stokes number is usually considered to be of order 1. Throughout this section we will denote by $\mathcal{O}(\beta^n)$, with $n \in \mathbb{N}$ any family of functions $(f^\beta)_{\beta\in]0,1[}$ such that $\dfrac{1}{\beta^n}f^\beta$ remains bounded in $L^\infty([0,\frac{T}{\beta}], H^r)$, for all $\beta \ll 1$ where the value of r will be determined accordingly throughout the chapter. It is noted that in the sequel we consider the Zakharov-Craig-Sulem formulation [34, 35, 94] of the Euler equations. Specifically, we consider the Euler equations in terms of the Dirichlet-Neumann operator as follows :

$$\begin{cases} \tilde{\eta}_{\tilde{t}} - \dfrac{1}{\beta}\mathcal{G}_\beta[\alpha\tilde{\eta}]\psi = 0\ , \\ \psi_{\tilde{t}} + \tilde{\eta} + \dfrac{\alpha}{2}|\nabla\psi|^2 - \dfrac{\alpha}{\beta}\dfrac{[\mathcal{G}_\beta[\alpha\tilde{\eta}]\psi + \alpha\beta\nabla\tilde{\eta}\cdot\nabla\psi]^2}{2(1+\alpha^2\beta|\nabla\tilde{\eta}|^2)} = 0\ . \end{cases} \tag{4.7}$$

Given a solution of this system, we reconstruct the potential $\tilde{\Phi}$ by solving the Laplace equation (4.8) below. More precisely, we introduce the trace of the velocity potential at the free surface, defined as

$$\psi = \tilde{\Phi}|_{\tilde{z}=1+\alpha\tilde{\eta}} ,$$

and the Dirichlet-Neumann operator $\mathcal{G}_\beta[\alpha\tilde{\eta}]\cdot$ as

$$\mathcal{G}_\beta[\alpha\tilde{\eta}]\psi = \partial_{\tilde{z}}\tilde{\Phi}|_{\tilde{z}=1+\alpha\tilde{\eta}} ,$$

with $\tilde{\Phi}$ solving the boundary value problem

$$\left\{ \begin{array}{l} \beta\partial_{\tilde{x}}^2\tilde{\Phi} + \beta\partial_{\tilde{y}}^2\tilde{\Phi} + \partial_{\tilde{z}}^2\tilde{\Phi} = 0 , \\ \partial_n\tilde{\Phi}|_{\tilde{z}=0} = 0 , \\ \tilde{\Phi}|_{\tilde{z}=1+\alpha\tilde{\eta}} = \psi . \end{array} \right. \tag{4.8}$$

We look for an asymptotic expansion of $\tilde{\Phi}$ of the form

$$\tilde{\Phi}^{app} = \sum_{j=0}^{N} \beta^j \tilde{\Phi}_j. \tag{4.9}$$

Plugging this expression into the boundary value problem (4.8) one can cancel the residual up to the order $\mathcal{O}(\beta^{N+1})$ provided that

$$\partial_{\tilde{z}}^2\tilde{\Phi}_j = -\partial_{\tilde{x}}^2\tilde{\Phi}_{j-1} - \partial_{\tilde{y}}^2\tilde{\Phi}_{j-1}, \quad j = 0, \dots, N , \tag{4.10}$$

(with the convention that $\tilde{\Phi}_{-1} = 0$), together with the boundary conditions

$$\left\{ \begin{array}{l} \tilde{\Phi}_j|_{\tilde{z}=1+\alpha\tilde{\eta}} = \delta_{0,j}\psi , \\ \partial_{\tilde{z}}\tilde{\Phi}_j|_{\tilde{z}=0} = 0 , \end{array} \right. \qquad j = 0, \dots, N , \tag{4.11}$$

(where $\delta_{0,j} = 1$ if $j = 0$ and 0 otherwise). Solving equation (4.10) with the boundary conditions (4.11) as in [62] one finds

$$\begin{aligned} \tilde{\Phi}_0 &= \psi , \\ \tilde{\Phi}_1 &= -\frac{1}{2}\tilde{z}^2\Delta\psi + \frac{1}{2}\Delta\psi + \alpha\tilde{\eta}\Delta\psi + \frac{1}{2}\alpha^2\tilde{\eta}^2\Delta\psi , \\ \tilde{\Phi}_2 &= \frac{1}{24}\tilde{z}^4\Delta^2\psi - \frac{1}{4}\tilde{z}^2\Delta^2\psi + \frac{5}{24}\Delta^2\psi + \frac{5}{6}\alpha\tilde{\eta}\Delta^2\psi - \frac{1}{2}\alpha\tilde{z}^2\Delta\tilde{\eta}\Delta\psi + \frac{1}{2}\alpha\Delta\tilde{\eta}\Delta\psi \\ &\quad - \alpha\tilde{z}^2\nabla\tilde{\eta}\cdot\nabla\Delta\psi + \alpha\nabla\tilde{\eta}\cdot\nabla\Delta\psi - \frac{1}{2}\alpha\tilde{z}^2\tilde{\eta}\Delta^2\psi + \mathcal{O}(\alpha^2) . \end{aligned}$$

Let $\tilde{U}, \tilde{V}$ be the dimensionless velocities at a dimensionless height θ ($0 \leq \theta \leq 1$) in the fluid column:

$$\tilde{U} = \tilde{\Phi}_{\tilde{x}}^{app}\big|_{\tilde{z}=\theta} = \psi_{\tilde{x}} - \frac{\beta}{2}(\theta^2 - 1)\Delta\psi_{\tilde{x}} + \mathcal{O}(\alpha\beta, \beta^2), \tag{4.12}$$

$$\tilde{V} = \tilde{\Phi}_{\tilde{y}}^{app}\big|_{\tilde{z}=\theta} = \psi_{\tilde{y}} - \frac{\beta}{2}(\theta^2 - 1)\Delta\psi_{\tilde{y}} + \mathcal{O}(\alpha\beta, \beta^2). \tag{4.13}$$

Remark 18. *A family of equations consistent with the original Boussinesq equation has been derived in the literature using different choices for the velocity. While other derivations use the averaged velocity or the velocity at the bottom, we will use the velocity variables at an arbitrary height* $0 \leq$

$\theta \leq 1$ defined by (4.12), (4.13). In the present re-scaling $\theta = 0$ corresponds to the bottom and $\theta = 1$ corresponds to the free-surface. In order to determine the evolution equations satisfied by the velocities $\tilde{U}$ and $\tilde{V}$, it is necessary to find a link between the free-surface potential ψ (as well as its derivatives), and the approximate potential $\tilde{\Phi}^{app}$ at $z = \theta$ (respectively its derivatives at $z = \theta$). We will introduce the following formula at height $0 \leq \theta \leq 1$:

$$\psi = \tilde{\Phi}^{app}\big|_{\tilde{z}=\theta} + \frac{\beta}{2}\left(\theta^2 - 1\right) \Delta\, \tilde{\Phi}^{app}\big|_{\tilde{z}=\theta} + \mathcal{O}(\alpha\beta, \beta^2),$$

This follows from the asymptotic expansion given by (4.9) where:

$$\psi = \left(1 - \frac{\beta}{2}(z^2 - 1)\right)^{-1} \tilde{\Phi}^{app} + \mathcal{O}(\alpha\beta, \beta^2).$$

Starting from equation (4.5c), we would like to find the equations satisfied by the velocities $\tilde{U}, \tilde{V}$. Deriving with respect to x (respectively y), one has the following equations at the free surface:

$$\tilde{\Phi}_{\tilde{t}\tilde{x}} + \alpha\left(\tilde{\Phi}_{\tilde{x}}\tilde{\Phi}_{\tilde{x}\tilde{x}} + \tilde{\Phi}_{\tilde{y}}\tilde{\Phi}_{\tilde{x}\tilde{y}} + \frac{1}{\beta}\tilde{\Phi}_{\tilde{z}}\tilde{\Phi}_{\tilde{x}\tilde{z}}\right) + \tilde{\eta}_{\tilde{x}} = 0, \tag{4.14}$$

$$\tilde{\Phi}_{\tilde{t}\tilde{y}} + \alpha\left(\tilde{\Phi}_{\tilde{x}}\tilde{\Phi}_{\tilde{x}\tilde{y}} + \tilde{\Phi}_{\tilde{y}}\tilde{\Phi}_{\tilde{y}\tilde{y}} + \frac{1}{\beta}\tilde{\Phi}_{\tilde{z}}\tilde{\Phi}_{\tilde{y}\tilde{z}}\right) + \tilde{\eta}_{\tilde{y}} = 0. \tag{4.15}$$

Replacing $\tilde{\Phi}^{app}$ in (4.14) and (4.15) respectively, we get:

$$\psi_{\tilde{t}\tilde{x}} + \alpha\left(\psi_{\tilde{x}}\psi_{\tilde{x}\tilde{x}} + \psi_{\tilde{y}}\psi_{\tilde{x}\tilde{y}}\right) + \tilde{\eta}_{\tilde{x}} = \mathcal{O}(\alpha\beta, \beta^2),$$
$$\psi_{\tilde{t}\tilde{y}} + \alpha\left(\psi_{\tilde{x}}\psi_{\tilde{x}\tilde{y}} + \psi_{\tilde{y}}\psi_{\tilde{y}\tilde{y}}\right) + \tilde{\eta}_{\tilde{y}} = \mathcal{O}(\alpha\beta, \beta^2).$$

Using Remark 18, the equations (4.14),(4.15) along with (4.5d) give the following Boussinesq system:

$$\tilde{U}_{\tilde{t}} + \tilde{\eta}_{\tilde{x}} + \frac{\beta}{2}(\theta^2 - 1)\Delta\tilde{U}_{\tilde{t}} + \alpha\left(\tilde{U}\tilde{U}_{\tilde{x}} + \tilde{V}\tilde{V}_{\tilde{x}}\right) = \mathcal{O}(\alpha\beta, \beta^2)\ , \tag{4.16a}$$

$$\tilde{V}_{\tilde{t}} + \tilde{\eta}_{\tilde{y}} + \frac{\beta}{2}(\theta^2 - 1)\Delta\tilde{V}_{\tilde{t}} + \alpha\left(\tilde{U}\tilde{U}_{\tilde{y}} + \tilde{V}\tilde{V}_{\tilde{y}}\right) = \mathcal{O}(\alpha\beta, \beta^2)\ , \tag{4.16b}$$

$$\tilde{\eta}_{\tilde{t}} + \tilde{U}_{\tilde{x}} + \tilde{V}_{\tilde{y}} + \frac{\beta}{2}\left[\theta^2 - \frac{1}{3}\right]\left(\Delta\tilde{U}_{\tilde{x}} + \Delta\tilde{V}_{\tilde{y}}\right) + \alpha\left((\tilde{\eta}\tilde{U})_{\tilde{x}} + (\tilde{\eta}\tilde{V})_{\tilde{y}}\right) = \mathcal{O}(\alpha\beta, \beta^2)\ . \tag{4.16c}$$

The motivation behind the derivation of a family of Boussinesq equations is the interest in improving the linear dispersive properties of the model. There have been several works on the subject [11, 28, 50, 72, 75] that lead to the derivation of consistent models with better frequency dispersion. We will proceed by a classical technique that consists of replacing the time derivative of the velocity by a space derivative of the free-surface in the higher order terms in (4.16a)-(4.16b). Recall that the velocity variables correspond to the velocity at an elevation $0 \leq \theta \leq 1$ instead of other choices that have been used in the velocity at the bottom or the averaged velocity. Observing that

$$\tilde{U}_{\tilde{t}} + \tilde{\eta}_{\tilde{x}} = \mathcal{O}(\alpha, \beta)\ , \tag{4.17}$$

$$\tilde{V}_{\tilde{t}} + \tilde{\eta}_{\tilde{y}} = \mathcal{O}(\alpha, \beta)\ , \tag{4.18}$$

$$\tilde{\eta}_{\tilde{t}} + \tilde{U}_{\tilde{x}} + \tilde{V}_{\tilde{y}} = \mathcal{O}(\alpha, \beta)\ , \tag{4.19}$$

we can write for all $\lambda, \mu \in \mathbb{R}$

$$\begin{aligned}\beta\ \Delta\left[\tilde{U}_{\tilde{x}}+\tilde{V}_{\tilde{y}}\right] &= \lambda\beta\ \Delta\left[\tilde{U}_{\tilde{x}}+\tilde{V}_{\tilde{y}}\right]+(1-\lambda)\beta\ \Delta\left[\tilde{U}_{\tilde{x}}+\tilde{V}_{\tilde{y}}\right]\\ &= \lambda\beta\ \Delta\left[\tilde{U}_{\tilde{x}}+\tilde{V}_{\tilde{y}}\right]-(1-\lambda)\beta\ \Delta\tilde{\eta}_{\tilde{t}}+\mathcal{O}(\alpha\beta,\beta^2)\ ,\end{aligned}$$

and also

$$\begin{aligned}\beta\ \Delta\tilde{U}_{\tilde{t}} &= \mu\beta\ \Delta\tilde{U}_{\tilde{t}}+(1-\mu)\beta\ \Delta\tilde{U}_{\tilde{t}}\\ &= \mu\beta\ \Delta\tilde{U}_{\tilde{t}}-(1-\mu)\beta\ \Delta\tilde{\eta}_{\tilde{x}}+\mathcal{O}(\alpha\beta,\beta^2)\ ,\end{aligned}$$

and similarly,

$$\beta\ \Delta\tilde{V}_{\tilde{t}} = \mu\beta\ \Delta\tilde{V}_{\tilde{t}}-(1-\mu)\beta\ \Delta\tilde{\eta}_{\tilde{y}}+\mathcal{O}(\alpha\beta,\beta^2)\ .$$

Substitution of these relations into (4.16a)–(4.16c) leads to the following general $a-b-c-d$ Boussinesq system

$$\tilde{U}_{\tilde{t}}+\tilde{\eta}_{\tilde{x}}+\alpha\left(\tilde{U}\tilde{U}_{\tilde{x}}+\tilde{V}\tilde{V}_{\tilde{x}}\right)+\beta a\Delta\tilde{\eta}_{\tilde{x}}-\beta b\Delta\tilde{U}_{\tilde{t}} = \mathcal{O}(\alpha\beta,\beta^2)\ , \tag{4.20a}$$

$$\tilde{V}_{\tilde{t}}+\tilde{\eta}_{\tilde{y}}+\alpha\left(\tilde{U}\tilde{U}_{\tilde{y}}+\tilde{V}\tilde{V}_{\tilde{y}}\right)+\beta a\Delta\tilde{\eta}_{\tilde{y}}-\beta b\Delta\tilde{V}_{\tilde{t}} = \mathcal{O}(\alpha\beta,\beta^2)\ , \tag{4.20b}$$

$$\tilde{\eta}_{\tilde{t}}+\tilde{U}_{\tilde{x}}+\tilde{V}_{\tilde{y}}+\alpha\left((\tilde{\eta}\tilde{U})_{\tilde{x}}+(\tilde{\eta}\tilde{V})_{\tilde{y}}\right)+\beta c\Delta\left(\tilde{U}_{\tilde{x}}+\tilde{V}_{\tilde{y}}\right)-\beta d\Delta\tilde{\eta}_{\tilde{t}} = \mathcal{O}(\alpha\beta,\beta^2)\ . \tag{4.20c}$$

where

$$\begin{aligned}&a=\frac{1}{2}(1-\theta^2)\mu, \quad b=\frac{1}{2}(1-\theta^2)(1-\mu),\\ &c=\frac{1}{2}\left(\theta^2-\frac{1}{3}\right)\lambda, \quad d=\frac{1}{2}\left(\theta^2-\frac{1}{3}\right)(1-\lambda)\ .\end{aligned} \tag{4.21}$$

After neglecting the high-order terms and writing the variables in dimensional form, system (4.20a)–(4.20c) is written as

$$U_t+g\eta_x+(UU_x+VV_x)+gh_0^2a\Delta\eta_x-h_0^2b\Delta U_t = 0\ , \tag{4.22a}$$

$$V_t+g\eta_y+(UU_y+VV_y)+gh_0^2a\Delta\eta_y-h_0^2b\Delta V_t = 0, \tag{4.22b}$$

$$\eta_t+h_0\left(U_x+V_y\right)+((\eta U)_x+(\eta V)_y)+h_0^3c\Delta\left(U_x+V_y\right)-h_0^2d\Delta\eta_t = 0\ . \tag{4.22c}$$

Remark 19. *One should note that by introducing another parameter ν in equation (4.18) (instead of choosing the same parameter μ in the decomposition for both equations (4.17) and (4.18)), the system (4.20a)-(4.20c) can be generalized to the $a-b-a_1-b_1-c-d$ system*

$$\begin{aligned}&\tilde{U}_{\tilde{t}}+\tilde{\eta}_{\tilde{x}}+\alpha\left(\tilde{U}\tilde{U}_{\tilde{x}}+\tilde{V}\tilde{V}_{\tilde{x}}\right)+\beta a\Delta\tilde{\eta}_{\tilde{x}}-\beta b\Delta\tilde{U}_{\tilde{t}} = \mathcal{O}(\alpha\beta,\beta^2)\ ,\\ &\tilde{V}_{\tilde{t}}+\tilde{\eta}_{\tilde{y}}+\alpha\left(\tilde{U}\tilde{U}_{\tilde{y}}+\tilde{V}\tilde{V}_{\tilde{y}}\right)+\beta a_1\Delta\tilde{\eta}_{\tilde{y}}-\beta b_1\Delta\tilde{V}_{\tilde{t}} = \mathcal{O}(\alpha\beta,\beta^2)\ ,\\ &\tilde{\eta}_{\tilde{t}}+\tilde{U}_{\tilde{x}}+\tilde{V}_{\tilde{y}}+\alpha\left((\tilde{\eta}\tilde{U})_{\tilde{x}}+(\tilde{\eta}\tilde{V})_{\tilde{y}}\right)+\beta c\Delta\left(\tilde{U}_{\tilde{x}}+\tilde{V}_{\tilde{y}}\right)-\beta d\Delta\tilde{\eta}_{\tilde{t}} = \mathcal{O}(\alpha\beta,\beta^2)\,,\end{aligned}$$

where

$$a_1=\frac{1}{2}(1-\theta^2)\nu, \quad b_1=\frac{1}{2}(1-\theta^2)(1-\nu).$$

Choosing distinct parameters μ and ν could be of interest when studying waves in the nearshore, where the dominant wave direction is approximately normal to the shoreline. However, in the present

work, we will stick to the four-parameter system (4.20a)-(4.20c).

In order to compute the associated mass, momentum, and energy and fluxes, we need expressions for the velocities and pressure. The velocity field can be easily computed using (4.12)–(4.13). The expression for the pressure is obtained from Bernoulli's equation [67,92] or equivalently by integrating the last equation in (4.1b) over the interval $[z,\eta]$ with $-h_0 \leq z \leq \eta$,

$$\mathring{\Phi}_t + \frac{1}{2}\left|\nabla_{X,z}\mathring{\Phi}\right|^2 = -\frac{P}{\rho} - gz + C. \tag{4.24}$$

We can find the constant C by evaluating the previous equation at the free surface $z=\eta$. Specifically, we find

$$C = \frac{P_{atm}}{\rho}\ , \tag{4.25}$$

where P_{atm} refers to the atmospheric pressure. We introduce the dynamic pressure by subtracting the hydrostatic pressure contribution:

$$P' = P - P_{atm} + gz\ , \tag{4.26}$$

which can be scaled using a typical wave amplitude by $gA\tilde{P}' = P'$, then

$$\begin{aligned}\tilde{P}' &= -\tilde{\Phi}_{\tilde{t}} - \frac{1}{2}\alpha(\tilde{\Phi}_{\tilde{x}}^2 + \tilde{\Phi}_{\tilde{y}}^2) - \frac{1}{2}\frac{\alpha}{\beta}(\tilde{\Phi}_{\tilde{z}}^2)\\ &= -\psi_{\tilde{t}} + \frac{\beta}{2}(\tilde{z}^2-1)\Delta\psi_{\tilde{t}} - \frac{\alpha}{2}\left[\psi_{\tilde{x}}^2 + \psi_{\tilde{y}}^2\right] + \mathcal{O}(\alpha\beta,\beta^2)\ .\end{aligned}$$

If we substitute $\tilde{\Phi}$ by $\tilde{\Phi}^{app}$ in (4.5c) we get:

$$\psi_{\tilde{t}} + \tilde{\eta} + \frac{\alpha}{2}|\nabla\psi|^2 = \mathcal{O}(\alpha\beta,\beta^2)\ . \tag{4.27}$$

Using the relation $\psi_{\tilde{x}\tilde{x}\tilde{t}} + \psi_{\tilde{y}\tilde{y}\tilde{t}} = \tilde{U}_{\tilde{x}\tilde{t}} + \tilde{V}_{\tilde{y}\tilde{t}} + \mathcal{O}(\beta)$ and (4.27), the scaled dynamic pressure becomes

$$\tilde{P}' = \tilde{\eta} + \frac{1}{2}\beta(\tilde{z}^2-1)[\tilde{U}_{\tilde{x}\tilde{t}} + \tilde{V}_{\tilde{y}\tilde{t}}] + \mathcal{O}(\alpha\beta,\beta^2)\ . \tag{4.28}$$

The total pressure in-terms of dimensional variables is then given by

$$P = P_{atm} - \rho g(z-\eta) + \frac{\rho}{2}\left[(z+h_0)^2 - h_0^2\right](U_{xt} + V_{yt}) + \mathcal{O}(\alpha\beta,\beta^2)\ . \tag{4.29}$$

The following section is devoted to a mathematically rigorous approach to understanding the validity of the two-dimensional Boussinesq system as an approximation of the water waves problem represented by the Euler equations (4.5) with a particular focus on justifiying approximate mass, momentum and energy balance laws similar to those presented in [1,2,61].

4.4 Preliminary Results

We first present results related to the formal approximations of the velocity potential. These results are necessary ingredients not only for the justification of the derivation of the Boussinesq systems but also for the justification of the mechanical balance laws that we will derive in the next section. Here, for the sake of simplicity, we assume that the Stokes number is equal to 1 ($\alpha = \beta$), so that we can work with a single small parameter α or β.

4.4.1 Flattening the Domain

The aim of this subsection is to present the method of "flattening the domain" of [67] so as to normalize the domain of the differential equations. Consider the boundary-value problem

$$\begin{cases} \beta\partial_{\tilde{x}}^2\omega + \beta\partial_{\tilde{y}}^2\omega + \partial_{\tilde{z}}^2\omega = R \text{ in } \Omega_t, \\ \omega|_{\tilde{z}=1+\alpha\tilde{\eta}} = 0, \\ \partial_n\omega|_{\tilde{z}=0} = 0, \end{cases} \tag{4.30}$$

with variable ω and a given smooth function R. We will give estimates on ω and ω_t that justify the results obtained in Theorem 8 and Theorem 9. To do that we transform the above system into an elliptic boundary value problem on a fixed domain. We assume that the water height is always bounded from below. In other words,

$$\exists h_{min} > 0, \qquad \forall X = (x, y) \in \mathbb{R}^2, \qquad 1 + \alpha\tilde{\eta} \geq h_{min}. \tag{4.31}$$

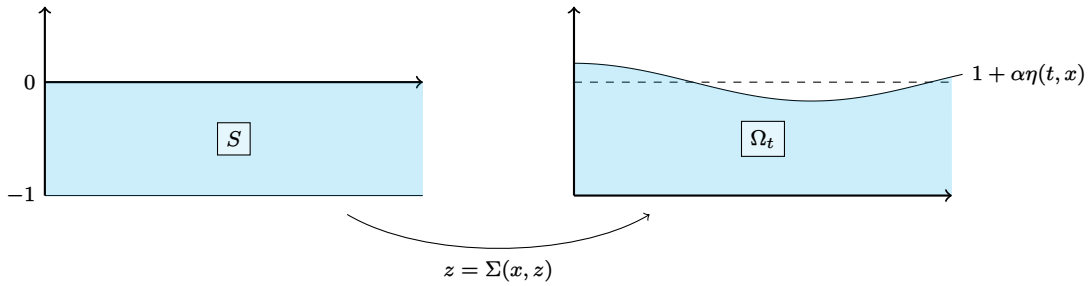

Figure 4.1: The transformation Σ in a two-dimensional setting

We transform the variable domain Ω_t into a flat strip $S = \mathbb{R}^2 \times (-1, 0)$ by introducing the following diffeomorphism (We choose here the most obvious diffeomorphism but there are other choices of regularizing diffeomorphisms that are useful for obtaining optimal regularity estimates, see Fig. 4.1):

$$\begin{aligned} \Sigma : \quad & S \longmapsto \Omega_t \\ & (X, z) \to \Sigma(X, z) = (X, (1 + \alpha\tilde{\eta})z + 1 + \alpha\tilde{\eta}) \end{aligned}$$

Then $\mathbf{w} = \omega \circ \Sigma$ and $\mathbf{R} = R \circ \Sigma$ satisfy the following elliptic boundary-value problem on the fixed domain S:

$$\begin{cases} \dfrac{1}{1 + \partial_{\tilde{z}}\sigma}\nabla_{X,z}^{\beta} \cdot P(\Sigma)\nabla_{X,z}^{\beta}\mathbf{w} = \mathbf{R} \text{ in } (-1, 0) \times \mathbb{R}^2 \\ \mathbf{w}|_{\tilde{z}=0} = 0 \\ e_z \cdot P(\Sigma)\nabla_{X,z}^{\beta}\mathbf{w}|_{\tilde{z}=-1} = 0. \end{cases} \tag{4.32}$$

Here we use the notations

$$\begin{aligned}\sigma(X,\tilde{z}) &= \alpha\tilde{\eta}\tilde{z}+1+\alpha\tilde{\eta},\\ P(\Sigma) &= I+Q(\Sigma),\\ Q(\Sigma) &= \begin{pmatrix} \partial_{\tilde{z}}\sigma I_2 & -\sqrt{\beta}\nabla\sigma \\ -\sqrt{\beta}\nabla\sigma^T & \frac{-\partial_{\tilde{z}}\sigma+\beta|\nabla\sigma|^2}{1+\partial_{\tilde{z}}\sigma} \end{pmatrix}.\end{aligned}$$

Theorem 7. *Let $\mathbf{R}\in H^{s,0}$, and let $\mathbf{w}\in H^{s+1,1}, s\geq 0$ solution of (4.32) such that assumption (4.31) is satisfied. Then there exists a constant*

$$C=C(h_{min}^{-1},\beta_{max},|\tilde{\eta}|_{H^{s+1\vee t_0+1}})$$

such that, for some $t_0>2$, the following estimate holds

$$|\Lambda^s\nabla^{\beta}_{X,z}\mathbf{w}|_{L^2}\leq C\,|\Lambda^s\mathbf{R}|_{L^2}\,,$$

where β_{max} is an upper bound of the shallowness parameter β.

Proof. We briefly sketch the proof for the case $s=0$. Multiplying the first equation of (4.32) by $\mathbf{w}$ and integrating yields

$$\int_S \nabla^{\beta}_{X,z}\mathbf{w}\cdot P(\Sigma)\nabla^{\beta}_{X,z}\mathbf{w}\,dXdz=\int_S \mathbf{R}\mathbf{w}\,dXdz.$$

Using the coercivity[1] of $P(\Sigma)$ along with the Poincaré inequality one gets the following estimate:

$$\begin{aligned}|\nabla^{\beta}_{X,z}\mathbf{w}|^2_{L^2} &\leq C(h_{min}^{-1},\beta_{\max},|\tilde{\eta}|_{H^{t_0+2}})|\mathbf{R}|_{L^2}|\mathbf{w}|_{L^2},\\ &\leq C\,|\mathbf{R}|_{L^2}|\nabla^{\beta}_{X,z}\mathbf{w}|_{L^2},\end{aligned}$$

which proves the theorem for $s=0$. For $s>0$, the reader may consult [67]. □

Theorem 8. *Let $(\eta^{Euler},\tilde{\Phi})$ be a regular solution of the Euler system (4.5) such that $(\eta^{Euler},\nabla\psi)\in H^s(\mathbb{R}^2)\times H^s(\mathbb{R}^2)$ with s large enough. Assume that the total water depth satisfies (4.31). Then, for $0<\tilde{t}<T/\beta$ we have,*

$$\begin{aligned}|\tilde{\Phi}^{app}_{\tilde{x}}-\tilde{\Phi}_{\tilde{x}}|_{L^\infty(\Omega_t)} &\lesssim \alpha\beta+\beta^2\ ,\\ |\tilde{\Phi}^{app}_{\tilde{y}}-\tilde{\Phi}_{\tilde{y}}|_{L^\infty(\Omega_t)} &\lesssim \alpha\beta+\beta^2\ ,\\ |\tilde{\Phi}^{app}_{\tilde{z}}-\tilde{\Phi}_{\tilde{z}}|_{L^\infty(\Omega_t)} &\lesssim \alpha\beta+\beta^2\ .\end{aligned}$$

Proof. Let $\omega=\tilde{\Phi}^{app}-\tilde{\Phi}$. Since $\tilde{\Phi}$ is a solution of (4.8) then its asymptotic expansion given by $\tilde{\Phi}^{app}$ satisfies

$$\begin{cases}\beta\partial^2_{\tilde{x}}\tilde{\Phi}^{app}+\beta\partial^2_{\tilde{y}}\tilde{\Phi}^{app}+\partial^2_{\tilde{z}}\tilde{\Phi}^{app}=r\,,\\ \tilde{\Phi}^{app}|_{\tilde{z}=1+\alpha\tilde{\eta}}=\psi\,,\\ \partial_n\tilde{\Phi}^{app}|_{\tilde{z}=0}=0\,,\end{cases}$$

where $r=\alpha\beta^2 r_1+\beta^3 r_2$ is a regular function in terms of $\tilde{z}$ and the derivatives of ψ.

[1]The coercivity of $P(\Sigma)$ follows from the choice of the diffeomorphism, and we have $\forall\theta\in\mathbb{R}^3$, $\forall(X,z)\in\mathcal{S}$, $\exists K>0/P(\Sigma)(X,z)\theta.\theta\geq K|\theta|^2$

Moreover, it is obvious that

$$\partial_n \omega|_{\tilde{z}=0} = 0 \quad \text{and} \quad \omega|_{\tilde{z}=1+\alpha\tilde{\eta}} = 0 \ .$$

Then ω satisfies the boundary-value problem (4.30) in the domain Ω_t:

$$\begin{cases} \beta\partial_{\tilde{x}}^2\omega + \beta\partial_{\tilde{y}}^2\omega + \partial_{\tilde{z}}^2\omega = r \ , \\ \partial_n\omega|_{\tilde{z}=0} = 0 \ , \\ \omega|_{\tilde{z}=1+\alpha\tilde{\eta}} = 0 \ . \end{cases} \tag{4.33}$$

Let Σ be the diffeomorphism defined in Section 4.4.1. Using Σ, we transform system (4.33) into a boundary value problem on a flat strip $S = (-1,0)\times\mathbb{R}^2$ and hence it follows by Theorem 7 that there exists a constant C such that:

$$|\Lambda^s \nabla^\beta_{X,z}\mathbf{w}|_{L^2} \leq C\,|\Lambda^s \mathbf{r}|_{L^2} \ ,$$

where $\mathbf{w} = \omega\circ\Sigma$ and $\mathbf{r} = r\circ\Sigma$. Using that $H^{s-1,1}(S) \hookrightarrow L^\infty((-1,0);H^{s-3/2}(\mathbb{R}^2))$ with the following estimate

$$\begin{aligned} |\nabla^\beta_{X,z}\mathbf{w}|_{H^{s-1,1}(S)} &= |\Lambda^{s-1}\nabla^\beta_{X,z}\mathbf{w}|_{L^2} + |\Lambda^{s-1}\nabla^\beta_{X,z}\partial_z\mathbf{w}|_{L^2} \\ &\leq C(h_{min}^{-1}, \beta_{max}, |\tilde{\eta}|_{H^{s+1}}, |Q|_{L^\infty H^s}, |Q|_{H^{s,1}}, |\partial_z Q|_{L^\infty H^s})|\Lambda^s\nabla^\beta_{X,z}\mathbf{w}|_{L^2} \end{aligned}$$

we get that for s large enough, we have $H^{s-3/2}(\mathbb{R}^2) \hookrightarrow L^\infty(\mathbb{R}^2)$. Therefore by using the above information we obtain

$$\begin{aligned} |\nabla^\beta_{X,z}\mathbf{w}|_{L^\infty(S)} &= \operatorname*{ess\,sup}_{(X,z)\in S} |\nabla^\beta_{X,z}\mathbf{w}| \\ &\leq |\nabla^\beta_{X,z}\mathbf{w}|_{L^\infty((-1,0);H^{s-3/2}(\mathbb{R}^2))} \\ &\leq C_1\alpha\beta^2 + C_2\beta^3, \end{aligned}$$

where C_1, C_2 are constants independent of β. The last estimate remains true on Ω_t for $\nabla^\beta_{X,z}\omega$. Since, after changing variables we have

$$\nabla^\beta_{X,z}\omega = \nabla^\beta_{X,z}\left(\mathbf{w}\circ\Sigma^{-1}\right) = J^T_{\Sigma^{-1}}\left(\nabla^\beta_{X,z}\mathbf{w}\circ\Sigma^{-1}\right),$$

where the coefficients of the Jacobian matrix $J^T_{\Sigma^{-1}}$ corresponding to the inverse transformation Σ^{-1}, are bounded. □

Remark 20. *The results of the above theorem can be extended to estimates of order $\mathcal{O}(\alpha^i\beta^j, \alpha^j\beta^i)$, $i,j \geq 1$.*

4.4.2 Expression for the pressure

The aim of this section is to find an approximation of the pressure defined in the context of the Euler equations. We begin with a useful remark derived from the Zakharov-Craig-Sulem equations (4.7), further details can be found in [67].

Lemma 10. *Let $\tilde{\eta} \in H^{s+1/2}(\mathbb{R}^2) \cap H^{t_0+2}(\mathbb{R}^2)$ with $s \geq 0$, $t_0 > 1$ satisfying (4.31). Then, the following mappings are continuous:*

$$\begin{aligned} \mathcal{G}_\beta[\alpha\tilde{\eta}] \colon \dot{H}^{s+1}(\mathbb{R}^2) &\to H^{s-1/2}(\mathbb{R}^2) \\ \psi &\mapsto \mathcal{G}_\beta[\alpha\tilde{\eta}]\psi \end{aligned} \tag{4.34}$$

$$\begin{aligned} \nu[\beta\tilde{\eta}] \colon \dot{H}^{s+1/2}(\mathbb{R}^2) &\to H^{s-1/2}(\mathbb{R}^2) \\ \psi &\mapsto \frac{[\mathcal{G}_\beta[\alpha\tilde{\eta}]\psi + \alpha\beta\nabla\tilde{\eta}\cdot\nabla\psi]^2}{2(1+\alpha^2\beta|\nabla\tilde{\eta}|^2)}. \end{aligned} \tag{4.35}$$

Theorem 9. *Under the same assumptions as those of Theorem 8 we have,*

$$|\tilde{\Phi}^{app}_{\tilde{t}} - \tilde{\Phi}_{\tilde{t}}|_{L^\infty(\Omega_t)} \lesssim \alpha\beta + \beta^2 \, .$$

Proof. Let $(\tilde{\eta}, \nabla\psi) \in H^s(\mathbb{R}^2) \times H^s(\mathbb{R}^2)$ with s large enough. As a result of Lemma 10 we have,

$$\frac{[\mathcal{G}_\beta[\alpha\tilde{\eta}]\psi + \alpha\beta\nabla\tilde{\eta}\cdot\nabla\psi]^2}{2(1+\alpha^2\beta|\nabla\tilde{\eta}|^2)} \in H^{s-1/2}(\mathbb{R}^2) \, .$$

In fact, this quantity appears in the expression of $\psi_{\tilde{t}}$ as shown in system (4.7):

$$\psi_{\tilde{t}} = -\tilde{\eta} - \frac{\alpha}{2}|\nabla\psi|^2 + \frac{\alpha}{\beta}\frac{[\mathcal{G}_\beta[\alpha\tilde{\eta}]\psi + \alpha\beta\nabla\tilde{\eta}\cdot\nabla\psi]^2}{2(1+\alpha^2\beta|\nabla\tilde{\eta}|^2)} \, .$$

Hence $\psi_{\tilde{t}} \in H^{s-1/2}(\mathbb{R}^2)$, and for s large enough we have $H^{s-1/2}(\mathbb{R}^2) \hookrightarrow L^\infty(\mathbb{R}^2)$.
We will use the notation $\omega = \tilde{\Phi}^{app} - \tilde{\Phi}$. Then ω satisfies the boundary value problem (4.33). In fact, if we denote $\phi = \tilde{\Phi} \circ \Sigma$ then ϕ satisfies:

$$\begin{cases} \nabla^\beta_{X,z} \cdot P(\Sigma)\nabla^\beta_{X,z}\phi = 0 \text{ in } \mathbb{R}^2 \times (-1,0) \\ \phi|_{\tilde{z}=0} = \psi \\ e_z \cdot P(\Sigma)\nabla^\beta_{X,z}\phi|_{\tilde{z}=-1} = 0 \end{cases} \tag{4.36}$$

We look for an approximate solution to the above system of the form

$$\phi^{app} = \sum_{j=0}^{N} \beta^j \phi_j.$$

By replacing ϕ^{app} in (4.36) and canceling higher order terms in β one obtains:

$$\frac{1}{H}\partial^2_{\tilde{z}}\phi_0 = 0, \qquad \phi_0|_{\tilde{z}=0} = \psi, \qquad \frac{1}{H}\partial_{\tilde{z}}\phi_0 = 0, \tag{4.37}$$

and for all $1 \leq j \leq n$,

$$\begin{cases} \dfrac{1}{H}\partial^2_{\tilde{z}}\phi_j = -A(\nabla, \partial_{\tilde{z}})\phi_{j-1}, \\ \phi_j|_{\tilde{z}=0} = \psi, \qquad \dfrac{1}{H}\partial_{\tilde{z}}\phi_j = 0. \end{cases} \tag{4.38}$$

The operator $A(\nabla, \partial_z)$ is given by

$$A(\nabla, \partial_{\tilde{z}})\bullet = \nabla \cdot (H\nabla \bullet) + \partial_{\tilde{z}}(\frac{|\nabla\sigma|^2}{H}\partial_{\tilde{z}}\bullet) - \nabla \cdot (\nabla\sigma\, \partial_{\tilde{z}}\bullet) - \partial_{\tilde{z}}(\nabla\sigma \cdot \nabla\bullet),$$

where $H(t,X) = 1 + \alpha\tilde{\eta}(t,X)$ denotes the water depth and $\sigma(t,X,\tilde{z}) = \alpha\tilde{\eta}(t,X)\tilde{z} + (1+\alpha\tilde{\eta}(t,X))$. Then, $\phi_0 = \psi$ and hence we can write:

$$\phi = \psi + h,$$

where h is function of $\tilde{z}$, $\tilde{\eta}$ and the derivatives of ψ. Thus, for $\mathbf{w} = \omega \circ \Sigma$ one has $\mathbf{w}_{\tilde{t}} \in H^{s-1/2}(\mathbb{R}^2)$. (One can choose initially $(\tilde{\eta}, \nabla\psi) \in H^{s+t_0}(\mathbb{R}^2) \times H^{s+t_0}(\mathbb{R}^2)$ with t_0 large enough.) Differentiating (4.32) with respect to time, and for the sake of brevity (we refer to [67, Lemma 5.4]) one readily obtains the following estimate:

$$|\Lambda^s \nabla^{\beta}_{X,z} \mathbf{w}_{\tilde{\mathbf{t}}}|_{L^2} \leq (\alpha\beta^2 + \beta^3) C(h_{min}^{-1}, \beta_{max}, |\tilde{\eta}|_{H^{s+t_0}}, |\nabla\psi|_{H^{s+t_0}}).$$

Using the embedding $H^{s-1,1} \hookrightarrow L^{\infty}((-1,0); H^{s-3/2})$ as in Theorem 8 one obtains the result. □

The approximation of the pressure term $\tilde{P}'$ is given by the following corollary.

Corollary 5. *Under the same assumptions as those of Theorem 8 we have,*

$$|\tilde{P}' - \tilde{Q}|_{L^{\infty}(\Omega_t)} \leq C(\alpha\beta + \beta^2), \tag{4.39}$$

where $\tilde{Q}$ is an approximation of the pressure defined by:

$$\tilde{Q} = \tilde{\eta} + \frac{1}{2}\beta(\tilde{z}^2 - 1)[\tilde{U}_{\tilde{x}\tilde{t}} + \tilde{V}_{\tilde{y}\tilde{t}}], \tag{4.40}$$

and C is a constant independent of β.

Proof. The definition of the pressure $\tilde{P}'$ is given in Section 4.3 by

$$\tilde{P}' = -\tilde{\Phi}_{\tilde{t}} - \frac{1}{2}\alpha\left(\tilde{\Phi}_{\tilde{x}}^2 + \tilde{\Phi}_{\tilde{y}}^2\right) - \frac{1}{2}\frac{\alpha}{\beta}\tilde{\Phi}_{\tilde{z}}^2. \tag{4.41}$$

The expression of the approximate pressure $\tilde{Q}$ follows directly by replacing $\tilde{\Phi}$ by $\tilde{\Phi}^{app}$ in the expression of $\tilde{P}'$ above (see equation (4.28) for details), and thus one has

$$\begin{aligned}\tilde{P}' - \tilde{Q} = &-\left(\tilde{\Phi}_{\tilde{t}} - \tilde{\Phi}^{app}_{\tilde{t}}\right) - \frac{1}{2}\alpha\left(\tilde{\Phi}_{\tilde{x}} - \tilde{\Phi}^{app}_{\tilde{x}}\right)\left(\tilde{\Phi}_{\tilde{x}} + \tilde{\Phi}^{app}_{\tilde{x}}\right) \\ &- \frac{1}{2}\alpha\left(\tilde{\Phi}_{\tilde{y}} - \tilde{\Phi}^{app}_{\tilde{y}}\right)\left(\tilde{\Phi}_{\tilde{y}} + \tilde{\Phi}^{app}_{\tilde{y}}\right) - \frac{1}{2}\frac{\alpha}{\beta}\left(\tilde{\Phi}_{\tilde{z}} - \tilde{\Phi}^{app}_{\tilde{z}}\right)\left(\tilde{\Phi}_{\tilde{z}} + \tilde{\Phi}^{app}_{\tilde{z}}\right).\end{aligned}$$

Given that $\tilde{\Phi}$ is regular enough and choosing s sufficiently large such that $H^s(\mathbb{R}^2) \hookrightarrow L^{\infty}(\mathbb{R}^2)$ implies that $\nabla_{X,z}\tilde{\Phi}^{app} \in L^{\infty}(\Omega_t)$ and consequently

$$\begin{aligned}|\nabla_{X,z}\tilde{\Phi} + \nabla_{X,z}\tilde{\Phi}^{app}|_{L^{\infty}(\Omega_t)} &\leq |\nabla_{X,z}\tilde{\Phi}|_{L^{\infty}(\Omega_t)} + |\nabla_{X,z}\tilde{\Phi}^{app}|_{L^{\infty}(\Omega_t)} \\ &\leq k\ ,\end{aligned}$$

for some constant k. Hence using the previous estimates, one can deduce the result. □

4.5 Derivation and justification of the mechanical balance laws

The Euler equations (4.5) conserve the total mass, momentum, and energy. However, this is not the case for most asymptotic models including the Boussinesq equations (4.16a)-(4.16b). Approximate expressions for the mass, momentum, and energy balances will be derived below and justified up to same order as that of the Boussinesq approximation. We refer to [57, 58] for similar results on the Korteweg-de Vries equation.

4.5.1 Mass balance

In this section, we establish the mass conservation properties of the general family of the two-dimensional Boussinesq systems (4.20a)–(4.20c), (4.21). The incompressibility of the fluid is expressed in terms of the velocity by equation (4.1a), which after integration and by using Leibniz rule yields

$$\frac{\partial}{\partial t}\int_{-h_0}^{\eta}\rho\,dz-\rho\eta_t+\frac{\partial}{\partial x}\int_{-h_0}^{\eta}\rho u\,dz-\rho u|_{z=\eta}\eta_x+\frac{\partial}{\partial y}\int_{-h_0}^{\eta}\rho v\,dz-\rho v|_{z=\eta}\eta_y+\int_{-h_0}^{\eta}\frac{\partial}{\partial z}(\rho w)\,dz=0\ .$$

However, without loss of generality we assume that $\rho=1$ which simplifies the previous relation. Since the vertical velocity at the bottom $w|_{(-h_0)}=0$ and the kinematic boundary condition at the free surface of the water is $\eta_t+\Phi_x\eta_x+\Phi_y\eta_y-\Phi_z=0\quad\text{on}\quad z=\eta(x,y,t)$, we have

$$\frac{\partial}{\partial t}\int_{-h_0}^{\eta}dz+\frac{\partial}{\partial x}\int_{-h_0}^{\eta}u\,dz+\frac{\partial}{\partial y}\int_{-h_0}^{\eta}v\,dz=0\ , \tag{4.42}$$

or equivalently in non-dimensional variables

$$\frac{\partial}{\partial\tilde t}\left(1+\alpha\tilde\eta\right)+\frac{\partial}{\partial\tilde x}\int_{\tilde z=0}^{1+\alpha\tilde\eta}\alpha\tilde\Phi_{\tilde x}\,d\tilde z+\frac{\partial}{\partial\tilde y}\int_{\tilde z=0}^{1+\alpha\tilde\eta}\alpha\tilde\Phi_{\tilde y}\,d\tilde z=0\ . \tag{4.43}$$

Substituting $\tilde\Phi$ by $\tilde\Phi^{app}$ and using Remark 18:

$$\begin{aligned}\frac{\partial}{\partial\tilde t}\left(1+\alpha\tilde\eta\right)+\frac{\partial}{\partial\tilde x}\left[\tilde U(\alpha+\alpha^2\tilde\eta)+\frac{\alpha\beta}{2}\left(\theta^2-\frac{1}{3}\right)\Delta\tilde U\right]\\+\frac{\partial}{\partial\tilde y}\left[\tilde V(\alpha+\alpha^2\tilde\eta)+\frac{\alpha\beta}{2}\left(\theta^2-\frac{1}{3}\right)\Delta\tilde V\right]=\mathcal{O}(\alpha\beta^2,\alpha^2\beta)\ .\end{aligned} \tag{4.44}$$

Finally, we obtain the differential balance equation

$$\tilde\eta_{\tilde t}+\tilde U_{\tilde x}+\tilde V_{\tilde y}+\alpha\left[(\tilde U\tilde\eta)_{\tilde x}+(\tilde V\tilde\eta)_{\tilde y}\right]+\frac{\beta}{2}\left(\theta^2-\frac{1}{3}\right)(\Delta\tilde U_{\tilde x}+\Delta\tilde V_{\tilde y})=\mathcal{O}(\alpha\beta,\beta^2)\ . \tag{4.45}$$

From (4.44) the non-dimensional mass and the non-dimensional mass fluxes are

$$\tilde M=1+\alpha\tilde\eta\ ,$$

$$\tilde q_{m_x}=\tilde U(\alpha+\alpha^2\tilde\eta)+\frac{\alpha\beta}{2}\left(\theta^2-\frac{1}{3}\right)\Delta\tilde U\ ,$$

$$\tilde{q}_{m_y} = \tilde{V}(\alpha + \alpha^2\tilde{\eta}) + \frac{\alpha\beta}{2}\left(\theta^2 - \frac{1}{3}\right)\Delta\tilde{V} \ .$$

Thus, the mass balance is

$$\frac{\partial}{\partial\tilde{t}}\tilde{M} + \frac{\partial}{\partial\tilde{x}}\tilde{q}_{m_x} + \frac{\partial}{\partial\tilde{y}}\tilde{q}_{m_y} = \mathcal{O}(\alpha\beta^2, \beta^3) \ . \tag{4.46}$$

In dimensional variables the quantities in mass balance equation (4.46) are the following:

$$\begin{aligned} M &= h_0 + \eta \ , \\ q_{m_x} &= U(h_0 + \eta) + h_0^3\frac{1}{2}\left(\theta^2 - \frac{1}{3}\right)\Delta U \ , \\ q_{m_y} &= V(h_0 + \eta) + h_0^3\frac{1}{2}\left(\theta^2 - \frac{1}{3}\right)\Delta V \ . \end{aligned}$$

Theorem 10. *Under the same assumptions as those of Theorem 8, there exists a constant C_1 independent of β such that the following approximate mass balance is satisfied:*

$$\begin{aligned} &\Big|\frac{\partial}{\partial\tilde{t}}(1 + \alpha\tilde{\eta}) + \frac{\partial}{\partial\tilde{x}}\left\{\tilde{U}(\alpha + \alpha^2\tilde{\eta}) + \frac{\alpha\beta}{2}\left(\theta^2 - \frac{1}{3}\right)\Delta\tilde{U}\right\} \\ &+ \frac{\partial}{\partial\tilde{y}}\left\{\tilde{V}(\alpha + \alpha^2\tilde{\eta}) + \frac{\alpha\beta}{2}\left(\theta^2 - \frac{1}{3}\right)\Delta\tilde{V}\right\}\Big|_{L^\infty(\mathbb{R}^2)} \leq C_1(\alpha\beta^2 + \alpha^2\beta) \ . \end{aligned}$$

Proof. It follows from (4.44) that the error term on the right-hand side can be expressed as a function of $\tilde{\eta}, \tilde{U}, \tilde{V}$ which can be in turn expressed in terms of $\tilde{\eta}$ and ψ. Hence taking s large enough such that $H^s(\mathbb{R}^2) \hookrightarrow L^\infty(\mathbb{R}^2)$, the error term is bounded. □

In the following section we find an approximate expression for momentum density and flux.

4.5.2 Momentum balance

For obtaining momentum balance, we first consider the Euler equations (4.1) written in terms of the pressure and velocity variables. Writing these equations in terms of the velocity potential $\boldsymbol{U} = \nabla_{X,z}\mathring{\Phi}$ and taking $\rho = 1$, we get the following equations

$$\begin{aligned} &\mathring{\Phi}_{xt} + (\mathring{\Phi}_x^2)_x + (\mathring{\Phi}_x\mathring{\Phi}_y)_y + (\mathring{\Phi}_x\mathring{\Phi}_z)_z + P_x = 0, \\ &\mathring{\Phi}_{yt} + (\mathring{\Phi}_y^2)_y + (\mathring{\Phi}_y\mathring{\Phi}_x)_x + (\mathring{\Phi}_y\mathring{\Phi}_z)_z + P_y = 0. \end{aligned}$$

Integrating over a fluid column and using the kinematic boundary condition (4.3) yields

$$\begin{aligned} &\frac{\partial}{\partial t}\int_{-h_0}^{\eta}\mathring{\Phi}_x\,dz + \frac{\partial}{\partial x}\int_{-h_0}^{\eta}\left(\mathring{\Phi}_x^2 + P\right)\,dz + \frac{\partial}{\partial y}\int_{-h_0}^{\eta}\mathring{\Phi}_x\mathring{\Phi}_y\,dz = 0, \\ &\frac{\partial}{\partial t}\int_{-h_0}^{\eta}\mathring{\Phi}_y\,dz + \frac{\partial}{\partial y}\int_{-h_0}^{\eta}\left(\mathring{\Phi}_y^2 + P\right)\,dz + \frac{\partial}{\partial x}\int_{-h_0}^{\eta}\mathring{\Phi}_y\mathring{\Phi}_x\,dz = 0. \end{aligned}$$

Expressing the above relations in non-dimensional variables (4.4) leads to the equations

$$\frac{\partial}{\partial\tilde{t}}\int_{\tilde{z}=0}^{1+\alpha\tilde{\eta}}\alpha\tilde{\Phi}_{\tilde{x}}\,d\tilde{z}+\frac{\partial}{\partial\tilde{x}}\int_{\tilde{z}=0}^{1+\alpha\tilde{\eta}}\left\{\alpha^2(\tilde{\Phi}_{\tilde{x}}^2)+\alpha\tilde{P}'-(\tilde{z}-1)\right\}\,d\tilde{z}+\frac{\partial}{\partial\tilde{y}}\int_{\tilde{z}=0}^{1+\alpha\tilde{\eta}}\alpha^2\tilde{\Phi}_{\tilde{x}}\tilde{\Phi}_{\tilde{y}}\,d\tilde{z}=0,$$
$$\frac{\partial}{\partial\tilde{t}}\int_{\tilde{z}=0}^{1+\alpha\tilde{\eta}}\alpha\tilde{\Phi}_{\tilde{y}}\,d\tilde{z}+\frac{\partial}{\partial\tilde{y}}\int_{\tilde{z}=0}^{1+\alpha\tilde{\eta}}\left\{\alpha^2(\tilde{\Phi}_{\tilde{y}}^2)+\alpha\tilde{P}'-(\tilde{z}-1)\right\}\,d\tilde{z}+\frac{\partial}{\partial\tilde{x}}\int_{\tilde{z}=0}^{1+\alpha\tilde{\eta}}\alpha^2\tilde{\Phi}_{\tilde{x}}\tilde{\Phi}_{\tilde{y}}\,d\tilde{z}=0.$$

Substituting non-dimensional velocity potentials $\tilde{\Phi}_{\tilde{x}}$ and $\tilde{\Phi}_{\tilde{y}}$ and $\tilde{P}'$ in terms of $\tilde{U}$ and $\tilde{V}$ gives the momentum balance equations:

$$\frac{\partial}{\partial\tilde{t}}\left\{(1+\alpha\tilde{\eta})\tilde{U}+\frac{\beta}{2}\left(\theta^2-\frac{1}{3}\right)\Delta\tilde{U}\right\}$$
$$+\frac{\partial}{\partial\tilde{x}}\left\{\tilde{\eta}+\alpha\tilde{U}^2+\frac{\alpha\tilde{\eta}^2}{2}-\frac{\beta}{3}(\tilde{U}_{\tilde{x}\tilde{t}}+\tilde{V}_{\tilde{y}\tilde{t}})+\frac{1}{2}\right\}+\frac{\partial}{\partial\tilde{y}}(\alpha\tilde{U}\tilde{V})=\mathcal{O}(\alpha\beta,\beta^2),\quad(4.47)$$

$$\frac{\partial}{\partial\tilde{t}}\left\{(1+\alpha\tilde{\eta})\tilde{V}+\frac{\beta}{2}\left(\theta^2-\frac{1}{3}\right)\Delta\tilde{V}\right\}$$
$$+\frac{\partial}{\partial\tilde{y}}\left\{\tilde{\eta}+\alpha\tilde{V}^2+\frac{\alpha\tilde{\eta}^2}{2}-\frac{\beta}{3}(\tilde{U}_{\tilde{x}\tilde{t}}+\tilde{V}_{\tilde{y}\tilde{t}})+\frac{1}{2}\right\}+\frac{\partial}{\partial\tilde{x}}(\alpha\tilde{U}\tilde{V})=\mathcal{O}(\alpha\beta,\beta^2).\quad(4.48)$$

If the terms of order $\mathcal{O}(\alpha\beta,\beta^2)$ are neglected, then the momentum balance equations written in dimensional variables take the following form:

$$\frac{\partial}{\partial t}\left\{(h_0+\eta)U+\frac{1}{2}\left(\theta^2-\frac{1}{3}\right)\Delta U\right\}+\frac{\partial}{\partial x}\left\{h_0U^2+\frac{g}{2}\left(h_0+\eta\right)^2-\frac{h_0^3}{3}(U_{xt}+V_{yt})\right\}+\frac{\partial}{\partial y}(h_0UV)=0,$$

$$\frac{\partial}{\partial t}\left\{(h_0+\eta)V+\frac{1}{2}\left(\theta^2-\frac{1}{3}\right)\Delta V\right\}+\frac{\partial}{\partial y}\left\{h_0V^2+\frac{g}{2}\left(h_0+\eta\right)^2-\frac{h_0^3}{3}(U_{xt}+V_{yt})\right\}+\frac{\partial}{\partial x}(h_0UV)=0.$$

The following theorem is a direct consequence of the estimates in Theorem 8 and Corollary 5.

Theorem 11. *Under the same assumptions as those of Theorem 8, there exists constants C_2, C_3 independent of β such that the following approximate momentum balances are satisfied:*

$$\Big|\frac{\partial}{\partial\tilde{t}}\left\{(1+\alpha\tilde{\eta})\tilde{U}+\frac{\beta}{2}\left(\theta^2-\frac{1}{3}\right)\Delta\tilde{U}\right\}$$
$$+\frac{\partial}{\partial\tilde{x}}\left\{\tilde{\eta}+\alpha\tilde{U}^2+\frac{\alpha}{2}\tilde{\eta}^2-\frac{1}{3}\beta(\tilde{U}_{\tilde{x}\tilde{t}}+\tilde{V}_{\tilde{y}\tilde{t}})+\frac{1}{2}\right\}+\frac{\partial}{\partial\tilde{y}}(\alpha\tilde{U}\tilde{V})\Big|_{L^\infty(\mathbb{R}^2)}\leq C_2(\alpha\beta+\beta^2),$$

$$\Big|\frac{\partial}{\partial\tilde{t}}\left\{(1+\alpha\tilde{\eta})\tilde{V}+\frac{\beta}{2}\left(\theta^2-\frac{1}{3}\right)\Delta\tilde{V}\right\}$$
$$+\frac{\partial}{\partial\tilde{y}}\left\{\tilde{\eta}+\alpha\tilde{V}^2+\frac{\alpha}{2}\tilde{\eta}^2-\frac{1}{3}\beta(\tilde{U}_{\tilde{x}\tilde{t}}+\tilde{V}_{\tilde{y}\tilde{t}})+\frac{1}{2}\right\}+\frac{\partial}{\partial\tilde{x}}(\alpha\tilde{U}\tilde{V})\Big|_{L^\infty(\mathbb{R}^2)}\leq C_3(\alpha\beta+\beta^2).$$

4.5.3 Energy balance

The exact energy balance satisfied by the Euler equations (4.5) reads

$$\frac{\partial}{\partial t}\left\{\frac{1}{2}\left|\nabla\Phi^2\right|+gz\right\}+\nabla\cdot\left\{\left(\frac{1}{2}\left|\nabla\Phi^2\right|+gz+P\right)\nabla\Phi\right\}=0.$$

Integrating over the a water column yields

$$\begin{aligned}\frac{\partial}{\partial t}\left\{\int_{-h_0}^{\eta}\frac{1}{2}\left|\nabla\Phi\right|^2\,dz+\int_0^{\eta}gz\,dz\right\}+\frac{\partial}{\partial x}\left\{\int_{-h_0}^{\eta}\left(\frac{1}{2}\left|\nabla\Phi\right|^2+gz+P\right)\Phi_x\right\}dz\\ +\frac{\partial}{\partial y}\left\{\int_{-h_0}^{\eta}\left(\frac{1}{2}\left|\nabla\Phi\right|^2+gz+P\right)\Phi_y\right\}dz=0.\end{aligned}$$

Using non-dimensional variables the last equation reads:

$$\begin{aligned}\frac{\partial}{\partial\tilde{t}}\left\{\int_{\tilde{z}=0}^{1+\alpha\tilde{\eta}}\frac{\alpha^2}{2}\left(\tilde{\Phi}_{\tilde{x}}^2+\tilde{\Phi}_{\tilde{y}}^2+\frac{1}{\beta}\tilde{\Phi}_{\tilde{z}}^2\right)d\tilde{z}+\int_{\tilde{z}=1}^{1+\alpha\tilde{\eta}}(\tilde{z}-1)\,d\tilde{z}\right\}\\ +\frac{\partial}{\partial\tilde{x}}\int_{\tilde{z}=0}^{1+\alpha\tilde{\eta}}\left\{\frac{\alpha^3}{2}\left(\tilde{\Phi}_{\tilde{x}}^2+\tilde{\Phi}_{\tilde{y}}^2+\frac{1}{\beta}\tilde{\Phi}_{\tilde{z}}^2\right)+\alpha\left(\tilde{z}-1\right)+\alpha^2\tilde{P}'+\alpha\left(1-\tilde{z}\right)\right\}\tilde{\Phi}_{\tilde{x}}\,d\tilde{z}\\ +\frac{\partial}{\partial\tilde{y}}\int_{\tilde{z}=0}^{1+\alpha\tilde{\eta}}\left\{\frac{\alpha^3}{2}\left(\tilde{\Phi}_{\tilde{x}}^2+\tilde{\Phi}_{\tilde{y}}^2+\frac{1}{\beta}\tilde{\Phi}_{\tilde{z}}^2\right)+\alpha\left(\tilde{z}-1\right)+\alpha^2\tilde{P}'+\alpha\left(1-\tilde{z}\right)\right\}\tilde{\Phi}_{\tilde{y}}\,d\tilde{z}=0.\end{aligned}$$

Substituting the expressions for $\tilde{\Phi}_{\tilde{x}}$ and $\tilde{\Phi}_{\tilde{y}}$ in terms of $\tilde{U}$ and $\tilde{V}$ and (4.28) leads to the energy balance equation

$$\begin{aligned}\frac{\partial}{\partial\tilde{t}}\left[\frac{1}{2}(\tilde{U}^2+\tilde{V}^2+\tilde{\eta}^2)+\frac{\beta}{2}\left(\theta^2-\frac{1}{3}\right)(\tilde{U}\Delta\tilde{U}+\tilde{V}\Delta\tilde{V})+\frac{\beta}{6}(\tilde{U}_{\tilde{x}}+\tilde{V}_{\tilde{y}})^2+\frac{\alpha\tilde{\eta}}{2}(\tilde{U}^2+\tilde{V}^2)\right]\\ +\frac{\partial}{\partial\tilde{x}}\left[\frac{\alpha}{2}(\tilde{U}^3+\tilde{V}^2\tilde{U})+\alpha\tilde{\eta}^2\tilde{U}+\tilde{\eta}\tilde{U}+\frac{\beta}{2}\left(\theta^2-\frac{1}{3}\right)\tilde{\eta}\Delta\tilde{U}-\frac{\beta}{3}\tilde{U}(\tilde{U}_{\tilde{x}\tilde{t}}+\tilde{V}_{\tilde{y}\tilde{t}})\right]\\ +\frac{\partial}{\partial\tilde{y}}\left[\frac{\alpha}{2}(\tilde{V}^3+\tilde{U}^2\tilde{V})+\alpha\tilde{\eta}^2\tilde{V}+\tilde{\eta}\tilde{V}+\frac{\beta}{2}\left(\theta^2-\frac{1}{3}\right)\tilde{\eta}\Delta\tilde{V}-\frac{\beta}{3}\tilde{V}(\tilde{U}_{\tilde{x}\tilde{t}}+\tilde{V}_{\tilde{y}\tilde{t}})\right]=\mathcal{O}(\alpha\beta,\beta^2)\ .\end{aligned}$$

Hence, the general form of the energy balance equation is

$$\frac{\partial}{\partial\tilde{t}}\tilde{E}+\frac{\partial}{\partial\tilde{x}}\tilde{q}_{E_x}+\frac{\partial}{\partial\tilde{y}}\tilde{q}_{E_y}=\mathcal{O}(\alpha\beta,\beta^2)\ . \tag{4.49}$$

The dimensional form of the quantities are

$$E=\frac{1}{2}h_0\left(U^2+V^2\right)+\frac{1}{2}h_0^3\left(\theta^2-\frac{1}{3}\right)(U\Delta U+V\Delta V)+\frac{1}{2}\left(U^2+V^2\right)\eta+\frac{1}{6}h_0^3\left(U_x+V_y\right)^2+\frac{1}{2}g\eta^2\ , \tag{4.50}$$

and

$$q_{E_x}=\frac{1}{2}h_0\left(U^3+UV^2\right)+gh_0\eta U+\frac{1}{2}gh_0^3\left(\theta^2-\frac{1}{3}\right)\eta\Delta U-\frac{1}{3}h_0^3U\left(U_{xt}+V_{yt}\right)+gU\eta^2, \tag{4.51}$$

$$q_{E_y} = \frac{1}{2}h_0\left(V^3 + U^2V\right) + gh_0\eta V + \frac{1}{2}gh_0^3\left(\theta^2 - \frac{1}{3}\right)\eta\Delta V - \frac{1}{3}h_0^3V\left(U_{xt} + V_{yt}\right) + gV\eta^2. \tag{4.52}$$

The following theorem follows by direct substitution of the estimates of Theorem 8 in the energy balance equation (4.49).

Theorem 12. *Under the same assumptions as those of Theorem 8, there exists a constant C_4 independent of β such that the following approximate energy balances is satisfied:*

$$\left|\frac{\partial}{\partial\tilde{t}}\tilde{E} + \frac{\partial}{\partial\tilde{x}}\tilde{q}_{E_x} + \frac{\partial}{\partial\tilde{y}}\tilde{q}_{E_y}\right|_{L^\infty(\mathbb{R}^2)} \leq C_4(\alpha\beta + \beta^2)\ ,$$

where $\tilde{E}$, $\tilde{q}_{E_x}$, and $\tilde{q}_{E_y}$ are given by:

$$\begin{aligned}
\tilde{E} &= \frac{1}{2}(\tilde{U}^2 + \tilde{V}^2 + \tilde{\eta}^2) + \frac{\beta}{2}\left(\theta^2 - \frac{1}{3}\right)(\tilde{U}\Delta\tilde{U} + \tilde{V}\Delta\tilde{V}) + \frac{\beta}{6}(\tilde{U}_{\tilde{x}} + \tilde{V}_{\tilde{y}})^2 + \frac{\alpha\tilde{\eta}}{2}(\tilde{U}^2 + \tilde{V}^2),\\
\tilde{q}_{E_x} &= \frac{\alpha}{2}(\tilde{U}^3 + \tilde{V}^2\tilde{U}) + \alpha\tilde{\eta}^2\tilde{U} + \tilde{\eta}\tilde{U} + \frac{\beta}{2}\left(\theta^2 - \frac{1}{3}\right)\tilde{\eta}\Delta\tilde{U} - \frac{\beta}{3}\tilde{U}(\tilde{U}_{\tilde{x}\tilde{t}} + \tilde{V}_{\tilde{y}\tilde{t}}),\\
\tilde{q}_{E_y} &= \frac{\alpha}{2}(\tilde{V}^3 + \tilde{U}^2\tilde{V}) + \alpha\tilde{\eta}^2\tilde{V} + \tilde{\eta}\tilde{V} + \frac{\beta}{2}\left(\theta^2 - \frac{1}{3}\right)\tilde{\eta}\Delta\tilde{V} - \frac{\beta}{3}\tilde{V}(\tilde{U}_{\tilde{x}\tilde{t}} + \tilde{V}_{\tilde{y}\tilde{t}}).
\end{aligned}$$

The above results justify the error bounds on the approximate mass, momentum, and energy conservation.

Bibliography

[1] A. Ali and H. Kalisch. Mechanical balance laws for Boussinesq models of surface water waves. *Journal of Nonlinear Science*, 22:371–398, 2012.

[2] A. Ali and H. Kalisch. On the formulation of mass, momentum and energy conservation in the KdV equation. *Acta Applicandae Mathematicae*, 133:113–131, 2014.

[3] S. Allgeyer, M.-O. Bristeau, D. Froger, R. Hamouda, V. Jauzein, A. Mangeney, J. Sainte-Marie, F. Souillé, and M. Vallée. Numerical approximation of the 3d hydrostatic Navier-Stokes system with free surface. *ESAIM: M2AN*, 53(6):1981–2024, 2019.

[4] E. Audusse. A multilayer Saint-Venant model: derivation and numerical validation. *Discrete and Continuous Dynamical Systems - B*, 5(2):189–214, 2005.

[5] E. Audusse, F. Bouchut, M.-O. Bristeau, and J. Sainte-Marie. Kinetic entropy inequality and hydrostatic reconstruction scheme for the Saint-Venant system. *Mathematics of Computation*, 85(302):2815–2837, 2016.

[6] E. Audusse, M.-O. Bristeau, M. Pelanti, and J. Sainte-Marie. Approximation of the hydrostatic Navier-Stokes system for density stratified flows by a multilayer model. Kinetic interpretation and numerical validation. *Journal of Computational Physics*, 230:3453–3478, 2011.

[7] E. Audusse, M.-O. Bristeau, B. Perthame, and J. Sainte-Marie. A multilayer Saint-Venant system with mass exchanges for Shallow Water flows. Derivation and numerical validation. *ESAIM: M2AN*, 45:169–200, 2011.

[8] Emmanuel Audusse, François Bouchut, Marie-Odile Bristeau, Rupert Klein, and Benoît Perthame. A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows. *SIAM J. Sci. Comput.*, 25(6):2050–2065, 2004.

[9] Emmanuel Audusse and Marie-Odile Bristeau. A well-balanced positivity preserving "second-order" scheme for shallow water flows on unstructured meshes. *J. Comput. Phys.*, 206(1):311–333, 2005.

[10] A.-J.-C. Barré de Saint-Venant. Théorie du mouvement non permanent des eaux, avec application aux crues des rivières et à l'introduction des marées dans leur lit. *Comptes rendus de l'Académie des Sciences Paris*, 73:147–154, 1871.

[11] Eric Barthélemy. Nonlinear shallow water theories for coastal waves. *Surveys in Geophysics*, 25:315–337, 07 2004.

[12] F. Berthelin and F. Bouchut. Relaxation to isentropic gas dynamics for a BGK system with single kinetic entropy. *Methods Appl. Anal.*, 9(2):313–327, 2002.

[13] P. L. Bhatnagar, E. P. Gross, and M. Krook. A model for collision processes in gases. I: Small amplitude processes in charged and neutral one-component systems. *Phys. Rev., II. Ser.*, 94:511–525, 1954.

[14] J. Bona, M. Chen, and J.-C. Saut. Boussinesq equations and other systems for small-amplitude long waves in nonlinear dispersive media. I: Derivation and linear theory. *Journal of Nonlinear Science*, 12(4), 2002.

[15] J. Bona, T. Colin, and D. Lannes. Long wave approximations for water waves. *Archive for Rational Mechanics and Analysis*, 178:373–410, 2005.

[16] Jerry L. Bona, Thierry Colin, and David Lannes. Long wave approximations for water waves. *Arch. Ration. Mech. Anal.*, 178(3):373–410, 2005.

[17] F. Bouchut. Construction of BGK models with a family of kinetic entropies for a given system of conservation laws. *J. Stat. Phys.*, 95(1-2):113–170, 1999.

[18] F. Bouchut. Entropy satisfying flux vector splittings and kinetic BGK models. *Numer. Math.*, 94(4):623–672, 2003.

[19] F. Bouchut, A. Mangeney-Castelnau, B. Perthame, and J.-P. Vilotte. A new model of Saint-Venant and Savage-Hutter type for gravity driven shallow water flows. *Comptes rendus de l'Académie des Sciences Paris*, 336(6):531–536, 2003.

[20] François Bouchut. *Nonlinear stability of finite volume methods for hyperbolic conservation laws and well-balanced schemes for sources.* Front. Math. Basel: Birkhäuser, 2004.

[21] François Bouchut and Xavier Lhébrard. Convergence of the kinetic hydrostatic reconstruction scheme for the Saint Venant system with topography. *Math. Comput.*, 90(329):1119–1153, 2021.

[22] J. Boussinesq. Théorie des ondes et des remous qui se propagent le long d'un canal rectangulaire horizontal, en communiquant au liquide contenu dans ce canal des vitesses sensiblement pareilles de la surface au fond. *J. Math. Pures. Appl.*, 17:55–108, 1872.

[23] Yann Brenier. Homogeneous hydrostatic flows with convex velocity profiles. *Nonlinearity*, 12(3):495–512, 1999.

[24] Didier Bresch, Alexandre Kazhikhov, and Jérôme Lemoine. On the two-dimensional hydrostatic Navier-Stokes equations. *SIAM J. Math. Anal.*, 36(3):796–814, 2004.

[25] M.-O. Bristeau, B. Di Martino, C. Guichard, and J. Sainte-Marie. Layer-averaged Euler and Navier-Stokes equations. *Communications in Mathematical Sciences*, 15(5):1221–1246, 2017.

[26] Marie-Odile Bristeau and Benoit Coussin. Boundary Conditions for the Shallow Water Equations solved by Kinetic Schemes. Research Report RR-4282, INRIA, 2001. Projet M3N.

[27] Marie-Odile Bristeau, Nicole Goutal, and Jacques Sainte-Marie. Numerical simulations of a non-hydrostatic shallow water model. *Comput. Fluids*, 47(1):51–64, 2011.

[28] F. Chazel, M. Benoit, A. Ern, and S. Piperno. A double-layer Boussinesq-type model for highly nonlinear and dispersive waves. *Proc. R. Soc. Lond., Ser. A, Math. Phys. Eng. Sci.*, 465(2108):2319–2346, 2009.

[29] A. A. Chesnokov. Exact solutions to the vortex shallow water equations. *Prikl. Mekh. Tekh. Fiz.*, 38(5):44–55, 1997.

[30] A. A. Chesnokov, G. A. El, Sergey L. Gavrilyuk, and M. V. Pavlov. Stability of shear shallow water flows with free surface. *SIAM Journal on Applied Mathematics*, 77(3):1068–1087, 2017.

[31] Alexander Chesnokov and V. Liapidevskii. Wave motion of an ideal fluid in a narrow open channel. *Journal of Applied Mechanics and Technical Physics*, 50:220–228, 2009.

[32] Alexander A. Chesnokov and Valery Yu. Liapidevskii. Shallow water equations for shear flows. In *Computational science and high performance computing IV*, volume 115 of *Notes Numer. Fluid Mech. Multidiscip. Des.*, pages 165–179. Springer, Berlin, 2011.

[33] F. Coron and B. Perthame. Numerical passage from kinetic to fluid equations. *SIAM J. Numer. Anal.*, 28(1):26–42, 1991.

[34] W. Craig and C. Sulem. Numerical simulation of gravity waves. *J. Comput. Phys.*, 108(1):73–83, 1993.

[35] W. Craig, C. Sulem, and P. L. Sulem. Nonlinear modulation of gravity waves: A region approach. *Nonlinearity*, 5(2):497–522, 1992.

[36] A. Decoene and J.-F. Gerbeau. Sigma transformation and ALE formulation for three-dimensional free surface flows. *International Journal for Numerical Methods in Fluids*, 59(4):357–386, 2009.

[37] Olivier Delestre, Carine Lucas, Pierre-Antoine Ksinant, Frédéric Darboux, Christian Laguerre, T.-N.-Tuoi Vo, François James, and Stéphane Cordier. SWASHES: a compilation of shallow water analytic solutions for hydraulic and environmental studies. *Int. J. Numer. Methods Fluids*, 72(3):269–300, 2013.

[38] Bernard Di Martino, Chourouk El Hassanieh, Edwige Godlewski, Julien Guillod, and Jacques Sainte-Marie. Hyperbolicity of a semi-Lagrangian formulation of the hydrostatic free-surface Euler system. arXiv:2308.15083, August 2023.

[39] V. Dougalis and D. Mitsotakis. Theory and numerical analysis of boussinesq systems: A review. In N. Kampanis, V. Dougalis, and J. Ekaterinaris, editors, *Effective Computational Methods in Wave Propagation*, pages 63–110. CRC Press, 2008.

[40] V. Dougalis, D. Mitsotakis, and J.-C. Saut. On some Boussinesq systems in two space dimensions: Theory and numerical analysis. *ESAIM: Math. Model. Num. Anal.*, 41:825–854, 2007.

[41] V. Dougalis, D. Mitsotakis, and J.-C. Saut. On initial-boundary value problems for a Boussinesq system of BBM-BBM type in a plane domain. *Discrete Contin. Dyn. Syst*, 23:1191–1204, 2009.

[42] V. Dougalis, D. Mitsotakis, and J.-C. Saut. Boussinesq systems of Bona-Smith type on plane domains: theory and numerical analysis. *J. Sci. Comp.*, 44(2):109–135, 2010.

[43] Vassilios A. Dougalis, Dimitrios E. Mitsotakis, and Jean-Claude Saut. On some Boussinesq systems in two space dimensions: Theory and numerical analysis. *ESAIM, Math. Model. Numer. Anal.*, 41(5):825–854, 2007.

[44] Chourouk El Hassanieh, Mathieu Rigal, and Jacques Sainte-Marie. Implicit kinetic schemes for the Saint-Venant system. hal-04048832, March 2023.

[45] Charles Fefferman. C^m extension by linear operators. *Annals of Mathematic*, 166(3):779–835, 2007.

[46] E. D. Fernandez-Nieto, M. Parisot, Y. Penel, and J. Sainte-Marie. A hierarchy of dispersive layer-averaged approximations of Euler equations for free surface flows. *Communications in Mathematical Sciences*, 16(5):1169–1202, 2018.

[47] S. Ferrari and F. Saleri. A new two-dimensional shallow water model including pressure effects and slow varying bottom topography. *ESAIM: M2AN*, 38(2):211–234, 2004.

[48] J.-F. Gerbeau and B. Perthame. Derivation of Viscous Saint-Venant System for Laminar Shallow Water; Numerical Validation. *Discrete and Continuous Dynamical Systems - B*, 1(1):89–102, 2001.

[49] A.E. Gill. *Atmosphere-Ocean Dynamics*. Number vol. 30 in Atmosphere-ocean Dynamics. Elsevier Science, 1982.

[50] Maurício F. Gobbi, James T. Kirby, and Ge Wei. A fully nonlinear Boussinesq model for surface waves. II: Extension to $O(kh)^4$. *J. Fluid Mech.*, 405:181–210, 2000.

[51] Edwige Godlewski and Pierre-Arnaud Raviart. *Numerical approximation of hyperbolic systems of conservation laws*, volume 118 of *Applied Mathematical Sciences*. Springer New York, 2nd edition edition, 2021.

[52] Laurent Gosse. *Computing qualitatively correct approximations of balance laws. Exponential-fit, well-balanced and asymptotic-preserving*, volume 2 of *SIMAI Springer Ser.* Milano: Springer, 2013.

[53] Laurent Gosse and Alain-Yves LeRoux. A well-balanced scheme designed for inhomogeneous scalar conservation laws. *Comptes Rendus de l'Académie des Sciences. Série I*, 323(5):543–546, 1996. in French.

[54] N. Goutal and J. Sainte-Marie. A kinetic interpretation of the section-averaged Saint-Venant system for natural river hydraulics. *Int. J. Numer. Methods Fluids*, 67(7):914–938, 2011.

[55] A.E. Green and P.M. Naghdi. A derivation of equations for wave propagation in water of variable depth. *J. Fluid Mech.*, 78:237–246, 1976.

[56] James M. Greenberg and Alain-Yves Le Roux. A well-balanced scheme for the numerical processing of source terms in hyperbolic equations. *SIAM Journal on Numerical Analysis*, 33(1):1–16, 1996.

[57] S. Israwi and H. Kalisch. Approximate conservation laws in the KdV equation. *Physics Letters A*, 383(9):854–858, 2019.

[58] S. Israwi and H. Kalisch. A mathematical justification of the momentum density function associated to the KdV equation. *Comptes Rendus. Mathématique*, 359(1):39–45, 2021.

[59] Samer Israwi, Henrik Kalisch, Theodoros Katsaounis, and Dimitrios Mitsotakis. A regularized shallow-water waves system with slip-wall boundary conditions in a basin: theory and numerical analysis. *Nonlinearity*, 35(1):750–786, 2022.

[60] Shi Jin and Lorenzo Pareschi. Asymptotic-preserving (ap) schemes for multiscale kinetic equations: a unified approach. In Heinrich Freistühler and Gerald Warnecke, editors, *Hyperbolic Problems: Theory, Numerics, Applications*, pages 573–582, Basel, 2001. Birkhäuser Basel.

[61] H. Kalisch, Z. Khorsand, and D. Mitsotakis. Mechanical balance laws for fully nonlinear and weakly dispersive water waves. *Physica D: Nonlinear Phenomena*, 333:243–253, 2016.

[62] B. Khorbatly, I. Zaiter, and S. Israwi. Derivation and well-posedness of the extended Green-Naghdi equations for flat bottoms with surface tension. *J. Math. Phys.*, 59(7):071501, 20, 2018.

[63] Bashar Khorbatly, Ralph Lteif, Samer Israwi, and Stéphane Gerbi. Mathematical modeling and numerical analysis for the higher order Boussinesq system. *ESAIM, Math. Model. Numer. Anal.*, 56(2):593–615, 2022.

[64] E. Yu. Knyazeva and A. A. Chesnokov. Stability criterion of shear fluid flow and the hyperbolicity of the long-wave equations. *J. Appl. Mech. Tech. Phys.*, 53(5):657–663, 2012.

[65] Igor Kukavica, Roger Temam, Vlad C. Vicol, and Mohammed Ziane. Local existence and uniqueness for the hydrostatic Euler equations on a bounded domain. *Journal of Differential Equations*, 250(3):1719–1746, 2011.

[66] D. Lannes and P. Bonneton. Derivation of asymptotic two-dimensional time-dependent equations for surface water wave propagation. *Physics of Fluids*, 21:016601, 2009.

[67] David Lannes. *The water waves problem. Mathematical analysis and asymptotics*, volume 188 of *Math. Surv. Monogr.* Providence, RI: American Mathematical Society (AMS), 2013.

[68] Philippe G. LeFloch. *Hyperbolic systems of conservation laws. The theory of classical and nonclassical shock waves.* Basel: Birkhäuser, 2002.

[69] C. D. Levermore and M. Sammartino. A shallow water model with eddy viscosity for basins with varying bottom topography. *Nonlinearity*, 14(6):1493–1515, 2001.

[70] P. L. Lions, B. Perthame, and E. Tadmor. Kinetic formulation of the isentropic gas dynamics and p-systems. *Commun. Math. Phys.*, 163(2):415–431, 1994.

[71] Pierre-Louis Lions, Benoît Perthame, and Panagiotis E. Souganidis. Existence and stability of entropy solutions for the hyperbolic systems of isentropic gas dynamics in Eulerian and Lagrangian coordinates. *Commun. Pure Appl. Math.*, 49(6):599–638, 1996.

[72] P. A. Madsen, H. B. Bingham, and Hua Liu. A new Boussinesq method for fully nonlinear waves from shallow to deep water. *J. Fluid Mech.*, 462:1–30, 2002.

[73] F. Marche. Derivation of a new two-dimensional viscous shallow water model with varying topography, bottom friction and capillary effects. *European Journal of Mechanics - B/Fluids*, 26:49–63, 2007.

[74] Nader Masmoudi and Tak Kwong Wong. On the H^s theory of hydrostatic Euler equations. *Archive for Rational Mechanics and Analysis*, 204(1):231–271, 2012.

[75] Okey Nwogu. Alternative form of boussinesq equations for nearshore wave propagation. *Journal of Waterway, Port, Coastal, and Ocean Engineering*, 119(6):618–638, 1993.

[76] B. Perthame. Boltzmann type schemes for gas dynamics and the entropy property. *SIAM J. Numer. Anal.*, 27(6):1405–1421, 1990.

[77] B. Perthame. Second-order Boltzmann schemes for compressible Euler equations in one and two space dimensions. *SIAM J. Numer. Anal.*, 29(1):1–19, 1992.

[78] B. Perthame and C. Simeoni. A kinetic scheme for the Saint-Venant system with a source term. *Calcolo*, 38(4):201–231, 2001.

[79] Benoit Perthame. An introduction to kinetic schemes for gas dynamics. In *An introduction to recent developments in theory and numerics for conservation laws. Proceedings of the international school, Freiburg/ Littenweiler, Germany, October 20–24, 1997*, pages 1–27. Berlin: Springer, 1999.

[80] Benoît Perthame. *Kinetic formulation of conservation laws*, volume 21 of *Oxf. Lect. Ser. Math. Appl.* Oxford: Oxford University Press, 2002.

[81] Michael Renardy. Ill-posedness of the hydrostatic Euler and Navier–Stokes equations. *Archive for Rational Mechanics and Analysis*, 194(3):877–886, 2009.

[82] Mathieu Rigal. *Low Froude regime and implicit kinetic schemes for the Saint-Venant system.* Phd thesis, Sorbonne Université, November 2022.

[83] F. Serre. Contribution à l'étude des écoulements permanents et variables dans les canaux. *La Houille blanche*, 8:374–872, 1953.

[84] Jack Sherman and Winifred J. Morrison. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *Annals of Mathematical Statistics*, 21(1):124–127, 1950.

[85] John R. Silvester. Determinants of block matrices. *The Mathematical Gazette*, 84(501):460–467, 2000.

[86] V. M. Teshukov. Long waves in an eddying barotropic liquid. *Journal of Applied Mechanics and Technical Physics*, 35(6):823–831, 1994.

[87] V. M. Teshukov. Simple waves on an ambient flow of an ideal incompressible fluid with free boundary. *Journal of Applied Mechanics and Technical Physics*, 38(2):211–218, 1997.

[88] V. M. Teshukov. Spatial stationary long waves in shear flows. *Prikl. Mekh. Tekh. Fiz.*, 45(2):28–39, 2004.

[89] V. M. Teshukov and M. M. Sterkhova. Characteristic properties of the system of equations of a shear flow with nonmonotonic velocity profile. *Journal of Applied Mechanics and Technical Physics*, 36(3):367–372, 1995.

[90] Vladimir Teshukov, Giovanni Russo, and Alexander Chesnokov. Analytical and numerical solutions of the shallow water equations for 2d rotational flows. *Mathematical Models and Methods in Applied Sciences*, 14(10):1451–1479, 2004.

[91] Vladimir Mikhailovich Teshukov. On the hyperbolicity of long wave equations. *Doklady Akademii Nauk SSSR*, 284(3):555–559, 1985. in Russian.

[92] G.B. Whitham. *Linear and nonlinear waves*, volume 42. John Wiley & Sons, New York, 1974.

[93] Yulong Xing and Chi-Wang Shu. A survey of high order schemes for the shallow water equations. *J. Math. Study*, 47(3):221–249, 2014.

[94] V. E. Zakharov. Stability of periodic waves of finite amplitude on the surface of a deep fluid. *Journal of Applied Mechanics and Technical Physics*, 9(2):190–194, 1968.

[95] V. E. Zakharov. Benney equations and quasiclassical approximation in the method of the inverse problem. *Functional Analysis and its Applications*, 14:89–98, 1980.

[96] V. Zeitlin, F. Bouchut, S. Medvedev, G. Reznik, and A. Stegner. *Nonlinear dynamics of rotating shallow water: methods and advances.* Elsevier, 2007.

Analysis and numerical approximation of some mathematical models of free-surface flows

Abstract

This thesis is dedicated to the study of some partial differential equations describing free-surface flows in fluid mechanics and it consists of three interrelated projects. The first project investigates the implementation of numerical schemes for the Saint-Venant system using a kinetic approach, with a primary focus on the one-dimensional case. By adopting an implicit-in-time kinetic approach, this work offers a computational advantage over traditional implicit schemes, since it presents an explicit expression for the inverse of the matrix. The implicit kinetic scheme preserves the positivity of the water height and satisfies an entropy inequality. The second contribution delves into the stability analysis of the hydrostatic Euler equations. A transformation is introduced to rewrite these equations as a generalized quasi-linear system with an integral operator, establishing equivalence under specific conditions. This transformation allows for deeper insights into the spectrum of the matrix operator. Furthermore, we propose an exact multi-layer $\mathbb{P}_0$-discretization, which could be used to solve numerically the transformed system and we analyse its spectrum. The third and final contribution is a work in progress aiming to provide a mathematical justification for the mechanical balance laws of the two-dimensional Boussinesq system. This system is widely used in nearshore zone applications and it is useful to assess its accuracy in terms of fundamental principles such as mass, momentum, and energy conservation. We give estimates to quantify the errors introduced by these approximations, offering valuable insights into the accuracy of the Boussinesq system.

Keywords: Free-surface flows, Euler equations, Numerical approximation, Implicit schemes, Boussinesq equations

---

Analyse et approximation numérique de quelques modèles mathématiques d'écoulements à surface libre

Résumé

Cette thèse est dédiée à l'étude de certaines équations aux dérivées partielles décrivant les écoulements à surface libre en mécanique des fluides et se compose de trois projets interconnectés. Le premier projet étudie l'implémentation de schémas numériques pour le système de Saint-Venant en utilisant une approche cinétique, en se concentrant principalement sur le cas unidimensionnel. En adoptant une approche cinétique implicite en temps, ce travail offre un avantage de calcul par rapport aux schémas implicites traditionnels, puisqu'il présente une expression explicite pour l'inverse de la matrice. Le schéma cinétique implicite préserve la positivité de la hauteur d'eau et satisfait une inégalité d'entropie. La deuxième contribution traite de l'analyse de la stabilité des équations d'Euler hydrostatiques. Une transformation est introduite pour réécrire ces équations comme un système quasi-linéaire généralisé avec un opérateur intégral, établissant l'équivalence dans des conditions spécifiques. Cette transformation permet de mieux comprendre le spectre de l'opérateur matriciel. En outre, nous proposons une discrétisation exacte multicouche de $\mathbb{P}_0$, qui pourrait être utilisée pour résoudre numériquement le système transformé et nous analysons son spectre. La troisième et dernière contribution est un travail en cours visant à fournir une justification mathématique des lois d'équilibre mécanique du système bidimensionnel de Boussinesq. Ce système est largement utilisé dans les applications des zones littorales et il est utile d'évaluer sa précision en termes de principes fondamentaux tels que la conservation de la masse, de la quantité de mouvement et de l'énergie. Nous donnons des estimations pour quantifier les erreurs introduites par ces approximations, offrant ainsi des indications précieuses sur la précision du système de Boussinesq.

Mots clés : Équations à surface libre, Équations d'Euler, Approximation numérique, Schémas implicites, Équations de Boussinesq

---

Laboratoire Jacques-Louis Lions – Sorbonne Université – Campus Pierre et Marie Curie – 4 place Jussieu – 75005 Paris – France
INRIA – 2 rue Simone Iff – 75012 Paris – France
Laboratoire de Mathématiques – Université Libanaise – Campus Rafic Hariri – Hadath – Al Chouf – Liban