Modeling, analysis and simulation of two geophysical flows

Léa Boittin

Sorbonne Université

Inria

Doctoral School **Ecole Doctorale Sciences Mathématiques de Paris Centre**

University Department **Inria Paris**

Thesis defended by **Léa Boittin**

Defended on **5th April, 2019**

In order to become Doctor from Sorbonne Université

Academic Field **Applied Mathematics**

Modeling, analysis and simulation of two geophysical flows

Sediment transport and variable density flows

Thesis supervised by Jacques Sainte-Marie

Committee members

Referees	Luca Formaggia	Professor at Politecnico di Milano
	Raphaèle Herbin	Professor at Université d'Aix-Marseille
Examiners	Claire Chainais-Hillairet	Professor at Université Lille 1
	Pascal Frey	Professor at Sorbonne Université
	Pauline Lafitte	Professor at CentraleSupélec
	Emmanuel Audusse	Associate Professor at Université Paris 13
	Martin Parisot	Junior Researcher at Inria
Supervisor	Jacques Sainte-Marie	Senior Researcher at Inria

Sorbonne Université

Inria

Doctoral School **Ecole Doctorale Sciences Mathématiques de Paris Centre**

University Department **Inria Paris**

Thesis defended by **Léa Boittin**

Defended on **5th April, 2019**

In order to become Doctor from Sorbonne Université

Academic Field **Applied Mathematics**

Modeling, analysis and simulation of two geophysical flows

Sediment transport and variable density flows

Thesis supervised by Jacques Sainte-Marie

Committee members

Referees	Luca Formaggia	Professor at Politecnico di Milano
	Raphaèle Herbin	Professor at Université d'Aix-Marseille
Examiners	Claire Chainais-Hillairet	Professor at Université Lille 1
	Pascal Frey	Professor at Sorbonne Université
	Pauline Lafitte	Professor at CentraleSupélec
	Emmanuel Audusse	Associate Professor at Université Paris 13
	Martin Parisot	Junior Researcher at Inria
Supervisor	Jacques Sainte-Marie	Senior Researcher at Inria

Sorbonne Université

Inria

École doctorale **Ecole Doctorale Sciences Mathématiques de Paris Centre**

Unité de recherche **Inria Paris**

Thèse présentée par **Léa Boittin**

Soutenue le **5 avril 2019**

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

Discipline **Mathématiques appliquées**

Modélisation, analyse et simulation de deux écoulements géophysiques

Transport sédimentaire et écoulement à densité variable

Thèse dirigée par Jacques Sainte-Marie

Composition du jury

Rapporteurs	Luca Formaggia	professeur au Politecnico di Milano
	Raphaèle Herbin	professeure à l'Université d'Aix-Marseille
Examinateurs	Claire Chainais-Hillairet	professeure à l'Université Lille 1
	Pascal Frey	professeur à Sorbonne Université
	Pauline Lafitte	professeure à CentraleSupélec
	Emmanuel Audusse	MCF à l'Université Paris 13
	Martin Parisot	chargé de recherche à l'Inria
Directeur de thèse	Jacques Sainte-Marie	directeur de recherche à l'Inria

Keywords: geophysical flows, sediment transport, non-local flux, variable density flow, multilayer model, numerical simulation

Mots clés : écoulements géophysiques, transport sédimentaire, flux non-local, écoulement à densité variable, modèle multicouches, simulation numérique

This thesis has been prepared at the following research units.

Inria Paris

2 rue Simone Iff
75012 Paris
France

☎ +33 1 80 49 40 00
Web Site http://inria.fr/

Laboratoire Jacques-Louis Lions

4 place Jussieu
75005 Paris
France

☎ +33 1 44 27 42 98
Web Site http://ljll.math.upmc.fr/

Every time I think I know what's going on, suddenly there's another layer of complications. I just want this damn thing solved.

John Scalzi, *The Last Colony*

Modeling, analysis and simulation of two geophysical flows
Sediment transport and variable density flows

Abstract

The present thesis deals with the modeling and numerical simulation of complex geophysical flows. Two processes are studied: sediment transport, and variable density flows. For both flows, the approach is the same. In each case, a reduced vertically-averaged model is derived from the 3D Navier-Stokes equations by making a specific asymptotic analysis. The models verify stability properties. Attention is paid to preserving these properties at the discrete level, in particular the entropy stability. The behavior of both models is illustrated numerically.
Concerning the sediment transport model, the sediment layer is first studied alone. Then, a coupled sediment-water model is presented and simulated. The influence of a viscosity term in the model for the sediment layer is investigated. Due to this viscosity term, the sediment flux is non-local. A transport threshold is added to the model. The water layer is modeled by the Shallow Water equations. Adding some non-locality to the model allows to simulate dune growth and propagation.
In the variable density flow model, the density is a function of one or several tracers such as temperature and salinity. The model derivation consists in removing the dependence of the density on the pressure. A layer-averaged formulation of the model is proposed, which is subsequently used to propose a numerical discretization. The numerical simulations emphasize the differences between this model and a model relying on the classical Boussinesq approximation.

Keywords: geophysical flows, sediment transport, non-local flux, variable density flow, multilayer model, numerical simulation

---

Modélisation, analyse et simulation de deux écoulements géophysiques
Transport sédimentaire et écoulement à densité variable

Résumé

Cette thèse traite de la modélisation et de la simulation numérique d'écoulements géophysiques complexes. Deux types d'écoulements sont étudiés, le transport sédimentaire par charriage et les écoulements à densité variable. La démarche suivie est la même pour les deux phénomènes. Dans chaque cas, un modèle réduit, moyenné suivant la verticale est dérivé à partir des équations de Navier-Stokes 3D en suivant une certaine asymptotique. Les modèles possèdent des propriétés de stabilité. Ces propriétés sont ensuite préservées au niveau discret, en particulier l'inégalité d'entropie.
En ce qui concerne le transport sédimentaire, la couche de sédiments est d'abord traitée seule, puis un modèle couplé pour les sédiments et l'eau est présenté et simulé. L'influence d'un terme de viscosité est étudiée. La présence du terme de viscosité rend le flux sédimentaire non-local. Un seuil pour le transport est introduit dans le modèle. L'eau est modélisée par les équations Shallow Water. L'ajout d'effets non-locaux permet de simuler la croissance et la propagation d'une dune.
Dans le modèle pour les écoulements à densité variable, la densité varie en fonction d'un ou plusieurs traceurs tels que la température et la salinité. La dérivation consiste à enlever la dépendance en pression dans la loi d'état du fluide. Une formulation moyennée suivant la verticale est proposée ; cette formulation est par la suite utilisée pour proposer une discrétisation. Les simulations font ressortir les différences entre le modèle étudié et un modèle classique reposant sur l'approximation de Boussinesq.

Mots clés : écoulements géophysiques, transport sédimentaire, flux non-local, écoulement à densité variable, modèle multicouches, simulation numérique

---

Inria Paris
2 rue Simone Iff – 75012 Paris – France

Contents

List of Figures

Acknowledgements/Remerciements

A 17 ans je voulais arrêter les maths.

A 18 ans je voulais arrêter les maths.

A 19 et 20 ans, pareil. D'ailleurs, j'ai presque réussi à arrêter pour de bon.

A 21 ans, j'ai été prise d'un doute, et vers 22-23 ans, j'ai fermement décidé de me remettre aux maths. Je remercie donc du fond du cœur tous les gens qui me font aimer les maths, la recherche en maths, et grâce à qui une thèse de mathématiques appliquées est une formidable aventure.

Rien n'aurait été possible sans mon directeur de thèse, Jacques, et mes encadrants, Martin et Emmanuel. Pour commencer, merci à vous de m'avoir fait confiance et de m'avoir recrutée. Mais ce n'était que le début. Je vous remercie de m'avoir fait découvrir un autre monde. Vous avez été disponibles, investis et passionnés. Vous m'avez soutenue tout au long de ma thèse, et aussi en ce qui concerne ma vie après la thèse - quand je passais des entretiens, j'étais convaincue que vous étiez moralement avec moi. J'ai la chance d'avoir eu des superviseurs dont j'admire les idées, que ce soit en termes de mathématiques ou de rédaction. Quand on commence une thèse, on fait confiance à ses encadrants par nécessité. Plus j'ai avancé dans ma thèse, et plus j'ai su pourquoi j'avais une grande confiance en mes encadrants. Merci !

Je suis très reconnaissante à Raphaèle Herbin et Luca Formaggia pour leur travail de relecture de ma thèse. J'ai apprécié le fait de recevoir des questions, commentaires et corrections de leur part. Merci à Claire Chainais, Pauline Lafitte et Pascal Frey d'avoir accepté de faire partie de mon jury.

J'exprime toute ma gratitude à l'indispensable Marie-Odile pour sa bonne humeur, son calme, ses très bons conseils et sa relecture d'une précision redoutable, ainsi que pour l'intérêt qu'elle a porté à ce que je vais devenir ensuite.

J'ai eu la chance de collaborer avec François Bouchut, notamment en ce qui concerne la limite bas-Mach pour la dérivation du modèle à densité variable. Ensemble, nous nous sommes acharnés sur la thermodynamique... François a également eu la patience de répondre à mes nombreuses questions. Ce fut très instructif. J'ai aussi bénéficié des commentaires pertinents d'Anne Mangeney.

Ringrazio il professore Luca Bonaventura. Luca, è tutta colpa tua, o tutto merito tuo, non so. Pensavo di smettere per sempre di fare matematica, ma sei venuto a trovarmi nel fondo dell'aula B.1.5 a Lecco. Sappiamo entrambi cos'è successo dopo.

Je remercie infiniment Julien, Eric et Alain du SIC, l'épatant service de support informatique de l'Inria. Ils m'ont épaulée à chacun de mes *quatre* crash informatiques et

m'ont toujours permis de me remettre à travailler en un temps record. (Je précise que les crash en question n'étaient pas de mon fait, ils étaient liés à un défaut de fabrication de ma machine.) Mention spéciale pour Julien qui connaissait ma configuration par cœur dès le troisième crash. Merci beaucoup à Maryse, qui maitrise parfaitement tous les mécanismes administratifs de l'Inria et répond toujours à mes sollicitations en un temps record !

A présent, j'aimerais remercier dans son ensemble la population du troisième étage du bâtiment A, à travers les équipes et à travers les âges. Merci à ceux qui étaient là quand j'ai commencé ma thèse, merci à ceux qui sont là au moment où je la finis, pour tout un tas de bons moments ensemble, pour vos conseils, votre soutien. Merci de faire de l'étage un endroit où il est si agréable d'être doctorant ! Je salue et remercie tout particulièrement mes camarades de bureau, j'ai nommé Fabien W. et Fabien S. Parce que a ça été un plaisir et une chance de faire toute ma thèse en compagnie de Fabien W., parce que j'ai beaucoup aimé travailler avec Fabien S. sur cette fichue thermodynamique, parce que les séjours à Erlangen et à Saint-Malo avec chacun de vous ont été très chouettes (malgré un orage mémorable essuyé à l'arrivée à Saint-Malo) parce que nos discussions du vendredi matin étaient un moment particulièrement sympathique, parce que nous avons parlé de science, de tout et surtout de n'importe quoi ! Merci aussi à l'équipe SERENA au quatrième étage.

Bien sûr, je dis merci à tous les membres de la très chaleureuse équipe ANGE. Quel plaisir d'avoir passé ces trois années avec vous, à l'Inria et en dehors ! Chaque année, participer à EGRIN avec vous a été un un événement.

Je remercie mes parents, qui croient en mois depuis le début, qui m'ont encouragée dans mes différents projets et qui acceptent le fait que j'aie régulièrement besoin de prendre l'air (c'est-à-dire de m'expatrier). Merci beaucoup à Clément qui m'aide à dédramatiser, à Kaki et à Dominique que j'ai pu voir à Paris, à Lecco, à Toulouse et à Bordeaux, et à Christine, qui a été présente pendant toutes ces années.

Merci à la famille Tissot pour, entre autres, de nombreux déjeuners sympathiques le dimanche, un accueil d'urgence très compréhensif (à cause de la crise dite "des souris dans l'appartement") et de très bonnes vacances à Noirmoutier - et merci d'avoir proposé de m'aider pour le pot de thèse, le déménagement, etc...

Je salue mes amis qui ont régulièrement eu le courage de me demander ce que contenait ma thèse.

Pour finir, merci à Olivier pour une infinité de choses. Un programme chargé nous attend : petit-déjeuner à Mayence, brunch à Paris et thé à Londres !

Introduction

Outline of the current chapter

Context

The expression "geophysical flow" refers to a number of natural flows. Rivers, landslides, pyroclastic flows, oceans, tsunamis, ice sheets, debris flows all fall under this category. For civil protection purposes, it is crucial to be able to forecast accurately the occurrence and magnitude of hazardous geophysical flows such as floods, pyroclastic flows and tsunamis. Forecasting floods has even been made mandatory by the European Union. The European Floods Directive (2007) "requires Member States to assess if all water courses and coast lines are at risk from flooding, to map the flood extent and assets and humans at risk in these areas and to take adequate and coordinated measures to reduce this flood risk." In [101], it was reported that floods have been the costliest catastrophe in Europe for the period 1950-2006. Moreover, designing mitigation measures becomes possible only if the phenomena are well characterized. Due to climate change, more and more extreme meteorological events occur, which in turn can trigger hazardous geophysical flows - intense rain can lead to a flash flood and more cyclones will mean more storm surges. Coastal phenomena are especially relevant since more than a half of the world population lives in coastal areas.

Some of the most famous events are the tsunamis of 2004 and 2011. The 2004 Indian Ocean earthquake and tsunami is one of the deadliest natural catastrophes in history, with more than 220,000 fatalities around the Indian Ocean. The 2011 Tohoku earthquake and

tsunami caused numerous fatalities and a very severe nuclear accident. Another famous example of deadly geophysical flow is the Armero tragedy (Colombia, 1985), in which 22,000 people were killed by lahars.

Europe is also affected. Recent events involving geophysical flows are for instance the floods of late May and early June 2013 in Central Europe, which caused 25 fatalities and 3.1 billion dollars insured losses, as reported in [104]. One can also mention the Rigopiano avalanche: on January 18, 2017, a large avalanche hit a hotel in Pescara (Italy), and caused 29 fatalities. In Norway, several landslides are expected to happen in the region of the mountain Mannen and the inhabitants regularly evacuate due to the threat. A first event can even trigger a second hazardous event. On April 7, 1934, a rockslide in the Tafjorden (a fjord near the village of Tafjord) created a tsunami which killed 40 people as it propagated in the fjord. Experts believe that a similar event can happen in the fjord Geiranger. A landslide could create a tsunami which would destroy several villages while advancing in the fjord.

Finally, events are recorded in France too. In June 2016, Paris and several surrounding cities were flooded for several days. The Aude département faced floods in October 2018. The damage is estimated to 220M€ and 15 people died. It had already been affected by floods in 1999.

Early warning systems play a crucial role when it comes to saving human lives. Assessing beforehand the consequences of potential events is necessary as well. In both cases, reliable and affordable numerical simulations can help tremendously.

However, the forecast of hazardous events is not the only motivation to study geophysical flows. As the preservation of the environment is a growing concern, the study of water flows is particularly relevant. For instance, one may wish to know how a hydraulic structure built on a river affects its flow and whether the morphology of the river as well as its chemical properties (e.g. turbidity) are affected. Studying geophysical flows can help to decide how to manage water resources. It is also a support activity for spatial planning. Geophysical flows are of interest to governments, NGOs, civil engineering companies, electricity producers, but also, when it comes to natural risks, insurance and reinsurance companies.

Geophysical flows are commonly described by shallow-flow models, that is to say, models which take into account the fact that the horizontal length of the flow is much bigger than its depth. Such models present a reduced complexity with respect to fully 3D models, which means that their numerical resolution is faster. One of their advantages is that they need only a fixed horizontal mesh - dealing with moving meshes is not necessary. These models can also be used to describe shallow flows which are not strictly speaking geophysical flows, for instance flows in pipes.

In shallow flow models, the horizontal velocity is approximated by a vertically constant velocity. But in several situations (large water depth, strong wind, high friction at the bottom), this approximation is not adapted. To overcome this limitation and describe fully 3D flows, one can resort to multilayer models. Multilayer models do not rely on the shallow flow approximation and necessitate only a 1D or 2D fixed mesh.

Much effort is now being dedicated by the scientific community to refining the current

models, and to improving their numerical resolution. It is then the role of mathematicians to propose models with good properties and robust numerical schemes.

In this context, the present thesis deals with two different geophysical flows. The first part of this work tackles the topic of sediment transport, for which a shallow flow model is used. In the second part, variable density flows are studied and described with a multilayer model. In what follows, further motivation for the study of sediment transport and variable density flows is given.

Sediment transport

Sediment transport occurs at very different time and space scales. Immersed sediment transport leads to the formation of ripples on the sand of the sea bed as the water advances and retreats. On geological time scales, sediment transport results in important modifications of the landscape, such as the creation of river braids, meanders and oxbow lakes. On medium space and time scales, phenomena such as the silting of dams and harbors as well as scour around bridge piles occur, which can lead to catastrophes and/or require costly mitigation measures. In 1987, the Schoharie Creek Bridge collapsed due to scour. In [90], the authors analyzed 36 historical cases of bridge failures and concluded that 64% of them were due to local scour. In [9], the case of the Loire river estuary is described. A channel was incised in the river bed so that large ships can reach the harbor of Nantes/Saint-Nazaire. At the point where the channel begins, the water depth increases suddenly, so that the water velocity decreases. Therefore, the sediments are deposited in the channel. Dredging the channel is then necessary. About $10^7 \mathrm{m}^3$ sediment are dredged each year in the Loire estuary, and dredging is a costly procedure. During floods, predicting correctly the bed aggradation is essential to be able to predict the water stage. For instance, in 1987, bed aggradation depths of 2m to 5m were measured in the in-town reach of the river Mallero, that is to say, the part of the Mallero that flows through the centre of Sondrio [110]. The peak discharge of the flash flood of 1987 was not catastrophic; it is because of sediment transport and bed aggradation that the city of Sondrio was indeed flooded. Moreover, the European Floods Directive of 2007 mentions the relevance of sediment transport for hazard assessment.

The present thesis is mainly concerned with sediment transport occurring under the action of water, but of course, aeolian sediment transport exists as well and is responsible, for instance, for the formation of dunes in deserts.

There are many challenges in sediment transport modeling and simulation. Of course, many difficulties arise from the variability of the parameters in a natural river. For instance, in a natural river reach, the sediment mixture is not homogeneous. But existing models may even fail to reproduce experiments performed in perfectly controlled laboratory conditions. Dune growth is one of the phenomena one wishes to capture. Avenues of research include an improved description of the water flow (with respect to the commonly used Shallow Water equations) and/or an improved description of the sediment flow. The models proposed should have good mathematical properties, and a stable

numerical solver should be designed.

Variable density flows

Stratified flows, i.e. flows which exhibit density variations in the vertical direction, are very frequent in nature. Oceans and lakes are examples of stratified fluids. The density of the water is influenced by its temperature, the presence of dissolved chemical species (such as salt in the ocean), and to a lesser extent, the pressure. Lakes are stratified due to temperature variations. In the summer, the surface of the lake is heated by the sun, therefore the water near the surface is warmer and lighter than the water at the bottom of the lake. A thermocline (a thin layer of water within which the temperature varies rapidly with depth) separates the surface water and the deep water. As winter approaches, the surface water becomes cooler and thereby denser; until the moment where it actually becomes denser than the deep water. Overturning occurs: the surface water sinks at the bottom of the lake. Stably stratified waters don't mix. As a result, the oxygen at the bottom is not renewed until overturning occurs. A famous case in France is that of the lagoon of Berre (*étang de Berre*). This small inland sea in the South of France communicates with the Mediterranean sea and originally received fresh water from three rivers only. In 1966, the lagoon started receiving a large volume of fresh water from the channel of an electricity plant. Thus, fresh water rested on top of salted water coming from the sea and the lagoon became stratified and eutrophic. The water supply from the artificial channel had to be reduced, and the lake is now consistently monitored. A 3D Boussinesq model is used to understand the functioning and forecast the evolution of the lagoon by the public authority in charge of the lagoon [128]. Density stratification processes are relevant for the modeling of natural aquatic systems because they define important factors for life.

Density variations induce gravity currents. In the lab, one can create a gravity current by putting next to each other in a box two fluids with different densities. This is called a lock-exchange experiment. The thermohaline circulation is one of many gravity currents found in nature - it is the large scale ocean circulation driven by density variations due to differences in salinity and temperature.

The existing models for variable density flows commonly rely on the Boussinesq approximation. They give good results in the case of small density variations. But such models do not conserve the water mass. They do not allow to simulate water expansion or contraction, while it is an important phenomenon, induced for instance by seasonal temperature variation. Again, proposing an improved water flow model endowed with good mathematical properties is a desirable goal. Due to the fully 3D nature of the phenomena, 3D models are needed. Proposing a computationally affordable model is one of the challenges scientists are confronting themselves with.

Contributions

The present thesis is concerned with the simulation of geophysical flows. Yet the aim of this work is not to provide realistic simulations of hazardous events such as those described above. We tackle here upstream problems, namely the derivation of new models and numerical schemes. Realistic test cases are out of the scope of this work.

The manuscript is divided in two parts. The first part is concerned with the modeling and simulation of sediment transport, while the second part deals with variable density flows.

The approach is the same in the two parts. In both cases, the objective is to describe and simulate a complex geophysical flow. The starting point is a 3D model, which we want to simplify. A thin-layer approximation is made - in the case of the variable density flows, the shallow approximation is not made, but the resulting model presents similarities with shallow flow models, because it resembles a superposition of shallow flow models. Moreover, asymptotics are made to investigate specific regimes. Simplified models are obtained. They are endowed with a dissipative energy balance and they preserve the positivity of the sediment depth/of the water depth. Finite-volume schemes are proposed. The schemes preserve the stability properties of the models at the discrete level.

Modeling and simulation of sediment transport

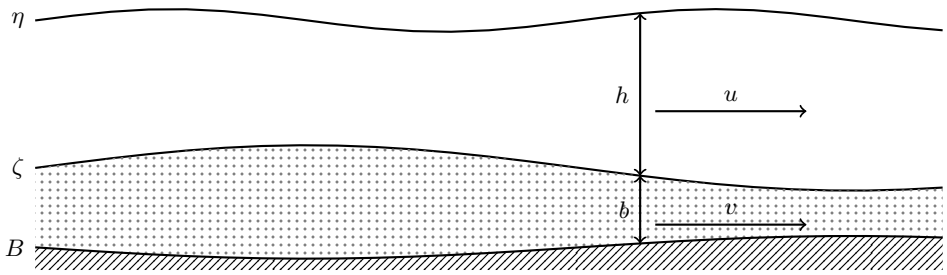


Figure 1 – Sediment and water layers

The system to be described is a stratified sediment-water system, see figure 1. Our main result consists in establishing a new model for bed load transport, taking into account the sediment viscosity, and in proposing a numerical scheme to simulate this model. A layer of water flows over a layer of movable sediment, itself flowing above a non-erodible substratum (the bedrock) $B(\mathbf{x})$. The interface between the sediment and water layers is $\zeta(\mathbf{x}, t)$ and the free surface is $\eta(\mathbf{x}, t)$. The sediment layer is considered as a homogeneous medium, i.e. it is seen as a kind of homogeneous mud, not a porous medium in which the water flows. The water and sediment layers are not miscible. The density of the water layer is ρ_w, that of the sediment layer is ρ_s. The ratio of the densities is $r = \frac{\rho_w}{\rho_s}$. The only sediment transport process modeled here is bed load transport. The system is considered to be shallow, and the time scales of interest range from medium to

long. To begin with, only the sediment layer is modeled. The water layer is taken into account only by means of a pressure p_ζ and a velocity u_ζ at the interface ζ. In classical bed load transport models, the solid flux is a closure relationship and it depends on some characteristics of the sediment (such as density, grain size) and on variables describing the flow in the water layer, the velocity for instance. Such models do not describe the mechanical behavior of the sediment layer. The present work aims, among other things, at proposing a model in which the rheology of the sediment layer is taken into account. Starting from the incompressible Navier-Stokes equations, several models are derived. The shallow flow approximation is made. Additionally, assumptions on the scalings of the friction coefficients and of the viscosity are made. Depending on the scaling of the parameters, several models can be obtained. Some of the models obtained already exist, such as the models in [131], [58] and the Exner model with a Grass law [68]. The model we are most interested in is new and reads

$$\begin{cases} \partial_t b + \nabla_{\mathbf{x}} \cdot (bv) = 0, \\ f_B(b, v, \tau) = \tau(b, v), \end{cases} \tag{1}$$

with $\tau(b, v)$ the shear stress

$$\tau\left(b, v\right) = -r\kappa_\zeta\left(v - u_\zeta\right) - b\nabla_{\mathbf{x}}\left(g\left(b + B\right) + \frac{p_\zeta}{\rho_s}\right) + \nabla_{\mathbf{x}} \cdot \left(2\mu_s b \mathrm{D}_{\mathbf{x}} v\right). \tag{2}$$

The force $f_B(b, v, \tau)$ is the effort due to the friction between the sediment layer and the substratum, its expression is

$$f_B\left(b, v, \tau\right) = \begin{cases} \left(\tau_c + \kappa_B \left\|v\right\|^\gamma\right) \dfrac{v}{\|v\|} & \text{if } \|\tau\| > \tau_c \\ \tau + v & \text{if } \|\tau\| \leq \tau_c \end{cases}$$

The magnitude of the critical shear stress is denoted by $\tau_c = \overline{\tau}\left(gb + \frac{p_\zeta}{\rho_s}\right)$. The parameter $\overline{\tau}(\mathbf{x})$ is the Coulomb coefficient. Throughout the manuscript, $\mathrm{D}_{\mathbf{x}}$ is the symmetric gradient $\mathrm{D}_{\mathbf{x}} = \frac{1}{2}(\nabla_{\mathbf{x}} + (\nabla_{\mathbf{x}})^t$. The unknowns are the thickness of the sediment layer b and the velocity in the sediment layer v. The parameters are the forcing velocity u_ζ and pressure p_ζ, the bottom friction $\kappa_B(x)$, the friction at the water-sediment interface κ_ζ, which can be any function of the forcing pressure p_ζ and the forcing velocity u_ζ, and the viscosity of the sediment layer μ_s. The first equation means that the mass of the sediment layer is conserved. The second equation is a force balance derived from a conservation of momentum equation. During the derivation, it emerges that the inertial terms are small compared to the leading order terms, so they are neglected and do not appear in the momentum balance. Thus the second equation simply means that the bottom friction is balanced by the friction at the interface, the pressure gradient and the viscous efforts. The definition of the bottom friction f_B imposes that there is a transport threshold. Indeed, if the value of the shear stress $||\tau||$ exerted on the sediment layer is below the magnitude of the critical shear stress τ_c, the solution of the momentum con-

servation equation is $v = 0$. Such a threshold allows to obtain non-flat stationary states. Naturally, if one chooses $\overline{\tau} = 0$, there is no threshold.

From the second equation of (1), one can get the value of the sediment velocity v. Assume for an instant that $\mu_s = 0$. Obtaining the value of v is then straightforward, and v depends only on *local* quantities. Now, take again $\mu_s > 0$. The viscosity term being a *nonlocal* term, v has a nonlocal expression, and then, in model (1), the sediment discharge bv is also nonlocal. Other nonlocal models for sediment transport exist, for instance [43], [35] (a brief description of these models can be found in section 1.3). Yet these models do not include a viscosity term; nonlocality is achieved by other means.

For smooth enough solutions, the model satisfies a dissipative balance for the mechanical energy. Moreover, assuming that the initial sediment thickness is positive, i.e. $b(\mathbf{x}, 0) \geqslant 0$, it can be proved that the sediment thickness remains positive as long as the velocity v remains bounded and b remains continuous.

A numerical scheme is proposed to solve (1). The scheme is designed in 1D only. A finite-volume, staggered-grid discretization is adopted. The sediment thickness b is discretized at the cell centres while the velocity v is discretized at the interfaces. The non-local model contains a "diffusion part": it is the gradient in (2). Indeed, for $\mu_s = 0$, the model (1) reduces to an advection-diffusion equation. At least the diffusion part of this type of equation should be discretized thanks to an implicit scheme (in the sense that the gradient of b is implicit), so that the computation is stable without imposing a restrictive (parabolic) condition on the time step. Because of this property of the limit model obtained when $\mu_s = 0$, an implicit discretization is used for the gradient of b in (2). The transport threshold is implemented. The velocity is obtained first, by solving a non-linear system. Then the sediment depth is updated. The proposed numerical schemes verify a dissipative balance for the discrete energy and the positivity of b is ensured at the discrete level.

A numerical scheme for the local model is proposed as well. Though it shares many features with the scheme for the non-local model, a different resolution strategy is adopted. When $\mu_s = 0$, it is possible to solve a system directly on the sediment depth b^{n+1} instead of v^{n+1}. This scheme is formally equivalent to the scheme for the non-local model when μ_s goes to zero, which shows the asymptotic-preserving property of the scheme for the non-local model.

In the test cases, the influence of the viscosity on the behavior of the solutions is evidenced. The possibility of obtaining non-flat stationary states due to the presence of a transport threshold is exhibited. A convergence test is performed.

Then, a coupled model for the water layer and the sediment layer is presented. The water layer is modeled by the Shallow Water equations. The coupled model satisfies a continuous energy balance. For the water layer, the staggered finite-volume scheme presented in [73] is adopted. The coupling strategy of the schemes for the two layers is proved not to create discrete entropy. Simulations of water flowing on a sediment dune are performed. Adding a viscosity term to the model allows to simulate dune growth and

propagation, whereas this is not possible with classical Shallow Water-Exner models.

The purpose of this chapter is not to provide another bed load transport formula. It does not either attempt to be competitive when it comes to matching experimental results - a comparison with experiments is beyond the scope of this work. The model should probably be enriched and complexified before a comparison with experiments is meaningful. The point here is to investigate the influence of a viscosity term in the solid flux - because of the viscosity term, the solid flux becomes non-local. Instead of deducing a formula from experiments, a model based on reasonable physics assumptions is derived. A numerical scheme is proposed, and the behavior of the model is illustrated numerically.

The second chapter of this part will be submitted as an article along with E. Audusse and M. Parisot under the title "On the Exner model and non-local approximations: modeling, analysis and numerical simulations."

The author of the present thesis was awarded a "Best PhD Student Poster Award" for this work at the conference CMWR XXII (http://cmwrconference.org) held in June 2018 in Saint-Malo, France.

Modeling and simulation of variable density flows

The system of interest is a water flow with free surface over variable topography. The water density ρ depends on the pressure and on one or two tracers (temperature T only or temperature T and salinity S). For the simulation of variable density flows in which the density variations are small, one typically resorts to the Boussinesq approximation. The density variations are neglected in the equations describing the flow except in the pressure term. Under this approximation, the water mass is not conserved. A 2D model which does not rely on this approximation was presented in [21]. The present work mainly aims at proposing and simulating a 3D model more rigorously derived and closer to physics than the model in [21].

First, the compressible Navier-Stokes equations are considered. The incompressible limit is performed so that the pressure dependence is removed from the state equation of the water. A low-Mach approximation of the Navier-Stokes equations is obtained. Two situations are considered: a situation where the temperature is the only tracer influencing the density $\rho = \rho(T^{eq})$ and another situation, in which $\rho = \rho(T^{eq}, S)$, which means that the effects from both the temperature and the salinity are taken into account. The temperature T^{eq} is the temperature in the incompressible limit. The models obtained for these two simulations are essentially the same model and the modeling approach can be extended to any number of tracers. The model without salinity reads

$$\begin{cases} \nabla.\mathbf{U} = -\frac{\rho'(T^{eq})}{\rho^2 c_p}(\nabla \cdot (\lambda \nabla T^{eq}) + \sigma : D(U)), \\ \frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{U}) = 0, \\ \frac{\partial \rho \mathbf{U}}{\partial t} + \nabla \cdot (\rho \mathbf{U} \otimes \mathbf{U}) + \nabla p = \rho \mathbf{g} + \nabla \cdot \sigma, \end{cases} \tag{3}$$

where $\mathbf{U}(t,x,y,z)=(u,v,w)^T$, σ is the deviatoric stress tensor, λ is the heat conductivity and c_p is the specific heat capacity at constant pressure. The fluid pressure is p and $\mathbf{g}=(0,0,-g)^T$ represents the gravity forces. Thermodynamic considerations are taken into account in the derivation so that the obtained models respect the second principle of thermodynamics. The energy equations of the incompressible models are close in spirit to that of the compressible Navier-Stokes system. An important feature of these models is that they do not use the Boussinesq approximation. They conserve the mass, and not the volume. Temperature variations induce expansion or contraction of the water body. The hydrostatic assumption is then made.

Next, layer-averaged models are proposed. In a first step, the Euler-Fourier model is vertically integrated. We denote by "Euler-Fourier" a model in which the density depends on a single tracer, typically the temperature T, and in which the viscous effects as well as the diffusion terms for the temperature are neglected. The layer-averaging procedure is performed as described in [38]. The layer-averaged Euler-Fourier model reads

$$
\begin{aligned}
&\frac{\partial h}{\partial t}+\sum_{\alpha=1}^{N}\nabla_{x,y}.(h_\alpha\mathbf{u}_\alpha)=0,\\
&\frac{\partial \rho_\alpha h_\alpha}{\partial t}+\sum_{\alpha=1}^{N}\nabla_{x,y}.(\rho_\alpha h_\alpha\mathbf{u}_\alpha)=\rho_{\alpha+1/2}G_{\alpha+1/2}-\rho_{\alpha-1/2}G_{\alpha-1/2},\qquad \alpha=1,...,N,\\
&\frac{\partial \rho_\alpha h_\alpha\mathbf{u}_\alpha}{\partial t}+\nabla_{x,y}.\left(\rho_\alpha h_\alpha\mathbf{u}_\alpha\otimes\mathbf{u}_\alpha\right)+\nabla_{x,y}\big(h_\alpha p_\alpha\big)=p_{\alpha+1/2}\nabla_{x,y}z_{\alpha+1/2}-p_{\alpha-1/2}\nabla_{x,y}z_{\alpha-1/2}\\
&\qquad\qquad +\mathbf{u}_{\alpha+1/2}\rho_{\alpha+1/2}G_{\alpha+1/2}-\mathbf{u}_{\alpha-1/2}\rho_{\alpha-1/2}G_{\alpha-1/2},\qquad \alpha=1,...,N,
\end{aligned}
$$

where the pressure terms p_α, $p_{\alpha+1/2}$ are given by

$$
p_\alpha=g\left(\frac{\rho_\alpha h_\alpha}{2}+\sum_{j=\alpha+1}^{N}\rho_j h_j\right)\qquad\text{and}\qquad p_{\alpha+1/2}=g\sum_{j=\alpha+1}^{N}\rho_j h_j.
$$

The quantity $G_{\alpha+1/2}$ (resp. $G_{\alpha-1/2}$) corresponds to mass exchange accross the interface $z_{\alpha+1/2}$ (resp. $z_{\alpha-1/2}$) and $G_{\alpha+1/2}$ is defined by

$$
G_{\alpha+1/2}\;=\;\sum_{j=1}^{\alpha}\left(\frac{\partial h_j}{\partial t}+\nabla_{x,y}.(h_j\mathbf{u}_j)\right)=-\sum_{j=1}^{N}\Big(\sum_{p=1}^{\alpha}l_p-\mathbb{1}_{j\leqslant\alpha}\Big)\nabla_{x,y}.(h_j\mathbf{u}_j),
$$

for $\alpha=1,\dots,N$. The velocities $\mathbf{u}_{\alpha+1/2}$ and the densities $\rho_{\alpha+1/2}$ at the interfaces are defined by

$$
v_{\alpha+1/2}=\begin{cases} v_\alpha & \text{if } G_{\alpha+1/2}\leqslant 0\\ v_{\alpha+1} & \text{if } G_{\alpha+1/2}>0\end{cases}
$$

for $v=\mathbf{u},\rho$. This model admits an energy balance with a third order rest term. The equilibria of the layer-averaged Euler-Fourier system are investigated in the case of an almost flat topography. They are found to be similar to those of the Euler system. In

particular, for the equilibrium to be stable, the fluid must be well-stratified, meaning that $\partial_z \rho|_{z_\alpha} < 0$ for $\alpha = 1, \dots, N$. In a second step, the Navier-Stokes-Fourier system is integrated. A simplified form of the rheology terms is used as in [6].

A finite-volume numerical scheme is then proposed for the layer-averaged Euler-Fourier system. The hydrostatic reconstruction technique [15] is used. It is shown that with any flux that is consistent with the semi-discrete in time Euler system and preserves the positivity of the water depth, the resulting scheme is well-balanced and preserves the non-negativity of the water depth. A maximum principle on the density is satisfied. In order to prove an in-cell entropy inequality in the case of a flat topography, a kinetic flux is adopted - the kinetic flux was already used in the context of the Shallow Water equations in [21]. The unsigned terms in the discrete entropy balance are third-order terms. A numerical scheme for the Navier-Stokes-Fourier system is proposed as well. The Euler part of the system is discretized as already done for the Euler-Fourier system, only the discretization of the diffusion terms must be specified. The viscosity terms in the momentum equation are discretized as in [6], that is to say, a classical $\mathbb{P}_1$ finite element type approximation with mass lumping is used. The temperature diffusion terms are discretized in the same manner. Though the schemes are written in the case where the temperature is the only tracer, they can easily be extended to the case where salinity is included as well.

The numerical scheme is validated. A convergence test towards the analytical solution proposed in [39] is made. A lock exchange simulation is made and compared to the experimental results given in [3]. A simple diffusion case is performed as well, for which an analytical solution is available under the Boussinesq assumption. The behavior of the Navier-Stokes-Fourier model is compared to that of the Boussinesq model in another diffusion test case, in which it is shown that starting from a well-stratified fluid and in the absence of exterior forcing, the two models do not reach the same thermal equilibrium.

The numerical scheme was implemented in the code Freshkiss3D [120], which has been developed by the ANGE team at Inria for several years (among others, E. Audusse, M.-O. Bristeau, D. Froger, J. Sainte-Marie and F. Souillé have been involved in the development).

Two articles will be submitted along with M.-O. Bristeau, F. Bouchut, A. Mangeney, J. Sainte-Marie and F. Souillé:

- *The Navier-Stokes system with temperature and salinity for free surface flows - Part I: Low-Mach approximation and layer-averaged formulation*
- *The Navier-Stokes system with temperature and salinity for free surface flows - Part II: Numerical scheme and validation*

The author of the present thesis was awarded a "Prix EGRIN 2018" (for young scientists) at the conference EGRIN 2018 (https://indico.math.cnrs.fr/event/3345/)

held in June 2018 in Le Lioran, France, for a presentation on this topic.

Outline of the conclusion and perspectives

Modeling and simulation of sediment transport

A new sediment transport model is derived from the incompressible Navier-Stokes equations. Its novelty lies in the introduction of a viscosity term in the solid flux, which implies that the solid flux is non-local. A transport threshold is included in the model. The analysis and numerical resolution of this model are challenging. A finite-volume scheme is proposed and a discrete dissipative energy balance is proved. The scheme is validated numerically and the influence of the viscosity term is shown.

A model for the complete sediment-water system is presented. The model for the sediment layer is coupled to the Shallow Water equations. A numerical strategy for the resolution of the coupled system is designed; an existing scheme is used for the water layer. Numerical simulations illustrate the behavior of the coupled system. Notably, the addition of viscosity allowed to obtain dune growth and propagation.

This work raises several questions, which are listed below.

- *Analysis of the continuous sediment model.* The regularity of the solutions of the model for the sediment layer (1) has not yet been investigated, while smooth enough solutions are necessary to obtain the positivity of the sediment depths. The analysis of the model (1) is difficult because this model is non-linear, and because no explicit formula is available for the inverse of the operator giving the velocity.

- *Stability of the numerical scheme.* The numerical experiments performed with several schemes show that the dissipation of the discrete energy was not enough to guarantee a stable computation.

- *Co-located finite volume scheme.* The choice is made to use here a staggered-grid finite volume approach. Proposing a numerical scheme relying on a co-located finite volume approach would be interesting. As there exist more solvers for the Shallow Water equations on a co-located approach, a larger choice would be available for the scheme in the water layer.

- *Threshold implementation.* The proposed numerical implementation of the threshold is not optimal. The fixed-point algorithm required many iterations to converge. A plausible explanation is that at each iteration, a new neighboring cell was set into motion. The higher the number of cells, the longer the process. A different implementation could certainly improve the performances of the algorithm. Inspiration can be taken from the literature on the simulation of contacts with a Coulomb friction law, or on the simulation of flows of viscoplastic materials [2]. Note that the main goal of this work is to study the viscosity term introduced in the model for the sediment layer, which is why we did not focus on the implementation of the transport threshold.

- *Non-newtonian rheology.* The sediment layer has been modeled as a Newtonian fluid. In practice, sediment is an immersed granular medium, for which a viscoplastic description would be more suitable (note that mud is also a viscoplastic fluid). Capturing the transition between the solid state and the liquid state is a first challenge. A possible description is the $\mu(I)$ rheology (see [9] for a description of the $\mu(I)$ rheology and the influence of the interstitial liquid on the rheology).

- *Thickness of the moveable sediment layer.* In the proposed model, the sediment layer moves "by slices": when motion occurs, a whole slice of sediment layer moves as a block, which means that all the available sediment is involved in the flow. Using an improved rheological description could give the depth of the sediment which indeed moves. In [93], the position of the solid/fluid interface in a viscoplastic flow is computed. The Drucker-Prager rheology is used.

- *Improved water flow description.* The present work focuses on the improvement of the description of the sediment layer. As mentioned in [61, 87], much better results can be obtained with respect to the classical Shallow Water-Exner model by using a more refined flow description. The modifications of the water flow induced by the evolution of the sediment layer should be better accounted for. One such flow description is offered by the Triple Deck approach [102, 105, 132]. Another possible description is provided by the multilayer approach [22].

Modeling and simulation of variable density flows

A 3D model for variable density flows is derived from the compressible Navier-Stokes equations. In a first step, the temperature variation is the only factor that influences the density. Then, density variations due to differences in temperature and salinity are considered. The model is obtained by performing the incompressible limit of the compressible Navier-Stokes equations: the fluid density depends on one or several tracers, but the dependence on the pressure is removed. This model is more rigorously derived than the model in [21]. It does not rely on the Boussinesq approximation and it is mass-conservative.

A layer-averaged formulation of the model is proposed. The stable equilibria of the Euler-Fourier model (where only the temperature is taken into account and where the heat diffusion and viscosity are neglected) are investigated. The layer-averaged models satisfy a dissipative energy balance.

Using the layer-averaged formulation, a finite-volume scheme is designed for the Euler-Fourier model. This scheme has several stability properties. The difficulty in proposing a scheme lies in the discretization of the non-conservative pressure terms. The scheme is validated in several test cases and the difference with the Boussinesq model is shown.

Several ideas for future work are listed below.

- *Other fluids and non-Newtonian rheology.* In this work, the water is modeled as a Newtonian fluid. One may wish to use the model proposed for another fluid for which a non-Newtonian description would be better suited. The modeling work

and numerical discretization are then more challenging. Non-Newtonian rheology descriptions include among others the Bingham model and the Herschel-Bulkley model.

- *Viscous dissipation and heat transfer.* Viscous dissipation affects the temperature distribution [26, 27]. In the proposed model, as both the viscous dissipation and temperature fluxes are modeled, we expect to be able to observe modifications of the temperature distribution due to shear stresses. As the viscosity of the water is low, this effect will be very small; it is more relevant in the case of fluids with high viscosity and low thermal conductivity.
- *Propagation of internal waves in a stratified ocean.* Internal waves are gravity waves that exist inside a stratified fluid. They occur when some fluid has been moved away from its equilibrium position. As equilibrium is restored, the fluid oscillates back and forth. For instance, internal waves can be generated by upwelling. In a lake, internal wave modeling is required to understand mixing and transport phenomena [76]. The proposed model and numerical scheme could be used to simulate internal waves.
- *Discrete entropy inequality with topography.* The discrete entropy inequality is stated in the case of a flat topography only. More work is needed to obtain a similar result with a variable topography. Note that in [17], in the case of the shallow water equations, a discrete entropy with an error term was obtained for the classical kinetic solver with the hydrostatic reconstruction. This error term is in the square of the topography jumps and tends to zero strongly as the space step tends to zero.

Common perspectives

We give below some perspectives related to both parts of this thesis.

- *Mass exchanges between the water and sediment layers.* In the proposed coupled model, the water and the sediment layers are non-miscible. In practice, some sediment (especially in the case of fine grains) can detach from the surface of the sediment layer and be transported as suspended load in the water layer. When the transport capacity of the water flow decreases, theses grains are deposited again on the sediment layer. Adding erosion and deposition terms to the model is a desirable goal. But more closure relations are needed for the erosion and deposition terms, or more modeling work should be done to define them.
- *Simulation of suspended load.* The proposed model could be coupled to a sediment transport model with erosion-deposition terms. Thus, a coupled sediment-water model including suspended load could be obtained, with the possibility to model high concentrations. Of course, if locally high concentrations of suspended load are to be modeled, the rheological properties of the water will be modified, and this must be accounted for too.

Part I

Sediment transport

1

A short review of existing models based on the Shallow Water equations

Outline of the current chapter

Multiple models for sediment transport exist. In [53], they are classified in several categories. Reduced Complexity Models can be used to explore general behaviors and investigate pattern formation, typically at very large scales. They are based on a discrete grid of cells in a 2D-plane, and generally include mass conservations equations for the sediment and the water and simplified descriptions of the sediment motion and of the water flow. Diffusion Models are used to investigate large-scale landscape evolution. For the simulation of rivers, models based on the Shallow Water equations coupled with the Exner equation are frequently used. They are adapted for time scales ranging from months to years. To simulate longer time scales, the unsteady and inertial terms are sometimes neglected. More details about this category of models is given below. The model developed in this chapter belongs to this category. For increased accuracy, one should move on to 3D modeling, and to turbulence modeling. In increasing order of accu-

racy, the last categories of models are the models using the unsteady Reynolds-averaged Navier-Stokes equations, Large-Eddy Simulation Models and Direct Numerical Simulations of the Navier-Stokes equations. Though more accurate, these models also have a much larger computational cost. They are therefore not adapted to the simulation of large-scale problems.

Sediment transport occurs in different modes. Bed load transport is what happens at the surface of the river bed: the grains move by saltation and reptation. The other sediment transport mode is the suspended load. Bed load transport occurs when the gravity forces acting on a grain are large enough to confine transport close to the surface of the river bed, while suspended load is observed when the hydrodynamic forces dominate [9]. As mentioned in [88], bed load transport is responsible for bank erosion, bed forms (dunes, ripples) and the rate at which a river incises relief. In other words, bed load transport has a large impact on the morphology of rivers. The present thesis is concerned with bed load only.

1.1 The Shallow Water equations

In this thesis, the water layer is modeled by the Shallow Water equations, also called Saint-Venant equations. The one-dimensional version of the Shallow Water equations was introduced by Adhémar Jean Claude Barré de Saint-Venant in 1871 [115]. The Shallow Water Equations are a system of hyperbolic partial differential equations describing "shallow" free surface flows. The word "shallow" means here that the longitudinal length scale of the phenomenon is much larger than the water depth. These equations are the model used in industrial software like HEC-RAS [121], MIKE HYDRO River [123], TELEMAC-MASCARET [127]. The two-dimensional formulation of the Shallow Water equations is

$$\partial_t h + \nabla_{\mathbf{x}} \cdot (h\mathbf{u}) = 0, \tag{1.1}$$

$$\partial_t (hu) + \nabla_{\mathbf{x}} . (h\mathbf{u} \otimes \mathbf{u} + g\frac{h^2}{2}\mathbf{Id}) = -gh\nabla z_b - S_f, \tag{1.2}$$

where h is the water depth, $\mathbf{u} = (u, v)^T$ the horizontal water velocity, z_b is the bottom topography and S_f is a friction term. The Shallow Water equations can be derived from the incompressible Navier-Stokes equations, which are widely used in fluid mechanics. The full derivation is performed in [63]. Since a similar technique is used later in the manuscript to derive the proposed model, we give a few details here. To get the 2D Shallow Water equations, a dimensionless Navier-Stokes system is used and scaling assumptions on the parameters are made. Let us introduce the characteristic horizontal length L and the characteristic height H. Under the shallow water assumption, we assume that the parameter ε defined by

$$\varepsilon = \frac{H}{L} \tag{1.3}$$

is very small, i.e. $\varepsilon \ll 1$. We also introduce the characteristic values for the horizontal velocity, vertical velocity, time and pressure

$$U = \sqrt{gH}, \quad W = \varepsilon U, \quad T = L/U, \quad P = U^2. \tag{1.4}$$

and we assume that

$$\frac{U^2}{gH} \approx 1.$$

The ratio U^2/gH is the square of the Froude number Fr. The Reynolds number is

$$Re = \frac{UL}{\mu}.$$

On the bottom, a Navier condition with a friction coefficient κ and a no-penetration condition are considered. Under the shallow water assumption $\varepsilon \ll 1$, the vertical acceleration terms are neglected, therefore the pressure is hydrostatic. Additionally, the friction and the viscosity are assumed to be small

$$\alpha = \varepsilon \alpha_0, \quad Re = \frac{Re_0}{\varepsilon}.$$

Integrating the resulting dimensionless system over the water depth gives the inviscid Shallow Water equations (1.1)-(1.2) with $S_f = \kappa u$. This system results from an approximation in $O(\varepsilon)$ of the Navier-Stokes equation.

1.2 The Exner equation

In coupled hydrodynamic-morphologic models where the water is modeled by the Shallow Water equations, the most classical model for sediment transport is the Exner equation [54]. The Exner equation models bed load transport, and it is a only a mass conservation equation - no further physical consideration is included. Let b be the sediment height over a non-erodible substratum B. The Exner equation states that

$$\partial_t b + \nabla . q_s = 0, \tag{1.5}$$

where q_s is a solid flux formula. The Exner equation can also be written with the deposition and erosion rates D and E

$$\partial_t b = \mathsf{D} - \mathsf{E} = -\nabla . q_s.$$

The results given by the Exner equation crucially depend on the choice of the solid flux formula q_s. These formulae frequently involve some of the following physical characteristics of the sediment: density, characteristic diameter; as well as characteristics of the water flow above the sediment layer, the velocity of the water flow for instance.

1.2.1 The transport threshold

Several formulae also involve a threshold for incipient motion. The transport threshold is controlled by a dimensionless number Θ called the Shields number [118], which is the ratio between the driving force and the stabilizing force acting on a sediment grain. A physical deduction of the Shields number is made in [9]. The driving force is the drag exerted by the water on the grain, while the stabilizing force is the grain's weight reduced by its buoyancy. The hydrodynamic force exerted by the water on a flat surface of the size of a grain is proportional to τD^2, where τ is the shear stress and D the grain diameter, while the weight of the grain reduced by its buoyancy is $(\rho_s - \rho_w)gD^3$. The Shields number is then defined by

$$\Theta = \frac{\tau}{(\rho_s - \rho_w)gD}. \tag{1.6}$$

Motion occurs when the threshold Shields number Θ_c, also called critical Shields number, is exceeded. To determine the threshold Shields number, experiments are necessary. Such experiments are reported in the original work of Shields [118], where the threshold Shields number is plotted against the particle Reynolds number. Determining the threshold Shields number raises several difficulties, such as deciding whether there is motion or not, and how long one should wait to detect the movement of a particle.

1.2.2 Some bed load transport formulae

Many bed load transport formulae are available from the literature. Some date back to the beginning of the 20^{th} century, while others are very recent, see for instance the formula developed in [112]. The following paragraph presents some of the most famous ones. The formulae presented below implicitly assume that transport occurs in the direction of the water flow velocity. The formula proposed by Grass [68] is one of the simplest. It does not involve a threshold. The solid flux is given by

$$q_s(u) = A_g u|u|^{m_g - 1}, \tag{1.7}$$

where the constants A_g and m_g are empirically determined. A_g takes into account the kinematic viscosity and the grain size and is such that $0 \leqslant A_g \leqslant 1$. The exponent m_g is a positive real number such that $1 \leqslant m_g \leqslant 4$, a commonly chosen value is $m_g = 3$. The Grass formula assumes that motion begins at the same time for the fluid and the sediment. It allows to build a coupled hyperbolic system of equations [45, 48]. When the Grass law is used, the Shallow Water-Exner system is always hyperbolic.

The Meyer-Peter & Müller formula [103] gives

$$Q_s = 8(\Theta - \Theta_c)^{3/2}, \tag{1.8}$$

with

$$Q_s = \frac{|q_s|}{\sqrt{g\left(\frac{\rho_s}{\rho_w} - 1\right)D^3}}.$$

This formula was derived from a fit of experimental laboratory data. The formula proposed by Ashida & Michiue [12] is

$$Q_s = 17(\Theta - \Theta_c)(\sqrt{\Theta} - \sqrt{\Theta_c}). \tag{1.9}$$

It results from a theoretical derivation and a fit of experimental data. Both the Meyer-Peter & Müller law (1.8) and the Ashida & Michiue law (1.9) give the same dependence $Q_s \propto \Theta^{3/2}$, so they provide similar predictions far from the threshold Θ_c. However, since equation (1.8) predicts $Q_s \propto (\Theta - \Theta_c)^{3/2}$ while equation (1.9) predicts $Q_s \propto (\Theta - \Theta_c)(\sqrt{\Theta} - \sqrt{\Theta_c})$, the two laws give different results close to the threshold [88].

Some stability analyses [62] have evidenced the role of the bed slope on bed load transport. Gravity tends to bring sediment grains from the top of dunes to their basis. This is a diffusive transport mechanism. Acknowledging gravity effects led to including the slope in solid flux formulae. Referring to ideas presented in [52], the authors of [133] thus use the formula

$$q_s = k|u|^m \left(\frac{u}{|u|} - c\nabla_{\mathbf{x}} b \right), \tag{1.10}$$

where k, m, c are constant parameters of the model.

All the formulae presented above are deterministic. But sediment transport seems to have some stochastic features [50], for instance due to turbulence in the water flow. Several authors present stochastic transport models, see [8], [82]. However, probabilistic models are beyond the scope of the present work and will not be discussed here.

1.2.3 Critics made to the bed load transport formulae

As mentioned in [64], none of the existing bed load transport formulae is universally accepted, or even strongly recommended for practical applications, though many performance assessments have been made - see the references in [64, 91]. Several shortcomings of classical bed load transport formulae can be identified. In a real-life application the sediment mixture is not homogeneous because the grains composing it have different sizes - and different shapes. The sediments in gravel bed rivers are poorly sorted [112]; however, many formulae only use one characteristic sediment diameter. Most of the formulae are derived from flume experimental data rather than from field measurements - it is the case of the Meyer-Peter & Müller formula. The authors of [64] highlight other assumptions inherent to the use of bed load formulae. Notably, the formulae also assume that the maximum possible amount of bed load is transported, regardless of material availability. In [64], the performance of twelve bed load transport formulae developed for gravel bed channel is tested, among which the Meyer-Peter & Müller formula and the Einstein formula . They use existing data for the flow conditions - they do not simulate the flow. Their conclusion is that none of the formulae is capable of generally predicting bed load transport in gravel bed rivers, and the errors can be very large. More recent studies report similar conclusions [91]. A formula may give good results in a case and totally fail in another case. The lack of field data to test the formulae as well as limited hydraulic data are some of the reasons which explain such poor performances; incomplete

knowledge and incorporation of the physics of bed load transport in the formulae also play a role. Naturally, the variability of the parameters inside a river makes the design of a formula very difficult.

1.3 Possible improvements to the Exner model

1.3.1 Necessity of a phase shift

Classical Shallow Water-Exner model do not manage to simulate dune growth. In these models, the maximum shear is in phase with the maximum of the river bed topography, and in many formulae, the shear directly controls the solid flux. Stability analyses indicate that a phase lag between the maximum shear stress and the maximum of the bed topography is necessary to obtain bed form amplification. This idea was pioneered by Kennedy [84] and studied thereafter in [51, 61, 87].

A possibility to obtain the phase lag is to directly insert it in the model in the form of a saturation length. As mentioned in [43], the solid flux does not adjust instantaneously to a change of shear stress. It adjusts with some temporal or spatial delay due to particle inertia or settling. There is a relaxation effect. The saturation length L_{sat} is the length needed for q_s to adjust to the value of the saturated flux q_{sat}, i.e. the flux when the erosion and deposition rates balance each other. It is controlled by the local value of the shear stress at the bed surface. In [9], the following relaxation equation is given for the flux q_s

$$L_{sat}\partial_x q_s = q_{sat} - q_s. \tag{1.11}$$

In [106], a theoretical expression for L_{sat} is given and compared with estimations of L_{sat} deduced from field observations. The predicted values match the observed values reasonably well - the predictions and observations agree within a factor of two. In [43], the length L_{sat} is interpreted as a deposition length $l_d = Vt_d$, related to the length traveled by a particle during a "flight". V is the velocity of a particle and t_d is a deposition time.

In [9], a linear stability analysis is performed. The river bed is sinusoidal. Equation (1.11) is used, and a shifted shear stress is prescribed. The result of the analysis is that the sinusoidal perturbations grow when the shear stress is "in advance" with respect to the topography (meaning that the maximum shear stress is upstream of the crest of the perturbation). Additionally, there is a cut-off wavelength below which the perturbations are stable. For the authors of [44], the cut-off comes from the fact that gravity effects dominate for short wavelengths, thus stabilizing the perturbations. The saturation length controls the size of the smallest perturbation which can develop [106]. Similarly, in [43], the cut-off length is linked to the deposition length l_d. Note that a dune cannot become infinitely large. For instance, slope effects limit its growth. When the slope becomes larger than the static friction angle, a small avalanche occurs. The reader is referred to [9] for a discussion of the nonlinear effects limiting dune growth.

1.3.2 Improvement of the flow model

Some works aim at obtaining an expression for the shifted shear stress without "manually" imposing the shift length. They do so by using an improved description of the water flow. An inherent limitation of the Shallow Water equations is that they poorly account for what happens in the water near the surface of the sediment layer. As the Shallow Water equations are depth-integrated equations, they do not make any distinction between the boundary layer, where the water velocity changes quickly, and the rest of the fluid. Yet interaction between the fluid and the sediment layer happens at the surface of the sediment layer. Refining the model in the water layer to better describe the zone near the sediment surface is then an interesting option. The Triple Deck model [102, 105, 132] is one of the models aiming at providing such a better description. As the name indicates, the water layer is divided in three "decks". In the "lower deck", a viscous problem is solved. The perturbations in the "lower deck" act on the "main deck" as a perturbation of the stream lines. The deflection of the stream lines is transmitted to the "upper deck", which is an ideal fluid layer. In the "upper deck", a pressure disturbance is created, and this pressure disturbance is transmitted back to the lower deck, which promotes the velocity perturbations. The Triple Deck theory is a more refined model than the Shallow Water equations, but it avoids resolving the full Navier-Stokes equations.

In [87], the water flow above an erodible bump is resolved using the Triple Deck theory. First, a linearized steady triple deck problem is solved. Then, a mass transport equation giving the solid flux is solved, and finally, the evolution of the topography is computed. The stability of the erodible bed is investigated both numerically and analytically; the results are in good agreement. Dune growth is obtained from initially small perturbations of the surface of the erodible material. A phase lag between the top of the sediment bump and the maximum shear stress is observed. Though the modeling simplifications are many, the importance of a good description of the boundary layer is clearly shown.

Another flow description was recently proposed in [81]. The water model consists in a layer of ideal fluid on top of a layer of viscous fluid. The two layers interact strongly with one another. Again, a phase lag between the shear stress and the topography is achieved. The authors of [81] compare the numerical results obtained with their model to numerical results obtained with the multilayer Saint-Venant model [22] and they report that the phase lag can also be achieved with the multilayer model.

In [61], a modified water flow description is used as well, and as before, the phase shift is obtained, not imposed. A non-local solid flux with an integral term is obtained. Linearizing the equations even gives an explicit expression for the shear stress, which features a non-local (integral) term

$$\tau_\zeta = f\rho_s v^2 \left(1 - b + \alpha \int_0^{+\infty} \xi^{-1/3} \partial_x b(x-\xi,t) d\xi\right),$$

where f is a dimensionless friction coefficient and α is a positive constant proportional

to the cubic root of the Reynolds number in the water flow. The expression of τ_ζ is then modified to take into account the effect of the bed slope and additional simplifications are performed. Finally, the equation giving the evolution of the sediment bed is

$$\partial_t b + \partial_x \left(\frac{b^2}{2} - \partial_x b + \int_0^{+\infty} \xi^{-1/3} \partial_x b(x-\xi,t) d\xi \right) = 0.$$

This model was later on studied in [5], [7]. A remarkable feature of this model is that it violates the maximum principle, which is necessary to obtain dune growth.

The refinement of the water flow model is out of the scope of the present work.

1.3.3 Improved description of the sediment layer

In this thesis, the choice was made to focus on the sediment layer. We present below several options for improving the description of the sediment layer.

Mechanics in the sediment layer

One can also notice that while the classical bed load transport formulae require the knowledge of the flow conditions in the water, they do not really provide any insight of the mechanics inside the sediment layer. In [131], the author investigates the case of sediment transport under dam-break flows, and he argues that for such highly erosive and transient flows, the sediment inertia should be explicitly accounted for. He then models the water-sediment system as a two-layer system, in which the transported sediment are described as a fluid. Both the water and the sediment are modeled by the inviscid Shallow Water Equations. The sediment and the layer have different densities and velocities. Comparisons with experiments of dam-break flows over erodible granular beds are made. The model was found to perform correctly when the density of the sediment layer is close to that of the water layer, but it did not give satisfactory results for heavier sediment layers. A similar model with three layers was later proposed in [25].

One of the ideas behind the model proposed in this thesis is indeed to provide a description of the mechanics of the sediment layer.

An example of derivation of the Shallow Water-Exner system

A formal deduction of the Shallow Water-Exner model is presented in [58]. Slope effects are incorporated in the definition of the solid flux, and the solid flux is obtained thanks to the derivation. For the sake of a later comparison, we present below some of the features of this work. Indeed, our approach is similar, though a different asymptotics is made, and the resulting model is of course different. In [58], there is no phase shift, while in our work, a phase shift between the maximum of the riverbed topography and the local maximum of the solid flux is obtained.

The Shallow Water-Exner system is derived from the Navier-Stokes equations through an asymptotic analysis. Between the water and mobile sediment layers, two friction laws are considered, a linear one and a quadratic one. The friction law considered between the static and mobile sediment layers is the Coulomb law. The sediment layer has an internal friction angle δ. Thus, after vertical integration, transport laws involving a threshold are obtained. The sediment layer described consists in a mobile sediment layer of thickness b lying on a sediment layer of thickness b_f which cannot move, but can exchange mass with the mobile sediment layer. The underlying bedrock, which is fixed and does not exchange mass with the sediment layer, is denoted by B. The density ratio is $r = \rho_w/\rho_s$. In the water layer, the classical Shallow Water equations are obtained. In the sediment layer, the fact that the sediment velocity is much smaller than the water velocity is taken into account when the scaling parameters are given. The order of magnitude of the sediment velocity is $U_s = \varepsilon^2 U$, where ε is the aspect ratio (1.3). The coupled model obtained reads

$$\begin{aligned}
&\partial_t h + \nabla_{\mathbf{x}}.(hu) = 0,\\
&\partial_t(hu) + \nabla_{\mathbf{x}}(hu \otimes u) + \frac{1}{2} g \nabla_{\mathbf{x}} h^2 + gh\nabla_{\mathbf{x}}(B + b + b_f) + \frac{gb}{r}\mathcal{P} = 0,\\
&\partial_t(b + b_f) + \nabla_{\mathbf{x}}\left(b v_b \sqrt{(1/r - 1)gD}\right) = 0,\\
&\partial_t b_f = -T_m,
\end{aligned}$$

with

$$\mathcal{P} = \nabla_{\mathbf{x}}(rh + b + b_f + B) + (1 - r)\mathrm{sgn}(v)\tan\delta.$$

v is the velocity in the mobile sediment layer. The term "sgn(v)" is to be understood as "direction of the vector v". The term $\frac{gb}{r}\mathcal{P}$ is actually a friction term. With the linear friction law at the water-sediment interface, the velocity v_b is

$$v_b^{(LF)} = \frac{1}{\sqrt{(1/r - 1)gD}} u - \frac{\vartheta}{1 - r}\mathcal{P},$$

while with the quadratic friction law, it is

$$v_b^{(QF)} = \frac{1}{\sqrt{(1/r - 1)gD}} u - \left(\frac{\vartheta}{1 - r}\right)^{1/2} |\mathcal{P}|^{1/2}\mathrm{sgn}(\mathcal{P}),$$

The quantity ϑ is

$$\vartheta = \frac{\Theta_c}{\tan\delta}.$$

The mass transfer between the static and mobile sediment layers is denoted by T_m. It is the difference between the erosion and deposition rates. The authors establish a link between the velocity v_b and classic transport laws with a threshold such as (1.8). Indeed, the Coulomb friction law between the static and mobile sediment layers introduces a threshold for the onset of motion. The authors first define a modified effective shear stress which takes into account slope effects. They proceed to defining an effective Shields

number, and then they obtain a threshold law. With the linear friction, the modified effective shear stress is

$$\tau_{eff}^{(LF)} = \frac{\vartheta D}{b}\tau^{(LF)} - g\rho_s \vartheta D \nabla_{\mathbf{x}}(rh + b + b_f + B)$$

Note that here it is not the bed slope which is taken into account, but the free surface slope. $\tau^{(LF)}$ is the shear exerted by the water at the water-sediment interface with a linear friction law. The velocity of the sediment layer is then

$$v_b^{(LF)} = \mathrm{sgn}(\tau_{eff}^{(LF)}) \left(\Theta_{eff}^{(LF)} - \Theta_c\right)_+ .$$

This model admits a dissipative balance for the mechanical energy. This is not generally the case for Shallow Water-Exner models. As shown in [137], in the general case, a mathematical entropy exists, but this entropy is not the mechanical energy.

Non-local effects in the sediment layer

A non-local sediment flux formula is proposed in [35]. The authors are concerned with the modeling of the sea bed on short time scales. The sea bed is considered to be a structure with low stiffness, and the fundamental assumption of the modeling approach is that the sea bed will adapt to the flow by some sort of minimal sand transport in order to minimise an energy expression. For simplicity, the approach is explained in one dimension. Let $J(b(x,t), h(x,t,b(x,t)))$ be the cost function to be minimized. For instance, J can contain a term linked to the energy of the water flow and a term forcing the sediment surface to stay close to its original shape. It is assumed that b will evolve so as to reduce the functional $J(b(x,t), h(x,t,b(x,t)))$. An evolution model from time t to time $t + \Delta t$ for the sediment surface can be given by

$$b(x, t + \Delta t) = b(x,t) - \Delta t \phi \nabla_b J(b(x,t), h(x,t,b(x,t))), \tag{1.12}$$

where ϕ is the "receptivity" of the bed, it can be linked to the porosity of the sediment layer. The notation ∇_b denotes the derivative with respect to b. In other words, the discrete form of the equation

$$\partial_t b + \phi \nabla_b J = 0$$

minimizes $J(b(x,t), h(x,t,b(x,t)))$. By analogy with the Exner equation (1.5), we have

$$\partial_x q_s = \phi \nabla_b J.$$

This implies a non-local definition for the solid flux q_s:

$$q_s(x,t) = q_s(-\infty, t) + \int_{-\infty}^{x} \phi \nabla_b J(a) da.$$

The shore line is positioned at $x = 0$ and the negative values of x correspond to points in the sea. One can safely assume that $q(-\infty, t) = 0$ and that $\nabla_b J(x) \to 0$ when $x \to -\infty$. The context is that of coastal modeling, and the influence of the water flow on the sediment bed decreases going away from the shore as the water depth increases. This expression is similar to the expression given in [61].

Using a non-local solid flux is one of the main ideas developed in this thesis.

1.3.4 Non-local models for other applications

We present here a few non-local models, used for other applications, that present some similarities with the non-local model (1). The behavior of the solution of (1) is probably very different from the solution of the non-local models below. Our point here is that the research articles on these non-local models highlight the difficulty of analyzing them and of proposing robust numerical schemes to solve them. There exist many non-local models. They are developed for a large variety of applications: chemotaxis, traffic flow, plasma physics...

In plasma physics, a non-local electron conduction model is used for the simulation of laser-driven Inertial Confinement Fusion experiments. While a local theory, the Spiter-Härm theory, was formerly used, some experiments have evidenced that the electron heat flow is non-local. This flux depends not only on the local conditions, but also on the portion of the temperature profile enclosed in a few hundreds of mean free paths. The heat flux is therefore expressed as

$$Q(t,x) = \int_{\mathbb{R}^3} W_\epsilon(x,x') Q_{SH}(t,x') dx',$$

which is the convolution of the Spitzer-Härm flux Q_{SH} with a kernel $W_\epsilon(x,x')$ that tends to a Dirac function as ϵ goes to 0. ϵ is the ratio of the velocity unit defined by the time and length scales over the thermal velocity. However, the choice of the kernel is not clear. One can try to describe the heat flux as the solution of a linear transport equation, see [117]. Following the ideas developed in [117], a non-local model for electron temperature is derived in [66]

$$\begin{cases} \partial_t \Theta + \frac{2}{3\rho} \partial_x Q = 0, \\ Q - \epsilon^2 \nu(\Theta) \partial_x^2 Q = -\kappa(\Theta) \partial_x \Theta, \end{cases} \tag{1.13}$$

where ν, κ are two smooth positive functions and ρ is the charge density. The total energy is conserved. This system resembles (1), though it is not as non-linear as (1). In (1.13), anti-diffusive effects may occur, i.e. in some parts of the domain the heat flux Q and the temperature gradient may have the same sign.

Another non-local model was developed by Patlak [108] and Keller and Segel [83]. It describes the space and time evolution of the density n of some cells attracted by a chemical having a concentration c. The chemotactic sensitivity χ is taken here constant.

The Patlak-Keller-Segel model reads

$$\partial_t n - \nabla_{\mathbf{x}}.(\nabla_{\mathbf{x}} n - \chi n \nabla_{\mathbf{x}} c) = 0, \tag{1.14}$$
$$\partial_t c - \Delta_{\mathbf{x}} c = n - c, \tag{1.15}$$

Equation (1.14) means that the cell density n diffuses and is advected at the velocity $\chi\nabla_{\mathbf{x}} c$, which is a non-local velocity because c is the solution of a non-local equation. The parabolic model (1.14),(1.15) admits blow-up solutions. Here, we are actually more interested in another form of the model where (1.15) is replaced by the elliptic equation

$$c - \Delta_{\mathbf{x}} c = n, \tag{1.16}$$

because (1.16) ressembles more the second equation of (1). Note that in (1.16), the concentration c instantly adapts to the variations of n. Let us take the gradient of equation (1.16) and denote by $\tilde{c}$ the gradient of the concentration c, we get

$$\tilde{c} - \Delta_{\mathbf{x}} \tilde{c} = \nabla_{\mathbf{x}} n,$$

which is very similar to the second equation of (1). Equation (1.14) can be rewritten as

$$\partial_t n - \nabla_{\mathbf{x}}.(\nabla_{\mathbf{x}} n - \chi n \tilde{c}) = 0,$$

which is the first equation of (1) with an extra diffusion term.

In [60], the main analysis results on (1.14),(1.16) are recalled. Assuming that the initial condition on the density is small enough, there exist at least a couple of nonnegative functions (n, c) and the global mass of cells is conserved with respect to time. Then, a finite-volume scheme is proposed and analyzed. For the time discretization, a fully implicit Euler scheme is used. Even if the scheme is implicit, a CFL condition (written here for the 1D scheme)

$$\chi \frac{\Delta t}{\Delta x} (\nabla_{i+1/2} c + \nabla_{i-1/2} c) < 1, \tag{1.17}$$

is needed for the advection part of equation (1.14). $\nabla_{i+1/2} c$ is the discrete gradient of the concentration c at the face $i + 1/2$. Note that we expect the model (1) to behave quite differently from the Patlak-Keller-Segel model. The system (1.14),(1.16) admits blow-up solutions, depending on the global initial mass of cells, while we rather expect the solutions of (1) to be quite smooth. The right-hand side of (1.16) is significantly different from that of the second equation of (1), it does not have the same sign.

The Stokes-Brinkman equations are a linear and stationary system of equations, they read [72]

$$\begin{cases} \nabla \cdot u = 0, \\ \mu \mathbf{K}^{-1} u - \tilde{\mu} \Delta u = -\nabla p, \end{cases} \tag{1.18}$$

where u is the fluid velocity, p is the pressure, μ is the viscosity and $\tilde{\mu}$ is an effective

viscosity, $\mathbf{K}$ is a permeability tensor. They allow to simulate incompressible free flow and incompressible flow in a porous domain without requiring interface modeling. The model (1) can be seen as a depth-integrated Stokes-Brinkman system with a free surface.

1.4 Numerical resolution of the Shallow Water-Exner system

In the present thesis, a numerical method for the resolution of the system (1) coupled to the Shallow Water equations is proposed. A finite-volume discretization is used. We propose here a short overview of the existing finite-volume schemes for the resolution of the Shallow Water-Exner system. The numerical treatment of the Shallow Water-Exner system is discussed only for classical, local solid flux formulae.

Two main approaches exist for solving the system. The first approach is said to be "uncoupled". The equations for the water layer and those for the sediment layer are solved separately. This approach is justified by the fact that in many cases, the water layer and the sediment bed evolve at very different time scales. The bed evolution is typically much slower, meaning that for instance, one could take a larger time step in the sediment layer than in the water layer. The hydrodynamic and morphodynamic unknowns are exchanged at some specific time instants only. This approach is adopted in the industrial codes [121, 123, 127]. However, when the characteristic time scales for the two layers are closed, the uncoupled approach is not suitable and stability issues appear. More specifically, the uncoupled approach cannot deal with supercritical flows [45].

In the "coupled" approach, the complete Shallow Water-Exner system is solved. A possible strategy is to approximate the eigenvalues of the Jacobian matrix of the full system. The computation of the eigenvalues is relatively easy when the solid flux is given by the law (1.7), but this is not the general case. This is the technique used in [94]. In [78], the authors use a flux-limited version of Roe's scheme to solve several formulations of the Shallow Water-Exner system. An approach based on the Roe scheme is adopted in [28] to propose a 1D and 2D scheme for unstructured grids. In [25], a relaxation approach is proposed. The water pressure and the sediment flux are relaxed; the computation of the eigenvalues is easy.

Finally, let us mention an intermediate approach introduced in [24]. A three-wave approximate Riemann solver for the Shallow Water-Exner system is proposed. The hydraulic and morphodynamic intermediate states are computed in a decoupled way, but the wave velocities of the full system are evaluated - they are approximate values of the eigenvalues of the Jacobian matrix of the Shallow Water-Exner system.

The majority of finite-volume schemes for the Shallow Water equations uses a colocated approach - the unknowns are discretized at the centers of the finite volumes. Yet schemes relying on a staggered approach have also been designed, see for instance [10, 75, 114]. The staggered grid is commonly referred to as one of the Arakawa grids, first introduced in [11]. The water depth is discretized at the centers of the cells, while the velocity (or the momentum) is discretized at the interfaces between the cells. The imple-

mentation of such methods is quite simple. (Note that the staggered-grid discretization is commonly chosen for finite difference schemes, see [42] for an example.) In [55], a staggered-grid scheme presented in [73] and based on the work in [75] is used to simulate the Shallow Water-Exner system.

The numerical scheme presented in this thesis uses a staggered-grid discretization. A coupled approach is adopted, yet the computation is stable. The numerical scheme for the sediment layer is implicit, and this is very different from what is classically done when simulating the Shallow Water-Exner system in the finite volume community.

2

A non-local sediment transport model

Outline of the current chapter

The contents of this chapter will be submitted under the form of an article along with E. Audusse and M. Parisot under the title "On the Exner model and non-local approximations: modeling, analysis and numerical simulations".

Abstract

The aim of this work is to model and simulate sediment transport under the action of water. A new model for the sediment layer is proposed. Its novelty lies in the presence of a viscosity term in the equation for the sediment velocity, which then makes the sediment flux non-local. A numerical scheme for the model of the sediment layer is designed and analyzed. Then, a coupled water-sediment model and a numerical strategy to simulate it are presented. Adding some non-locality leads to dune growth and propagation.

2.1 Overview of the water-sediment system

2.1.1 Bilayer Navier-Stokes equations

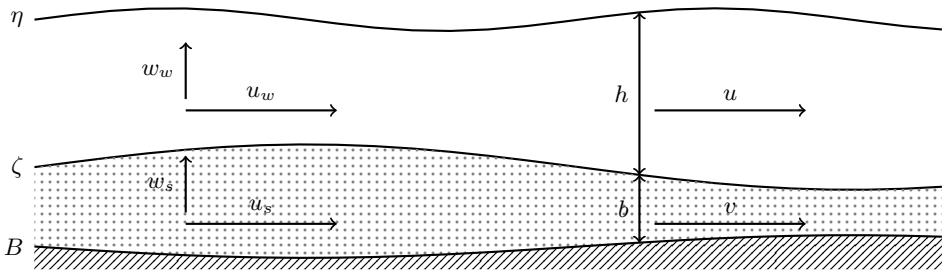


Figure 2.1 – Description of the unknowns in the stratified sediment-water system. left: Navier-Stokes unknowns, right: shallow water unknowns.

We consider a Cartesian coordinate system where $\mathbf{x} \in \mathbb{R}^d$, $d \in \{1, 2\}$, is the coordinate in the horizontal plane, $z \in \mathbb{R}$ is the coordinate in the vertical direction and $t \in \mathbb{R}_+$ is the time coordinate. In this work, we focus on the modeling of the coupled water-sediment system in the regime of bedload transport. We assume that due to gravity, the flow is well-stratified, i.e. there exists an interface $\zeta(\mathbf{x}, t)$ that splits the flow in two parts, see Figure 2.1 : below this interface and above a non-erodible substratum $B(\mathbf{x})$ from now on referred to as the sediment layer, only the sediment phase is present ; above the interface and below the free surface $\eta(\mathbf{x}, t)$ from now on referred to as the water layer, only the water phase is present. In addition, we assume that the flow in both layers can be modeled as a continuous media and more precisely with the incompressible Navier-Stokes equations. Let us precise some notations. In each layer $k \in \{s, w\}$ (s stands for sediment whereas w stands for water), the horizontal velocity is denoted by $u_k(\mathbf{x}, z, t) \in \mathbb{R}^d$, the vertical velocity by $w_k(\mathbf{x}, z, t) \in \mathbb{R}$ and the pressure by $p_k(\mathbf{x}, t)$. The governing equations in each layer read

$$\begin{aligned} \nabla_{\mathbf{x}} \cdot u_k + \partial_z w_k &= 0 \ , \\ \rho_k \left(\partial_t u_k + (u_k \cdot \nabla_{\mathbf{x}}) u_k + w_k \partial_z u_k\right) &= -\nabla_{\mathbf{x}} p_k + \nabla_{\mathbf{x}} \cdot (2\mu_k \, \mathrm{D}_{\mathbf{x}} u_k) \\ &\quad + \partial_z \left(\mu_k \left(\partial_z u_k + \nabla_{\mathbf{x}} w_k\right)\right) \ , \\ \rho_k \left(\partial_t w_k + (u_k \cdot \nabla_{\mathbf{x}}) w_k + w_k \partial_z w_k\right) &= -\partial_z p_k - g\rho_k + \partial_z \left(2\mu_k \partial_z w_k\right) \\ &\quad + \nabla_{\mathbf{x}} \cdot \left(\mu_k \left(\nabla_{\mathbf{x}} w_k + \partial_z u_k\right)\right) \ , \end{aligned} \tag{2.1}$$

where the symmetric gradient is used $\mathrm{D}_{\mathbf{x}} u = \frac{\nabla_{\mathbf{x}} u + (\nabla_{\mathbf{x}} u)^t}{2}$. The fluid $k \in \{s, w\}$ is characterized by its density $\rho_k \in \mathbb{R}_+^*$ and its viscosity $\mu_k \in \mathbb{R}_+^*$ fixed. The free surface and the water-sediment interface are respectively governed by the kinematic equations

$$\partial_t \eta + u_{w|_{z=\eta}} \cdot \nabla_{\mathbf{x}} \eta - w_{w|_{z=\eta}} = 0 \qquad \text{and} \qquad \partial_t \zeta + u_{s|_{z=\zeta}} \cdot \nabla_{\mathbf{x}} \zeta - w_{s|_{z=\zeta}} = 0.$$

The no-penetration condition is assumed at the water-sediment interface and at the

substratum

$$\partial_t \zeta + u_{w|_{z=\zeta}} \cdot \nabla_{\mathbf{x}} \zeta - w_{w|_{z=\zeta}} = 0 \qquad \text{and} \qquad u_{s|_{z=B}} \cdot \nabla_{\mathbf{x}} B - w_{s|_{z=B}} = 0.$$

For each surface $\xi \in \{B, \zeta, \eta\}$, we define the normal by

$$N_\xi = \frac{1}{\sqrt{1 + |\nabla_{\mathbf{x}} \xi|^2}} \begin{pmatrix} -\partial_x \xi \\ -\partial_y \xi \\ 1 \end{pmatrix}$$

and a base of tangent vectors by

$$T_\xi^1 = \frac{1}{\sqrt{1 + |\partial_x \xi|^2}} \begin{pmatrix} 1 \\ 0 \\ \partial_x \xi \end{pmatrix} \qquad \text{and} \qquad T_\xi^2 = \frac{1}{\sqrt{1 + |\partial_y \xi|^2}} \begin{pmatrix} 0 \\ 1 \\ \partial_y \xi \end{pmatrix} .$$

The viscosity tensor is defined for $k \in \{s, w\}$ by

$$\Sigma_k = 2\mu_k \begin{pmatrix} \mathrm{D}_{\mathbf{x}}\, u_k & \dfrac{\nabla_{\mathbf{x}} w_k + \partial_z u_k}{2} \\ \dfrac{\nabla_{\mathbf{x}} w_k + \partial_z u_k}{2} & \partial_z w_k \end{pmatrix}$$

and the stress tensor is assumed to be continuous at the free surface and at the water-sediment interface, i.e.

$$\text{and} \qquad \begin{array}{l} \left(-p_{w|_{z=\eta}} Id + \Sigma_{w|_{z=\eta}}\right) N_\eta = 0 \\ \left(-p_{s|_{z=\zeta}} Id + \Sigma_{s|_{z=\zeta}}\right) N_\zeta = \left(-p_{w|_{z=\zeta}} Id + \Sigma_{w|_{z=\zeta}}\right) N_\zeta. \end{array}$$

For the sake of simplicity, the atmospheric pressure was set to zero. To close the system, friction laws must be given at the water-sediment interface and at the substratum

$$\begin{array}{ll} \left(\left(-p_{w|_{z=\zeta}} Id + \Sigma_{w|_{z=\zeta}}\right) N_\zeta\right) \cdot T_\zeta = \kappa_\zeta \mathbf{U}_\zeta^\star \cdot T_\zeta, & \forall T_\zeta \in \mathrm{vect}\{T_\zeta^1, T_\zeta^2\} \\ \left(\left(-p_{s|_{z=B}} Id + \Sigma_{s|_{z=B}}\right) N_B\right) \cdot T_B = F_B, & \forall T_B \in \mathrm{vect}\{T_B^1, T_B^2\} \end{array}$$

where the shear reads $\mathbf{U}_\zeta^\star = \mathbf{U}_{w|_{z=\zeta}} - \mathbf{U}_{s|_{z=\zeta}}$ with the velocities $\mathbf{U}_k = (u_k, w_k)^t$. Details about the friction force at substratum f_B are given in §2.1.2. The friction coefficient $\kappa_\zeta > 0$ at the interface and the friction force at the substratum F_B can be functions of the horizontal space variable, the surrounding pressure and the shear, i.e.

$$\text{and} \qquad \begin{array}{rcl} \kappa_\zeta : & \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}_+ \\ & \left(p_{w|_{z=\zeta}}, \mathbf{U}_\zeta^\star\right) \mapsto \kappa_\zeta \left(p_{w|_{z=\zeta}}, \mathbf{U}_\zeta^\star\right) \\ F_B : & \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}_+ \\ & \left(\mathbf{x}, p_{s|_{z=B}}, \mathbf{U}_{s|_{z=B}}\right) \mapsto F_B \left(\mathbf{x}, p_{s|_{z=B}}, \mathbf{U}_{s|_{z=B}}\right). \end{array}$$

The system (2.1) must be completed with initial data

$$u_k(\mathbf{x}, z, 0) = u_k^0(\mathbf{x}, z) \ , \qquad w_k(\mathbf{x}, z, 0) = w_k^0(\mathbf{x}, z) \ ,$$
$$\zeta(\mathbf{x}, 0) = \zeta^0(\mathbf{x}) \qquad \text{and} \qquad \eta(\mathbf{x}, 0) = \eta^0(\mathbf{x}).$$

that have to satisfy the compatibility condition

$$\begin{aligned} \nabla_{\mathbf{x}} u_k^0 + \partial_z w_k^0 &= 0 \ , \\ (u^0_{w|_{z=\zeta}} - u^0_{s|_{z=\zeta}}) \nabla_{\mathbf{x}} \zeta^0 - (w^0_{w|_{z=\zeta}} - w^0_{s|_{z=\zeta}}) &= 0 \ , \\ u^0_{w|_{z=B}} . \nabla_{\mathbf{x}} B - w^0_{w|_{z=B}} &= 0 \ . \end{aligned}$$

2.1.2 Introduction of a threshold for the onset of motion

Classical laws of solid sediment transport used in the context of hydraulic engineering have a threshold for incipient motion, and are a power function of the difference between a shear stress and a critical shear stress. The Meyer-Peter and Müller law [103] and van Rijn law [113] are examples of such laws involving a threshold. In the following, we show that a threshold for the onset of motion can also be introduced in the models (2.3) by means of the operator $F_B\left(\mathbf{x}, p_{s|_{z=B}}, \mathbf{U}_{s|_{z=B}}\right)$ by defining it with a Coulomb friction law.

The Coulomb friction law, classically used for contacts between solids, claims that at each point $\mathbf{x}$ either the contact is static, i.e. $v(\mathbf{x}) = 0$, or the friction force is on a cone and its direction is opposed to the velocity. In particular, its magnitude does not depend on the velocity. On the other hand, the friction law for a contact between fluid layers is usually proportional to a power of the relative velocity, see for instance the Manning-Strickler law [96]. Since the sediment layer is a mixture that is neither really a fluid nor a solid, we propose that the friction is a combination of the two classical frictions, i.e.

$$F_B(\mathbf{x}, p, \mathbf{U}) = \begin{cases} \left(\overline{\tau} p + \kappa_B \|\mathbf{U}\|^{\gamma}\right) \dfrac{\mathbf{U}}{\|\mathbf{U}\|} & \text{if } \mathbf{U} \neq 0 \\ \min\left(1, \dfrac{\overline{\tau}}{\|\tau\|} p\right) \tau & \text{else.} \end{cases}$$

The parameter $\overline{\tau}(\mathbf{x}) \geq 0$ is the Coulomb friction coefficient, that goes to zero if the layer is almost fluid. The parameter $\kappa_B(\mathbf{x}) > 0$ is the Strickler's coefficient, that goes to zero if the layer is almost solid. The parameter $0 < \gamma \approx 1$ depends on the regime of the flow, usually $\gamma = 1$ for a laminar flow and $\gamma = 2$ for a turbulent flow. The shear stress at the surface of the substratum τ is the resultant of the forces on a column of sediment except the friction at the substratum, i.e.

$$\tau = \kappa_\zeta \mathbf{U}_\zeta^\star - \int_B^\zeta \left(\nabla_{\mathbf{x}} p_s - \nabla_{\mathbf{x}} \cdot \left(2\mu_k \, \mathrm{D}_{\mathbf{x}} \, u_k\right) + \partial_z \left(\mu_k \left(\partial_z u_k + \nabla_{\mathbf{x}} w_k\right)\right)\right) \mathrm{d}z.$$

Even if the model (2.1) to describe the sediment-water system neglects several physical processes such as suspended sediment, water exchanged between the sediment layer and

the water layer, as well as the possibility for the sediment to have a complex rheology and to contain grains of several sizes, it is already too complex to be simulated at the scale of a river or of a harbor. To carry out large-scale simulations, a reduced model is needed. Introducing reduced models for the sediment layer is the purpose of the next section.

2.2 The sediment layer integrated model

This section focuses on the sediment layer, without considering the water layer. More precisely, the action of water on sediments is solely due to water pressure and water velocity at the sediment-water interface, respectively referred to in this section by $p_\zeta = p_{w|_{z=\zeta}}$ and $\mathbf{U}_\zeta = \mathbf{U}_{w|_{z=\zeta}}$. This is a preliminary step before considering the coupled water and sediment layers. Different models are derived for the sediment layer considering different parameters. Then, a numerical scheme is presented and validated with numerical experiments.

2.2.1 Vertically averaged models

To derive reduced models using asymptotic arguments, we introduce the following characteristic values of the system

$$\begin{array}{llll} t = T\widetilde{t}\ , & \mathbf{x} = L\widetilde{\mathbf{x}}\ , & z = H\widetilde{z} + B_0\ , & \zeta = H\widetilde{\zeta} + B_0\ , \quad B = H\widetilde{B} + B_0\ , \\ u_s = U\widetilde{u}_s\ , & w_s = W\widetilde{w}_s\ , & p_s = P\widetilde{p}_s\ , & \\ \bar{\tau} = C\Upsilon\ , & \kappa_B = \dfrac{K_B}{U^{\gamma-1}}\widetilde{\kappa}_{\widetilde{B}}\ , & \kappa_\zeta = K_\zeta\widetilde{\kappa}_{\widetilde{\zeta}}\ , & \end{array}$$

with $T = \frac{L}{U}$, H and L respectively the characteristic time and characteristic horizontal and vertical dimensions of the problem, U, W and P respectively the characteristic values of the horizontal and vertical velocities and the pressure, K_ξ, $\xi \in \{B, \zeta\}$ the characteristic values of the friction parameters. A vertical reference level B_0 has been introduced. From now on, we assume that the surfaces are smooth enough, so that the variations of the surfaces are respectively characterized by

$$\nabla_{\mathbf{x}} B\left(\mathbf{x}\right) = \frac{H}{L}\nabla_{\widetilde{\mathbf{x}}}\widetilde{B}\left(\widetilde{\mathbf{x}}\right) \quad \text{and} \quad \nabla_{\mathbf{x}}\zeta\left(\mathbf{x}\right) = \frac{H}{L}\nabla_{\widetilde{\mathbf{x}}}\widetilde{\zeta}\left(\widetilde{\mathbf{x}}\right).$$

The flow is therefore characterized by the following dimensionless numbers

$$\begin{array}{lll} \varepsilon = \dfrac{H}{L}\ , & F_r = \dfrac{U}{\sqrt{gH}}\ , & R_e = \dfrac{\rho_s U L}{2\mu_s}\ , \\ r = \dfrac{\rho_w}{\rho_s}\ , & \Theta_B = \dfrac{K_B U}{\rho_s g H}\ , & \Theta_\zeta = \dfrac{K_\zeta U_\zeta^\star}{\rho_w g H}\ , \end{array} \tag{2.2}$$

respectively named the numbers of shallowness, the Froude number, the Reynolds number, the density ratio and what we call the large scale Shields numbers at the substratum and at the sediment-water interface. The shear celerity $U_\zeta^\star$ is the characteristic value

of the velocity difference at the sediment-water interface $\mathbf{U}_\zeta^\star = \mathbf{U}_\zeta - \mathbf{U}_{s|_{z=\zeta}}$. The large scale Shields number at the sediment-water interface can be linked to the classical Shields number, i.e. $\theta_\zeta = \frac{K_\zeta U_\zeta^\star}{(\rho_s-\rho_w)gD}$, by introducing the characteristic grain diameter D. More precisely $\Theta_\zeta = \frac{1-r}{r}\frac{D}{H}\theta_\zeta$.

In the following, several vertically-integrated models are formally derived depending on the scaling of the dimensionless numbers, in particular the Reynolds number and the large scale Shields numbers. The models can be summarized by the following general model

$$\begin{aligned} \partial_{\widetilde{t}}\widetilde{b} + \varepsilon^\alpha \nabla_{\widetilde{\mathbf{x}}} \cdot \left(\widetilde{b}\widetilde{v}\right) &= 0 \ , \\ (1-\alpha)\,\beta \left(\partial_{\widetilde{t}}\left(\widetilde{b}\widetilde{v}\right) + \nabla_{\widetilde{\mathbf{x}}} \cdot \left(\widetilde{b}\widetilde{v}\otimes\widetilde{v}\right)\right) + \widetilde{F}_{\widetilde{B}}\left(\widetilde{v}\right) &= \widetilde{\tau} \end{aligned} \tag{2.3}$$

where the dimensionless friction at the surface of the substratum reads

$$\widetilde{F}_{\widetilde{B}}\left(\widetilde{b},\widetilde{v},\widetilde{\tau},\widetilde{p}_{\widetilde{\zeta}}\right) = \begin{cases} \left(\Upsilon\left(\widetilde{b}+r\widetilde{p}_{\widetilde{\zeta}}\right) + \widetilde{\kappa}_{\widetilde{B}}\left\|\widetilde{v}\right\|^\gamma\right)\dfrac{\widetilde{v}}{\left\|\widetilde{v}\right\|} & \text{if } \widetilde{v}\neq 0, \\ \min\left(1, \dfrac{\Upsilon\left(\widetilde{b}+r\widetilde{p}_{\widetilde{\zeta}}\right)}{\left\|\widetilde{\tau}\right\|}\right)\widetilde{\tau} & \text{else,} \end{cases} \tag{2.4}$$

and the dimensionless shear stress reads

$$\widetilde{\tau} = -r\widetilde{\kappa}_{\widetilde{\zeta}}\left(\varepsilon^\alpha\widetilde{v} - \widetilde{u}_{\widetilde{\zeta}}\right) - \beta\widetilde{b}\nabla_{\widetilde{\mathbf{x}}}\left(\widetilde{b}+\widetilde{B}+r\widetilde{p}_{\widetilde{\zeta}}\right) + \omega\nabla_{\widetilde{\mathbf{x}}}\cdot\left(2\widetilde{b}\mathrm{D}_{\widetilde{\mathbf{x}}}\widetilde{v}\right). \tag{2.5}$$

The parameters $(\alpha,\beta,\omega)\in\{0,1\}^3$ depend on the scaling, see Propositions 1 and 2 and activate the following physical processes. The parameter α corresponds to the scale of the ratio between the solid velocity and the forcing velocity u_ζ. For $\alpha=0$, the order of magnitude of the solid velocity is the same as that of the forcing velocity, while for $\alpha=1$, the solid velocity is one order smaller than the forcing velocity. The parameter β reflects the impact of the gravity on the motion of the sediment layer i.e. when $\beta=0$, the gravity term can be neglected whereas when $\beta=1$ it cannot. Similarly, the parameter ω reflects the impact of the viscosity of the sediment layer, i.e. when $\omega=0$, the viscosity term can be neglected whereas when $\omega=1$ it cannot.

The following models were derived assuming that $F_r=1$. The proofs and the models are still valid for lower Froude numbers. However, low Froude regimes are beyond the scope of this work. For details about the low Froude limit, see [85, 86, 136].

Remark 1. In equation (2.3) and if $\alpha=1$, the term $\varepsilon^\alpha\widetilde{\kappa}_{\widetilde{\zeta}}\widetilde{v}$ is of the order of the modelling error. However, in the perspective of the derivation of the coupled system, it is kept to ensure energy dissipation, see Proposition 2.3.1.

Models with local sediment discharge

Let us now introduce the models with large Reynolds number. We derive three models corresponding to three different asymptotic regimes. In all these models, the solid flux

is local. The model Proposition 1.i) corresponds to the Exner model with a Grass-type flux [68], where the solid velocity is proportional to the forcing velocity. The model Proposition 1.ii) corresponds to the bilayer model. This model was previously used to describe the sediment transport in particular in [138]. In model Proposition 1.iii) the gravity effect remains and the solid velocity is one order smaller than in the other models in Proposition 1. The importance of the gravity term was first evidenced in [62, 95]. A model very similar to model Proposition 1.iii) was derived in [58] assuming a different scaling, i.e. while in [58], the dimensionless horizontal velocity in the sediment was assumed to be of the order of ε^2, we obtain the order of magnitude of $\widetilde{u}_s$ during our derivation, and this order of magnitude is ε.

Proposition 1. *Assume that* $F_r = 1$, $1 > r = O(1)$ *and one of the following scalings*

i) $R_e = 1$, $\Theta_\zeta = 1$, $\Theta_B = 1$ *and* $C = 1$ *then* $(\alpha, \beta, \omega) = (0, 0, 0)$,

ii) $R_e = \varepsilon^{-1}$, $\Theta_\zeta = \varepsilon$, $\Theta_B = \varepsilon$ *and* $C = 1$ *then* $(\alpha, \beta, \omega) = (0, 1, 0)$,

iii) $R_e = 1$, $\Theta_\zeta = \varepsilon$, $\Theta_B = 1$ *and* $C = \varepsilon$ *then* $(\alpha, \beta, \omega) = (1, 1, 0)$.

Then the system (2.3) *with the initial condition*

$$\widetilde{b}(\widetilde{\mathbf{x}}, 0) = \widetilde{\zeta}^0(\widetilde{\mathbf{x}}) - \widetilde{B}(\widetilde{\mathbf{x}}), \tag{2.6}$$

and, in the case of model Proposition 1.ii) only, the additional initial condition

$$\widetilde{v}(\widetilde{\mathbf{x}}, 0) = \widetilde{v}^0(\widetilde{\mathbf{x}}),$$

and where $\widetilde{p}_{\widetilde{\zeta}} = \frac{p_\zeta}{\rho_w U^2}$, $\widetilde{u}_{\widetilde{\zeta}} = \frac{u_\zeta}{U}$, *is an approximation of the Navier-Stokes system for the sediment layer with the following modeling errors*

$$\left|\widetilde{\zeta} - \widetilde{B} - \widetilde{b}\right| = O\left(\varepsilon^{1+\alpha}\right), \qquad |\widetilde{u}_s - \varepsilon^\alpha \widetilde{v}| = O\left(\varepsilon^{1+\alpha}\right). \tag{2.7}$$

Proof. First, the mass conservation is obtained classically by integration the divergence free equation (the first equation of (2.1)). In addition, it yields that $W = \varepsilon U$ and from the vertical momentum equation (the third equation of (2.1)), it classically yields that $\frac{P}{\rho_s U^2} = O(1)$.

Now, let us consider the main terms of the horizontal momentum equation, of the stress continuity conditions at the surface of the substratum and at the sediment-water interface. In the cases ii) and i), we write the following auxiliary problem

$$\begin{aligned} \partial^2_{\widetilde{z}} \widetilde{u}_s &= O(\varepsilon), \\ \partial_{\widetilde{z}} \widetilde{u}_s &= O(\varepsilon), && \text{at } \widetilde{z} = \widetilde{\zeta} \\ \partial_{\widetilde{z}} \widetilde{u}_s &= O(\varepsilon), && \text{at } \widetilde{z} = \widetilde{B}. \end{aligned}$$

These equations impose that the vertical variations of $\widetilde{u}_s$ are of size ε, i.e.

$$\widetilde{u}_s\left(\widetilde{\mathbf{x}}, \widetilde{z}, \widetilde{t}\right) = \overline{u}_{s,0}\left(\widetilde{\mathbf{x}}, \widetilde{t}\right) + O(\varepsilon).$$

Let us now focus on the pressure. Since $\partial_{\widetilde{z}}\widetilde{u}_s = O(\varepsilon)$, integrating the vertical momentum equation leads to the hydrostatic relation, i.e.

$$\widetilde{p}_s = \widetilde{\zeta} - \widetilde{z} + r\widetilde{p}_{\widetilde{\zeta}} + O(\varepsilon). \tag{2.8}$$

Integrating the horizontal momentum equation between B and ζ yields

$$\begin{aligned} &\partial_{\widetilde{t}}\left(\left(\widetilde{\zeta}-\widetilde{B}\right)\overline{u}_{s,0}\right) + \nabla_{\widetilde{\mathbf{x}}}\cdot\left(\left(\widetilde{\zeta}-\widetilde{B}\right)\overline{u}_{s,0}\otimes\overline{u}_{s,0}\right) + \frac{\widetilde{\zeta}-\widetilde{B}}{F_r^2}\nabla_{\widetilde{\mathbf{x}}}\left(\widetilde{\zeta}+r\widetilde{p}_{\widetilde{\zeta}}\right) \\ &= -\frac{\Theta_B}{\varepsilon F_r^2}\widetilde{F}_{\widetilde{B}}\left(\overline{u}_{s,0}\right) - \frac{\Theta_\zeta}{\varepsilon F_r^2}r\widetilde{\kappa}_{\widetilde{\zeta}}\left(\overline{u}_{s,0}-\widetilde{u}_{\widetilde{\zeta}}\right) + \frac{1}{R_e}\nabla_{\widetilde{\mathbf{x}}}\cdot\left(\left(\widetilde{\zeta}-\widetilde{B}\right)\mathrm{D}_{\widetilde{\mathbf{x}}}\overline{u}_{s,0}\right) + O(\varepsilon). \end{aligned} \tag{2.9}$$

This gives the result considering the scaling.

In the case iii), we write the following auxiliary problem

$$\begin{aligned} \partial^2_{\widetilde{z}}\widetilde{u}_s &= O(\varepsilon^2), \\ \partial_{\widetilde{z}}\widetilde{u}_s &= O\left(\varepsilon^2\right), && \text{at } \widetilde{z}=\widetilde{\zeta} \\ \partial_{\widetilde{z}}\widetilde{u}_s &= \varepsilon\widetilde{\kappa}_{\widetilde{B}}\widetilde{u}_s + O(\varepsilon^2), && \text{at } \widetilde{z}=\widetilde{B}. \end{aligned}$$

It yields that $\widetilde{u}_s = O(\varepsilon)$ and $\widetilde{u}_s$ does not depend on z up to the order ε^2, i.e.

$$\widetilde{u}_s\left(\widetilde{\mathbf{x}},\widetilde{z},\widetilde{t}\right) = \varepsilon\overline{u}_{s,1}\left(\widetilde{\mathbf{x}},\widetilde{t}\right) + O\left(\varepsilon^2\right). \tag{2.10}$$

Since $\partial_{\widetilde{z}}\widetilde{u}_s = O\left(\varepsilon^2\right)$, integrating the vertical momentum equation leads to the hydrostatic relation (2.8) (up to the order ε^2). Next, we integrate the horizontal momentum equation between $\widetilde{B}$ and $\widetilde{\zeta}$; we get

$$\begin{aligned} &\frac{\Theta_B}{F_r^2}\widetilde{F}_{\widetilde{B}}\left(\overline{u}_{s,1}\right) - \frac{\varepsilon}{R_e}\nabla_{\widetilde{\mathbf{x}}}\cdot\left(\left(\widetilde{\zeta}-\widetilde{B}\right)\mathrm{D}_{\widetilde{\mathbf{x}}}\overline{u}_{s,1}\right) \\ &\quad = -\frac{\widetilde{\zeta}-\widetilde{B}}{F_r^2}\nabla_{\widetilde{\mathbf{x}}}\left(\widetilde{\zeta}+r\widetilde{p}_{\widetilde{\zeta}}\right) - \frac{\Theta_\zeta}{\varepsilon F_r^2}r\widetilde{\kappa}_{\widetilde{\zeta}}\left(\varepsilon\overline{u}_{s,1}-\widetilde{u}_{\widetilde{\zeta}}\right) + O(\varepsilon). \end{aligned} \tag{2.11}$$

Using the orders of magnitude of the Reynolds number $Re = 1$ and of the sediment-water interface large-scale Shields number $\Theta_\zeta = \varepsilon$ allows to further simplify equation (2.11). Thus, the result is obtained.

Note that the scaling of the velocity (2.10) might be not satisfied by the initial condition $\widetilde{u}_s^0 = \frac{u_s^0}{U}$. Introducing the short time $\sigma = \frac{\widetilde{t}}{\varepsilon^2}$, the previous auxiliary problem becomes

$$\begin{aligned} \partial_\sigma\widetilde{u}_s - \frac{1}{R_e}\partial^2_{\widetilde{z}}\widetilde{u}_s &= O\left(\varepsilon^2\right), \\ \partial_{\widetilde{z}}\widetilde{u}_s &= O\left(\varepsilon^2\right), && \text{at } \widetilde{z}=\widetilde{\zeta}, \\ \partial_{\widetilde{z}}\widetilde{u}_s &= O(\varepsilon), && \text{at } \widetilde{z}=\widetilde{B}. \end{aligned}$$

therefore $\widetilde{u}_s$ becomes of the order of ε within a characteristic time of the order of $O(\varepsilon^2)$. □

Models with non-local sediment discharge

The main drawback of the Exner models derived in Proposition 1 is that they do not take into account the viscosity in the sediment layer. To derive models involving an operator accounting for the viscosity, the product between the Reynolds number and the Shields number at the surface of the substratum must be of the order of ε. These models, presented in Proposition 2, are viscous versions of the Spinewine model [138], the Grass model [68] and the model in [58].

Proposition 2. *Assume that* $F_r = 1$, $1 > r = O(1)$ *and one of the following scalings*

i) $R_e = \varepsilon$, $\Theta_\zeta = 1$, $\Theta_B = 1$ *and* $C = 1$ *then* $(\alpha, \beta, \omega) = (0, 0, 1)$,

ii) $R_e = 1$, $\Theta_\zeta = \varepsilon$, $\Theta_B = \varepsilon$ *and* $C = 1$ *then* $(\alpha, \beta, \omega) = (0, 1, 1)$,

iii) $R_e = \varepsilon$, $\Theta_\zeta = \varepsilon$, $\Theta_B = 1$ *and* $C = \varepsilon$ *then* $(\alpha, \beta, \omega) = (1, 1, 1)$.

Then the system (2.3) *with the initial condition* (2.6) *and where* $\widetilde{p}_{\widetilde{\zeta}} = \frac{p_{w|_{z=\zeta}}}{P}$, $\widetilde{u}_{\widetilde{\zeta}} = \frac{u_{w|_{z=\zeta}}}{U}$, *is an approximation of the Navier-Stokes sediment layer with the modeling errors* (2.7).

Proof. The proof is similar to that of Proposition 1. However the auxiliary problem coming from the horizontal momentum equation and the boundary condition reads

$$\begin{aligned} &\partial^2_{\widetilde{z}} \widetilde{u}_s = O\left(\varepsilon^2\right), \\ &\partial_{\widetilde{z}} \widetilde{u}_s = O\left(\varepsilon^2\right), && \text{at } \widetilde{z} = \widetilde{\zeta} \\ &\partial_{\widetilde{z}} \widetilde{u}_s = O\left(\varepsilon^2\right), && \text{at } \widetilde{z} = \widetilde{B} \end{aligned}$$

thus $\widetilde{u}_s$ does not depend on z up to the order ε^2, i.e.

$$\widetilde{u}_s\left(\widetilde{\mathbf{x}}, \widetilde{z}, \widetilde{t}\right) = \overline{u}_{s,0}\left(\widetilde{\mathbf{x}}, \widetilde{t}\right) + \varepsilon \overline{u}_{s,1}\left(\widetilde{\mathbf{x}}, \widetilde{t}\right) + O\left(\varepsilon^2\right).$$

As in the proof of Proposition 1, the initial condition which does not necessary satisfy this scaling vanishes within a characteristic time of the order of $O\left(\varepsilon^2\right)$. Integrating the vertical momentum equation allows to obtain that the pressure is hydrostatic (2.8). The horizontal momentum equation vertically integrated between B and ζ is again equation (2.9).

The first two results i) and ii) are direct simplifications of (2.9) considering the scaling. For the scaling iii), the main terms of (2.9) read

$$\widetilde{\kappa}_{\widetilde{B}} \overline{u}_{s,0} - \nabla_{\widetilde{\mathbf{x}}} \cdot \left(\left(\widetilde{\zeta} - \widetilde{B} \right) \mathrm{D}_{\widetilde{\mathbf{x}}} \overline{u}_{s,0} \right) = 0$$

and it follows that $\overline{u}_{s,0} = 0$. The next term reads

$$\left(\widetilde{\zeta} - \widetilde{B} \right) \nabla_{\widetilde{\mathbf{x}}} \left(\widetilde{\zeta} + \widetilde{p}_{\widetilde{\zeta}} \right) = -\widetilde{F}_{\widetilde{B}} \left(\overline{u}_{s,1} \right) + \nabla_{\widetilde{\mathbf{x}}} \cdot \left(\left(\widetilde{\zeta} - \widetilde{B} \right) \mathrm{D}_{\widetilde{\mathbf{x}}} \overline{u}_{s,1} \right) + \widetilde{\kappa}_{\widetilde{\zeta}} \widetilde{u}_{\widetilde{\zeta}}.$$

As previously mentioned, the term $-\varepsilon \widetilde{\kappa}_{\widetilde{\zeta}} \overline{u}_{s,1}$ is added to the right hand side without modification of the modeling error, see Remark 1. □

In the following, we will focus on model (2.3) in the case where $(\alpha, \beta, \omega) = (1, 1, 1)$, i.e. Proposition 2.iii). From the physical point of view, this model seems interesting because throughout its derivation, it emerges that the solid velocity is much smaller that the water velocity, which corresponds to observations, and it takes into account viscous effects in the sediment layer. From the mathematical point of view, this model is much more complex than it seems. Even if this model can naively be considered as simple as the model with inertia $(\alpha, \beta, \omega) = (0, 1, 1)$, for which numerical strategies are already proposed in the literature [1, 33, 138], it is not the case. In particular it is well known that the numerical scheme for the case with inertia $(\alpha, \beta, \omega) = (0, 1, 1)$ does not give satisfactory results in the regime where the inertia terms are negligible, see [29], and a first step is to propose a numerical scheme for the asymptotic model without inertia. Let us first rewrite it in the dimensional framework by multiplying the equation for the conservation of mass by $\frac{HU}{L}$ and the momentum balance by $\frac{HU^2}{L}$. It reads

$$\begin{cases} \partial_t b + \nabla_{\mathbf{x}} \cdot (bv) = 0 \ , \\ \qquad f_B\left(b, v, \tau\right) = \tau\left(b, v\right) \end{cases} \tag{2.12}$$

with τ a dimensional version of (2.5), i.e.

$$\tau\left(b, v\right) = -r\kappa_\zeta\left(v - u_\zeta\right) - b\nabla_{\mathbf{x}}\left(g\left(b + B\right) + \frac{p_\zeta}{\rho_s}\right) + \nabla_{\mathbf{x}} \cdot \left(2\mu_s b \mathrm{D}_{\mathbf{x}} v\right) \tag{2.13}$$

and $f_B\left(b, v, \tau\right)$ is a dimensional version of (2.4). In our case, where the inertia of the sediment layer is neglected, the friction at the substratum can be reformulated as

$$f_B\left(b, v, \tau\right) = \begin{cases} \left(\tau_c + \kappa_B \left\|v\right\|^\gamma\right) \dfrac{v}{\left\|v\right\|} & \text{if } \left\|\tau\right\| > \tau_c(b, p_\zeta) \\ \tau + v & \text{if } \left\|\tau\right\| \leq \tau_c(b, p_\zeta) \end{cases}$$

with the critical shear stress $\tau_c\left(b, p_\zeta\right) = \overline{\tau}\left(gb + \frac{p_\zeta}{\rho_s}\right)$. Note that for $\mu_s = 0$ the solid velocity reads

$$v = \left(\frac{\left\|\tau\right\| - \tau_c}{\kappa_B}\right)_+^{\frac{1}{\gamma}} \frac{\tau}{\left\|\tau\right\|}.$$

We recover a power law with a threshold similar to the laws used by hydraulic engineers, see [103, 113].

In particular, let us denote by $\hat{b}$ the solution of the system without viscosity $\mu_s = 0$. Neglecting the term $\kappa_\zeta v$ (which is indeed negligible as shown in the derivation), the shear stress τ does not depend on the solid velocity v, thus it can be computed explicitly. Keeping the term $\kappa_\zeta v$, it is no more possible to obtain an explicit formula for the solid velocity because it is solution of the non-linear problem

$$\kappa_B \left\|\hat{v}\right\|^\gamma + r\kappa_\zeta \left\|\hat{v}\right\| - \left(\left\|\hat{\tau}\right\| - \tau_c\left(\hat{b}, p_\zeta\right)\right)_+ = 0 \qquad \text{and} \qquad \hat{v} = \left\|\hat{v}\right\| \frac{\hat{\tau}}{\left\|\hat{\tau}\right\|} \tag{2.14}$$

with

$$\hat{\tau} = r\kappa_\zeta u_\zeta - \hat{b}\nabla_{\mathbf{x}}\left(g\left(\hat{b}+B\right)+\frac{p_\zeta}{\rho_s}\right).$$

However, as long as $\gamma > 0$ and using a monotonicity argument, it is clear that there exists an unique positive solution $\|\hat{v}\|$ to the problem (2.14).

In the case of a viscous sediment layer $\mu_s > 0$, the well-posedness is not as clear because τ is a non-local function of the solid velocity v. A few properties of the model (2.12) are presented in §2.2.1 but the full analysis of model (2.12) is out of the scope of the present paper. The computation of the solution of (2.12) is not at all trivial and this work will focus on this point further below in §2.2.2. The analysis (i.e. the proof of the existence and uniqueness of the solution) is let as a perspective of the current work, see §2.5.

Analysis

In this section, the main physical properties of model (2.12) are shown in order to improve the relevance of this model. In particular, an energy balance is proved.

Proposition 3. *Assume that the initial condition is positive, i.e.* $b^0(\mathbf{x}) \geq 0$*, and let* $b \in C^0$*. As long as the solid velocity stays bounded, i.e.* $v \in L^\infty$*, the solution is positive, i.e.* $b(\mathbf{x},t) \geq 0$*.*

Proof. This result is a classical result coming from the continuity equation. Multiplying the continuity equation by $\mathbb{1}_{b_-}$, with $\mathbb{1}_\phi$ is the indicator of the support of ϕ and $b_- = \min(0,b)$, we get

$$\partial_t \|b_-\|_{L^1} = \int_{\mathbb{R}} \left|bv\cdot\nabla_{\mathbf{x}}\mathbb{1}_{b_-}\right| \mathrm{d}x \leq \int_{\mathbb{R}} \left|b\nabla_{\mathbf{x}}\mathbb{1}_{b_-}\right| \mathrm{d}x \, \|v\|_{L^\infty}.$$

Assuming that v is bounded, the right hand side vanishes by continuity of b. We conclude by considering the initial condition. □

Before stating the energy balance, we introduce the function $\mathcal{E}$, giving the potential energy of a column of fluid of height h placed upon a topography at elevation B

$$\mathcal{E}(h,B) = gh\left(\frac{h}{2}+B\right). \tag{2.15}$$

Proposition 4. *For smooth enough solutions, the mechanical energy of* (2.12) *satisfies the following energy balance*

$$\begin{aligned}\partial_t\left(\mathcal{E}_s + b\tfrac{p_\zeta}{\rho_s}\right) + \nabla_{\mathbf{x}}\cdot\left(\left(g\left(b+B\right)+\tfrac{p_\zeta}{\rho_s}\right)bv\right) - \nabla_{\mathbf{x}}\cdot\left(2\mu_s bv\cdot D_{\mathbf{x}}v\right)\\ = b\partial_t\frac{p_\zeta}{\rho_s} - v\cdot f_B - r\kappa_\zeta v\cdot(v-u_\zeta) - 2\mu_s b(D_{\mathbf{x}}v):(D_{\mathbf{x}}v),\end{aligned}$$

with $\mathcal{E}_s = \mathcal{E}(b,B)$*.*

Proof. Let us multiply the continuity equation of (2.12) by $g\,(b+B)+\frac{p_\zeta}{\rho_s}$, while the equation on the velocity is multiplied by v. Combining the two equations gives the result. □

The mechanical energy of the sediment layer is made only of its potential energy. Note that the estimate of Proposition 4 is a dissipation law in the sense that without forcing, i.e. $u_\zeta = \partial_t p_\zeta = 0$, the total mechanical energy decreases, i.e. $\partial_t \int_{\mathbb{R}^d} \mathcal{E}_s \,\mathrm{d}\mathbf{x} \leq 0$, since the friction term is dissipative, i.e. $v \cdot f_B \geq 0$ and $(\mathrm{D}_\mathbf{x} v) : (\mathrm{D}_\mathbf{x} v) \geqslant 0$ because the matrix $\mathrm{D}_\mathbf{x} v$ is symmetric.

2.2.2 Numerical scheme

In the present section, we propose a numerical strategy to solve (2.12) in one dimension. A staggered grid discretization is used to limit the size of the stencil of the viscosity operator. Such a staggered grid discretization was already used for the SWE in [75, 114] and for the shallow water-Exner system in [55]. A linear system is solved to obtain the velocity at the faces between the control volumes. This strategy was already proposed in a simpler case in [66] and a multidimensional version [65], to approximate the solution of the Schurtz-Nicolaï model [117] in plasma physics.

Let us consider a Cartesian grid of points $x_{i+1/2} = \delta_x \left(i + \frac{1}{2}\right)$ with $\delta_x > 0$ the constant space step. The numerical unknown b_i^{n+1} is the approximation of the sediment thickness b averaged in a cell $]x_{i-1/2}, x_{i+1/2}[$ at time $t^{n+1} = t^n + \delta_t^n$, where δ_t^n is an adaptative time step defined later on. In addition, $v_{i+1/2}^n$ is an approximation of the solid velocity v at the interface $x_{i+1/2}$ and at time t^n. For readability purposes, the following centered discrete operators are used

$$\partial_{i+1/2}^\delta : \begin{array}{ccc} \mathbb{R}^{N_x} & \to & \mathbb{R} \\ (\psi_j)_{1\leq j\leq N_x} & \mapsto & \dfrac{\psi_{i+1}-\psi_i}{\delta_x} \end{array} \quad \text{and} \quad \partial_i^\delta : \begin{array}{ccc} \mathbb{R}^{N_f} & \to & \mathbb{R} \\ \left(\psi_{j+1/2}\right)_{1\leq j\leq N_f} & \mapsto & \dfrac{\psi_{i+1/2}-\psi_{i-1/2}}{\delta_x} \end{array}$$

with N_x the number of cells and N_f the number of interfaces of the grid. Let us set the forcing terms $u_{i+1/2}^n$, p_i^n and $p_{i+1/2}^n$ respectively defined by the values (or an approximation of) $u_\zeta\left(x_{i+1/2}, t^n\right)$, $p_\zeta\left(x_i, t^n\right)$ and $p_\zeta\left(x_{i+1/2}, t^n\right)$. The discrete friction coefficient at the substratum-sediment interface reads $\kappa_{B,i+1/2} = \kappa_B\left(x_{i+1/2}\right)$. The discrete friction coefficient at the interface depends on the velocities u_ζ, v and on the pressure at the interface p_ζ, i.e.

$$\kappa_{\zeta,i+1/2}^n = \kappa_\zeta\left(p_{i+1/2}^n, u_{i+1/2}^n - v_{i+1/2}^n\right). \tag{2.16}$$

Numerical scheme for the model with non-local sediment discharge

Let us focus on the numerical resolution of the non-local model (2.12). Indeed, the main objective of the present work is to characterize the behavior of system (2.12). As will be shown later, the scheme for the non-local model degenerates towards a scheme for the

model with local sediment discharge as the viscosity μ_s goes to 0. This is a desirable property, since the Exner system with a local sediment discharge has shown its relevance in many situations, see the examples given in [53] and [135].

We choose the following discretization for the continuity equation of (2.12)

$$b_i^{n+1} = b_i^n - \delta_t^n \partial_i^\delta (b^n v^{n+1}). \tag{2.17}$$

The choice of an implicit discretization is motivated in Remark 3. Details about the computation of the numerical solution are given below. A discretization of the second equation of (2.12) reads

$$f_{B,i+1/2}^{n+1} = \tau_{i+1/2}^n \tag{2.18}$$

The friction at the surface of the substratum is

$$f_{B,i+1/2}^{n+1} = \begin{cases} \tau_{c,i+1/2}^n \text{sign}\left(v_{i+1/2}^{n+1}\right) + \kappa_{B,i+1/2} |v_{i+1/2}^{n+1}|^{\gamma-1} v_{i+1/2}^{n+1} & \text{if } \|\tau_{i+1/2}^{n+1}\| > \tau_{c,i+1/2}^n, \\ \tau_{i+1/2}^n + v_{i+1/2}^{n+1} & \text{if } \|\tau_{i+1/2}^{n+1}\| \leqslant \tau_{c,i+1/2}^n \ . \end{cases}$$

The critical shear stress is defined by

$$\tau_{c,i+1/2}^n = \tau_c(b_{i+1/2}^n, p_{\zeta,i+1/2}^n). \tag{2.19}$$

and the effective shear stress is defined by

$$\tau_{i+1/2}^{n+1} = -r\kappa_{\zeta,i+1/2}^n (v_{i+1/2}^{n+1} - u_{\zeta,i+1/2}^{n+1}) - b_{i+1/2}^n \partial_{i+1/2}^\delta \phi^{n+1} + \partial_{i+1/2}^\delta (2\mu_s b^n \partial^\delta v^{n+1}). \tag{2.20}$$

The discrete potential is

$$\phi_i^{n+1} = g(b_i^{n+1} + B_i) + \frac{p_{\zeta,i}^{n+1}}{\rho_s}. \tag{2.21}$$

An upwind reconstruction is adopted for the sediment depth at the interface

$$b_{i+1/2}^n = \begin{cases} b_i^n & \text{if } v_{i+1/2}^{n+1} > 0, \\ \frac{b_i^n + b_{i+1}^n}{2} & \text{if } v_{i+1/2}^{n+1} = 0, \\ b_{i+1}^n & \text{if } v_{i+1/2}^{n+1} < 0. \end{cases} \tag{2.22}$$

Upwinding is required to ensure the positivity and the entropy stability of the solution under an hyperbolic CFL condition, see Proposition 5. More precisely, we will see that the positivity of the thickness b_i^n is ensured at convergence if the time step satisfies the following implicit CFL condition

$$\lambda_s^{n+1} \delta_t^n \leqslant \frac{\delta_x}{2}, \tag{2.23}$$

with

$$\lambda_s^{n+1} = \max_{1 \leq i \leq Nf} \left(v_{i+1/2}^{n+1}\right). \tag{2.24}$$

Proposition 5. *Assume that the initial condition is non-negative, i.e.* $b_i^0 \geqslant 0$*, and that the time step satisfies the CFL condition* (2.23)*. Then the solution of the scheme* (2.17)*,* (2.18) *is non-negative, i.e.* $b_i^n \geqslant 0$*, and the solution satisfies a dissipation law of the mechanical energy*

$$\begin{aligned}
&\frac{\left(\mathcal{E}_{s,i}^{n+1} + b_i^{n+1}\frac{p_{\zeta,i}^{n+1}}{\rho_s}\right) - \left(\mathcal{E}_{s,i}^{n} + b_i^{n}\frac{p_{\zeta,i}^{n}}{\rho_s}\right)}{\delta_t^n} + \partial_i^\delta\left(\phi^{n+1}b^n v^{n+1} - \mathcal{M}^{n+1}\right) \\
&\leqslant \frac{b_i^n}{\rho_s}\left(\frac{p_{\zeta,i}^{n+1} - p_{\zeta,i}^{n}}{\delta_t^n}\right) - \frac{v_{i+1/2}^{n+1}}{2} f_{B,i+1/2}^{n+1} - r\frac{\kappa_{\zeta,i+1/2}^n}{2} v_{i+1/2}^{n+1}(v_{i+1/2}^{n+1} - u_{\zeta,i+1/2}^{n+1}) \\
&\qquad - \frac{v_{i-1/2}^{n+1}}{2} f_{B,i-1/2}^{n+1} - r\frac{\kappa_{\zeta,i-1/2}^n}{2} v_{i-1/2}^{n+1}(v_{i-1/2}^{n+1} - u_{\zeta,i-1/2}^{n+1}) \\
&\qquad -2\mu_s \frac{b_{i+1}^n\left(\partial_{i+1}^\delta v^{n+1}\right)^2 + 2b_i^n\left(\partial_i^\delta v^{n+1}\right)^2 + b_{i-1}^n\left(\partial_{i-1}^\delta v^{n+1}\right)^2}{4}
\end{aligned} \tag{2.25}$$

with

$$\mathcal{E}_{s,i}^n = \mathcal{E}(b_i^n, B_i),$$

and

$$(\mathcal{M}^{n+1})_{i+1/2} = 2\mu_s \frac{b_{i+1}^n(v_{i+1/2}^{n+1} + v_{i+3/2}^{n+1})\partial_{i+1}^\delta v^{n+1} + b_i^n(v_{i-1/2}^{n+1} + v_{i+1/2}^{n+1})\partial_i^\delta v^{n+1}}{4}.$$

Proof. Assume that at the time iteration n, the solution is non-negative. Under the CFL condition (2.23), the non-negativity of b_i^{n+1} is a classical property of the upwind scheme.

Let us now focus on the mechanical energy. Equation (2.17) is multiplied by ϕ_i^{n+1}. We get

$$\begin{aligned}
\frac{\left(\mathcal{E}_{s,i}^{n+1} + b_i^{n+1}\frac{p_{\zeta,i}^{n+1}}{\rho_s}\right) - \left(\mathcal{E}_{s,i}^{n} + b_i^{n}\frac{p_{\zeta,i}^{n}}{\rho_s}\right)}{\delta_t^n} &+ \partial_i^\delta\left(\phi^{n+1}b^n v^{n+1}\right) = -\frac{(b_i^{n+1} - b_i^n)^2}{2\delta_t^n} + \frac{b_i^n}{\rho_s}\left(\frac{p_{\zeta,i}^{n+1} - p_{\zeta,i}^{n}}{\delta_t^n}\right) \\
&+ b_{i+1/2}^n v_{i+1/2}^{n+1}\frac{\partial_{i+1/2}^\delta \phi^{n+1}}{2} + b_{i-1/2}^n v_{i-1/2}^{n+1}\frac{\partial_{i-1/2}^\delta \phi^{n+1}}{2},
\end{aligned} \tag{2.26}$$

where the reconstruction $\phi_{i+1/2}^{n+1}$ is defined as

$$\phi_{i+1/2}^{n+1} = \frac{\phi_i^{n+1} + \phi_{i+1}^{n+1}}{2}.$$

An expression for the mechanical work term is obtained by multiplying equation (2.18)

by $v_{i+1/2}^{n+1}$

$$r\kappa_{\zeta,i+1/2}^{n}\left(v_{i+1/2}^{n+1}\right)^2+v_{i+1/2}^{n+1}f_{B,i+1/2}^{n+1}-v_{i+1/2}^{n+1}\partial_{i+1/2}^{\delta}(2\mu_s b^n\partial^{\delta}v^{n+1})$$
$$-r\kappa_{\zeta,i+1/2}^{n}u_{\zeta,i+1/2}^{n+1}v_{i+1/2}^{n+1}=-b_{i+1/2}^{n}v_{i+1/2}^{n+1}\partial_{i+1/2}^{\delta}\phi^{n+1},$$

which is then substituted in equation (2.26). Now, it remains to prove that the term $\frac{1}{2}(v_{i+1/2}^{n+1}\partial_{i+1/2}^{\delta}(2\mu_s b^n\partial^{\delta}v^{n+1})+v_{i-1/2}^{n}\partial_{i-1/2}^{\delta}(2\mu_s b^n\partial^{\delta}v^{n+1}))$ is indeed dissipative. Expanding this term and using the identity $a(a-b)=a^2/2-b^2/2+(a-b)^2/2$ allows to write

$$\begin{aligned}
&\frac{1}{2}\left(v_{i+1/2}^{n+1}\partial_{i+1/2}^{\delta}(2\mu_s b^n\partial^{\delta}v^{n+1})+v_{i-1/2}^{n}\partial_{i-1/2}^{\delta}(2\mu_s b^n\partial^{\delta}v^{n+1})\right)\\
&\quad=\frac{\mu_s}{2\delta_x^2}\Big(b_{i+1}^{n}\left((v_{i+3/2}^{n+1})^2-(v_{i+1/2}^{n+1})^2-(v_{i+3/2}^{n+1}-v_{i+1/2}^{n+1})^2\right)\\
&\qquad+b_{i}^{n}\left((v_{i+1/2}^{n+1})^2-(v_{i-1/2}^{n+1})^2-(v_{i+1/2}^{n+1}-v_{i-1/2}^{n+1})^2\right)\\
&\qquad-b_{i}^{n}\left((v_{i+1/2}^{n+1})^2-(v_{i-1/2}^{n+1})^2+(v_{i+1/2}^{n+1}-v_{i-1/2}^{n+1})^2\right)\\
&\qquad-b_{i-1}^{n}\left((v_{i-1/2}^{n+1})^2-(v_{i-3/2}^{n+1})^2+(v_{i-1/2}^{n+1}-v_{i-3/2}^{n+1})^2\right)\Big)
\end{aligned}$$

and then

$$\begin{aligned}
&\frac{1}{2}\left(v_{i+1/2}^{n+1}\partial_{i+1/2}^{\delta}(2\mu_s b^n\partial^{\delta}v^{n+1})+v_{i-1/2}^{n+1}\partial_{i+1/2}^{\delta}(2\mu_s b^n\partial^{\delta}v^{n+1})\right)\\
&\quad=2\frac{\mu_s}{\delta_x}\left(b_{i+1}\frac{v_{i+1/2}^{n+1}+v_{i+3/2}^{n+1}}{4}\partial_{i+1}^{\delta}v^{n+1}-b_{i-1}\frac{v_{i-1/2}^{n+1}+v_{i+1/2}^{n+1}}{4}\partial_{i-1}^{\delta}v^{n+1}\right)\\
&\qquad-2\mu_s\frac{b_{i+1}^{n}\left(\partial_{i+1}^{\delta}v^{n+1}\right)^2+2b_{i}^{n}\left(\partial_{i}^{\delta}v^{n+1}\right)^2+b_{i-1}\left(\partial_{i-1}^{\delta}v^{n+1}\right)^2}{4}.
\end{aligned}$$

□

Note that the computation of $v_{i+1/2}^{n+1}$ is not obvious, since it depends on b_i^{n+1} which has yet to be computed. Equations (2.17) and (2.18) form a system with N_f+N_x unknowns. This system can be reduced to a system with only N_f unknowns by replacing b_i^{n+1} in (2.18) using the scheme (2.17), which gives

$$\begin{aligned}
&(r\kappa_{\zeta,i+1/2}^{n}+\kappa_{B,i+1/2}|v_{i+1/2}^{n+1}|^{\gamma-1})v_{i+1/2}^{n+1}-\partial_{i+1/2}^{\delta}(2\mu_s b^n\partial^{\delta}v^{n+1})-g\delta_t^{n}b_{i+1/2}^{n}(\partial^{\delta}(b^n v^{n+1}))\\
&\qquad=-gb_{i+1/2}^{n}\partial_{i+1/2}^{\delta}\phi^{n+1}-\tau_{c,i+1/2}^{n}sign(v_{i+1/2}^{n+1}),\quad(2.27)
\end{aligned}$$

if $|\tau_{i+1/2}^{n+1}|>\tau_{c,i+1/2}^{n}$, and $v_{i+1/2}^{n+1}=0$ otherwise. The system described by (2.27) is actually nonlinear, because the reconstruction $b_{i+1/2}^{n}$ and the shear $\tau_{i+1/2}^{n}$ depend on $v_{i+1/2}^{n+1}$. In practice, a fixed-point method is used. The presence of the threshold makes the system stiff, hence the choice of the Newton fixed-point method. The velocity is initialized

with $v_{i+1/2}^{n,0} = 0$ for all $1 \leqslant i \leqslant N_f$. At convergence, $v_{i+1/2}^{n+1} = \lim_{q\to+\infty} v_{i+1/2}^{n,q}$. The convergence criterion used to estimate the convergence of the iterative process is based on the l^∞-norm of the variation of the solution $b^{n,q}$ between two successive iterations. At each iteration, the shear stress $\tau_{i+1/2}^{n,q}$ is compared to the critical shear stress $\tau_{c,i+1/2}^{n,q}$. If $||\tau_{i+1/2}^{n,q}|| \leqslant \tau_{c,i+1/2}^{n,q}$, the corresponding lines in the matrix and in the right-hand side of the system are modified. In the line i of the matrix, the diagonal coefficient is set to 1 while the others are set to 0. The line i in the right-hand side is set to 0. At each iteration, the time step is estimated using the relation

$$\delta_t^{n,q} = \frac{\lambda \delta_x}{2 v_{i+1/2}^{n,q}},$$

with $v_{i+1/2}^{n,q}$ the velocity at the iteration q and with a given $\lambda \leqslant 1$ to satisfy the CFL condition (2.23).

Remark 2. Initializing the fixed-point iteration with $v_{i+1/2}^{n,0} = v_{i+1/2}^{n}$ is likely to improve the performance of the fixed-point method. Yet, this could also introduce memory effects in the solution which are not present in the second equation of (2.12).

Scheme for the local sediment discharge

In this paragraph, we show how the scheme for the non-local model degenerates towards a scheme for the local model when the viscosity goes to zero. The transport threshold $\overline{\tau}$ is set to 0 for the sake of readability. A hat is put on the variables of the local scheme to distinguish them from those of the non-local scheme. When μ_s goes to 0, the scheme (2.17), (2.18) is formally equivalent to the scheme

$$\hat{b}_i^{n+1} = \hat{b}_i^n + \delta_t^n \partial_i^\delta \left(g \frac{(\hat{b}^n)^2}{r\hat{\kappa}_\zeta^n + \kappa_B |\hat{v}^{n+1}|^{\gamma-1}} \partial_i^\delta \hat{\phi}^{n+1} - \hat{b}^n \left(\frac{r\hat{\kappa}_\zeta^n u_\zeta^{n+1}}{r\hat{\kappa}_\zeta^n + \kappa_B |\hat{v}^{n+1}|^{\gamma-1}} \right) \right), \quad (2.28)$$

where the velocity $\hat{v}_{i+1/2}^{n+1}$ is naturally defined by

$$\hat{v}_{i+1/2}^{n+1} = -g \frac{\hat{b}_{i+1/2}^n}{r\hat{\kappa}_{\zeta,i+1/2}^n + \kappa_{B,i+1/2} |\hat{v}_{i+1/2}^{n+1}|^{\gamma-1}} \partial_{i+1/2}^\delta \hat{\phi}^{n+1} + \left(\frac{r\hat{\kappa}_{\zeta,i+1/2}^n u_{\zeta,i+1/2}^{n+1}}{r\hat{\kappa}_{\zeta,i+1/2}^n + \kappa_{B,i+1/2} |\hat{v}_{i+1/2}^{n+1}|^{\gamma-1}} \right)$$

The discrete variables $\hat{\kappa}_{\zeta,i+1/2}^n$, $\hat{\tau}_{c,i+1/2}^n$, $\hat{\tau}_{i+1/2}^n$ $\hat{\phi}_i^{n+1}$ are respectively defined as in the equations (2.16), (2.19), (2.20), (2.21). The reconstruction $\hat{b}_{i+1/2}^n$ is an upwind reconstruction defined as in (2.22).

A system is solved for the unknown $\hat{b}^{n+1}$, which is the one quantity in which we are interested, while in the non-local scheme a system is solved for the velocity v^{n+1}. Again, as equation (2.28) describes a non-linear system, a fixed-point method is used. To initialize the method, we set $\hat{v}_{i+1/2}^{n,0} = 0$ for all $1 \leqslant i \leqslant N_f$.

Remark 3. Equation (2.28) can be seen as a discretization of a nonlinear convection-diffusion equation

$$\partial_t \hat{b} + \partial_x \left(\hat{A}\hat{b} - \hat{D}\partial_x \hat{\phi} \right) = 0, \tag{2.29}$$

with

$$\hat{A} = \frac{r\kappa_\zeta u_\zeta}{r\hat{\kappa}_\zeta + \kappa_B ||\hat{v}||^{\gamma-1}},$$
$$\hat{D} = g\frac{\hat{b}^2}{r\hat{\kappa}_\zeta + \kappa_B |\hat{v}||^{\gamma-1}}.$$

It is well-known that at least the diffusion part of the equation has to be discretized with an implicit scheme to be stable without a restrictive (parabolic) condition on the time step.

The scheme for the local sediment discharge (2.28) satisfies Proposition 5 with $\mu_s = 0$. Moreover, the scheme for the local sediment discharge is stable under the CFL condition (2.23). Yet the CFL condition (2.23) is not the classical CFL condition of an advection-diffusion equation such as (2.28), see §2.4.

Boundary conditions

We briefly discuss here the implementation of several types of boundary conditions. The computational domain is made of N_x cells. The boundary conditions on b^n are implemented by means of ghost cells numbered 0 and N_x+1. As regards v^{n+1}, boundary edges indexed by $1/2$, $N_x + 1/2$ are used. Moreover, depending on the type of boundary condition chosen for v^{n+1}, the matrix defined by (2.18) is modified. The boundary conditions used in practice depend on the test case and will be specified for each test case.

2.2.3 Numerical validation

In this section, the behavior of the numerical scheme described in §2.2.2 is illustrated. For all the test cases, except when indicated otherwise, the parameters are set to $g = 9.81$, $\gamma = 1$, $\mu_s = 0.5$, $\overline{\tau} = 0$. The friction coefficients $\kappa_{B,i+1/2}$ and $\kappa_{\zeta,i+1/2}$ are constant in space. For the bottom friction coefficient, $\kappa_{B,i+1/2} = 1$ for all $1 \leqslant i \leqslant N_f$. The length of the domain is 1.

Synthetic forcing

The convergence of the scheme presented in §2.2.2 towards an analytical solution is studied. Though no analytical solution exists in the general case, one may recover an analytical solution by imposing the adequate forcing. More precisely, considering $p_\zeta = 0$, we look for the forcing u_ζ such that the thickness b is stationary, i.e. $b = \beta(x)$. The continuity equation implies that $\beta v = Q$, with Q a constant. The fact that $Q \neq 0$ implies

that the shear stress is larger than the threshold value and the second equation of (2.12) gives an expression for the forcing velocity

$$u_\zeta = \frac{1}{r\kappa_\zeta}\left(g\tau_c\beta + \kappa_B\left(\frac{Q}{\beta}\right)^\gamma + r\kappa_\zeta\frac{Q}{\beta} + g\beta(\beta' + B') + 2\mu_s Q\frac{\beta\beta'' - (\beta')^2}{2}\right).$$

This strategy is applied to assess the convergence of the numerical scheme. We set

$$\beta(x) = 1 + 0.1\sin(2\pi x) \qquad \text{and} \qquad Q = 1.$$

The friction coefficient κ_ζ is constant in space and in time, its value is $\kappa_\zeta = 10^{-3}$. The bottom is flat, i.e. $B(x) = 0$ and the initial condition is $b^0 = 1$. Dirichlet boundary conditions are imposed: the values imposed at the boundaries are those of the analytical solution, i.e. $b_0^n = 1 + 0.1\sin\left(2\pi\left(\frac{-\delta_x}{2}\right)\right)$, $b_{Nx+1}^n = 1 + 0.1\sin\left(2\pi\left(L + \frac{\delta_x}{2}\right)\right)$.

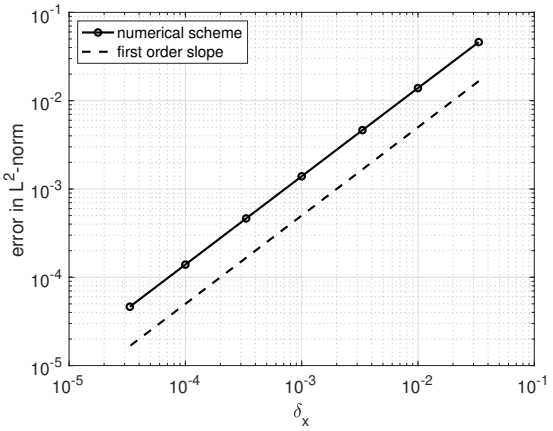


Figure 2.2 – §2.2.3 Convergence towards an analytical solution

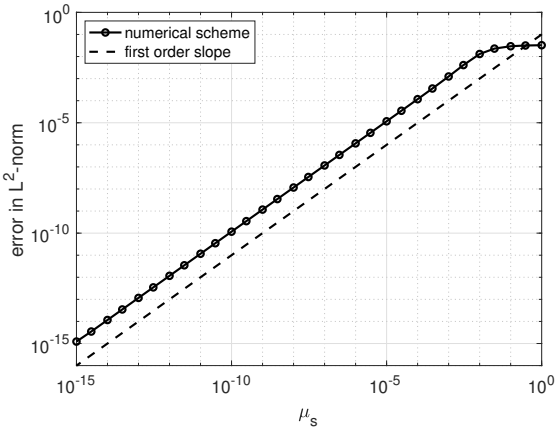


Figure 2.3 – §2.2.3 Convergence towards the local scheme

In Figure 2.2, the errors in L^2-norm are plotted at $T = 20$ for several values of δ_x. The time T is long enough for the numerical solution to reach the stationary regime. As expected because of the use of an upwind reconstruction in the continuity equation, the convergence order of the scheme is 1.

Asymptotic behavior

As explained in §2.2.2, the scheme (2.17), (2.18) converges to the scheme (2.28) as μ_s goes to 0. In this section, the convergence of (2.17), (2.18) towards (2.28) with respect to μ_s is studied numerically. There is no forcing term: $u_\zeta = 0, p_\zeta = 0$. Coherently, the friction at the interface is $\kappa_\zeta = 0$. The bottom is flat, i.e. $B(x) = 0$. The initial condition

is described by

$$b^0(x) = \begin{cases} 1 & \text{if} \quad x < 0.4 \\ 1.1 & \text{if} \quad 0.4 \leqslant x \leqslant 0.6 \\ 1 & \text{if} \quad x > 0.6 \end{cases} \tag{2.30}$$

Wall boundary conditions are imposed. The solutions given by the non-local scheme (2.17), (2.18) for decreasing viscosity values are compared to the solution yielded by (2.28) at the time $T = 2 \times 10^{-3}$. The final time T is chosen such that the sediment bump is not entirely flat at T even for the lower viscosity values, which allows to quantify the differences between the solutions. The space step is $\delta_x = 10^{-3}$. To prevent the error due to the time discretization from degrading the comparison between the two schemes, a time step $\delta_t = 10^{-5}$ is imposed. With this value for the time step, the CFL condition (2.23) is always satisfied - in particular, this value is a valid choice for the lower viscosity values. The results are shown on Figure 2.3. As μ_s goes to 0, the solution of the non-local scheme (2.17), (2.18) converges towards the solution of the local scheme with order 1.

Influence of the viscosity on the shape of the solution

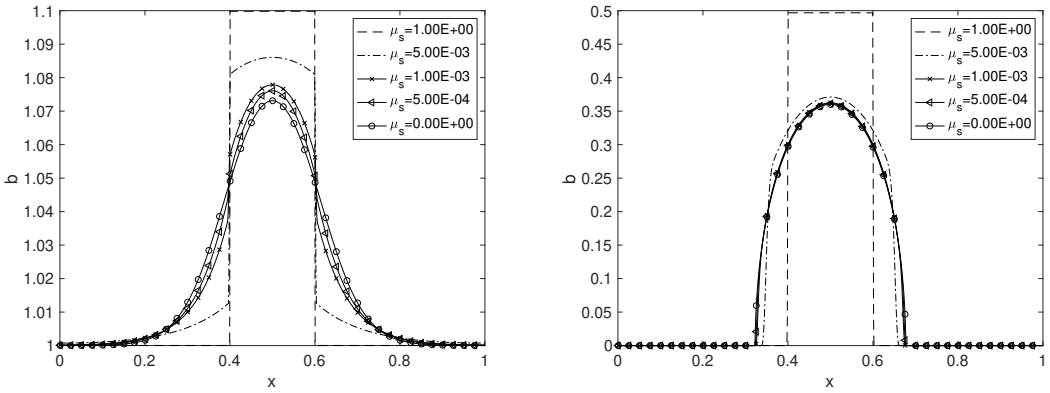
The influence of the viscosity on the behavior of the sediment layer is illustrated. Two cases are investigated, corresponding to two different initial conditions. Solutions are computed for various values of the viscosity μ_s. For all the viscosity values, the space step is $\delta_x = 10^{-3}$. The time step is variable, it is determined using the CFL condition (2.23). In the first case, the initial condition is described by equation (2.30). In the second case, a higher sediment bump is set on a dry bed

$$b^0(x) = \begin{cases} 0 & \text{if} \quad x < 0.4, \\ 0.5 & \text{if} \quad 0.4 \leqslant x \leqslant 0.6, \\ 0 & \text{if} \quad x > 0.6. \end{cases} \tag{2.31}$$

Thus the ability of the scheme (2.17), (2.18) to deal with dry fronts is assessed. The results are shown on Figure 2.4. These results are converged.

When the viscosity is small, the sediment bump quickly becomes flat and smooth. For a very high viscosity, the initial condition is almost preserved; a longer simulation time would be needed to see a change in the shape of the sediment layer. In the case of a wet bed, the higher the viscosity, the longer the discontinuities in the sediment bump are preserved. In the case of a dry bed, the fronts between the wet and dry zones are very sharp.

Note that the scheme (2.17), (2.18) is able to deal with dry fronts. In particular, no oscillations are produced.



(a) Initial condition (2.30), $T = 4 \times 10^{-4}$ (b) Initial condition (2.31), $T = 5 \times 10^{-3}$

Figure 2.4 – §2.2.3 Influence of the viscosity on the shape of the solution - $\delta_x = 10^{-3}$.

Non-flat stationary state

The effect of the introduction of a threshold for the onset of motion is illustrated. The value of the critical shear stress is set to $\overline{\tau} = 1$. The bottom has a constant slope of 0.1, it is described by

$$B(x) = 0.5 - 0.1x.$$

The initial condition is given by

$$b^0(x) = \begin{cases} 0.1 & \text{if} \quad x < 0.4, \\ 0.2 & \text{if} \quad 0.4 \leqslant x \leqslant 0.6, \\ 0.1 & \text{if} \quad x > 0.6. \end{cases}$$

The space step is $\delta_x = 10^{-3}$. Two different simulations are run, one with $\mu_s = 0$ and the other with $\mu_s = 0.5$. The two simulations are run with a constant time step $\delta_t = 10^{-6}$, which is small enough for the CFL condition to be satisfied in both cases. This value is also small enough for the numerical scheme not to "miss" the stationary state. The presence of a threshold allows to obtain non-flat stationary states. Thus angles of repose (which are a property of granular materials) appear in the numerical solutions. The final states are shown on Figure 2.5. Note that the numerical solutions do not reach their final states at the same time. Reaching the final state takes much longer when the viscosity is $\mu_s = 0.5$. The transients of the numerical solutions are not the same either - this was to be expected, given the results in §2.2.3. Yet the stationary state does not seem to depend on the viscosity. It is characterized by $|\tau(b, 0)| \leqslant \tau_c(b, p_\zeta)$. With $\mu_s = 0.5$, the shear stress $\tau(b, 0)$ is equal to the threshold (for the points which have moved during the simulation), while for $\mu_s = 0$, it is slightly below, see Figure 2.6.

The implementation of the threshold raises many theoretical and numerical difficulties which are beyond the scope of the present work. For a discussion of these problems, see

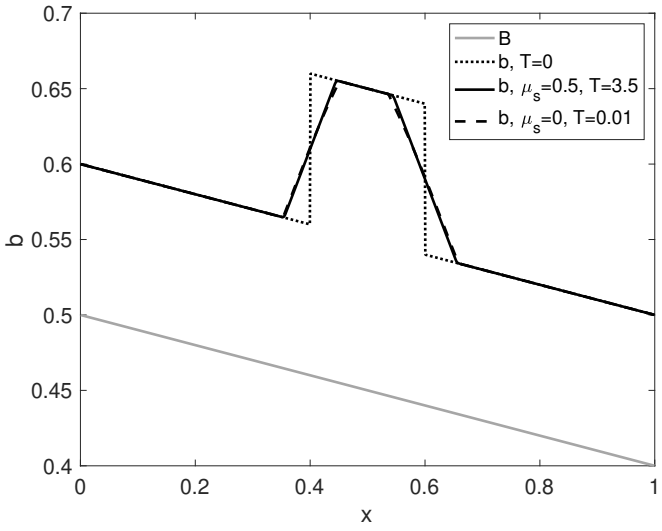


Figure 2.5 – S2.2.3 Non-flat stationary state

for instance [2].

2.3 Coupled water and sediment system

We are now interested in the modeling and simulation of the whole system, made of a water layer and a sediment layer. A derivation for the coupled system is presented, as well as a numerical strategy for solving it. Numerical results are shown.

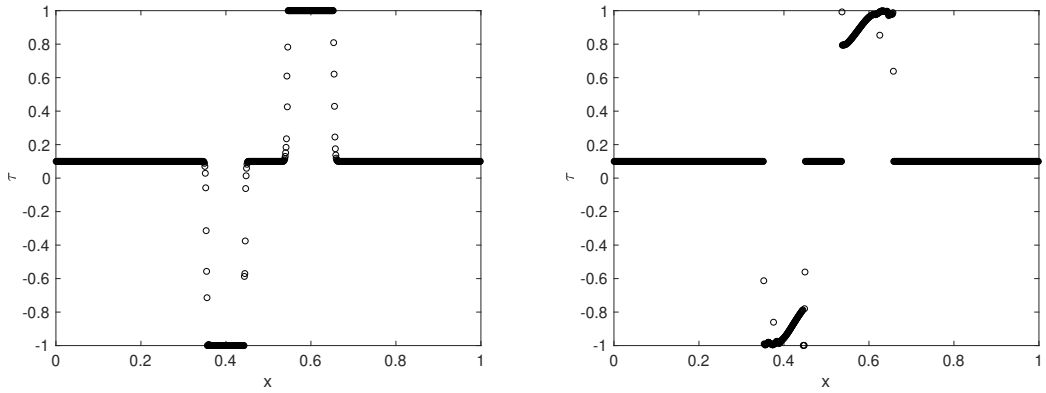
2.3.1 Modeling of the coupled system

The derivation of the models for the sediment layer from the Navier-Stokes equations was already performed in §2.2.1. In order to derive the coupled system, the dimensionless numbers introduced in (2.2) are again considered. Additionally, the Reynolds number in the water layer is introduced

$$Re_w = \frac{\rho_w U L}{2\mu_w},$$

with μ_w the viscosity of the water. For Re_w small enough and one of the sets of parameters described in Proposition 1.ii), Proposition 1.iii), Proposition 2.ii), Proposition 1.iii), the coupled model made of

$$\begin{aligned} \partial_{\widetilde{t}}\widetilde{h} + \nabla_{\widetilde{\mathbf{x}}}\cdot\left(\widetilde{h}\widetilde{u}\right) &= 0\ , \\ \partial_{\widetilde{t}}\left(\widetilde{h}\widetilde{u}\right) + \nabla_{\widetilde{\mathbf{x}}}\cdot\left(\widetilde{h}\widetilde{u}\otimes\widetilde{u} + \frac{\widetilde{h}^2}{2}Id\right) + \widetilde{h}\nabla_{\widetilde{\mathbf{x}}}\left(\widetilde{b}+\widetilde{B}\right) &= -\mathbb{1}_{\widetilde{b}>0}\widetilde{\kappa}_{\widetilde{\zeta}}(\widetilde{u}-\widetilde{v}) - \mathbb{1}_{\widetilde{b}=0}\widetilde{\kappa}_R\widetilde{u}\ , \end{aligned} \tag{2.32}$$



(a) $\mu_s = 0.5$ (b) $\mu_s = 0$

Figure 2.6 – §2.2.3 $\tau(b, 0)$

in the water layer and (2.3) in the sediment layer with the initial conditions

$$\begin{aligned} \widetilde{h}\left(\widetilde{\mathbf{x}}, 0\right) &= \widetilde{\eta}^0\left(\widetilde{\mathbf{x}}\right) - \widetilde{\zeta}\left(\widetilde{\mathbf{x}}\right), \\ \widetilde{u}(\widetilde{\mathbf{x}}, 0) &= \frac{1}{\widetilde{\eta}^0\left(\widetilde{\mathbf{x}}\right) - \widetilde{\zeta}^0\left(\widetilde{\mathbf{x}}\right)} \int_{\widetilde{\eta}^0(\widetilde{\mathbf{x}})}^{\widetilde{\zeta}^0(\widetilde{\mathbf{x}})} \widetilde{u}_w\left(\widetilde{\mathbf{x}}, z\right)^0 dz \ , \end{aligned}$$

and (2.6) is derived from the Navier-Stokes equations with the modeling errors

$$\left|\widetilde{\eta} - \widetilde{\zeta} - \widetilde{h}\right| = O\left(\varepsilon\right) \ , \qquad \left|\widetilde{u}_w - \widetilde{u}\right| = O\left(\varepsilon\right). \tag{2.33}$$

in the water layer and (2.7) in the sediment layer. In the model (2.3) for the sediment layer, we have $\widetilde{u}_{\widetilde{\zeta}} = \widetilde{u}$ and $\widetilde{p}_{\widetilde{\zeta}} = \widetilde{h}$. The derivation of the SWE in the water layer is the classical one [63].

Remark 4. Choosing one of the scalings Proposition 1.ii), iii), Proposition 2.ii), iii) is necessary so that the SWE can be obtained in the water layer. A low friction $\Theta_\zeta = \varepsilon$ at the sediment-water interface is required.

Remark 5. Note that the indicator function $\mathbb{1}_{\widetilde{b}>0}$ was added to the friction term $\widetilde{\kappa}_{\widetilde{\zeta}}\left(\widetilde{u} - \widetilde{v}\right)$ in the model for the water layer. Physically, when the sediment layer vanishes, the water is directly in contact with the bedrock. The friction between the water and the rock in the absence of sediment is modeled by the term $\mathbb{1}_{\widetilde{b}=0}\widetilde{\kappa}_R\widetilde{u}$. In what follows, for the sake of lightness, this subtlety on the modeling of the friction is omitted and the friction is simply written $\widetilde{\kappa}_{\widetilde{\zeta}}\left(\widetilde{u} - \widetilde{v}\right)$.

Remark 6. The orders of approximations are different in the two layers. How- ever, the fact that the order of approximation in the water layer is ε does not prevent the approximation from being of the order of ε^2 in the sediment layer. Indeed, in the second equation of (2.1), u is multiplied by κ_ζ, which is of the order of ε.

Once again, we focus on model Proposition 2.iii) in the sediment layer. From now on, we work with the dimensional bilayer system with model Proposition 2.iii) in the sediment layer. The coupled model is

$$\begin{cases} \partial_t h + \nabla_{\mathbf{x}} \cdot (hu) = 0, \\ \partial_t (hu) + \nabla_{\mathbf{x}} \cdot \left(hu \otimes u + g\dfrac{h^2}{2} Id \right) = -gh\nabla_{\mathbf{x}}(b+B) - \kappa_\zeta (u-v), \end{cases} \tag{2.34}$$

coupled with (2.12). The dimensional expression of τ is given by (2.13) with $u_\zeta = u$ and $p_\zeta = \rho_w gh$.

Energy balance of the coupled system

In order to be coherent with physics, the system (2.34),(2.12) must dissipate the mechanical energy.

Proposition 6. *For smooth enough solutions, the mechanical energy of system* (2.34),(2.12) *satisfies the following energy balance*

$$\partial_t (\mathcal{K} + \mathcal{E}_{tot}) + \nabla_{\mathbf{x}} \cdot (\mathcal{K}u + h\phi_w u + b\phi_s v) - \nabla_{\mathbf{x}} \cdot (2\mu_s b v \cdot D_{\mathbf{x}} v) \\ = -v \cdot f_B - r\kappa_\zeta |u-v|^2 - 2\mu_s b (D_{\mathbf{x}} v) : (D_{\mathbf{x}} v),$$

where $\mathcal{K} = \dfrac{1}{2} rh|u|^2$ *is the kinetic energy of the water layer and* $\mathcal{E}_{tot} = \left(grh \left(\dfrac{h}{2} + b + B \right) + gb \left(\dfrac{b}{2} + B \right) \right)$ *is the potential energy of the whole system. We have set the potentials* $\phi_w = rg(h+b+B)$ *and* $\phi_s = g(b + rh + B)$.

Proof. The classical energy balance for the Shallow Water equations with a moving topography is

$$\partial_t (\mathcal{E}_w + \mathcal{K} + ghb) + \nabla_{\mathbf{x}} \cdot (\mathcal{K}u + \phi_w u) = -gh\partial_t b - \kappa_\zeta (u-v),$$

with $\mathcal{E}_w = \mathcal{E}(h, B)$. Summing this equation multiplied by r with the energy balance for the sediment layer given in Proposition 4 gives the result. To rearrange the potential energy terms, the following computation is performed

$$r\partial_t (\mathcal{E}_w + \mathcal{K} + ghb) + \partial_t (\mathcal{E}_s + rgbh) - rgb\partial_t h - rgh\partial_t b = \partial_t (r\mathcal{E}_w + \mathcal{E}_s + rgbh) = \partial_t \mathcal{E}_{tot}$$

□

2.3.2 Numerical strategy for the coupled system

The numerical scheme for the bilayer system is presented. We have already introduced a scheme for the sediment layer. We look for a scheme in the water layer which can be coupled to the scheme (2.17)-(2.18) for the sediment layer. Since the scheme (2.17)-(2.18) is staggered, we look for a staggered scheme for the water layer. Moreover, the

terms $u_{i+1/2}, v_{i+1/2}$ involved in the friction terms between the sediment and the water must be taken at the same time in the scheme for the water layer and in the scheme for the sediment layer, that is to say, at time t^{n+1}. Note that an implicit friction term is advantageous for the stability of the scheme in the water layer.

To avoid dealing with large systems, the following choice is made. In the momentum equation for the water layer, the topography is taken at the time t^n (that is to say, the contribution from the sediment depth is taken at the time t^n); while in the sediment layer, the pressure contribution coming from the water layer is taken at the time t^{n+1}. In the scheme for the sediment layer, the pressure term $p^{n+1}_{\zeta,i+1/2}/\rho_s$ is replaced by its expression $p^{n+1}_{\zeta,i+1/2}/\rho_s = rgh^{n+1}_{i+1/2}$.

The scheme described in [73, 75] and used for the simulation of the Exner-Shallow Water system in [55] meets the design requirements presented at the beginning of the current section. It reads

$$h_i^{n+1} = h_i^n - \delta_t^n \partial_i^\delta \mathcal{F}^{h,n}, \tag{2.35}$$

$$\begin{aligned} h^{n+1}_{c,i+1/2} u^{n+1}_{i+1/2} &= h^n_{c,i+1/2} u^n_{i+1/2} - \delta_t^n \partial^\delta_{i+1/2} \mathcal{Q}^n \\ &- g\delta_t^n h^{n+1}_{c,i+1/2} \partial^\delta_{i+1/2} \zeta^n - \delta_t^n \kappa_{\zeta,i+1/2} \left(u^{n+1}_{i+1/2} - v^{n+1}_{i+1/2} \right), \end{aligned} \tag{2.36}$$

with the following notations

$$h^n_{c,i+1/2} = \frac{h_i^n + h_{i+1}^n}{2},$$

$$\mathcal{F}^{h,n}_{i+1/2} = h^n_{i+1/2} u^n_{i+1/2}, \qquad h^n_{i+1/2} = \begin{cases} h_i^n & \text{if } u^n_{i+1/2} > 0, \\ \frac{h_i^n + h_{i+1}^n}{2} & \text{if } u^n_{i+1/2} = 0, \\ h^n_{i+1} & \text{if } u^n_{i+1/2} < 0, \end{cases}$$

$$\mathcal{Q}_i^n = q_i^n u_i^n + g \frac{(h_i^{n+1})^2}{2},$$

with

$$q_i^n = \frac{1}{2} \left(\mathcal{F}^{h,n}_{i+1/2} + \mathcal{F}^{h,n}_{i-1/2} \right), \quad u_i^n = \begin{cases} u^n_{i-1/2} & \text{if } q_i^n > 0, \\ \frac{u^n_{i-1/2} + u^n_{i+1/2}}{2} & \text{if } q_i^n = 0, \\ u^n_{i+1/2} & \text{if } q_i^n < 0. \end{cases}$$

The CFL condition for the water layer is

$$\lambda_w^n \delta_t^n \leqslant \frac{\delta_x}{2}, \tag{2.37}$$

with

$$\lambda_w^n = \max_{1 \leqslant i \leqslant N_x} \left(\frac{|q_i^n|}{h_i^n} + \sqrt{g h_i^n} \right). \tag{2.38}$$

As the friction term is implicit, equations (2.18) and (2.36) are coupled. However, the resolution of a large system to find the vector (u^{n+1}, v^{n+1}) is actually not necessary. One can resort to a splitting strategy. Let the momentum equation in the water be computed without friction first, and the water velocity be corrected with the friction term afterwards. The system (2.35), (2.36) is recast into two systems

$$h_i^{n^*} = h_i^n + \delta_t^n S_{h,i}^n, \tag{2.39}$$

$$h_{c,i+1/2}^{n^*} u_{i+1/2}^{n^*} = h_{c,i+1/2}^n u_{i+1/2}^n + \delta_t^n S_{hu,i+1/2}^n, \tag{2.40}$$

with

$$S_{h,i}^n = -\partial_i^\delta \mathcal{F}^{h,n}, \tag{2.41}$$

$$S_{hu,i+1/2}^n = -\partial_{i+1/2}^\delta \mathcal{Q}^n - g^n h_{c,i+1/2}^{n+1} \partial_{i+1/2}^\delta \zeta^n, \tag{2.42}$$

and

$$h_i^{n+1} = h_i^{n^*}, \tag{2.43}$$

$$h_{c,i+1/2}^{n+1} u_{i+1/2}^{n+1} = h_{c,i+1/2}^{n^*} u_{i+1/2}^{n*} - \delta_t^n \kappa_{\zeta,i+1/2} \left(u_{i+1/2}^{n+1} - v_{i+1/2}^{n+1} \right), \tag{2.44}$$

with $h_{c,i+1/2}^{n^*} = h_{c,i+1/2}^{n+1}$. The variable $u_{i+1/2}^{n*}$ is computed explicitly. Equation (2.44) gives an expression for $u_{i+1/2}^{n+1}$ as a function of $u_{i+1/2}^{n*}$ and $v_{i+1/2}^{n+1}$ which can be substituted in (2.18). Thus, a system is solved for v^{n+1} only. The scheme (2.39), (2.40) is a forward Euler scheme using the sources $S_{h,i}^n, S_{hu,i+1/2}^n$, which depend only on h^n, u^n, b^n and B.

Theoretically, we do not know if the time step prescribed by the CFL condition in the water layer will satisfy the CFL condition in the sediment layer as well. (In practice, this is the case in many situations because the velocity in the water layer is typically much larger than that in the sediment layer.) The management of the time step is described in the pseudo-algorithm below. The function *Euler* called in the algorithm is defined immediately below.

Discrete energy balance for the coupled system

We prove that the way we couple a scheme for the water layer with our scheme for the sediment layer does not create entropy.

Proposition 7. *Let a scheme (WS) for the water layer such that the discrete entropy inequality*

$$\frac{\mathcal{E}(h_i^{n+1}, \hat{B}_i) - \mathcal{E}(h_i^n, \hat{B}_i)}{\delta_t^n} + \frac{1}{\delta_t^n} \left(\frac{\mathcal{K}_{i+1/2}^{n+1} + \mathcal{K}_{i-1/2}^{n+1}}{2} - \frac{\mathcal{K}_{i+1/2}^n + \mathcal{K}_{i-1/2}^n}{2} \right) + \partial_i^\delta \mathcal{J} \leqslant R_i - \kappa_\zeta u_{i+1/2}^{n+1} (u_{i+1/2}^{n+1} - v_{i+1/2}^{n+1}) \tag{2.45}$$

Algorithme 1 : Time loop for the coupled scheme

while $(t < T)$ **do**
- $\lambda_w^n \leftarrow (2.38)$;
- $S_h^n \leftarrow (2.41)$; $S_{hu}^n \leftarrow (2.42)$;
- **while** $error(b^{n,q}) > maximum\_tolerance$ **do**
  - $h^{n,q} = Euler(\delta_t^{n,q}, h^n, S_h^n)$; $(hu)^{n,q} = Euler(\delta_t^{n,q}, (hu)^n, S_h^n)$;
  - $v^{n,q} \leftarrow (2.18)$;
  - $b^{n,q} \leftarrow (2.17)$;
  - $\lambda_s^{n,q} \leftarrow (2.24)$;
  - $\delta_t^{n,q} = 1/\max\left(\frac{1}{T-t}, \frac{\lambda_w^n}{\delta_x}, \frac{\lambda_s^{n,q}}{\delta_x}\right)$;
  - compute $error(b^{n,q})$;
- $h^{n+1} \leftarrow h^{n,q}$;
- $b^{n+1} \leftarrow b^{n,q}$;
- $(hu)^{n+1} \leftarrow (2.44)$;

Algorithme 2 : *Euler*

Input : ϕ, δ_t^n, S^n

$\phi \leftarrow \phi + \delta_t^n S^n$

holds, where

$$\mathcal{E}(h_i^n, \hat{B}_i) = g h_i^n \left(\frac{h_i^n}{2} + \hat{B}_i\right)$$

is the potential energy, $\hat{B}_i$ *is the topography seen by the water,*

$$\mathcal{K}_{i+1/2}^n = h_{i+1/2}^n (u_{i+1/2})^2$$

is the kinetic energy at the face $i+1/2$ *and at the time* t^n*,* $\mathcal{J}_{i+1/2}$ *is an energy flux and* R_i^n *is a rest that goes to 0 as* δ_t^n *goes to 0 under the CFL condition respected by the scheme (WS). Then the total discrete entropy of the scheme for the bilayer system satisfies*

$$\begin{aligned}
&\frac{1}{\delta_t^n}\left(\mathcal{E}_{tot,i}^{n+1} + \frac{\mathcal{K}_{i+1/2}^{n+1} + \mathcal{K}_{i-1/2}^{n+1}}{2} - \left(\mathcal{E}_{tot,i}^{n} + \frac{\mathcal{K}_{i+1/2}^{n} + \mathcal{K}_{i-1/2}^{n}}{2}\right)\right) \\
&\quad + \partial_i^\delta\left(r\mathcal{J} + \phi^{n+1} b^n v^{n+1}\right) - \partial_i \mathcal{M}^{n+1} \leqslant -\frac{v_{i+1/2}^{n+1}}{2} f_{B,i+1/2}^{n+1} - \frac{v_{i-1/2}^{n+1}}{2} f_{B,i-1/2}^{n+1} \\
&\qquad - \frac{r}{2}\kappa_{\zeta,i+1/2}(u_{i+1/2}^{n+1} - v_{i+1/2}^{n+1})^2 - \frac{r}{2}\kappa_{\zeta,i-1/2}(u_{i-1/2}^{n+1} - v_{i-1/2}^{n+1})^2 \\
&\qquad\quad - 2\mu_s\left(\frac{b_{i+1}^n\left(\partial_{i+1}^\delta v^{n+1}\right)^2 + 2b_i^n\left(\partial_i^\delta v^{n+1}\right)^2 + b_{i-1}^n\left(\partial_{i-1}^\delta v^{n+1}\right)^2}{4}\right),
\end{aligned}$$

where $\mathcal{E}^n_{tot,i}$ *is the total potential energy defined by*

$$\mathcal{E}^n_{tot,i} = r\mathcal{E}(h^n_i, b^n_i + B_i) + \mathcal{E}(b^n_i, B_i),$$

and $f^{n+1}_{B,i+1/2}$ *and* $\mathcal{M}^{n+1}_{i+1/2}$ *are defined as in Proposition 5.*

Proof. First, a discrete entropy for the scheme (WS) with variable-in-time topography must be obtained. The discrete inequality (2.45) is given for a topography that is constant in time. §2.3.2, we said that the contribution of the sediment depth to the topography term in the scheme for the water layer is taken at time t^n. Therefore, (2.45) immediately holds for $\hat{B}^n_i = b^n_i + B_i$. Then, noticing that

$$\mathcal{E}(h^{n+1}_i, b^{n+1}_i + B_i) = \mathcal{E}(h^{n+1}_i, b^n_i + B_i) + gh^{n+1}_i(b^{n+1}_i - b^n_i),$$

we get

$$\frac{\mathcal{E}(h^{n+1}_i, b^{n+1}_i + B_i) - \mathcal{E}(h^n_i, b^n_i + B_i)}{\delta^n_t} + \frac{1}{\delta^n_t}\left(\frac{\mathcal{K}^{n+1}_{i+1/2} + \mathcal{K}^{n+1}_{i-1/2}}{2} - \frac{\mathcal{K}^n_{i+1/2} + \mathcal{K}^n_{i-1/2}}{2}\right) + \partial^\delta_i \mathcal{J} \leqslant R_i - \kappa_\zeta u^{n+1}_{i+1/2}(u^{n+1}_{i+1/2} - v^{n+1}_{i+1/2}) + gh^{n+1}_i(b^{n+1}_i - b^n_i). \quad (2.46)$$

To conclude the proof, equation (2.46) is multiplied by r and summed to the discrete energy balance for the sediment layer (2.25). The potential energy terms $\mathcal{E}^{n+1}_{tot,i}$, $\mathcal{E}^n_{tot,i}$ are obtained by performing the following calculation

$$r\mathcal{E}(h^{n+1}_i, b^{n+1}_i + B_i) - r\mathcal{E}(h^{n+1}_i, b^n_i + B_i) + \mathcal{E}^{n+1}_{s,i} - \mathcal{E}^n_{s,i} - rgh^{n+1}_i(b^{n+1}_i - b^n_i) - rgb^n_i(h^{n+1}_i - h^n_i) = \mathcal{E}^{n+1}_{tot,i} - \mathcal{E}^n_{tot,i}$$

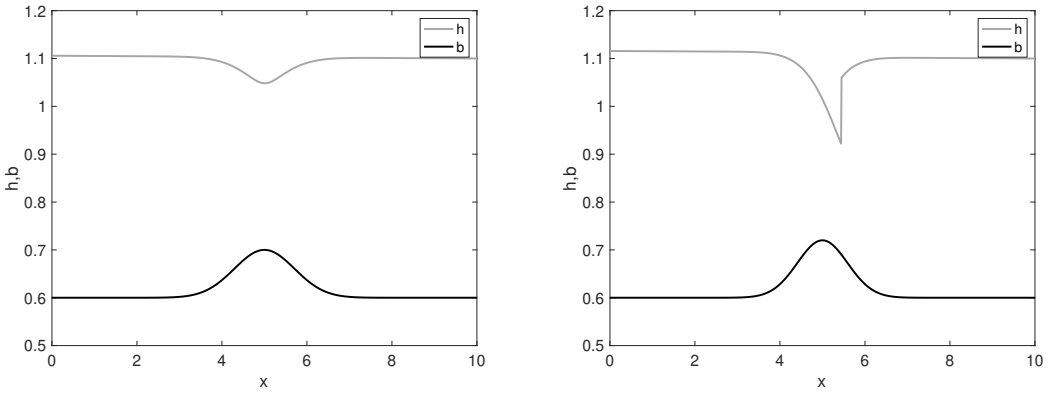
□

The scheme described in [73, 75] verifies the hypothesis in Proposition 7.

2.3.3 Numerical results for the coupled system

In what follows, the behavior of the coupled system is illustrated. The length of the domain is $L = 10$. The surface of the bedrock is $B(x) = 0.5$. Unless specified otherwise, the left boundary condition in the water is $(hu)(0,t) = 0.5$ and the right boundary condition is $h(L,t) = 0.5$. In the sediment layer, the boundary conditions on the left are $b(0,t) = 0.1$ and $\partial_x v(0,t) = 0$. The boundary conditions on the right are $\partial_x b(L,t) = 0$ and $\partial_x v(L,t) = 0$. Except in the second test case in §2.3.3, the initial shape of the sediment dune is given by

$$b^0(x) = 0.1\left(1 + e^{-(x-5)^2}\right). \quad (2.47)$$



(a) Bed described by (2.47)

(b) Bed described by (2.48)

Figure 2.7 – Initially subcritical and transcritical flows

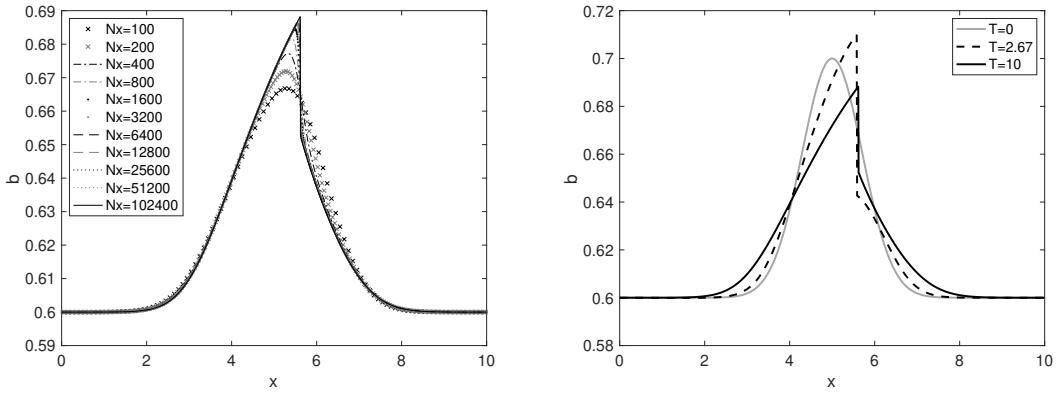
These boundary conditions and initial sediment shape are those described in [55]. Again, $\kappa_{B,i+1/2} = 1$ for all $i = 1, \dots, N_f$. The friction coefficient at the water-sediment interface is also constant in space. Its value is $\kappa_{\zeta,i+1/2} = 2 \times 10^{-3}$ for all $i = 1, \dots, N_f$. As in §2.2.3, $g = 9.81$, $\gamma = 1$ and there is no threshold for the onset of motion, i.e. $\overline{\tau} = 0$.

To initialize the water layer, we run the simulation with a fixed sediment layer until the water reaches the steady state. The steady flow over the water bump is shown on Figure 2.7(a). With the mentioned boundary and initial conditions, the steady water flow is fluvial.

Dune growth test

In this section, we study the influence of the viscosity term on the evolution of a sediment dune under an initially subcritical water flow. This is a classical test case. The difference with existing models such as the Grass model and that in [58] is shown. Two different studies are performed: one with $\mu_s = 0$, and one with $\mu_s = 0.5$.

A convergence study is performed with $\mu_s = 0$. The sediment profiles obtained at $T = 10$ for an increasing number of cells N_x are shown on Figure 2.8. For the converged solutions, the dune steepens and grows slightly. The sediment accumulates on the downstream side of the bump and a shock seems to appear in the sediment layer at the beginning of the solution. Yet, there is no hydraulic jump in the water layer, the flow remains fluvial everywhere in the domain. The growth of the dune is due to a transient occurring when the sediment layer becomes deformable. Indeed, during the initialization phase for the water, the sediment layer is rigid and the water layer does not transmit energy to it. When the sediment layer becomes erodible, it suddenly receives energy from the water layer. Then, after a short time, the height of the sediment bump decreases due to the diffusion process. This phenomenon is difficult to catch numerically. Large values of N_x are necessary to observe the initial dune growth.



(a) Sediment profiles at $T = 10$, $\mu_s = 0$

(b) Time evolution, $N_x = 102400$

Figure 2.8 – §2.3.3 Dune growth test, $\mu_S = 0$

The influence of the viscosity on the time evolution of the dune is assessed. The viscosity is $\mu_S = 0.5$. Then, a convergence study is performed with $\mu_s = 0.5$. The convergence is harder to obtain than with $\mu_s = 0$, in the sense that a larger number of cells N_s is needed to correctly approximate the solution.

The converged solutions are shown on Figure 2.11(a) and have the following behavior. When the simulation starts, the dune begins to grow. Then, a hydraulic jump appears in the water layer and the dune sharpens, evolving into a peak under the hydraulic jump and a smaller bump following it. The sediment peak moves downstream along with the hydraulic jump and the phenomenon maintains itself. The sediment velocity is much lower than that of the water; it is the deformation in the sediment layer that moves as fast as the hydraulic jump. A zoom on the hydraulic jump and on the sediment underneath reveals that the sediment profile brutally changes under the shock in the water layer, see Figure 2.11(b). The profiles obtained in the sediment layer are very sharp. But they do not result from a shock in the sediment layer: multiple cells are involved in the peak, see Figure 2.11(b).

Figure 2.9 shows the sediment profiles obtained as the mesh is refined. For the coarsest meshes used, i.e. for $N_x \leqslant 400$, the numerical solution fails to capture the correct behavior: no hydraulic jump is created. The dune grows a bit on the downstream side and then is eroded. Neither the peak height nor the peak velocity are correct and the solutions appear shifted with respect to the reference solution. The evolution of the error in norm L^2 in time and space as a function of N_x is shown on Figure 2.10. This error is given by the formula

$$\|x^n - x_{ref}\|_{l^2_\delta} = \left(\frac{1}{LT} \Delta T_{out} \sum_{n=1}^{N_{out}} \left(\delta_x \sum_{k=1}^{N_x} |x_k^n - x_{ref}|^2 \right) \right)^{1/2},$$

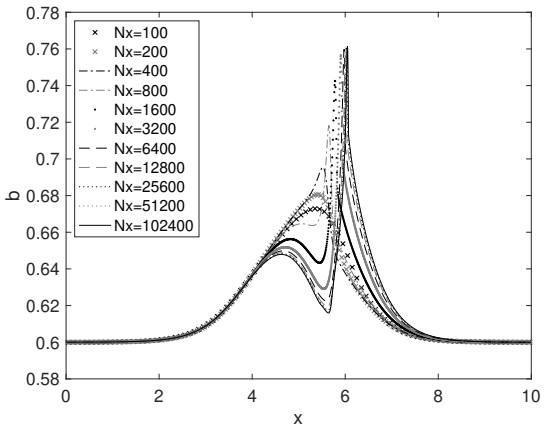


Figure 2.9 – §2.3.3 Sediment profiles at $T = 10$, $\mu_S = 0.5$

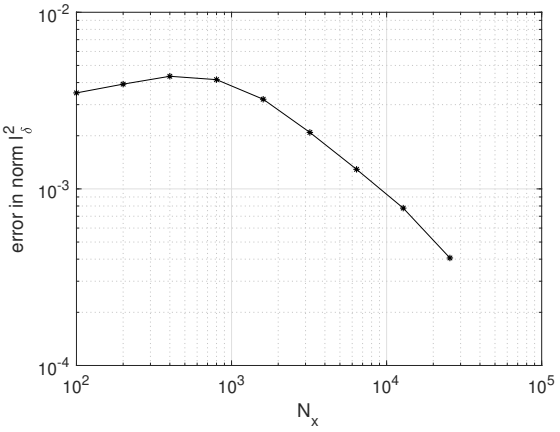


Figure 2.10 – §2.3.3 Error in norm l^2_δ, $\mu_S = 0.5$

where ΔT_{out} is the time interval between two outputs and N_{out} the number of outputs. The values of the numerical solution and of the reference solution are designated by x_k and x_{ref} respectively. The fact that the order of convergence is not 1 is due to the presence of shocks in the numerical solution for the water. The number $N_x = 25600$ is enough to catch the correct behavior of the solution.

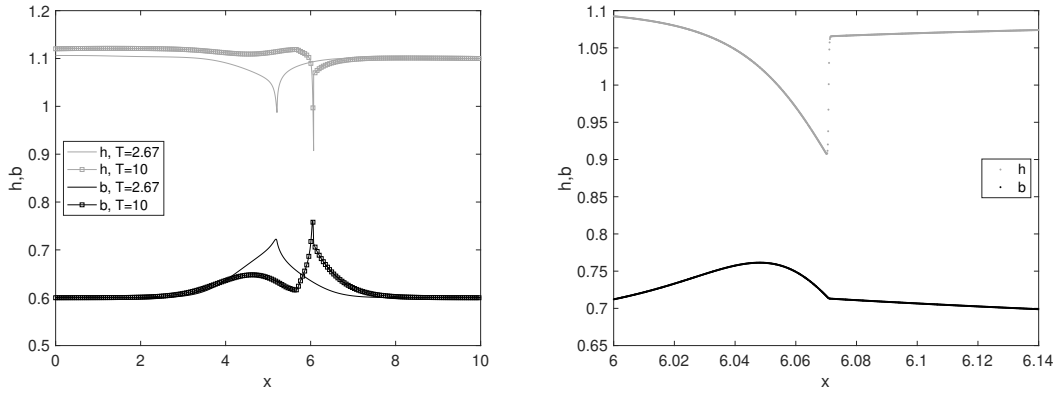
The same simulation is performed with higher viscosity values. For higher viscosity values, the sediment profiles obtained are less sharp and steep. The velocity of the sediment layer is lower. For viscosity values of the order of 10, the bump seems not to move at all: the phenomenon is too slow to be observed at this time scale.

To emphasize the difference between the non-local model, the Grass model and the local model (similar to the model in [58]), the solid fluxes are plotted for each model. The sediment fluxes are plotted for the numerical solutions described in §2.3.3 with $N_x = 25600$. Logically enough, the local flux is plotted for the numerical solution computed with $\mu_s = 0$ and the non-local flux is plotted for the numerical solution computed with $\mu_s = 0.5$. The Grass flux is plotted for the solution computed with $\mu_s = 0$. The fluxes are plotted at the beginning of the simulations ($T = 0.67$), that is to say, before the dunes have become significantly different from the initial condition. The Grass formula gives the following solid flux

$$q_s = A_g |u_w|^{m-1} u_w,$$

with A_g and m two constants and q_s the sediment flux. Typically, $m = 3$. The constant A_g can be determined by identification with the Meyer-Peter and Müller formula. Yet, as the computation A_g involves the Strickler coefficient, determining its value is irrelevant in our case. To illustrate what the Grass flux for this sediment bump would be, we take $m = 3$ and we simply plot $|u_w|^2 u_w$, see Figure 2.12(a). The sediment flux is maximal at the top of the sediment bump.

The solid flux bv for the local model is plotted on Figure 2.12(b). Due to the presence



(a) Time evolution

(b) Zoom on the hydraulic jump, $T = 10$

Figure 2.11 – §2.3.3 Dune evolution, $\mu_S = 0.5$

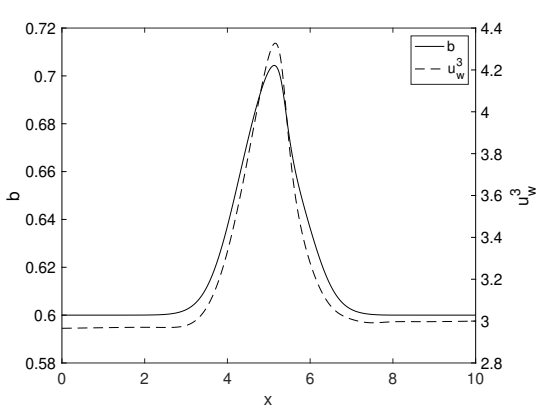
of the slope term, the monotonicity of the flux is very different with respect to the Grass flux. Yet the local maximum of the flux coincides with the peak of the sediment bump. The additional effect of the viscosity term is illustrated on Figure 2.12(c). The local maximum of the flux is shifted with respect to the peak of the sediment bump, it is located upstream. From the plots of the fluxes, it is not easy to predict the evolution of the dunes. Yet, the fact that they are very different from the plot of the Grass flux - and from each other - provides some insight about why the behaviors subsequently observed are complex.

Numerous authors, among which the authors of [61, 87], have argued that such a shift between the shear stress and the peak of the sediment bump is necessary for the dune to grow. In these works, the solid flux directly depends on the shear stress τ, but in the local and non-local models, the solid flux is bv, which is why we are interested in the shift of bv with respect to the sediment bump. While in [61, 87] the shift between the maximum solid flux and the sediment bump was achieved via an improved description of the water layer, in the present work, it is obtained thanks to the addition of the viscosity term.

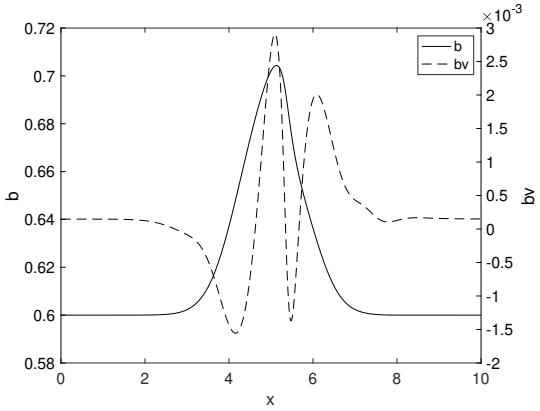
Sensitivity with respect to some parameters

In this section, we assess the sensitivity of the behavior of the sediment layer with respect to some parameters, namely the flow rate imposed in the water on the left boundary and the initial condition. The viscosity is $\mu_s = 0$. The numerical solutions are computed with $N_x = 25600$.

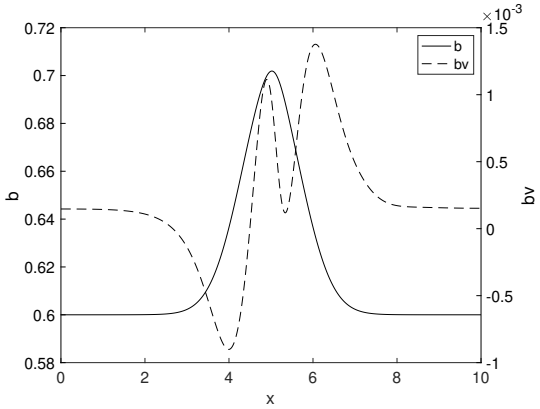
The sensitivity with respect to the flow rate is assessed first. The results are shown on Figure 2.13(a). For $(hu)(0,t) = 0.477$, the dune grows, a hydraulic jump appears and the dune propagates. For $(hu)(0,t) = 0.476$, the dune sharpens but does not manage to grow and does not propagate. A small change in the inflow boundary condition can



(a) Grass flux, $\mu_s = 0$

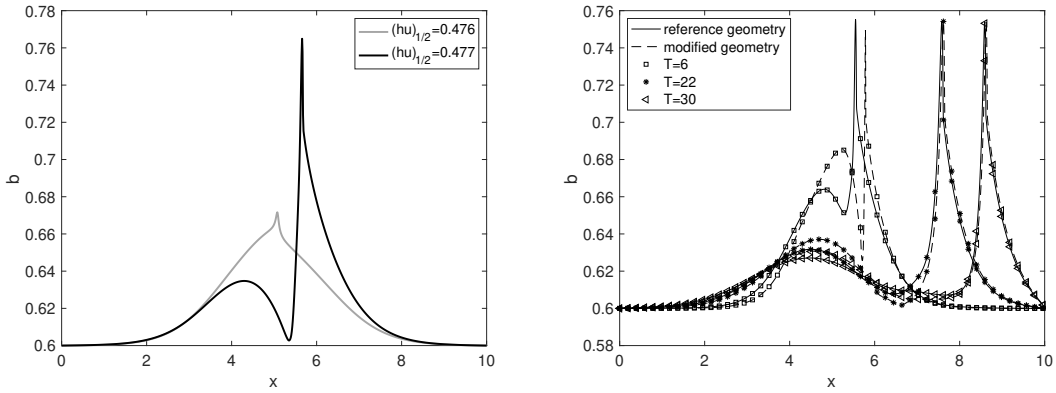


(b) Local flux, $\mu_s = 0$



(c) Non-local flux, $\mu_s = 0$

Figure 2.12 – §2.3.3 Comparison of the fluxes at $T = 0.67$



(a) Sediment profiles for different values of $(hu)(0,t)$ - $\mu_s = 0.5$, $T = 18$

(b) Comparison of bed evolution

Figure 2.13 – §2.3.3 Parameter sensitivity, $\mu_S = 0.5$

induce a very different behavior, which means that the evolution of the sediment layer is very sensitive to the inflow boundary condition.

Then, the behavior of the sediment dune when the steady-state solution in the water is transcritical is illustrated. The initial shape of the dune is now given by

$$b^0(x) = 0.1\left(1 + 1.2e^{-\frac{(x-5)^2}{0.8333\sqrt{2}}}\right). \tag{2.48}$$

which means that the sediment dune is initially higher and narrower than the one described in equation (2.47), though both dunes have the same mass. When the initial shape of the dune is given by equation (2.48), the stationary solution in the water presents a hydraulic jump, see Figure 2.48, contrarily to what happens when the initial condition is given by (2.47). The numerical solution is computed with $N_x = 25600$ and compared with the solution obtained in §2.3.3 with the same number of cells and a viscosity $\mu_s = 0.5$. Figure 2.13(b) shows the behaviors of the dunes initially described by (2.47) (solid line, with the caption "reference geometry") and (2.48) (dashed line, with the caption "modified geometry"). At the beginning of the simulations, the two dunes have very different profiles. As time advances, their shapes become similar. The front peaks overlap, while the smaller bumps upstream are different. Once the shapes have become similar, they remain similar as the dunes propagate. This shows the continuity of the evolution of the sediment layer with respect to the initial condition in the water layer.

2.4 Other numerical schemes

In this section, we present and investigate the behavior of other numerical schemes for the sediment layer. The rationale for the design of different numerical schemes is the

following:

- Even if the main objective of this work is to illustrate the influence of the viscosity operator on the behavior of the sediment layer, in practice, the flow in the non-local model should not be very different from the flow in the local model. The local model can be reformulated as a nonlinear convection-diffusion equation. It is then a natural idea to use this structure to design a scheme for the local model, and then try to extend this scheme for the non-local model. The scheme presented in §2.2.2 does not rely on the features of the local model.

- In the scheme (2.17), (2.18), the link between the continuity equation and the equation on the velocity is poorly exploited. Indeed, the continuity equation is discretized regardless of the nature of the velocity - the fact that the velocity v is itself a function of the sediment depth b is not used.

Therefore, we present here numerical schemes relying on the properties of the local model. For reasons that will be explained below, these numerical schemes are not satisfactory. We merely report the attempts that we have made. A scheme (2.50) for the local model is first presented. It is then immediately extended into a scheme (2.53) for the non-local model. A dissipative entropy balance is proved for the scheme for the non-local model. The same proof can be used to show that the scheme for the local model satisfies a dissipative entropy balance.

2.4.1 A scheme for the local model

As in §2.2.2 the hat variables are used for the local model. The local model is reformulated as a nonlinear convection-diffusion equation

$$\partial_t \hat{b} + \partial_x \left(\hat{b}\hat{V}_a + \hat{b}\hat{V}_d \right) = 0 \tag{2.49}$$

with the advection velocity

$$\hat{V}_a\left(\hat{b}\right) = \begin{cases} \dfrac{r\kappa_\zeta}{r\kappa_\zeta + \kappa_B ||\hat{v}||^{\gamma-1}} u_\zeta & \text{if } \hat{\tau} > \tau_c\left(\hat{b}, p_\zeta\right) \\ 0 & \text{if } \hat{\tau} \leq \tau_c\left(\hat{b}, p_\zeta\right) \end{cases},$$

and the diffusion velocity

$$\hat{V}_d\left(\hat{b}\right) = -\tilde{D}\partial_x\hat{\phi} \qquad \text{with} \qquad \tilde{D}\left(\hat{b}\right) = \begin{cases} \dfrac{\hat{b}}{r\kappa_\zeta + \kappa_B ||\hat{v}||^{\gamma-1}} & \text{if } \hat{\tau} > \tau_c\left(\hat{b}, p_\zeta\right) \\ 0 & \text{if } \hat{\tau} \leq \tau_c\left(\hat{b}, p_\zeta\right). \end{cases}$$

This formulation is slightly different from (2.29). Here, we insist on the fact that the velocity $\hat{v}$ can be split into two velocities, that is to say, $\hat{v} = \hat{V}_a + \hat{V}_d$. For the reason explained in remark 3, the diffusion part of the equation is discretized with an implicit

scheme. A discretization of (2.49) reads

$$\begin{aligned} \hat{b}_i^{n\star} &= \hat{b}_i^n - \delta_t^n \partial_i^\delta \left(\hat{b}_a^n \hat{V}_a^n \right) \\ \hat{b}_i^{n+1} &= \hat{b}_i^{n\star} - \delta_t^n \partial_i^\delta \left(\hat{b}_d^{n\star} \hat{V}_d^{n+1} \right) \end{aligned} \tag{2.50}$$

The upwind reconstruction according to the advection velocity is denoted by $\hat{b}^n_{a,i+1/2}$ i.e.

$$\hat{b}^n_{a,i+1/2} = \begin{cases} \hat{b}_i^n & \text{if } V^n_{a,i+1/2} > 0 \\ \frac{\hat{b}_i^n + \hat{b}_{i+1}^n}{2} & \text{if } V^n_{a,i+1/2} = 0 \\ \hat{b}_{i+1}^n & \text{if } V^n_{a,i+1/2} < 0 \end{cases}.$$

The upwind reconstruction according to the diffusion velocity is denoted by $\hat{b}^n_{d,i+1/2}$ i.e.

$$\hat{b}^n_{d,i+1/2} = \begin{cases} \hat{b}_i^n & \text{if } V^n_{d,i+1/2} > 0 \\ \frac{\hat{b}_i^n + \hat{b}_{i+1}^n}{2} & \text{if } V^n_{d,i+1/2} = 0 \\ \hat{b}_{i+1}^n & \text{if } V^n_{d,i+1/2} < 0 \end{cases}.$$

The discrete velocity and the diffusion parameters are defined by

$$\hat{V}^n_{a,i+1/2} = \frac{r\hat{\kappa}^n_{\zeta,i+1/2}}{r\hat{\kappa}^n_{\zeta,i+1/2} + \kappa_{B,i+1/2}|\hat{\mathcal{V}}^{n+1}_{i+1/2}|^{\gamma-1}} u^{n+1}_{i+1/2}$$

$$\text{and} \qquad \hat{D}^n_{i+1/2} = \frac{\hat{b}^n_{a,i+1/2}}{r\hat{\kappa}^n_{\zeta,i+1/2} + \kappa_{B,i+1/2}|\hat{\mathcal{V}}^{n+1}_{i+1/2}|^{\gamma-1}}$$

if $\hat{\tau}^n_{i+1/2} > \hat{\tau}^n_{c,i+1/2}$, while if $\hat{\tau}^n_{i+1/2} \leqslant \hat{\tau}^n_{c,i+1/2}$ they are both zero. The effective shear stress $\hat{\tau}^n_{i+1/2}$ is defined by

$$\hat{\tau}^n_{i+1/2} = -r\kappa_{\zeta,i+1/2}(\hat{\mathcal{V}}^n_{i+1/2} - u^{n+1}_{\zeta,i+1/2}) - \hat{b}^n_{a,i+1/2}\partial^\delta_{i+1/2}\hat{\phi}^{n+1}$$

and the critical shear stress by

$$\hat{\tau}^n_{c,i+1/2} = \tau_c(\hat{b}^n_{a,i+1/2}, p^n_{\zeta,i+1/2}).$$

The diffusion velocity is defined by

$$\hat{V}^{n+1}_{d,i+1/2} = -\frac{\hat{b}^n_{a,i+1/2}}{\hat{b}^{n\star}_{d,i+1/2}}\hat{D}^n_{i+1/2}\partial^\delta_{i+1/2}\hat{\phi}^{n+1} - \frac{\hat{b}^n_{a,i+1/2}}{\hat{b}^{n\star}_{d,i+1/2}} \frac{\hat{\tau}^n_{c,i+1/2}\text{sign}(\hat{\mathcal{V}}^n_{i+1/2})}{r\hat{\kappa}^n_{i+1/2} + \kappa_{B,i+1/2}|\hat{\mathcal{V}}^{n+1}_{i+1/2}|^{\gamma-1}}, \tag{2.51}$$

with the discrete potential

$$\hat{\phi}_i^{n+1} = g(b_i^{n+1} + B) + \frac{p_{\zeta,i}^{n+1}}{\rho_s}.$$

We have introduced above the apparent velocity

$$\hat{\mathcal{V}}_{i+1/2}^n = \frac{\hat{b}_{a,i+1/2}^n \hat{V}_{a,i+1/2}^n + \hat{b}_{d,i+1/2}^{n\star} \hat{V}_{d,i+1/2}^{n+1}}{\hat{b}_{a,i+1/2}^n}.$$

The motivation for the definition of $\hat{\mathcal{V}}_{i+1/2}^n$ is given in the proof of proposition 9. While the presence of the coefficient $\hat{b}_{a,i+1/2}^n / \hat{b}_{d,i+1/2}^{n\star}$ in the definition of the diffusion velocity (2.51) may seem artificial, it is actually necessary to obtain the correct discrete energy balance (in the sense that it is consistent with the discrete energy balance stated in proposition 4). More precisely, the presence of the same coefficient is necessary in the case of the scheme for the nonlocal sediment discharge, see proposition 9.

Proposition 8. *Assume that the initial condition is non-negative* $\hat{b}_i^0 \geq 0$ *and the time step satisfies the following CFL condition*

$$\max_i \left(\left| \hat{V}_{a,i+1/2}^n \right| \right) \delta_t^n \leq \frac{\delta_x}{2}. \tag{2.52}$$

Then the solution of the scheme (2.49) *is non-negative, i.e.* $\hat{b}_i^n \geq 0$.

Proof. Under the CFL condition (2.52), the non-negativity of the right-hand side of (2.50) is a classical result of the up-wind scheme. Then the non-negativity follow since the matrix of the system defined by (2.50) is an M-matrix. □

Note that other choices for $b_{d,i+1/2}^{n*}$ would be possible. For instance, one could argue that a centered reconstruction is also be a valid choice. It is the most natural discretization for a non-linear diffusion equation, and the matrix of the system defined by (2.50) is an M-matrix independently from the choice of the reconstruction $b_{d,i+1/2}^{n*}$. However, in the next paragraph, the scheme (2.50) will be extended for the model with non-local sediment discharge, and the upwind reconstruction is the only one with which we are able to prove the positivity.

2.4.2 Extension for the non-local model

Let us now focus on the numerical resolution of the non-local model (2.12). Mimicking the IMEX scheme (2.50), the sediment velocity is split into two components, the advection part and the diffusion part. However, the diffusion step is computed as an advection with a diffusion velocity defined further, i.e.

$$\begin{aligned} b_i^{n\star} &= b_i^n - \delta_t^n \partial_i^\delta \left(b_a^n V_a^n \right) \\ b_i^{n+1} &= b_i^{n\star} - \delta_t^n \partial_i^\delta \left(b_d^{n\star} V_d^{n+1} \right) \end{aligned} \tag{2.53}$$

where $b^n_{a,i+1/2}$ is an upwind (with respect to $V^n_{a,i+1/2}$) reconstruction of the thickness at the face $i+1/2$

$$b^n_{a,i+1/2} = \begin{cases} b^n_i & \text{if } V^n_{a,i+1/2} > 0 \\ \frac{b^n_i + b^n_{i+1}}{2} & \text{if } V^n_{a,i+1/2} = 0 \\ b^n_{i+1} & \text{if } V^n_{a,i+1/2} < 0, \end{cases}$$

and $b^n_{d,i+1/2}$ is an upwind reconstruction of the thickness at the face $i+1/2$ with respect to $V^{n+1}_{d,i+1/2}$.

$$b^{n\star}_{d,i+1/2} = \begin{cases} b^{n\star}_i & \text{if } V^{n+1}_{d,i+1/2} > 0 \\ \frac{b^n_i + b^n_{i+1}}{2} & \text{if } V^{n+1}_{d,i+1/2} = 0 \\ b^{n\star}_{i+1} & \text{if } V^{n+1}_{d,i+1/2} < 0 \end{cases}.$$

The velocities are solutions of the following problems

$$M^n V^n_a = W^n_a \tag{2.54}$$

$$\text{and} \qquad M^n \tilde{V}^{n+1}_d = W^{n+1}_d \tag{2.55}$$

with the vectors $\tilde{V}^{n+1}_d = \left(\frac{b^{n\star}_{d,i+1/2}}{b^n_{a,i+1/2}} V^{n+1}_{d,i+1/2}\right)_{1\le i\le N_f}$, $V^n_a = \left(V^n_{a,i+1/2}\right)_{1\le i\le N_f}$ and for $k \in \{a,d\}$, $W^n_k = \left(W^n_{k,i+1/2}\right)_{1\le i\le N_f}$. For the definition of the problems (2.54), (2.55), the threshold for motion must be taken into account. We define the discrete effective shear stress as

$$\tau^n_{i+1/2} = -r\kappa_{\zeta,i+1/2}(\mathcal{V}^n_{i+1/2} - u^{n+1}_{\zeta,i+1/2}) - b^n_{a,i+1/2}\partial^\delta_{i+1/2}\phi^{n+1} + \partial^\delta_{i+1/2}(2\mu_s b\partial^\delta \mathcal{V})$$

with

$$\phi^{n+1}_i = \left(g\left(b^{n+1}_i + B_i\right) + \frac{p^n_{\zeta,i}}{\rho_s}\right)$$

and the apparent velocity

$$\mathcal{V}^n_{i+1/2} = \frac{b^n_{a,i+1/2}V^n_{a,i+1/2} + b^{n\star}_{d,i+1/2}V^{n+1}_{d,i+1/2}}{b^n_{a,i+1/2}}.$$

The critical shear stress at the face $i+1/2$ is

$$\tau^n_{c,i+1/2} = \tau_c(b^n_{a,i+1/2}, p^n_{\zeta,i+1/2}).$$

The notation $f_{B,i+1/2}$ is introduced to denote the quantity

$$f^n_{B,i+1/2} = \begin{cases} \tau^n_{c,i+1/2}\text{sign}\left(\mathcal{V}^n_{i+1/2}\right) + \kappa_{B,i+1/2}|\mathcal{V}^{n+1}_{i+1/2}|^{\gamma-1}\mathcal{V}^n_{i+1/2} & \text{if } \|\tau^n_{i+1/2}\| > \tau^n_{c,i+1/2} \\ \tau^n_{i+1/2} + \mathcal{V}^n_{i+1/2} & \text{if } \|\tau^n_{i+1/2}\| \leqslant \tau^n_{c,i+1/2} \end{cases}.$$

In the case where the effective shear stress exceeds the threshold, i.e. $\|\tau^n_{i+1/2}\| > \tau^n_{c,i+1/2}$, the right-hand-sides of the problems (2.54), (2.55) are respectively defined by

$$\begin{aligned} & W^n_{a,i+1/2} = r\kappa^n_{\zeta,i+1/2} u^{n+1}_{i+1/2} \\ \text{and} \qquad & W^{n+1}_{d,i+1/2} = -b^n_{a,i+1/2}\partial^\delta_{i+1/2}\phi^{n+1} - \tau^n_{c,i+1/2}\text{sign}\left(\mathcal{V}^n_{i+1/2}\right) \ , \end{aligned}$$

The matrix M^n is such that for $V = \left(V_{i+1/2}\right)_{1\leq i\leq N_f}$, we have

$$(M^n V)_{i+1/2} = \left(r\kappa^n_{\zeta,i+1/2} + \kappa_{B,i+1/2}|V^{n+1}_{i+1/2}|^{\gamma-1}\right) V_{i+1/2} - \partial^\delta_{i+1/2}\left(2\mu_s b^n \partial^\delta V\right).$$

In the case where the effective shear stress is below the threshold, i.e. $\|\tau^n_{i+1/2}\| \leqslant \tau^n_{c,i+1/2}$, the right-hand-sides of the problems (2.54), (2.55) are both zero and the line i of the matrix M^n is such that

$$(M^n)_{i,j} = \delta_{i,j},$$

with $\delta_{i,j}$ the Kronecker symbol.

The shear stresses $\tau^n_{c,i+1/2}$, $\tau^n_{i+1/2}$ are estimated first. Then the velocity $V^n_{a,i+1/2}$ is estimated solving a first N_f-linear system, leading to a time step estimation $\delta^n_t = \frac{\lambda\delta_x}{2V^n_{a,i+1/2}}$ with a given $\lambda \leq 1$ to satisfy (2.56). A first approximation of the thickness is given by

$$b^{n\star}_i = b^n_i - \delta^n_t \partial^\delta_i \left(b^n_a V^n_a\right).$$

The second equation of (2.53) and (2.55) form a $(N_f + N_x)$ non-linear system. As in §2.2.2, the system is reduced to a system with only N_f unknowns by replacing b^{n+1}_i in (2.55) using the second equation of (2.53). Again, a fixed-point method is required to solve this system and because of the stiffness of the problem, the Newton method is chosen.

Proposition 9. *Assume that the initial condition is non-negative $b^0_i \geq 0$ and the time step satisfies the following CFL condition*

$$\max_i \left(\left|V^n_{a,i+1/2}\right|, \left|V^n_{d,i+1/2}\right|\right)\delta^n_t \leq \frac{\delta_x}{2}. \tag{2.56}$$

Then the asymptotic solution ($q \to \infty$) of the scheme (2.53) is non-negative, i.e. $b^n_i \geq 0$

and the solution satisfies a dissipation law of the mechanical energy

$$\begin{aligned}
&\frac{\mathcal{E}_{s,i}^{n+1}-\mathcal{E}_{s,i}^{n}}{\delta_t^n}+\partial_i^\delta\left(\phi^{n+1}\hat{b}_a^n\mathcal{V}^n-\mathcal{M}\right)\\
\leqslant\ &\frac{b_i^n}{\rho_s}\left(\frac{p_{\zeta,i}^{n+1}-p_{\zeta,i}^{n}}{\delta t^n}\right)-\frac{\mathcal{V}_{i+1/2}^n}{2}f_{B,i+1/2}^n-r\frac{\kappa_{\zeta,i+1/2}}{2}\mathcal{V}_{i+1/2}^n(\mathcal{V}_{i+1/2}^n-u_{i+1/2}^{n+1})\\
&-\frac{\mathcal{V}_{i-1/2}^n}{2}f_{B,i-1/2}^n-r\frac{\kappa_{\zeta,i-1/2}}{2}\mathcal{V}_{i-1/2}^n(\mathcal{V}_{i-1/2}^n-u_{i-1/2}^{n+1})\\
&-2\mu_s\left(\frac{b_{i+1}^n\left(\partial_{i+1}^\delta\mathcal{V}^n\right)^2+2b_i^n\left(\partial_i^\delta\mathcal{V}^n\right)^2+b_{i-1}^n\left(\partial_{i-1}^\delta\mathcal{V}^n\right)^2}{4}\right)
\end{aligned}$$

with $\mathcal{E}_{s,i}^n$, $\mathcal{M}_{i+1/2}$ *defined as in Proposition 5 (in the expression of* $\mathcal{M}_{i+1/2}$*, the velocity* $\mathcal{V}^n$ *replaces* v^{n+1}*).*

Proof. Assume that at time iteration n, the solution is non-negative, i.e. $b_i^n\geq 0$. Under the CFL condition (2.56), the non-negativity of $b_i^{n\star}$ is a classical property of the upwind scheme. $b_i^{n\star}$ being non-negative, the same property ensures that b_i^{n+1} is non-negative.

Let us now focus on the mechanical energy. Each of the equations of (2.53) is multiplied by ϕ_i^{n+1} and the sum is made. We get

$$\begin{aligned}
\frac{\mathcal{E}_{s,i}^{n+1}-\mathcal{E}_{s,i}^{n}}{\delta_t^n}+\partial_i^\delta\left(\phi^{n+1}b_a^n\mathcal{V}^n\right)=&-\frac{(b_i^{n+1}-b_i^n)^2}{2\delta_t^n}+\frac{b_i^n}{\rho_s}\left(\frac{p_{\zeta,i}^{n+1}-p_{\zeta,i}^{n}}{\delta t^n}\right)\\
&+b_{a,i+1/2}^n\mathcal{V}_{i+1/2}^n\frac{\partial_{i+1/2}^\delta\phi^{n+1}}{2}+b_{a,i-1/2}^n\mathcal{V}_{i-1/2}^n\frac{\partial_{i-1/2}^\delta\phi^{n+1}}{2}.
\end{aligned}\tag{2.57}$$

Multiplying equation (2.55) by $\mathcal{V}_{i+1/2}^n$ immediately gives

$$\begin{aligned}
&(r\kappa_{\zeta,i+1/2}^n+\kappa_{B,i+1/2}^n)\tilde{V}_{d,i+1/2}^{n+1}-\mathcal{V}_{i+1/2}^n\partial_{i+1/2}^\delta(2\mu_s b^n\partial^\delta\tilde{V}_d^{n+1})\\
&\qquad=-b_{a,i+1/2}^n\mathcal{V}_{i+1/2}^n\partial_{i+1/2}^\delta\phi^{n+1}-\tau_{c,i+1/2}^n\mathrm{sign}(\mathcal{V}_{i+1/2}^n)\mathcal{V}_{i+1/2}^n.
\end{aligned}\tag{2.58}$$

In the left-hand side of (2.58), we make the square of the weighted velocity $\mathcal{V}_{i+1/2}^n$ appear

$$\begin{aligned}
&(r\kappa_{\zeta,i+1/2}^n+\kappa_{B,i+1/2}^n)\left(\mathcal{V}_{i+1/2}^n\right)^2-\mathcal{V}_{i+1/2}^n\partial_{i+1/2}^\delta(2\mu_s b^n\partial^\delta\mathcal{V}^n)-\mathcal{V}_{i+1/2}^n(M^nV_a^n)_i\\
&\qquad=-b_{a,i+1/2}^n\mathcal{V}_{i+1/2}^n\partial_{i+1/2}^\delta\phi^{n+1}-\tau_{c,i+1/2}^n\mathrm{sign}(\mathcal{V}_{i+1/2}^n)\mathcal{V}_{i+1/2}^n.
\end{aligned}\tag{2.59}$$

Using (2.54) finally gives the following expression for $b_{a,i+1/2}^n\mathcal{V}_{i+1/2}^n\partial_{i+1/2}^\delta\phi^{n+1}$

$$\begin{aligned}
&r\kappa_{\zeta,i+1/2}^n\left(\mathcal{V}_{i+1/2}^n\right)^2+\mathcal{V}_{i+1/2}^nf_{B,i+1/2}^n-\mathcal{V}_{i+1/2}^n\partial_{i+1/2}^\delta(2\mu_s b^n\partial^\delta\mathcal{V}^n)\\
&\qquad-r\kappa_{\zeta,i+1/2}^nu_{i+1/2}^{n+1}\mathcal{V}_{i+1/2}^n=-b_{a,i+1/2}^n\mathcal{V}_{i+1/2}^n\partial_{i+1/2}^\delta\phi^{n+1},
\end{aligned}$$

which is then substituted in (2.57).

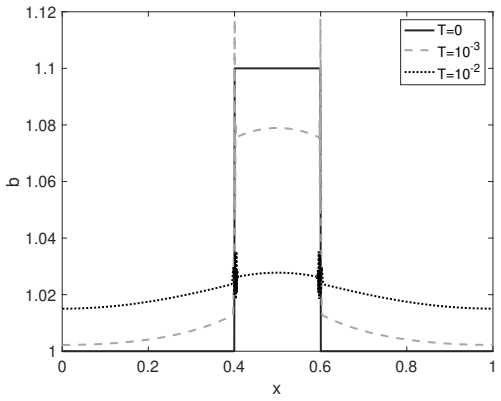


Figure 2.14 – §2.4 Numerical instabilities.

The term $\frac{1}{2}(\mathcal{V}^n_{i+1/2}\partial^\delta_{i+1/2}(2\mu_s b^n \partial^\delta \mathcal{V}^n) + \mathcal{V}^n_{i-1/2}\partial^\delta_{i+1/2}(2\mu_s b^n \partial^\delta \mathcal{V}^n))$ is dissipative. The calculations needed to prove it are exactly those performed in the proof of Proposition 5. The result and its proof are still valid if $\|\tau^n_{i+1/2}\| \leqslant \tau^n_{c,i+1/2}$ or $\|\tau^n_{i-1/2}\| \leqslant \tau^n_{c,i-1/2}$. □

Remark 7. Under the CFL condition (2.52) the scheme (2.50) for the local model satisfies proposition 9 with $\mu_s = 0$ for all $1 \leqslant i \leqslant N_f$. Note that in the case of the local scheme, a CFL condition based only the advection velocity is sufficient.

The scheme (2.53), (2.54), (2.55) turned out to be unstable in practice. The instabilities are shown on Figure 2.14. Spurious oscillations are created. For short times, the oscillations exceed the initial condition. The oscillations eventually disappear.

A possible explanation is the following. The diffusion velocity actually computed (i.e., the velocity that is the solution of the non-linear system) is $\tilde{V}^{n+1}_d$. The velocity V^{n+1}_d is recovered by computing $V^{n+1}_d = b^n_a / b^{n*}_d \tilde{V}^{n+1}_d$. In the second equation of (2.53), the velocity V^{n+1}_d is multiplied by b^{n*}_d. Note that $b^{n*}_d V^{n+1}_d = b^n_a \tilde{V}^{n+1}_d$, so in some sense, the upwinding with respect to V_d is lost. In any case, it is lost for $\mu_s = 0$. Note that for $\mu_s = 0$, this upwinding is actually not necessary.

A tempting solution to preserve the upwinding with respect to V^{n+1}_d is to replace (2.55) by

$$M^n V^{n+1}_d = W^{n+1}_d, \tag{2.60}$$

and to define the apparent velocity as

$$\mathcal{V}'^n_{i+1/2} = \frac{b^n_{a,i+1/2} V^n_{a,i+1/2} + b^{n\star}_{d,i+1/2} V^{n+1}_{d,i+1/2}}{b^{n\star}_{d,i+1/2}}. \tag{2.61}$$

In practice, the scheme (2.53), (2.54), (2.60) is stable. Yet it does not satisfy a dissipative balance for the discrete entropy. Rest terms $\mathcal{R}_{i-1/2}$, $\mathcal{R}_{i+1/2}$ appear in the right-hand side

of the discrete entropy balance

$$\begin{aligned}
&\frac{\mathcal{E}_{s,i}^{n+1}-\mathcal{E}_{s,i}^{n}}{\delta t^n}+\partial_i^\delta\left(\phi^{n+1}\hat{b}_a^n\mathcal{V}'^n-\mathcal{M}\right)\\
\leqslant\frac{b_i^n}{\rho_s}\left(\frac{p_{\zeta,i}^{n+1}-p_{\zeta,i}^{n}}{\delta t^n}\right)&-\frac{\mathcal{V}'^n_{i+1/2}}{2}f^n_{B,i+1/2}-r\frac{\kappa_{\zeta,i+1/2}}{2}\mathcal{V}'^n_{i+1/2}\left(\mathcal{V}'^n_{i+1/2}-\frac{b^n_{a,i+1/2}}{b^{n\star}_{d,i+1/2}}u^{n+1}_{i+1/2}\right)\\
&-\frac{\mathcal{V}'^n_{i-1/2}}{2}f^n_{B,i-1/2}-r\frac{\kappa_{\zeta,i-1/2}}{2}\mathcal{V}'^n_{i-1/2}\left(\mathcal{V}'^n_{i-1/2}-\frac{b^n_{a,i-1/2}}{b^{n\star}_{d,i-1/2}}u^{n+1}_{i-1/2}\right)\\
&-2\mu_s\frac{b^n_{i+1}\left(\partial^\delta_{i+1}\mathcal{V}'^n\right)^2+2b^n_i\left(\partial^\delta_i\mathcal{V}'^n\right)^2+b^n_{i-1}\left(\partial^\delta_{i-1}\mathcal{V}'^n\right)^2}{4}-\mathcal{R}_{i+1/2}-\mathcal{R}_{i-1/2}
\end{aligned}$$

The rest term $\mathcal{R}_{i+1/2}$ is

$$\begin{aligned}
\mathcal{R}_{i+1/2} &= 2b^n_{i+1}V^n_{a,i+1}\left(\frac{b^n_{a,i+3/2}}{b^{n\star}_{d,i+3/2}}-\frac{b^n_{a,i+1/2}}{b^{n\star}_{d,i+1/2}}\right)-2b^n_iV^n_{a,i}\left(\frac{b^n_{a,i+1/2}}{b^{n\star}_{d,i+1/2}}-\frac{b^n_{a,i-1/2}}{b^{n\star}_{d,i-1/2}}\right)\\
&+ b^n_{i+1}\left(V^n_{a,i+3/2}-V^n_{a,i+1/2}\right)\left(\frac{b^n_{a,i+3/2}}{b^{n\star}_{d,i+3/2}}-\frac{b^n_{a,i+1/2}}{b^{n\star}_{d,i+1/2}}\right)\\
&+ b^n_i\left(V^n_{a,i+1/2}-V^n_{a,i-1/2}\right)\left(\frac{b^n_{a,i+1/2}}{b^{n\star}_{d,i+1/2}}-\frac{b^n_{a,i-1/2}}{b^{n\star}_{d,i-1/2}}\right),
\end{aligned}$$

and its sign cannot be determined.

If one replaces (2.54) by

$$M^n\tilde{V}^n_a=W_a,$$

with $\tilde{V}^n_{a,i+1/2}=(b^n_{a,i+1/2}/b^{n\star}_{d,i+1/2})V^n_{a,i+1/2}$, and defines the apparent velocity as in (2.61), a dissipative balance for the discrete entropy is obtained but the upwinding with respect to u_ζ is lost in the continuity equation, because $b^n_{a,i+1/2}V^n_{a,i+1/2}=b^{n\star}_{d,i+1/2}\tilde{V}^n_{a,i+1/2}$.

Moreover, when the velocity is split, the implementation of the transport threshold is not clear. Our guess is that the term $-\tau^n_{c,i+1/2}\mathrm{sign}(\mathcal{V}^n_{i+1/2})$ should be in the right-hand side of the equation for the diffusion velocity. This is certainly the right choice in the absence of forcing term, when the only process involved is diffusion. In such a case, putting $-\tau^n_{c,i+1/2}\mathrm{sign}(\mathcal{V}^n_{i+1/2})$ in the right-hand side of the equation for the advection velocity artificially creates a non-zero advection velocity, and the stationary state obtained is wrong. In a case of pure diffusion, including $-\tau^n_{c,i+1/2}\mathrm{sign}(\mathcal{V}^n_{i+1/2})$ in W_d^{n+1} gave the correct stationary state. In a case with both advection and diffusion, the computation turned out to be stable, yet it was hard to understand whether the stationary state obtained was the correct one. In the absence of viscosity, the criterion defining the stationary state should be $||\kappa_\zeta u_\zeta - gb\nabla b|| \leqslant gb\tau_c$. For the stationary solution, the quantity $||\kappa_\zeta u_\zeta - gb\nabla b||$ is visibly below $gb\tau_c$ everywhere in the domain.

The conclusions regarding these numerical experiments are the following:

- Discrete entropy dissipation is not enough to ensure the stability of the computa-

tion. Another property is needed to better characterize the stability of the computation.

- Upwinding with respect to the gradient of the surface of the sediment layer is necessary.

- All the attempts to extend the scheme (2.50) for the local system into a scheme for the non-local system have failed. The scheme (2.17), (2.18) was therefore adopted.

2.5 Conclusions and perspectives

The sediment layer is first modeled alone. Starting from the incompressible Navier-Stokes equations, several models are derived. The shallow flow approximation is made. Different models are obtained depending on the scaling of the physical parameters. A transport threshold is included in the modeling approach. The classical Exner model with a Grass law is one of the models we recover. Yet, the most interesting model derived is the one in which the solid flux depends is influenced by the gradient of the pressure and by a viscosity term. Due to the presence of the viscosity term, this model is non-local. The present work subsequently focuses on this model. It is briefly analyzed. The positivity of the sediment depth is trivial in the inviscid case and easy to obtain for smooth solutions in the viscous case. Moreover, for smooth enough solutions, the model satisfies a dissipative balance for the mechanical energy.

A numerical scheme for the sediment layer is designed. This scheme is positive and satisfies a dissipative balance for the discrete entropy. Its convergence is checked. The influence of the viscosity on the behavior of the numerical solutions is clearly visible. The transport threshold is implemented and the possibility to obtain non-flat stationary states is shown.

Other numerical schemes are briefly discussed. While they seem reasonable because they rely on the properties of the asymptotic system obtained when μ_s goes to zero, they are not satisfactory because they are unstable and/or they dp not satisfy a discrete dissipative entropy balance.

Then, a coupled model for the water layer and the sediment layer is presented. The water layer is modeled by the Shallow Water equations. The coupled model satisfies a dissipative energy balance. Then, using for the water layer the scheme described in [73, 75], a numerical strategy for the coupled system is proposed and tested. The coupling strategy does not create entropy. Simulations of water flowing above a sediment dune are performed. The viscosity enables the growth of the sediment dune. The dune maintains itself and is dragged downstream by a hydraulic jump in the water layer. A possible explanation for the dune growth is that the non-local flux exhibits a local maximum shifted upstream with respect to the top of the sediment bump. The evolution of the sediment layer is quite robust with respect to the initial geometry of the sediment layer and the initial condition in the water layer, while it is very sensitive to the inflow condition in the water layer.

From the theoretical point of view, a more thorough analysis of the non-local system could be done. The regularity of the solutions has not yet been investigated, while it is crucial to obtain the positivity of the sediment depth.

The physical relevance of the results is beyond the scope of the present work, the purpose of which was chiefly to investigate the effect of the viscosity term. The simulations of the coupled system could be confronted with experimental data. This would require fitting the parameters κ_B and μ_s. In this work, a constant κ_ζ was used for the simulations, but one could try to take a water-sediment friction term of the Chézy or Manning type.

Several challenges are still to be addressed in the modeling. In this work, the sediment layer is described as a Newtonian fluid, but non-Newtonian rheologies, for instance a Drucker-Prager rheology, are appealing. Moreover, in the description we have proposed, no mass exchanges occur between the sediment and water layers. This is coherent with the fact that we are dealing with bed load, but the suspension and deposition of grains is a question of relevance. In [21], a layerwise discretized model of the water for density stratified flows is presented. Including mass exchanges between the sediment layer and the water and adopting a layerwise-discretized approach to simulate density variations in the water is a challenging objective.

List of main symbols used in Chapter 2

Symbol	Description
$\mathbf{x}$	Coordinate in the horizontal plane
z	Coordinate in the vertical direction
t	Time
g	Gravitational acceleration
η	Free surface elevation
ζ	Elevation of water-sediment interface
B	Bedrock elevation
N_ξ	Normal vector to surface ξ, $\xi \in \{B, \zeta, \eta\}$
T_{xi}	Tangent vector to surface ξ, $\xi \in \{B, \zeta, \eta\}$
h	Water depth
b	Sediment depth
u_w	Horizontal water velocity vector in Navier-Stokes model
u_s	Horizontal sediment velocity vector in Navier-Stokes model
w_w	Vertical water velocity in Navier-Stokes model
w_s	Vertical sediment velocity in Navier-Stokes model
Σ_w	Viscosity tensor in water layer
Σ_s	Viscosity tensor in sediment layer
u	Horizontal water velocity vector in vertically integrated model
v	Horizontal sediment velocity vector in vertically integrated model
u_ζ	Forcing velocity on sediment surface
p_w	Pressure in water layer
p_s	Pressure in sediment layer
p_ζ	Forcing pressure on sediment surface
ρ_w	Water density
ρ_s	Sediment density
r	Density ratio of water layer with respect to sediment layer

Symbol	Description
μ_w	Water viscosity
μ_s	Sediment viscosity
κ_ζ	Friction coefficient at sediment-water interface
κ_B	Friction coefficient at sediment-bedrock interface
γ	Exponent in fluid friction law at sediment-bedrock interface
$\bar{\tau}$	Coulomb friction coefficient
τ	Shear stress exerted on sediment layer
τ_c	Critical shear stress
f_B	Friction effort at the surface of the substratum
ε	Shallowness parameter
Fr	Froude number
Re	Reynolds number in sediment layer
Re_w	Reynolds number in water layer
Θ_ζ	Large-scale Shields number at sediment-water interface
Θ_B	Large-scale Shields number at substratum
L	Characteristic value of horizontal dimension
H	Characteristic value of vertical dimension
T	Characteristic value of time
U	Characteristic value of horizontal velocity
W	Characteristic value of vertical velocity
P	Characteristic value of pressure
B_0	Vertical reference level
K_ζ	Characteristic value of friction coefficient at sediment-water interface
K_B	Characteristic value of friction coefficient at sediment-bedrock interface
C	Characteristic value of Coulomb friction coefficient
α	Scaling ratio between sediment velocity and forcing velocity
β	Impact of gravity on sediment motion
ω	Impact of viscosity on sediment motion
$\mathcal{E}_{tot}$	Potential energy of water-sediment system
$\mathcal{E}_w$	Potential energy of water layer
$\mathcal{E}_s$	Potential energy of sediment layer
$\mathcal{K}$	Kinetic energy of water layer
ϕ_w	Potential of water layer
ϕ_s	Potential of sediment layer
δ_t	Time step (constant)
δ_t^n	Time step at step n
δ_x	Space step
N_x	Number of cells in computational domain
N_f	Number of faces in computational domain
$\partial^\delta_{i+1/2}$	Centered finite difference at cell interface
∂^δ_i	Centered finite difference at cell center

Part II

The Navier-Stokes system with temperature and salinity for free-surface flows

3

Low-Mach approximation & layer-averaged formulation

Outline of the current chapter

The contents of this chapter will be submitted under the form of an article along with M.-O. Bristeau, F. Bouchut, A. Mangeney, J.Sainte-Marie and F. Souillé under the title "The Navier-Stokes system with temperature and salinity for free surface flows - Part I: Low-Mach approximation and layer-averaged formulation".

Abstract

In this paper, we are interested in free surface flows where density variations coming e.g. from temperature or salinity differences play a significant role. Starting from the compressible Navier-Stokes system, we derive the so-called Navier-Stokes-Fourier system in an incompressible context (the density does not depend on the fluid pressure). The low-Mach scaling is used. The case where the density depends only on the temperature is studied first. Then the variations of the

fluid density with respect to the temperature and the salinity are considered. We also give a layer-averaged formulation of the obtained models. Such a formulation is very useful for the numerical analysis and numerical approximations of the models that are presented in a companion paper.

Keywords: Navier-Stokes equations, compressible and incompressible fluids, free surface flows, variable density flows, low-Mach approximation, layer-averaged formulation

3.1 Introduction

In oceans and lakes, one of the predominant driving forces is the difference in density, caused by salinity and temperature variations (increasing the salinity and lowering the temperature of a fluid both increase its density) [111]. Oceans and lakes are stratified: the water density varies along the vertical direction. In the present work, we aim at describing and simulating variable density flows with free surface.

The density variations are usually small and a common assumption in the case of geophysical flows is the Boussinesq approximation [36]. It is widely used to simplify the Navier-Stokes equations with variable density and consists in ignoring density variations in momentum conservation equations except in the buoyancy force term. The consequences of the Boussinesq approximation are listed in [13]. The Boussinesq approximation is the basis of many ocean models such as ICON [122], NEMO [125] and POM [126]. Under the Boussinesq approximation, the density can be defined as a function of any given tracer, the temperature or a pollutant for instance. In the case of ocean water the density is a function of the pressure, the temperature and the salinity.

Several authors have shown the benefits of taking into account non-Boussinesq effects in lake and ocean models, either for the propagation of internal waves [130] or for sea level variations induced by expansion/contraction processes [69, 92, 100]. Moreover situations where the stratification due to temperature and/or salinity can be broken due to external forcing terms (e.g. the wind during upwelling phenomena) are common. This results in a mixing of fresh/cold/salted waters and hence isopycnal models where fluids of different densities are considered as non-miscible fluids are not relevant anymore [46, 71].

Consequently, approaches to take into account the non-Boussinesq effects have been developed. A possibility is to adopt pressure coordinates [77, 129]. The non-Boussinesq equations written in pressure coordinates are isomorphic to the Boussinesq equations in z-coordinates, which allows to use the same algorithm for the non-Boussinesq model as for the Boussinesq model. So far, however, this approach does not seem popular in ocean modeling, though it is available in the code MITgcm [124]. In [13], a non-hydrostatic non-Boussinesq model is presented. A non-hydrostatic pressure anomaly is related to a compressible (non-Boussinesq) density anomaly. Yet the authors of [13] are primarily concerned with the simulation of acoustic waves and the steric effect is not investigated numerically. This model has recently been included in CROCO [119].

The approach chosen here to propose a non-Boussinesq model is different. Starting from the compressible Navier-Stokes equations, we propose a formulation of the Navier-Stokes-Fourier system in the context of an incompressible fluid. (The term "Fourier" refers to Fourier's law of thermal conduction). For most geophysical flows, the water

can be considered as incompressible in the sense that the variation of its density with respect to the fluid pressure is small. In the incompressible model, the acoustic waves are no longer present. This is advantageous from the computational point of view because a restrictive condition must be imposed on the time step when the acoustic waves are included [107]. The incompressible limit of the Navier-Stokes-Fourier system has been extensively studied, see [4], [56] and the references therein. The incompressible limit is a low-Mach approximation of the Navier-Stokes-Fourier system. We pay close attention that the obtained models do not violate the second principle of thermodynamics. The Navier-Stokes-Fourier model derived in this paper does not rely on the Boussinesq approximation. Instead of conserving the volume of fluid and not the mass (which is what the Boussinesq approximation implies), the derived model strickly conserves the mass and is enriched by the thermohaline dilatation effects i.e. the volume is no longer conserved. In the second section of this paper, a layer-averaged model is derived from the Navier-Stokes-Fourier system. Vertically averaged multilayer models [6, 21, 23, 38] are a way to describe stratified flows and to overcome the limitations inherent to isopycnal models. The model proposed here is close to the one in [21], yet it is more rigorously derived here, and closer to physics. Moreover, the model in the present work is 3D while it was only 2D in [21]. We also prove that the obtained multilayer models (one for the Euler-Fourier system and one for the Navier-Stokes-Fourier system) satisfy an energy balance and we show that the stable equilibria of the multilayer model for the Euler system with variable density are those of the classical Euler system. Moreover, the multilayer approach does not require moving meshes. As the equations obtained on each layer are similar to the classical one-layer Shallow Water equations, we can use the existing robust and accurate techniques developed for the Shallow Water equations. A numerical scheme and numerical test cases are presented in a companion paper [31].

The paper is organized as follows. The incompressible Navier-Stokes-Fourier and Euler-Fourier models are derived in section 3.2. In section 3.3, the multilayer formulations are given and the properties of the multilayer models are analyzed.

3.2 The 3d Navier-Stokes-Fourier system

3.2.1 The compressible Navier-Stokes-Fourier system

We consider the classical compressible Navier-Stokes system describing a free surface gravitational flow over a bottom topography $z_b(x,y)$,

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{U}) = 0, \tag{3.1}$$

$$\frac{\partial (\rho \mathbf{U})}{\partial t} + \nabla \cdot (\rho \mathbf{U} \otimes \mathbf{U}) + \nabla p - \nabla \cdot \sigma = \rho \mathbf{g}, \tag{3.2}$$

$$\frac{\partial}{\partial t}\left(\rho \frac{|\mathbf{U}|^2}{2} + \rho e\right) + \nabla \cdot \left((\rho \frac{|\mathbf{U}|^2}{2} + \rho e + p - \sigma)\mathbf{U}\right) = -\nabla \cdot Q_T + \rho \mathbf{g}.\mathbf{U}, \tag{3.3}$$

where $\mathbf{U}(t,x,y,z) = (u,v,w)^T$ is the velocity, ρ is the density, p is the fluid pressure, σ is the viscosity stress and $\mathbf{g} = (0,0,-g)^T$ represents the gravity forces. The internal specific energy is denoted by e, the temperature by T. The heat flux Q_T obeys the Fourier law $Q_T = -\lambda \nabla T$, hence the name "Navier-Stokes-Fourier", λ being the heat conductivity. The quantity ∇ denotes $\nabla = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}\right)^T$. In the following, we will also use the notations $\mathbf{u}$ and $\nabla_{x,y}$, $\mathbf{u}(t,x,y,z) = (u,v)^T$ is the horizontal velocity and $\nabla_{x,y}$ corresponds to the projection of ∇ on the horizontal plane i.e. $\nabla_{x,y} = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}\right)^T$. The square norm of the velocity vector is $|\mathbf{U}|^2 = u^2 + v^2 + w^2$.

We consider a free surface flow (see Fig. 3.1), therefore we assume

$$z_b(x,y) \leqslant z \leqslant \eta(t,x,y) := h(t,x,y) + z_b(x,y)$$

with $z_b(x,y)$ the bottom elevation and $h(t,x,y)$ the water depth.

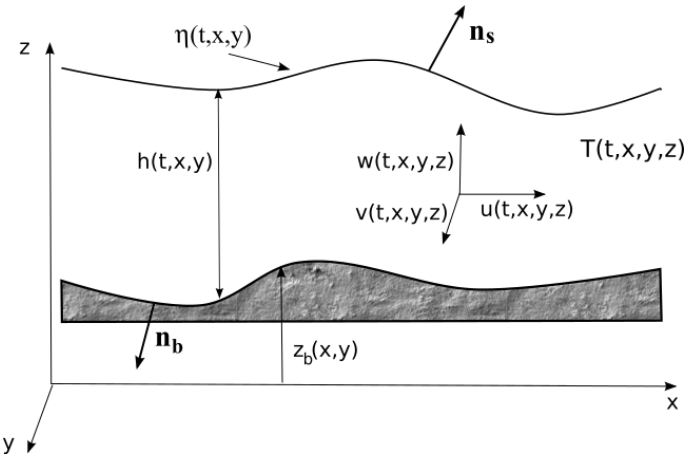


Figure 3.1 – Flow domain with water height $h(t,x,y)$, free surface $\eta(t,x,y)$ and bottom $z_b(x,y)$.

The term $-\rho g w$ in (3.3) prevents this equation from being directly a local energy conservation law. Nevertheless one can write it in terms of the gravitational potential energy, $\rho g w = \partial_t(\rho g z) + \nabla \cdot (\rho g z \mathbf{U})$, which leads to a conservative equation. The integration of this term is performed below, see Remark 1. For the sake of simplicity we work here with the local energy equation (3.3).

Regarding constitutive equations, we assume that the fluid is Newtonian i.e. the viscous part of the Cauchy stress depends linearly on the velocity gradient. Hence the stress tensor $\mathbf{\Sigma}$ is given by

$$\mathbf{\Sigma} \equiv -p\mathbb{1} + \sigma = -p\mathbb{1} + \zeta \nabla \cdot \mathbf{U}\mathbb{1} + 2\mu D(\mathbf{U}).$$

where μ is the viscosity coefficient, ζ is the second viscosity and $D(\mathbf{U}) = (\nabla \mathbf{U} +$

$(\nabla \mathbf{U})^T)/2$.

Among the thermodynamic variables ρ, p, T, e, only two of them are independent. This implies in particular that we have an equation of state under the form

$$f(\rho, T, p) = 0. \tag{3.4}$$

The thermodynamic variables are linked by the identity

$$de = \frac{p}{\rho^2} d\rho + T ds, \tag{3.5}$$

where s is the specific entropy of the fluid. Classically, in order to have a good entropy structure one has to assume that $-s$ is a convex function of $1/\rho, e$. In section 3.2.4 the case for which there is an additional thermodynamic variable S, the salinity, is described.

Energy equations can be deduced from the above equations. Multiplying (3.2) by $\mathbf{U}$ yields the kinetic energy equation

$$\frac{\partial}{\partial t}\left(\rho \frac{|\mathbf{U}|^2}{2}\right) + \nabla \cdot \left((\rho \frac{|\mathbf{U}|^2}{2} + p - \sigma)\mathbf{U}\right) = p\nabla \cdot \mathbf{U} - \sigma : D(\mathbf{U}) + \rho \mathbf{g}.\mathbf{U}. \tag{3.6}$$

Subtracting (3.6) to (3.3) gives the equation for the internal energy

$$\frac{\partial \rho e}{\partial t} + \nabla \cdot (\rho e \mathbf{U}) = -p\nabla \cdot \mathbf{U} + \sigma : D(\mathbf{U}) - \nabla \cdot Q_T,$$

or equivalently

$$\rho \frac{de}{dt} = -p\nabla \cdot \mathbf{U} + \sigma : D(\mathbf{U}) - \nabla \cdot Q_T, \tag{3.7}$$

with the classical notation $d/dt \equiv \partial/\partial t + \mathbf{U} \cdot \nabla$. We can write the continuity equation (3.1) as

$$\rho \frac{d\rho}{dt} + \rho^2 \nabla \cdot \mathbf{U} = 0. \tag{3.8}$$

With the thermodynamic relation (3.5) one can write $ds = de/T - (p/T\rho^2)d\rho$, thus multiplying (3.7) by $1/T$ and (3.8) by $-p/T\rho^2$ we obtain

$$\rho \frac{ds}{dt} = \frac{1}{T}\sigma : D(\mathbf{U}) - \frac{1}{T}\nabla \cdot Q_T.$$

This can be written also

$$\frac{\partial \rho s}{\partial t} + \nabla \cdot (\rho s \mathbf{U}) = \frac{1}{T}\sigma : D(\mathbf{U}) - \nabla \cdot \frac{Q_T}{T} - Q_T \cdot \frac{\nabla T}{T^2}, \tag{3.9}$$

which gives the increase with time of $\int \rho s$, the second principle of thermodynamics.

3.2.2 Boundary conditions

Bottom and free surface

Let $\mathbf{n}_b$ and $\mathbf{n}_s$ be the unit outward normals at the bottom and at the free surface respectively, defined by (see Fig 3.1)

$$\mathbf{n}_b = \frac{1}{\sqrt{1+|\nabla_{x,y} z_b|^2}} \begin{pmatrix} \nabla_{x,y} z_b \\ -1 \end{pmatrix}, \qquad \mathbf{n}_s = \frac{1}{\sqrt{1+|\nabla_{x,y} \eta|^2}} \begin{pmatrix} -\nabla_{x,y} \eta \\ 1 \end{pmatrix}.$$

On the bottom we prescribe an impermeability condition

$$\mathbf{U} \cdot \mathbf{n}_b = 0, \tag{3.10}$$

and a friction condition given e.g. by a Navier law

$$(\boldsymbol{\Sigma} \cdot \mathbf{n}_b) \cdot \mathbf{t}_i = -\kappa \mathbf{U} \cdot \mathbf{t}_i, \quad i = 1, 2, \tag{3.11}$$

with κ a Navier coefficient and $(\mathbf{t}_i, i = 1, 2)$ two tangential vectors. For some applications, we rather use more specific friction laws and the equation (3.11) is then replaced by

$$(\Sigma \cdot \mathbf{n}_b) \cdot \mathbf{t}_i = -\kappa(h, \mathbf{U}) \cdot \mathbf{t}_i, \quad i = 1, 2,$$

with $\kappa(h, \mathbf{U}).\mathbf{U} \geqslant 0$. On the free surface, we use the kinematic boundary condition

$$\frac{\partial \eta}{\partial t} + \mathbf{u}(t, x, y, \eta) \cdot \nabla_{x,y} \eta - w(t, x, y, \eta) = 0, \tag{3.12}$$

and the no stress condition

$$\boldsymbol{\Sigma} \cdot \mathbf{n}_s = -p^a(t, x, y)\mathbf{n}_s + W(t, x, y)\mathbf{t}_s, \tag{3.13}$$

where $p^a(t, x, y)$, $W(t, x, y)$ are two given quantities, p^a (resp. W) mimics the effects of the atmospheric pressure (resp. the wind blowing at the free surface) and $\mathbf{t}_s$ is a given unit horizontal vector. Throughout the paper $p^a = cst$, $W = 0$. For the temperature, Neumann or Dirichlet boundary conditions can be taken, see subsection 3.2.5.

Remark 1. Computing the quantity $\int_{z_b}^{\eta} z \left(\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{U}) \right) dz$ and using the boundary conditions (3.12), (3.10) one finds

$$\frac{\partial}{\partial t} \int_{z_b}^{\eta} \rho z dz + \nabla_{x,y} \cdot \int_{z_b}^{\eta} \rho z \mathbf{u} dz = \int_{z_b}^{\eta} \rho w dz,$$

which is the integrated local conservation of gravitational potential energy.

Fluid boundaries and solid walls

On solid walls we prescribe a slip condition (neglecting the viscosity)

$$\mathbf{U}.\mathbf{n} = 0,$$

coupled with an homogeneous Neumann condition

$$\frac{\partial \mathbf{u}}{\partial \mathbf{n}} = 0,$$

$\mathbf{n}$ being the outward normal to the considered wall.

In this paper we consider fluid boundaries where we neglect the viscosity and on which we prescribe zero, one or two of the following conditions depending on the type of the flow (fluvial or torrential): water level $h + z_b(x, y)$ given, flux $h\mathbf{U}$ given.

The system is completed with some initial conditions

$$h(0, x, y) = h^0(x, y), \quad \rho(0, x, y) = \rho^0(x, y), \quad \mathbf{U}(0, x, y, z) = \mathbf{U}^0(x, y, z).$$

3.2.3 The incompressible limit

In this section, the incompressible limit of the compressible Navier-Stokes equations is performed. As already mentioned in the introduction, one of the motivations for this limit is that the density of the water varies very little with pressure variations, and removing acoustic waves from the model is advantageous from the computational point of view. Therefore, we now consider the equation of state of the fluid (3.4) under the form

$$\tilde{f}\big(\rho, T, \varepsilon(p - p_{ref})\big) = 0, \tag{3.14}$$

where $\varepsilon \ll 1$ is a small parameter and with p_{ref} a reference pressure constant in space and time. In other words we have assumed the particular form for the pressure

$$p = p_{ref} + \frac{p_0}{\varepsilon}, \tag{3.15}$$

with $p_0(\rho, T)$ having no small scale.

Remark 2. When writing equation (3.14), we assume that the density of the water depends very weakly on the pressure, and this is true in practice. A possible equation of state for seawater (involving the salinity S, which we will include later on in section 3.2.4) is to write (3.14) as

$$\rho(S, T, p) = \frac{\rho(S, T, p_{ref})}{1 - \frac{p - p_{ref}}{K(S, T, p - p_{ref})}},$$

where in the fraction $(p - p_{ref})/K(S, T, p - p_{ref})$, the denominator is very large with

respect to the numerator, so that this law could actually be written

$$\rho(S,T,p) = \frac{\rho(S,T,p_{ref})}{1-\varepsilon(p-p_{ref})}.$$

This law was published in [134], where values of the density ρ at different pressures and constant S, T are also given. One can see that the density varies slowly with respect to the pressure.

Remark 3. One could consider that the reference pressure p_{ref} varies in time, for instance because of changes in the boundary conditions of the system - p_{ref} adapts to temperature fluxes and mass fluxes at the boundaries. Here, for the sake of simplicity, we consider that p_{ref} is constant in space and in time. Yet the model derivation should not be significantly different with $p_{ref} = p_{ref}(t)$.

Taking into account the thermodynamic identity (3.5) and due to (3.15) it is necessary to consider the following rescaling for e and s,

$$e + \frac{p_{ref}}{\rho} = \frac{e_0}{\varepsilon}, \qquad s = \frac{s_0}{\varepsilon}, \quad \text{with} \quad de_0 = \frac{p_0}{\rho^2}d\rho + Tds_0. \tag{3.16}$$

The incompressible limit is performed by letting ε go to 0. As p is the physical pressure, it has to remain finite. Therefore according to (3.15) at the limit we get $p_0(\rho, T) = 0$. In other words

$$T = T^{eq}(\rho),$$

or equivalently $\rho = \rho(T^{eq})$. The superscript ^{eq} is used for the quantities at equilibrium, i.e. quantities constrained by the relation $p_0(\rho, T) = 0$.

We have the following result.

Proposition 1. *The system*

$$\nabla \cdot \mathbf{U} = -\frac{\rho'(T^{eq})}{\rho^2 c_p}\nabla \cdot (\lambda \nabla T^{eq}), \tag{3.17}$$

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{U}) = 0, \tag{3.18}$$

$$\frac{\partial (\rho \mathbf{U})}{\partial t} + \nabla \cdot (\rho \mathbf{U} \otimes \mathbf{U}) + \nabla p - \nabla \cdot \sigma = \rho \mathbf{g}, \tag{3.19}$$

with the relation $T = T^{eq}(\rho)$ and where p is a Lagrange multiplier, is the formal limit of the system (3.1)-(3.3), with (3.15), (3.16) as ε goes to 0. The energy balance verified by (3.17)-(3.19) is

$$\frac{\partial}{\partial t}\left(\rho\frac{|\mathbf{U}|^2}{2} - p_{ref} + \rho\frac{e_0^{eq}}{\varepsilon}\right) + \nabla \cdot \left(\left(\rho\frac{|\mathbf{U}|^2}{2} - p_{ref} + \rho\frac{e_0^{eq}}{\varepsilon} + p - \sigma\right)\mathbf{U}\right)$$
$$= \nabla \cdot \left(\frac{\lambda_0}{\varepsilon}\nabla T^{eq}\right) + \rho \mathbf{g}.\mathbf{U} + (p - p_{ref})\nabla \cdot \mathbf{U} - \sigma : D(\mathbf{U}). \tag{3.20}$$

The energy balance (3.20) is expressed in the rescaled variables defined by (3.15), (3.16) in order to show clearly the order of magnitude of each term. The quantities e_0^{eq} and λ_0 are defined further below.

Proof. We first rewrite (3.1) under the form

$$\rho \frac{d\rho}{dt} = -\rho^2 \nabla \cdot \mathbf{U}. \tag{3.21}$$

When $\varepsilon \to 0$ we get $\rho = \rho(T^{eq})$, and we can multiply (3.21) by $dT^{eq}/d\rho$ to get an equation for the temperature

$$\rho \frac{dT^{eq}}{dt} = -\rho^2 \frac{dT^{eq}}{d\rho} \nabla \cdot \mathbf{U}. \tag{3.22}$$

In the sequel, we consider that for the heat conduction λ and the fluid viscosity μ, we are in the following asymptotic regime,

$$\lambda = \frac{\lambda_0}{\varepsilon}, \quad \text{and} \quad \mu \sim 1.$$

Hence, multiplying (3.7) by ε and taking the limit, we get according to (3.16)

$$\rho \frac{de_0^{eq}}{dt} = \nabla \cdot (\lambda_0 \nabla T^{eq}). \tag{3.23}$$

Writing then with (3.15), (3.16) the enthalpy $H \equiv e + p/\rho = H_0/\varepsilon$ with $H_0 = e_0 + p_0/\rho$, one gets that the rescaled enthalpy at equilibrium H_0^{eq} and internal energy at equilibrium are equal

$$H_0^{eq} = e_0^{eq},$$

and $H_0^{eq} = H_0^{eq}(T^{eq})$. The differential relation linking H_0^{eq} and T_{eq} can then be written

$$dH_0^{eq} = c_{p_0} dT^{eq},$$

where $c_{p_0} = \varepsilon c_p$ is the rescaled heat capacity at constant pressure. We thus get a second equation for the temperature

$$\rho c_{p_0} \frac{dT^{eq}}{dt} = \nabla \cdot (\lambda_0 \nabla T^{eq}). \tag{3.24}$$

The compatibility condition between (3.22) and (3.24) is given by (3.17), where we have replaced the scaled quantities c_{p_0} and λ_0 by their physical values εc_p and $\varepsilon\lambda$ respectively. The pressure p in (3.19) can finally be interpreted as a Lagrange multiplier for the equation (3.17). The momentum equation (3.19) together with the mass equation (3.21) gives again the kinetic energy equation (3.6). Adding it to (3.23) divided by ε and to trivial terms in p_{ref} finally gives the energy balance (3.20). □

Remark 4. At the limit, the thermodynamic identity (3.16) becomes

$$de_0^{eq} = T^{eq} ds_0^{eq}.$$

From (3.23) we obtain the equation for the evolution of the entropy

$$\frac{\partial}{\partial t}(\rho s_0^{eq}) + \nabla \cdot (\rho s_0^{eq} \mathbf{U}) - \frac{1}{T^{eq}} \nabla \cdot (\lambda_0 \nabla T^{eq}) = 0.$$

Written in the conservative/dissipative form, this gives

$$\frac{\partial}{\partial t}(\rho s_0^{eq}) + \nabla \cdot (\rho s_0^{eq} \mathbf{U}) - \nabla \cdot \left(\lambda_0 \frac{\nabla T^{eq}}{T^{eq}} \right) = \lambda_0 \frac{|\nabla T^{eq}|^2}{(T^{eq})^2},$$

which shows that in accordance with the second law of thermodynamics, the total entropy $\int \rho s_0^{eq}$ can only increase.

The energy balance (3.20) exhibits discrepancies of the order ε with respect to the original energy balance (3.3). Namely, the terms $(p - p_{ref})\nabla \cdot \mathbf{U} - \sigma : D(\mathbf{U})$ are of size ε with respect to the leading order terms and they are not present in equation (3.3). In order to get an energy balance closer in spirit to the original compressible equations, we go one step further and some corrections of size ε are incorporated into the system.

Proposition 2. *The system*

$$\nabla \cdot \mathbf{U} = -\frac{\rho'(T^{eq})}{\rho^2 c_p} (\nabla \cdot (\lambda \nabla T^{eq}) + \sigma : D(\mathbf{U})), \tag{3.25}$$

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{U}) = 0, \tag{3.26}$$

$$\frac{\partial (\rho \mathbf{U})}{\partial t} + \nabla \cdot (\rho \mathbf{U} \otimes \mathbf{U}) + \nabla p - \nabla \cdot \sigma = \rho \mathbf{g}, \tag{3.27}$$

with the relation $T = T^{eq}(\rho)$ and where p is a Lagrange multiplier, is an approximation of order ε of the formal limit (3.17)-(3.19) of the system (3.1)-(3.3) with (3.15), (3.16). The system (3.25)-(3.27) satisfies the energy balance equation

$$\begin{aligned} \frac{\partial}{\partial t} \left(\rho \frac{|\mathbf{U}|^2}{2} - p_{ref} + \rho \frac{e_0}{\varepsilon} \right) + \nabla \cdot \left(\left(\rho \frac{|\mathbf{U}|^2}{2} - p_{ref} + \rho \frac{e_0}{\varepsilon} + p - \sigma \right) \mathbf{U} \right) \\ = \nabla \cdot \left(\frac{\lambda_0}{\varepsilon} \nabla T^{eq} \right) + \rho \mathbf{g}.\mathbf{U}, \end{aligned} \tag{3.28}$$

where e_0 is an independent (rescaled) energy variable such that $e_0 - e_0^{eq} = O(\varepsilon)$.

Proof. The proof is similar to that of proposition 1. At the limit, the variables are constrained by the relation $\rho = \rho(T^{eq})$. To obtain an energy balance close to (3.3), one must define an independent energy variable e_0 ($e_0 \neq e_0^{eq}$) satisfying

$$\rho \frac{de_0}{dt} = \nabla \cdot (\lambda_0 \nabla T^{eq}) - \varepsilon (p - p_{ref}) \nabla \cdot \mathbf{U} + \varepsilon \sigma : D(\mathbf{U}). \tag{3.29}$$

With this correction, the energy balance is now (3.28). The internal energy at equilibrium e_0^{eq} is taken to satisfy (3.29) but without the term $\varepsilon(p-p_{ref})\nabla\cdot\mathbf{U}$, which implies that $e_0 - e_0^{eq}$ is of order ε. Recalling that $H_0^{eq} = e_0^{eq}$, this corresponds to the equation for the enthalpy at equilibrium H_0^{eq}

$$\rho\frac{dH_0^{eq}}{dt} = \nabla\cdot(\lambda_0\nabla T^{eq}) + \varepsilon\sigma : D(\mathbf{U}). \tag{3.30}$$

Eq. (3.30) appears as a correction of (3.23). Notice that if the correction $-\varepsilon(p-p_{ref})\nabla\cdot\mathbf{U}$ were incorporated, the model obtained would be even more accurate, but it would contain derivatives of p, which makes the equations much more difficult to handle since p is a Lagrange multiplier. Here we restrict ourselves to the correction $\varepsilon\sigma : D(\mathbf{U})$ to keep the model simple.

From equation (3.30), we get a corrected equation on the temperature

$$\rho c_{p_0}\frac{dT^{eq}}{dt} = \nabla\cdot(\lambda_0\nabla T^{eq}) + \varepsilon\sigma : D(\mathbf{U}).$$

Consequently, the compatibility constraint (in the rescaled variables) becomes

$$\nabla\cdot\mathbf{U} = -\frac{\rho'(T^{eq})}{\rho^2 c_{p_0}}(\nabla\cdot(\lambda_0\nabla T^{eq}) + \varepsilon\sigma : D(\mathbf{U})),$$

which is (3.25). In physical variables, the equation for the temperature can also be written

$$\rho c_p\frac{dT^{eq}}{dt} = -\nabla\cdot Q_T + \sigma : D(\mathbf{U}).$$

□

Remark 5. The entropy equation for model (3.25)-(3.27) is

$$\frac{\partial}{\partial t}(\rho s_0^{eq}) + \nabla\cdot(\rho s_0^{eq}\mathbf{U}) - \nabla\cdot\left(\lambda_0\frac{\nabla T^{eq}}{T^{eq}}\right) = \lambda_0\frac{|\nabla T^{eq}|^2}{(T^{eq})^2} + \frac{\varepsilon}{T^{eq}}\sigma : D(\mathbf{U}).$$

As $\sigma : D(\mathbf{U}) \geqslant 0$, we obtain again that the total entropy $\int \rho s_0^{eq}$ can only increase. Here there is no discrepency with the original entropy equation (3.9).

Remark 6. In models (3.17)-(3.19) and (3.25)-(3.27), the temperature T^{eq} is no longer an independent variable of the system. It is recovered by inverting the equation of state $\rho = \rho(T^{eq})$.

Remark 7. The model (3.25)-(3.27) is very similar to what is classically obtained when the low-Mach limit of the Navier-Stokes equations with thermal conduction is taken, see for instance [107]. The limit is usually performed by expanding the variables of the system in power series of the Mach number. The method we have used here is different. As we have given ourselves only a generic equation of state, we cannot express the Mach number, let alone make it appear in the equations. Instead, the result is obtained via a rescaling of the variable part of the pressure.

Remark 8. An example of equation of state to which the previous asymptotics can be applied is the stiffened gas law [74]

$$p = (\gamma - 1)\rho e - \gamma p_\infty, \quad \frac{c_p}{\gamma} T = e - \frac{p_\infty}{\rho},$$

with the constraint $e - p_\infty/\rho > 0$, where $\gamma > 1$, $p_\infty > 0$, $c_p > 0$ are constants. Here the entropy is given by $s = \frac{c_p}{\gamma} \log(T/\rho^{\gamma-1})$. Then the scaling assumptions (3.15), (3.16) are satisfied when

$$p_\infty = -p_{ref} + \frac{p_{\infty 0}}{\varepsilon}, \qquad c_p = \frac{c_{p_0}}{\varepsilon},$$

with $p_{\infty 0}$ and c_{p_0} constants independent of ε. At equilibrium we get the relation $\rho T^{eq} = \frac{\gamma}{\gamma-1}\frac{p_{\infty 0}}{c_{p_0}}$, T^{eq} is inversely proportional to ρ.

Because of the stability inherited from the energy balance (3.28), that is consistent with (3.3), in the sequel we consider the system (3.25)-(3.27) instead of (3.17)-(3.19).

Remark 9. In [21] the authors focused on the following 2D (x, z) model

$$\frac{\partial \rho}{\partial t} + \frac{\partial \rho u}{\partial x} + \frac{\partial \rho w}{\partial z} = 0, \tag{3.31}$$

$$\frac{\partial \rho u}{\partial t} + \frac{\partial \rho u^2}{\partial x} + \frac{\partial \rho uw}{\partial z} + \frac{\partial p}{\partial x} = \frac{\partial \sigma_{xx}}{\partial x} + \frac{\partial \sigma_{xz}}{\partial z}, \tag{3.32}$$

$$\frac{\partial p}{\partial z} = -\rho g + \frac{\partial \sigma_{zx}}{\partial x} + \frac{\partial \sigma_{zz}}{\partial z}, \tag{3.33}$$

$$\rho = \rho(T), \tag{3.34}$$

$$\frac{\partial \rho T}{\partial t} + \frac{\partial \rho u T}{\partial x} + \frac{\partial \rho w T}{\partial z} = \frac{\lambda}{c_p}\frac{\partial^2 T}{\partial x^2} + \frac{\lambda}{c_p}\frac{\partial^2 T}{\partial z^2}. \tag{3.35}$$

Rewriting equation (3.35) in the non-conservative form, we get

$$\rho \frac{\partial T}{\partial t} + \rho u \frac{\partial T}{\partial x} + \rho w \frac{\partial T}{\partial z} = \frac{\lambda}{c_p}\left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial z^2}\right).$$

Multiplying this equation by $\rho'(T)/\rho$ gives an equation for ρ

$$\frac{\partial \rho}{\partial t} + u\frac{\partial \rho}{\partial x} + w\frac{\partial \rho}{\partial z} = \frac{\rho'(T)}{\rho}\frac{\lambda}{c_p}\left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial z^2}\right). \tag{3.36}$$

Finally, subtracting (3.31) to (3.36) and rearranging the terms gives a compatibility condition similar to (3.25) for a constant λ

$$\frac{\partial u}{\partial x} + \frac{\partial w}{\partial z} = \frac{\rho'(T)}{\rho^2 c_p}\lambda\left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial z^2}\right).$$

This shows that the model (3.31)-(3.35) is very close to a hydrostatic version of model (3.25)-(3.27).

3.2.4 The Navier-Stokes-Fourier system with salinity

We now consider the situation where the fluid density depends on the temperature T and on another internal variable, the salinity S. This is the case of sea water. The compressible Navier-Stokes-Fourier system with temperature and salinity can be written

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{U}) = 0, \tag{3.37}$$

$$\frac{\partial (\rho \mathbf{U})}{\partial t} + \nabla \cdot (\rho \mathbf{U} \otimes \mathbf{U}) + \nabla p - \nabla \cdot \sigma = \rho \mathbf{g}, \tag{3.38}$$

$$\frac{\partial}{\partial t}\left(\rho \frac{|\mathbf{U}|^2}{2} + \rho e\right) + \nabla \cdot \left(\left(\rho \frac{|\mathbf{U}|^2}{2} + \rho e + p - \sigma\right)\mathbf{U}\right) = -\nabla \cdot \mathbf{F}^T + \rho \mathbf{g}.\mathbf{U}, \tag{3.39}$$

$$\frac{\partial (\rho S)}{\partial t} + \nabla \cdot (\rho S \mathbf{U}) = -\nabla \cdot \mathbf{F}^S. \tag{3.40}$$

The local mass and momentum conservation equations are identical to (3.1) and (3.2), whereas the energy equation is slightly modified: the heat flux is now $\mathbf{F}^T$. The conservation equation on the mass fraction of chlorides S can also be written as

$$\rho \frac{dS}{dt} = -\nabla \cdot \mathbf{F}^S, \tag{3.41}$$

with $\mathbf{F}^S$ the salt flux. According to [80], the molecular fluxes of heat and salt $\mathbf{F}^T$ and $\mathbf{F}^S$ are expressed in terms of the thermodynamic Onsager forces related to the entropy equation (3.45) below,

$$\mathbf{F}^S = A \nabla \left(\frac{-\mu_S}{T}\right) + B \nabla \left(\frac{1}{T}\right), \tag{3.42}$$

$$\mathbf{F}^T = B \nabla \left(\frac{-\mu_S}{T}\right) + C \nabla \left(\frac{1}{T}\right), \tag{3.43}$$

where A, B and C are three independent coefficients to be specified later, and μ_S is the chemical potential of seawater. The equation of state of the fluid is

$$f(\rho, T, S, p) = 0,$$

and the thermodynamic identity now reads

$$de = \frac{p}{\rho^2} d\rho + T ds + \mu_S dS. \tag{3.44}$$

A natural assumption for the hyperbolic structure of the model is that $-s$ is a convex function of $1/\rho, e, S$. From (3.39), we get the equation on the internal energy

$$\rho \frac{de}{dt} = -p \nabla \cdot \mathbf{U} + \sigma : D(\mathbf{U}) - \nabla \cdot \mathbf{F}^T.$$

Let us explain how the formulas (3.42), (3.43) lead to the second law of thermodynamics. The equation for the entropy is obtained using the thermodynamic identity (3.44) combined to the mass and and salinity equations (3.8), (3.41),

$$\rho\left(\frac{\partial s}{\partial t}+\mathbf{U}\cdot\nabla s\right)=\frac{1}{T}\left(\sigma:D(\mathbf{U})-\nabla\cdot\mathbf{F}^T+\mu_S\nabla\cdot\mathbf{F}^S\right),$$

that can be written under conservative/dissipative form

$$\frac{\partial\rho s}{\partial t}+\nabla\cdot(\rho s\mathbf{U})=\frac{1}{T}\sigma:D(\mathbf{U})-\nabla\cdot\left(\frac{1}{T}\mathbf{F}^T-\frac{\mu_S}{T}\mathbf{F}^S\right)+\mathbf{F}^T\cdot\nabla\left(\frac{1}{T}\right)-\mathbf{F}^S\cdot\nabla\left(\frac{\mu_S}{T}\right). \quad (3.45)$$

Substituting the expressions (3.42), (3.43) in the right-hand side of (3.45), we obtain the following quadratic form for the nonconservative terms

$$\mathbf{F}^T\cdot\nabla\left(\frac{1}{T}\right)-\mathbf{F}^S\cdot\nabla\left(\frac{\mu_S}{T}\right)=C\left|\nabla\left(\frac{1}{T}\right)\right|^2-2B\nabla\left(\frac{1}{T}\right)\cdot\nabla\left(\frac{\mu_S}{T}\right)+A\left|\nabla\left(\frac{\mu_S}{T}\right)\right|^2. \quad (3.46)$$

For this quadratic form to be nonnegative, the three constraints are $A>0$, $C>0$ and $AC>B^2$. With these constraints the expressions (3.42), (3.43) of $\mathbf{F}^S$ and $\mathbf{F}^T$ can be written in terms of the gradients of the salinity S, temperature T and pressure p (as in [80], equations (B.26) and (B.27)) by writing $\mu_S=\mu_S(T,S,p)$ and assuming $\partial_S\mu_S>0$,

$$\mathbf{F}^S=-\rho k^S\left(\nabla S+\frac{\partial_p\mu_S}{\partial_S\mu_S}\nabla p\right)-\left(\frac{\rho k^S T}{\partial_S\mu_S}\partial_T\left(\frac{\mu_S}{T}\right)+\frac{B}{T^2}\right)\nabla T, \quad (3.47)$$

$$\mathbf{F}^T=-\rho c_p k^T\nabla T+\frac{B\partial_S\mu_S}{\rho k^S T}\mathbf{F}^S, \quad (3.48)$$

where $k^T>0$ and $k^S>0$ are the thermal and molecular diffusivities of salt, related to A, B, C by

$$A=\frac{\rho k^S T}{\partial_S\mu_S},\qquad C=\rho c_p k^T T^2+\frac{B^2}{A}. \quad (3.49)$$

The free coefficients are thus now k^S, k^T and B. Note that $\mathbf{F}^T$ in (3.48) is written as a gradient of T (as in the case where the temperature is the only tracer), plus another term, due to the presence of salt. Using (3.49) and (3.42), the quadratic form (3.46) can be rewritten

$$C\left|\nabla\left(\frac{1}{T}\right)\right|^2-2B\nabla\left(\frac{1}{T}\right)\cdot\nabla\left(\frac{\mu_S}{T}\right)+A\left|\nabla\left(\frac{\mu_S}{T}\right)\right|^2=\rho c_p k^T T^2\left|\nabla\frac{1}{T}\right|^2+\frac{1}{A}\left|\mathbf{F}^S\right|^2, \quad (3.50)$$

which shows that it is indeed nonnegative. Thus with (3.45) the total entropy $\int\rho s$ can only increase, in accordance with the second principle of thermodynamics.

We now perform the incompressible limit as in section 3.2.3. We introduce the equa-

tion of state of the fluid under the form

$$\tilde{f}\big(\rho, T, S, \varepsilon(p - p_{ref})\big) = 0, \tag{3.51}$$

with $\varepsilon \ll 1$. We rescale p as in (3.15). Taking into account the thermodynamic identity (3.44), e, s are rescaled as in (3.16) and μ_S scales as $1/\varepsilon$, which yields

$$p = p_{ref} + \frac{p_0}{\varepsilon}, \quad e + \frac{p_{ref}}{\rho} = \frac{e_0}{\varepsilon}, \quad s = \frac{s_0}{\varepsilon}, \quad \mu_S = \frac{\mu_{S0}}{\varepsilon}, \tag{3.52}$$

$$\text{with} \quad de_0 = \frac{p_0}{\rho^2} d\rho + T ds_0 + \mu_{S0} dS. \tag{3.53}$$

As $\varepsilon \to 0$, the finiteness of p yields the equilibrium relation $p_0(\rho, T, S) = 0$ or equivalently $T = T^{eq}(\rho, S)$.

Proposition 3. *The system*

$$\nabla \cdot \mathbf{U} = \frac{1}{\rho^2 c_p} \left(\frac{\partial \rho}{\partial T^{eq}}\right)_S \left((T^{eq})^2 \left(\frac{\partial \left(\mu_S / T^{eq}\right)}{\partial T^{eq}}\right)_S \nabla \cdot \mathbf{F}^S + \nabla \cdot \mathbf{F}^T - \sigma : D(\mathbf{U}) \right) + \frac{1}{\rho^2} \left(\frac{\partial \rho}{\partial S}\right)_{T^{eq}} \nabla \cdot \mathbf{F}^S, \tag{3.54}$$

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{U}) = 0, \tag{3.55}$$

$$\frac{\partial (\rho \mathbf{U})}{\partial t} + \nabla \cdot (\rho \mathbf{U} \otimes \mathbf{U}) + \nabla p - \nabla \cdot \sigma = \rho \mathbf{g}, \tag{3.56}$$

$$\frac{\partial (\rho S)}{\partial t} + \nabla \cdot (\rho S \mathbf{U}) = -\nabla \cdot \mathbf{F}^S, \tag{3.57}$$

$$\mathbf{F}^S = -\rho k^S \nabla S - \left(\frac{\rho k^S T}{\partial_S \mu_S} \partial_T \left(\frac{\mu_S}{T}\right) + \frac{B}{T^2} \right) \nabla T, \tag{3.58}$$

$$\mathbf{F}^T = -\rho c_p k^T \nabla T + \frac{B \partial_S \mu_S}{\rho k^S T} \mathbf{F}^S, \tag{3.59}$$

with $T = T^{eq}(\rho, S)$ and where p is a Lagrange multiplier, is an approximation of order ε of the formal limit of the system (3.37)-(3.40), (3.47), (3.48), (3.52).

Proof. At the limit there remain only two independent thermodynamic variables, and one can take T and S. We consider that $\mathbf{F}^S$ is bounded but $\mathbf{F}^T \sim 1/\varepsilon$. With (3.42), (3.43) this means that $A \sim \varepsilon$, $B \sim 1$, $C \sim 1/\varepsilon$, and using (3.49) that

$$k^S \sim 1, \quad c_p k^T \sim 1/\varepsilon.$$

At equilibrium the term in ∇p in (3.47) disappears in the expression of $\mathbf{F}^S$ since μ_S depends only weakly on p (this is a consequence of the scaling assumption (3.51)) giving (3.58), (3.59). The mass conservation equation (3.37), the momentum equation (3.38) and the salinity equation (3.40) are unchanged. Considering the enthalpy $H = e + p/\rho = H_0/\varepsilon$, the equation for the rescaled enthalpy at equilibrium $H_0^{eq} = e_0^{eq}$ becomes, using

the correction as in section 3.2.3,

$$\rho \frac{dH_0^{eq}}{dt} = -\varepsilon \nabla \cdot \mathbf{F}^T + \varepsilon \sigma : D(\mathbf{U}). \tag{3.60}$$

One can write at equilibrium

$$dH^{eq} = \left(\frac{\partial H^{eq}}{\partial T^{eq}}\right)_S dT^{eq} + \left(\frac{\partial H^{eq}}{\partial S}\right)_{T^{eq}} dS. \tag{3.61}$$

Combining (3.60) (written in the physical variables) with (3.61) gives

$$\rho \left(\frac{\partial H^{eq}}{\partial T^{eq}}\right)_S \frac{dT^{eq}}{dt} - \left(\frac{\partial H^{eq}}{\partial S}\right)_{T^{eq}} \nabla \cdot \mathbf{F}^S + \nabla \cdot \mathbf{F}^T - \sigma : D(\mathbf{U}) = 0. \tag{3.62}$$

We have similarly for the density

$$d\rho = \left(\frac{\partial \rho}{\partial T^{eq}}\right)_S dT^{eq} + \left(\frac{\partial \rho}{\partial S}\right)_{T^{eq}} dS.$$

The quantities $\left(\frac{\partial \rho}{\partial T^{eq}}\right)_S$ and $\left(\frac{\partial \rho}{\partial S}\right)_{T^{eq}}$ are known from the equation of state of salted water, and using (3.21), (3.41) we deduce another equation on the temperature

$$-\rho^2 \nabla \cdot \mathbf{U} = \left(\frac{\partial \rho}{\partial T^{eq}}\right)_S \rho \frac{dT^{eq}}{dt} - \left(\frac{\partial \rho}{\partial S}\right)_{T^{eq}} \nabla \cdot \mathbf{F}^S. \tag{3.63}$$

Combining (3.62) with (3.63) gives an expression for $\rho^2 \nabla \cdot \mathbf{U}$

$$\rho^2 \nabla \cdot \mathbf{U} = \frac{\left(\frac{\partial \rho}{\partial T^{eq}}\right)_S}{\left(\frac{\partial H^{eq}}{\partial T^{eq}}\right)_S} \left(- \left(\frac{\partial H^{eq}}{\partial S}\right)_{T^{eq}} \nabla \cdot \mathbf{F}^S + \nabla \cdot \mathbf{F}^T - \sigma : D(\mathbf{U}) \right) + \left(\frac{\partial \rho}{\partial S}\right)_{T^{eq}} \nabla \cdot \mathbf{F}^S, \tag{3.64}$$

that generalizes (3.25). We recall that by definition $c_p = \left(\frac{\partial H}{\partial T}\right)_{S,p}$, thus $c_p = c_{p_0}/\varepsilon$ and at equilibrium

$$c_{p_0} = \left(\frac{\partial H_0}{\partial T^{eq}}\right)_S, \tag{3.65}$$

which enables to express the denominator in (3.64). Note that we obtain (3.64) without using the thermodynamic identity (3.44), and without involving μ_S. Next in (3.64) it remains to express $\partial H^{eq}/\partial S$. One has at equilibrium (see [80], equations (A.11.1) and (A.11.2))

$$\left(\frac{\partial H^{eq}}{\partial S}\right)_{T^{eq}} = \mu_S - T^{eq} \frac{\partial \mu_S}{\partial T^{eq}} = -(T^{eq})^2 \frac{\partial}{\partial T^{eq}} (\mu_S / T^{eq}), \tag{3.66}$$

thus finally (3.64) gives (3.54). The relation (3.66) can be deduced from the limit of the

thermodynamic identity (3.52). At equilibrium we have

$$ds_0^{eq} = \frac{dH_0^{eq}}{T^{eq}} - \frac{\mu_{S0}}{T^{eq}} dS, \tag{3.67}$$

which we reformulate as

$$d\left(s_0^{eq} - \frac{H_0^{eq}}{T^{eq}}\right) = \frac{H_0^{eq}}{(T^{eq})^2} dT^{eq} - \frac{\mu_{S0}}{T^{eq}} dS.$$

The left-hand side is an exact differential form, therefore we can write that the two cross derivatives with respect to T, S and S, T are equal. It yields

$$\frac{\partial}{\partial T^{eq}}\left(-\frac{\mu_{S0}}{T^{eq}}\right) = \frac{\partial}{\partial S}\left(\frac{H_0^{eq}}{(T^{eq})^2}\right),$$

which gives (3.66). □

Remark 10. Because of (3.67), (3.60), the entropy equation (3.45) is still valid for our model (3.54)-(3.59), and the quadratic form on the right-hand side takes the form (3.50).

Remark 11. A criterion of well-posedness of our incompressible system (3.54)-(3.59) can be derived as follows. We write that the second-order terms in the coupled S and T^{eq} equations (3.57), (3.62) give a diffusion matrix with positive eigenvalues. With (3.58), (3.59) we have at equilibrium (using (3.66))

$$\mathbf{F}^S = -\rho k^S \nabla S - E \nabla T, \qquad \text{with } E = \frac{\rho k^S T}{\partial_S \mu_S} \partial_T\left(\frac{\mu_S}{T}\right) + \frac{B}{T^2},$$

$$\begin{aligned}\mathbf{F}^T - \left(\frac{\partial H^{eq}}{\partial S}\right)_T \mathbf{F}^S &= -\rho c_p k^T \nabla T + \left(\frac{B \partial_S \mu_S}{\rho k^S T} + T^2 \partial_T(\mu_S/T)\right) \mathbf{F}^S \\ &= -\rho c_p k^T \nabla T + E \frac{T \partial_S \mu_S}{\rho k^S} \mathbf{F}^S.\end{aligned}$$

The diffusion matrix of the system is thus, taking into account (3.65),

$$\begin{pmatrix} \rho k^S & E \\ E\frac{T\partial_S \mu_S}{c_p} & E^2 \frac{T \partial_S \mu_S}{\rho k^S c_p} + \rho k^T \end{pmatrix}.$$

We obtain positive eigenvalues under the natural conditions also mentioned in [80]

$$k^S > 0, \quad k^T > 0, \quad \partial_S \mu_S > 0, \quad c_p > 0.$$

Note that for the particular choice of B such that $E = 0$ we have a diagonal diffusion matrix in S, T.

3.2.5 The Euler-Fourier system

Neglecting the fluid viscosity, the Navier-Stokes-Fourier system (3.25)-(3.27) derived in paragraph 3.2.3 reads (for the sake of simplicity we consider that the density only depends on the temperature)

$$\nabla \cdot \mathbf{U} = -\frac{\rho'(T)}{\rho^2 c_p} \nabla \cdot (\lambda \nabla T), \tag{3.68}$$

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{U}) = 0, \tag{3.69}$$

$$\frac{\partial (\rho \mathbf{U})}{\partial t} + \nabla \cdot (\rho \mathbf{U} \otimes \mathbf{U}) + \nabla p = \rho \mathbf{g}. \tag{3.70}$$

For the sake of lightness, the exponent ^{eq} is dropped in this part and in the rest of the present document. The system (3.68)-(3.70) is completed with the boundary conditions (3.10), (3.12) and

$$p(t, x, y, \eta) = p^a(t, x, y),$$

the previous condition coming from (3.13) when the viscosity vanishes. Boundary conditions for the temperature also have to be considered, we can choose either Neumann or Dirichlet conditions namely at the bottom

$$\lambda \nabla T . \mathbf{n}_b = F T_b^0, \tag{3.71}$$

or

$$T_b = T_b^0, \tag{3.72}$$

and at the free surface

$$\lambda \nabla T . \mathbf{n}_s = F T_s^0, \tag{3.73}$$

or

$$T_s = T_s^0, \tag{3.74}$$

where FT_b^0, FT_s^0 are two given temperature fluxes and T_b^0, T_s^0 are two given temperatures. Since $\rho = \rho(T)$, no further conditions are necessary.

We call the model (3.68)-(3.70) the Euler-Fourier system. For simplicity, the hydrostatic assumption and the Boussinesq approximation are presented below for this system.

3.2.6 The hydrostatic assumption

The hydrostatic assumption consists in neglecting the vertical acceleration of the fluid

$$\rho \left(\frac{\partial w}{\partial t} + \frac{\partial uw}{\partial x} + \frac{\partial vw}{\partial y} + \frac{\partial w^2}{\partial z} \right) \approx 0,$$

see [37, 70, 98] for the analysis of hydrostatic models and for their asymptotic derivation [14, 59, 97]. This implies that the pressure reads

$$p = \int_z^\eta \rho g dz.$$

Therefore, the hydrostatic approximation of the system (3.68)-(3.70) consists in the model

$$\nabla \cdot \mathbf{U} = -\frac{\rho'(T)}{\rho^2 c_p} \nabla \cdot (\lambda \nabla T), \tag{3.75}$$

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{U}) = 0, \tag{3.76}$$

$$\frac{\partial (\rho \mathbf{u})}{\partial t} + \nabla_{x,y} \cdot (\rho \mathbf{u} \otimes \mathbf{u}) + \frac{\partial (\rho \mathbf{u} w)}{\partial z} + \nabla_{x,y} \int_z^\eta \rho g dz = 0, \tag{3.77}$$

completed with the boundary conditions (3.10), (3.12)

3.2.7 The Boussinesq assumption

In geophysical water flows, density variations are often considered as small and this allows justifying the Boussinesq assumption, which consists in considering the density variations only in the gravitational forces. More precisely, assuming

$$\rho = \rho(T) = \rho_0 + f(T),$$

with $f(T) \ll \rho_0$ and $f'(T)$ small leads to writing the incompressible hydrostatic Euler system (3.81)-(3.83) under the form

$$\nabla \cdot \mathbf{U} = 0, \tag{3.78}$$

$$\rho_0 c_p \frac{\partial T}{\partial t} + \nabla \cdot (T \mathbf{U}) = \nabla \cdot (\lambda \nabla T), \tag{3.79}$$

$$\rho_0 \left(\frac{\partial \mathbf{u}}{\partial t} + \nabla_{x,y} \cdot (\mathbf{u} \otimes \mathbf{u}) + \frac{\partial (\mathbf{u} w)}{\partial z} \right) + \nabla_{x,y} \int_z^\eta \rho g dz = 0, \tag{3.80}$$

where the density variations only appear on the gravitational forces. Notice that whereas in (3.75), the divergence of the velocity field equals the dilatation due to the temperature effects, the Boussinesq assumption implies a divergence free condition in (3.78) in order to obtain an energy balance.

The Boussinesq assumption is valid in various regimes [69, 99] but

- it does not ensure a conservation of the kinetic energy since $\rho_0 \frac{|\mathbf{u}|^2}{2}$ is conserved instead of $\rho \frac{|\mathbf{u}|^2}{2}$,
- for long time phenomena (sloshing, wave propagation,...) significant differences appear when the Boussinesq assumption is made, see [21, paragraph 6.2].

In this work, the Boussinesq assumption is not done and some remarks about its validity are given in the following paragraph, see also [69, 99].

3.3 The layer-averaged models

In this section we propose a layer-averaged formulation of the Navier-Stokes-Fourier system (3.25)-(3.27) i.e. a flow where the density only depends on a single tracer, typically the temperature T. In a first step we neglect the viscous effects within the fluid and the diffusion terms for the temperature. Therefore we consider the incompressible and hydrostatic Euler system with variable density and free surface defined by (3.81)-(3.83).

$$\nabla \cdot \mathbf{U} = 0, \tag{3.81}$$

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{U}) = 0, \tag{3.82}$$

$$\frac{\partial (\rho \mathbf{u})}{\partial t} + \nabla_{x,y} \cdot (\rho \mathbf{u} \otimes \mathbf{u}) + \frac{\partial (\rho \mathbf{u} w)}{\partial z} + \nabla_{x,y} \int_z^\eta \rho g dz = 0. \tag{3.83}$$

Then in paragraph 3.3.2, the dissipative terms will be considered.

3.3.1 The layer-averaged Euler system with variable density

In order to describe and simulate complex flows where the velocity field cannot be approximated by its vertical mean, multilayer models have been developed [14, 19, 20, 34, 40, 41]. Unfortunately these models are physically relevant for non miscible fluids. In [21, 22, 57, 116], some authors have proposed a simpler and more general formulation for multilayer model with mass exchanges between the layers. The obtained model has the form of a conservation law with source terms. The layer-averaged approximation of the 3d Navier-Stokes system with constant density is studied in [6]. Compared to the constant density case, when considering the density variations, additional source terms appear, see remark 12. Notice that in [21] the hydrostatic Navier-Stokes equations with variable density is tackled but only in the 2d context.

With respect to commonly used Euler or Navier-Stokes approximations, the appealing features of the proposed multilayer approach are the easy handling of the free surface, which does not require moving meshes (e.g. [47]), and the possibility to take advantage of robust and accurate numerical techniques developed in extensive amount for classical one-layer Saint-Venant equations.

We consider a discretization of the fluid domain by layers (see Fig. 3.2) where the layer α contains the points of coordinates (x, y, z) with $z \in L_\alpha(t, x, y) = (z_{\alpha-1/2}, z_{\alpha+1/2})$ and $\{z_{\alpha+1/2}\}_{\alpha=1,\dots,N}$ is defined by

$$\begin{cases} z_{\alpha+1/2}(t,x,y) = z_b(x,y) + \sum_{j=1}^{\alpha} h_j(t,x,y), \quad \alpha \in [0,\dots,N], \\ h_\alpha(t,x,y) = z_{\alpha+1/2}(t,x,y) - z_{\alpha-1/2}(t,x,y) = l_\alpha h(t,x,y), \end{cases} \tag{3.84}$$

and $\sum_{\alpha=1}^{N} l_\alpha = 1$.

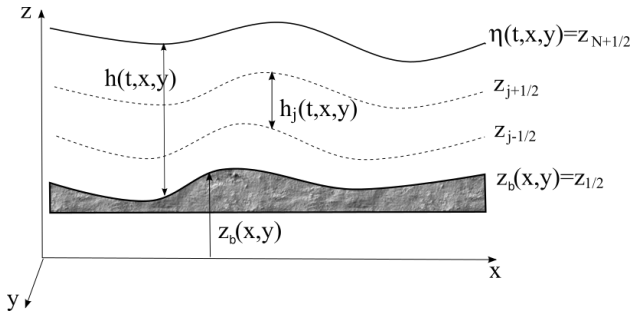


Figure 3.2 – Notations for the layerwise discretization.

The layer-averaging process for the 2d hydrostatic Euler and Navier-Stokes systems is precisely described in the paper [38] with a general rheology and in [6] for the 3d Navier-Stokes system with constant density, the reader can refer to it. In the following, we present a Galerkin type approximation of the Euler system also leading to a layer-averaged version of the Euler system.

Using the notations (4.2.1), let us consider the space $\mathbb{P}_{0,h}^{N,t}$ of piecewise constant functions defined by

$$\mathbb{P}_{0,h}^{N,t} = \left\{ \mathbb{1}_{z\in L_\alpha(t,x,y)}(z), \quad \alpha \in \{1,\dots,N\} \right\}, \tag{3.85}$$

where $\mathbb{1}_{z\in L_\alpha(t,x,y)}(z)$ is the characteristic function of the layer $L_\alpha(t,x,y)$. Using this formalism, the projection of ρ, u, v and w on $\mathbb{P}_{0,h}^{N,t}$ is a piecewise constant function defined by

$$X^N(t,x,y,z,\{z_\alpha\}) = \sum_{\alpha=1}^{N} \mathbb{1}_{[z_{\alpha-1/2},z_{\alpha+1/2}]}(z) X_\alpha(t,x,y), \tag{3.86}$$

for $X \in (\rho,u,v,w)$. When the quantities $\{\rho_\alpha(t,x,y)\}_{\alpha=1,\dots,N}$ are known, if the function $T \mapsto \rho(T)$ is invertible, it is possible to recover the temperature using the formula

$$T^N(t,x,z,\{z_\alpha\}) = \sum_{\alpha=1}^{N} \mathbb{1}_{[z_{\alpha-1/2},z_{\alpha+1/2}]}(z) \rho^{-1}(\rho_\alpha(t,x,y)).$$

In the following, we no more handle variables corresponding to vertical means of the solution of the Euler equations (3.81)-(3.83) and we adopt notations inherited from (3.86).

The three following propositions hold.

Proposition 4. *Using the space $\mathbb{P}_{0,h}^{N,t}$ defined by (3.85) and the decomposition (3.86), the Galerkin approximation of the incompressible and hydrostatic Euler equations (3.81)-*

(3.83),(3.10),(3.12) *leads to the system*

$$\frac{\partial h}{\partial t} + \sum_{\alpha=1}^{N} \nabla_{x,y} \cdot (h_\alpha \mathbf{u}_\alpha) = 0, \tag{3.87}$$

$$\frac{\partial \rho_\alpha h_\alpha}{\partial t} + \nabla_{x,y} \cdot (\rho_\alpha h_\alpha \mathbf{u}_\alpha) = \rho_{\alpha+1/2} G_{\alpha+1/2} - \rho_{\alpha-1/2} G_{\alpha-1/2}, \qquad \alpha = 1, ..., N, \tag{3.88}$$

$$\begin{aligned}\frac{\partial \rho_\alpha h_\alpha \mathbf{u}_\alpha}{\partial t} + \nabla_{x,y} \cdot (\rho_\alpha h_\alpha \mathbf{u}_\alpha \otimes \mathbf{u}_\alpha) + \nabla_{x,y}\big(h_\alpha p_\alpha\big) &= p_{\alpha+1/2} \nabla_{x,y} z_{\alpha+1/2} - p_{\alpha-1/2} \nabla_{x,y} z_{\alpha-1/2} \\ &+ \mathbf{u}_{\alpha+1/2} \rho_{\alpha+1/2} G_{\alpha+1/2} - \mathbf{u}_{\alpha-1/2} \rho_{\alpha-1/2} G_{\alpha-1/2}, \qquad \alpha = 1, ..., N,\end{aligned} \tag{3.89}$$

where the presure terms p_α, $p_{\alpha+1/2}$ *are given by*

$$p_\alpha = g\left(\frac{\rho_\alpha h_\alpha}{2} + \sum_{j=\alpha+1}^{N} \rho_j h_j\right) \qquad \text{and} \qquad p_{\alpha+1/2} = g \sum_{j=\alpha+1}^{N} \rho_j h_j. \tag{3.90}$$

The quantity $G_{\alpha+1/2}$ *(resp.* $G_{\alpha-1/2}$*) corresponds to mass exchange accross the interface* $z_{\alpha+1/2}$ *(resp.* $z_{\alpha-1/2}$*) and* $G_{\alpha+1/2}$ *is defined by*

$$G_{\alpha+1/2} = \sum_{j=1}^{\alpha} \left(\frac{\partial h_j}{\partial t} + \nabla_{x,y} \cdot (h_j \mathbf{u}_j)\right) = -\sum_{j=1}^{N} \Big(\sum_{p=1}^{\alpha} l_p - \mathbb{1}_{j \leqslant \alpha}\Big) \nabla_{x,y} \cdot (h_j \mathbf{u}_j), \tag{3.91}$$

for $\alpha = 1, \ldots, N$*. The velocities* $\mathbf{u}_{\alpha+1/2}$ *and the densities* $\rho_{\alpha+1/2}$ *at the interfaces are defined by*

$$v_{\alpha+1/2} = \begin{cases} v_\alpha & \text{if } G_{\alpha+1/2} \leqslant 0 \\ v_{\alpha+1} & \text{if } G_{\alpha+1/2} > 0 \end{cases} \tag{3.92}$$

for $v = \mathbf{u}, \rho$*.*

Remark 12. In the constant density case, the integration of the pressure term gives

$$\int_{z_{\alpha-1/2}}^{z_{\alpha+1/2}} \nabla_{x,y} p dz = \nabla_{x,y} \left(\rho_0 g \frac{h h_\alpha}{2}\right) + \rho_0 g h_\alpha \nabla_{x,y} z_b,$$

which is the sum of a conservative term and a source term depending on the given topography z_b. In the variable density case, the integration of the pressure term yields the terms

$$\int_{z_{\alpha-1/2}}^{z_{\alpha+1/2}} \nabla_{x,y} p dz = \nabla_{x,y}(h_\alpha p_\alpha) - p_{\alpha+1/2} \nabla_{x,y} z_{\alpha+1/2} + p_{\alpha-1/2} \nabla_{x,y} z_{\alpha-1/2}.$$

Note that $z_{\alpha+1/2}$, $z_{\alpha-1/2}$ are not given data, they depend on the unknown h. Therefore, the pressure source terms are more difficult to handle in the variable density case.

The smooth solutions of (3.87),(3.89) satisfy an energy balance and we have the following proposition.

Proposition 5. *The system (3.87),(3.89) admits, for smooth solutions, the energy balance*

$$\begin{aligned}
&\frac{\partial}{\partial t}E_\alpha + \nabla_{x,y}\cdot\left(\mathbf{u}_\alpha\left(E_\alpha + h_\alpha p_\alpha\right)\right)\\
&\quad= \left(\rho_{\alpha+1/2}\frac{|\mathbf{u}_{\alpha+1/2}|^2}{2} + g\rho_{\alpha+1/2}z_{\alpha+1/2}\right)G_{\alpha+1/2} + p_{\alpha+1/2}\big(G_{\alpha+1/2} - \frac{\partial z_{\alpha+1/2}}{\partial t}\big)\\
&\quad\quad - \left(\rho_{\alpha-1/2}\frac{|\mathbf{u}_{\alpha-1/2}|^2}{2} + g\rho_{\alpha-1/2}z_{\alpha-1/2}\right)G_{\alpha-1/2} - p_{\alpha-1/2}\big(G_{\alpha-1/2} - \frac{\partial z_{\alpha-1/2}}{\partial t}\big)\\
&\quad\quad -\frac{1}{2}\left(\rho_{\alpha+1/2}|\mathbf{u}_{\alpha+1/2}-\mathbf{u}_\alpha|^2 + gh_\alpha(\rho_{\alpha+1/2}-\rho_\alpha)\right)G_{\alpha+1/2}\\
&\quad\quad +\frac{1}{2}\left(\rho_{\alpha-1/2}|\mathbf{u}_{\alpha-1/2}-\mathbf{u}_\alpha|^2 - gh_\alpha(\rho_{\alpha-1/2}-\rho_\alpha)\right)G_{\alpha-1/2},
\end{aligned} \tag{3.93}$$

with

$$E_\alpha = \rho_\alpha\frac{h_\alpha|\mathbf{u}_\alpha|^2}{2} + \frac{\rho_\alpha g}{2}\big(z_{\alpha+1/2}^2 - z_{\alpha-1/2}^2\big). \tag{3.94}$$

The sum of Eqs. (3.93) for $\alpha = 1,\dots,N$ gives the energy balance

$$\begin{aligned}
&\frac{\partial}{\partial t}\sum_{\alpha=1}^{N}E_\alpha + \sum_{\alpha=1}^{N}\nabla_{x,y}\cdot\mathbf{u}_\alpha\left(E_\alpha + h_\alpha p_\alpha\right)\\
&\quad= -\sum_{\alpha=1}^{N}\rho_{\alpha+1/2}\frac{|\mathbf{u}_{\alpha+1}-\mathbf{u}_\alpha|^2}{2}|G_{\alpha+1/2}|\\
&\quad\quad -\frac{g}{2}\sum_{\alpha=1}^{N}\left(h_\alpha(\rho_{\alpha+1/2}-\rho_\alpha) + h_{\alpha+1}(\rho_{\alpha+1/2}-\rho_{\alpha+1})\right)G_{\alpha+1/2}.
\end{aligned} \tag{3.95}$$

In the energy balance (3.95), the first line of the right hand side is non positive due to the upwinding (3.92). Concerning the second, it is a third order term since we have

$$h_\alpha(\rho_{\alpha+1/2}-\rho_\alpha) + h_{\alpha+1}(\rho_{\alpha+1/2}-\rho_{\alpha+1}) \approx h_\alpha^3\left.\frac{\partial^2\rho}{\partial z^2}\right|_\alpha = \mathcal{O}(l_\alpha^3).$$

It is noticeable that, thanks to the kinematic boundary condition at each interface, the vertical velocity is no more a variable of the system (3.89). This is an advantage of this formulation over the hydrostatic model where the vertical velocity is needed in the momentum equation (3.77) and is deduced from the incompressibility condition (3.75). Even if the vertical velocity w no more appears in the model (3.87)-(3.89), it can be obtained as follows.

Proposition 6. *The piecewise constant approximation of the vertical velocity w satisfying Eq. (3.86) is given by*

$$w_\alpha = k_\alpha - z_\alpha\nabla_{x,y}\cdot\mathbf{u}_\alpha \tag{3.96}$$

with

$$k_1 \;=\; \nabla_{x,y}\cdot(z_b\mathbf{u}_1), \qquad k_{\alpha+1}=k_\alpha+\nabla_{x,y}\cdot\big(z_{\alpha+1/2}(\mathbf{u}_{\alpha+1}-\mathbf{u}_\alpha)\big).$$

The quantities $\{w_\alpha\}_{\alpha=1}^N$ *are obtained only using a post-processing of the variables governing the system* (3.87)-(3.89).

Notice that relation (3.96) *is equivalent to*

$$\frac{\partial z_\alpha}{\partial t}+\mathbf{u}_\alpha.\nabla_{x,y}z_\alpha = w_\alpha+\frac{G_{\alpha+1/2}+G_{\alpha-1/2}}{2}, \tag{3.97}$$

and using (3.99) *is also equivalent to*

$$\frac{\partial}{\partial t}\left(\rho_\alpha\frac{z_{\alpha+1/2}^2-z_{\alpha-1/2}^2}{2}\right)+\nabla_{x,y}\cdot\left(\rho_\alpha\frac{z_{\alpha+1/2}^2-z_{\alpha-1/2}^2}{2}\mathbf{u}_\alpha\right)=\rho_\alpha h_\alpha w_\alpha$$
$$+(z_\alpha\rho_{\alpha+1/2}+\rho_\alpha\frac{h_\alpha}{2})G_{\alpha+1/2}-(z_\alpha\rho_{\alpha-1/2}-\rho_\alpha\frac{h_\alpha}{2})G_{\alpha-1/2}, \tag{3.98}$$

see [38, paragraph 4.2].

Proof of prop. 4. Considering the divergence free condition (3.81), using the decomposition (3.86) and the space of test functions (3.85), we consider the quantity

$$\int_{\mathbb{R}}\mathbf{1}_{z\in L_\alpha(t,x,y)}\nabla\cdot\mathbf{U}^N dz=0,$$

with $\mathbf{U}^N=(u^N,v^N,w^N)^T$. Simple computations give

$$0=\int_{\mathbb{R}}\mathbf{1}_{z\in L_\alpha(t,x,y)}\nabla\cdot\mathbf{U}^{\mathbf{N}}dz=\frac{\partial h_\alpha}{\partial t}+\frac{\partial}{\partial x}\int_{z_{\alpha-1/2}}^{z_{\alpha+1/2}}u\,dz+\frac{\partial}{\partial y}\int_{z_{\alpha-1/2}}^{z_{\alpha+1/2}}v\,dz-G_{\alpha+1/2}+G_{\alpha-1/2},$$

leading to

$$\frac{\partial h_\alpha}{\partial t}+\nabla_{x,y}\cdot(h_\alpha\mathbf{u}_\alpha)=G_{\alpha+1/2}-G_{\alpha-1/2}, \tag{3.99}$$

with $G_{\alpha\pm1/2}$ defined by

$$G_{\alpha+1/2}=\frac{\partial z_{\alpha+1/2}}{\partial t}+\mathbf{u}_{\alpha+1/2}\cdot\nabla_{x,y}z_{\alpha+1/2}-w_{\alpha+1/2}.$$

The sum for $\alpha=1,\dots,N$ of the above relations gives Eq. (3.87) where the kinematic boundary conditions (3.10),(3.12) corresponding to

$$G_{1/2}=G_{N+1/2}=0, \tag{3.100}$$

have been used. Similarly, the sum for $j=1,\dots,\alpha$ of the relations (3.99) with (3.100) gives the expression (3.91) for $G_{\alpha+1/2}$.

Now we consider the Galerkin approximation of Eqs. (3.82),(3.83) i.e. the quantities

$$\int_{\mathbb{R}} \mathbf{1}_{z\in L_\alpha(t,x,y)} \left(\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{U}) dz \right) dz = 0,$$

and

$$\int_{\mathbb{R}} \mathbf{1}_{z\in L_\alpha(t,x,y)} \left(\frac{\partial (\rho^N \mathbf{u}^N)}{\partial t} + \nabla_{x,y} \cdot (\rho^N \mathbf{u}^N \otimes \mathbf{u}^N) + \frac{\partial (\rho^N \mathbf{u}^N w^N)}{\partial z} + \nabla_{x,y} \int_z^\eta \rho^N g dz \right) dz = 0,$$

leading, after simple computations, to Eqs. (3.88),(3.89). □

Proof of prop. 5. In order to obtain (3.93) we multiply Eq. (3.88) by $gz_\alpha - |\mathbf{u}_\alpha|^2/2$ and Eq. (3.89) by $\mathbf{u}_\alpha$, we sum the two obtained equations and we perform simple manipulations. More precisely, the momentum equation along the x axis multiplied by u_α reads

$$\left(\frac{\partial}{\partial t}(\rho_\alpha h_\alpha u_\alpha) + \frac{\partial}{\partial x}\left(\rho_\alpha h_\alpha u_\alpha^2 + h_\alpha p_\alpha\right) + \frac{\partial}{\partial y}\left(\rho_\alpha h_\alpha u_\alpha v_\alpha\right) \right) u_\alpha = \\ \left(p_{\alpha+1/2} \frac{\partial z_{\alpha+1/2}}{\partial x} - p_{\alpha-1/2} \frac{\partial z_{\alpha-1/2}}{\partial x} + u_{\alpha+1/2} G_{\alpha+1/2} - u_{\alpha-1/2} G_{\alpha-1/2} \right) u_\alpha.$$

The pressure terms are treated separately from the other terms. The previous equation is rewritten as

$$I_{u,\alpha} + \frac{\partial}{\partial x}(h_\alpha u_\alpha p_\alpha) = I_{p,u,\alpha},$$

with

$$I_{u,\alpha} = \left(\frac{\partial}{\partial t}(\rho_\alpha h_\alpha u_\alpha) + \frac{\partial}{\partial x}\left(\rho_\alpha h_\alpha u_\alpha^2\right) + \frac{\partial}{\partial y}\left(\rho_\alpha h_\alpha u_\alpha v_\alpha\right) - u_{\alpha+1/2} G_{\alpha+1/2} + u_{\alpha-1/2} G_{\alpha-1/2} \right) u_\alpha,$$

and

$$I_{p,u,\alpha} = h_\alpha p_\alpha \frac{\partial u_\alpha}{\partial x} + p_{\alpha+1/2} u_\alpha \frac{\partial z_{\alpha+1/2}}{\partial x} - p_{\alpha-1/2} u_\alpha \frac{\partial z_{\alpha-1/2}}{\partial x}.$$

Using (3.88) multiplied by $-u_\alpha^2/2$, the term $I_{u,\alpha}$ becomes

$$I_{u,\alpha} = \frac{\partial}{\partial t}\left(\frac{\rho_\alpha h_\alpha u_\alpha^2}{2}\right) + \frac{\partial}{\partial x}\left(u_\alpha \frac{\rho_\alpha h_\alpha u_\alpha^2}{2}\right) + \frac{\partial}{\partial y}\left(v_\alpha \frac{\rho_\alpha h_\alpha u_\alpha^2}{2}\right) \\ - \rho_{\alpha+1/2} \frac{u_{\alpha+1/2}^2}{2} G_{\alpha+1/2} + \rho_{\alpha-1/2} \frac{u_{\alpha-1/2}^2}{2} G_{\alpha-1/2} \\ + \rho_{\alpha+1/2} \frac{(u_{\alpha+1/2} - u_\alpha)^2}{2} G_{\alpha+1/2} - \rho_{\alpha-1/2} \frac{(u_{\alpha-1/2} - u_\alpha)^2}{2} G_{\alpha-1/2}. \quad (3.101)$$

When the second component of Eq. (3.89) is multiplied by v_α, we write in a similar

manner

$$I_{v,\alpha} + \frac{\partial}{\partial y}(h_\alpha v_\alpha p_\alpha) = I_{p,v,\alpha},$$

and a similar expression is obtained for $I_{v,\alpha}$.

The pressure terms $I_{p,u,\alpha}, I_{p,v,\alpha}$ are handled together. First we notice that

$$p_{\alpha+1/2} = p_\alpha - \frac{\rho_\alpha g h_\alpha}{2}, \quad \text{and} \quad p_{\alpha-1/2} = p_\alpha + \frac{\rho_\alpha g h_\alpha}{2},$$

so that

$$\begin{aligned} I_{p,u,\alpha} + I_{p,v,\alpha} &= h_\alpha p_\alpha \nabla_{x,y} \cdot \mathbf{u}_\alpha + p_{\alpha+1/2}\mathbf{u}_\alpha \cdot \nabla_{x,y} z_{\alpha+1/2} - p_{\alpha-1/2}\mathbf{u}_\alpha \cdot \nabla_{x,y} z_{\alpha-1/2} \\ &= p_\alpha \nabla_{x,y} \cdot (h_\alpha \mathbf{u}_\alpha) - g\rho_\alpha h_\alpha \mathbf{u}_\alpha \cdot \nabla_{x,y} z_\alpha. \end{aligned}$$

Using (3.99), the sum of the pressure terms becomes

$$I_{p,u,\alpha} + I_{p,v,\alpha} = p_\alpha \left(G_{\alpha+1/2} - G_{\alpha-1/2} - \frac{\partial h_\alpha}{\partial t} \right) - g\rho_\alpha h_\alpha \mathbf{u}_\alpha \cdot \nabla_{x,y} z_\alpha, \tag{3.102}$$

or equivalently

$$\begin{aligned} I_{p,u,\alpha} + I_{p,v,\alpha} = p_{\alpha+1/2} G_{\alpha+1/2} - p_{\alpha-1/2} G_{\alpha-1/2} - p_\alpha \frac{\partial h_\alpha}{\partial t} \\ + g\rho_\alpha \frac{h_\alpha}{2}\left(G_{\alpha+1/2} - G_{\alpha-1/2}\right) - g\rho_\alpha h_\alpha \mathbf{u}_\alpha \cdot \nabla_{x,y} z_\alpha. \end{aligned}$$

Finally, we obtain

$$\begin{aligned} I_{p,u,\alpha} + I_{p,v,\alpha} = p_{\alpha+1/2} \left(G_{\alpha+1/2} - \frac{\partial z_{\alpha+1/2}}{\partial t} \right) - p_{\alpha-1/2} \left(G_{\alpha-1/2} - \frac{\partial z_{\alpha-1/2}}{\partial t} \right) \\ + g\rho_\alpha \frac{h_\alpha}{2}\left(G_{\alpha+1/2} - G_{\alpha-1/2}\right) - g\rho_\alpha h_\alpha \mathbf{u}_\alpha \cdot \nabla_{x,y} z_\alpha - g\rho_\alpha h_\alpha \frac{\partial z_\alpha}{\partial t}, \end{aligned} \tag{3.103}$$

where

$$z_{\alpha\pm1/2} = z_\alpha \pm \frac{h_\alpha}{2}$$

has been used. Next, we multiply Eq. (3.88) by $g z_\alpha$ and we arrange the right-hand-side to get

$$\begin{aligned} g z_\alpha \frac{\partial(\rho_\alpha h_\alpha)}{\partial t} + g z_\alpha \nabla_{x,y} \cdot (\rho_\alpha h_\alpha) = g z_{\alpha+1/2} \rho_{\alpha+1/2} G_{\alpha+1/2} - g z_{\alpha-1/2} \rho_{\alpha-1/2} G_{\alpha-1/2} \\ - g \frac{h_\alpha}{2}\left(\rho_{\alpha+1/2} G_{\alpha+1/2} + \rho_{\alpha-1/2} G_{\alpha-1/2}\right) \end{aligned} \tag{3.104}$$

Summing

$$I_{u,\alpha} + I_{v,\alpha} + \nabla_{x,y} \cdot (h_\alpha \mathbf{u}_\alpha p_\alpha) = I_{p,u,\alpha} + I_{p,v,\alpha} \tag{3.105}$$

with Eq. (3.104) gives the result.

Finally summing the relations (3.93) for $\alpha = 1, \ldots, N$ gives (3.95) which completes the proof. □

Proof of prop. 6. Using the boundary condition (3.10), an integration from z_b to z of the divergence free condition (3.81) easily gives

$$w = -\nabla_{x,y} \cdot \int_{z_b}^{z} \mathbf{u}\, dz.$$

Replacing formally in the above equation $\mathbf{u}$ (resp. w) by $\mathbf{u}^N$ (resp. w^N) defined by (3.86) and performing an integration over the layer L_1 of the obtained relation yields

$$h_1 w_1 = -\int_{z_b}^{z_{3/2}} \nabla_{x,y} \cdot \int_{z_b}^{z} \mathbf{u}_1\, dz dz_1 = h_1 \nabla_{x,y} \cdot (z_b \mathbf{u}_1) - \frac{z_{3/2}^2 - z_b^2}{2} \nabla_{x,y} \cdot \mathbf{u}_1,$$

i.e.

$$w_1 = \nabla_{x,y} \cdot (z_b \mathbf{u}_1) - z_1 \nabla_{x,y} \cdot \mathbf{u}_1,$$

corresponding to (3.96) for $\alpha = 1$. A similar computation for the layers $L_2, \ldots, L_N$ proves the result (3.96) for $\alpha = 2, \ldots, N$.

A more detailed version of this proof is given in [38].

Notice that performing computations similar to those depicted in prop. 4 we obtain that

$$0 = \int_{z_\alpha}^{z_{\alpha+1/2}} \nabla \cdot \mathbf{U}^{\mathbf{N}} dz = w_{\alpha+1/2} - w_\alpha + \frac{h_\alpha}{2} \nabla_{x,y} \cdot \mathbf{u}_\alpha + (\mathbf{u}_\alpha - \mathbf{u}_{\alpha+1/2}) \cdot \nabla_{x,y} z_{\alpha+1/2}, \quad (3.106)$$

and

$$0 = \int_{z_{\alpha-1/2}}^{z_\alpha} \nabla \cdot \mathbf{U}^{\mathbf{N}} dz = w_\alpha - w_{\alpha-1/2} + \frac{h_\alpha}{2} \nabla_{x,y} \cdot \mathbf{u}_\alpha + (\mathbf{u}_{\alpha-1/2} - \mathbf{u}_\alpha) \cdot \nabla_{x,y} z_{\alpha-1/2}, \quad (3.107)$$

and the two relations (3.106),(3.107) are consistent with the definition (3.96) in the sense that the sum of Eqs. (3.106),(3.107) gives (3.99) whereas the substraction of Eqs. (3.106) and (3.107) gives (3.97). Finally, equation (3.98) is rewritten as

$$\begin{aligned} \rho_\alpha h_\alpha \frac{\partial z_\alpha}{\partial t} + z_\alpha \frac{\partial \rho_\alpha h_\alpha}{\partial t} + \rho_\alpha h_\alpha \mathbf{u}_\alpha \cdot \nabla_{x,y} z_\alpha + z_\alpha \nabla_{x,y} \cdot (\rho_\alpha h_\alpha \mathbf{u}_\alpha) = \rho_\alpha h_\alpha w_\alpha \\ + \left(z_\alpha \rho_{\alpha+1/2} + \rho_\alpha \frac{h_\alpha}{2} \right) G_{\alpha+1/2} - \left(z_\alpha \rho_{\alpha-1/2} - \rho_\alpha \frac{h_\alpha}{2} \right) G_{\alpha-1/2} \end{aligned}$$

and simplified into

$$\rho_\alpha h_\alpha \frac{\partial z_\alpha}{\partial t} + \rho_\alpha h_\alpha \mathbf{u}_\alpha \cdot \nabla_{x,y} z_\alpha = \rho_\alpha h_\alpha w_\alpha + \rho_\alpha h_\alpha \frac{G_{\alpha+1/2} + G_{\alpha-1/2}}{2}$$

using (3.99). Dividing by $\rho_\alpha h_\alpha$ gives equation (3.97). □

Proposition 7. *For equally distributed layers, the static equilibria of system* (3.87)-(3.89) *verify*

$$\begin{aligned} \nabla_{x,y}\tilde{\rho}_{\alpha+1/2} - \left.\frac{\partial \rho}{\partial z}\right|_{z_{\alpha+1/2}} \nabla_{x,y} z_{\alpha+1/2} = 0, \\ \nabla_{x,y}\eta = O\left(\frac{1}{N}\right), \end{aligned} \tag{3.108}$$

with $\tilde{\rho}_{\alpha+1/2} = \dfrac{\rho_{\alpha+1}+\rho_\alpha}{2}$. *The first relation in* (3.108) *can be re-interpreted as*

$$\nabla_{x,y}\tilde{\rho}_{z_{\alpha+1/2}} = 0.$$

Remark 13. For N large enough, the free surface is almost flat and the conditions (3.108) correspond to the static equilibria of the Euler system.
Moreover, for N large enough and $\|\nabla_{x,y} z_b\|$ small, all the interfaces between the layers are flat and we get $\nabla_{x,y}\tilde{\rho}_{\alpha+1/2} = 0$. Such a condition allows "checkerboard modes" for the density. However, as explained in the proof of proposition 8, these checkerboard modes are not stable equilibria.

Proof of prop. 7. Inserting $u_\alpha = 0$ and replacing all the time derivatives by 0 in system (3.87)-(3.89) gives

$$\nabla_{x,y}(h_\alpha p_\alpha) = p_{\alpha+1/2}\nabla_{x,y} z_{\alpha+1/2} - p_{\alpha-1/2}\nabla_{x,y} z_{\alpha-1/2}, \quad \alpha = 1,\dots,N,$$

which we simplify to get

$$\nabla_{x,y} p_\alpha = -g\rho_\alpha \nabla_{x,y} z_\alpha, \quad \alpha = 1,\dots,N. \tag{3.109}$$

For $\alpha = N$, equation (3.109) becomes

$$g\frac{h_N}{2}\nabla_{x,y}\rho_N = -g\rho_N \nabla_{x,y}\eta.$$

Note that the left-hand side of the previous relation depends on the number of layers N. Assuming that the layers have the same size, we get $g\frac{h}{2N}\nabla_{x,y}\rho_N = -g\rho_N\nabla_{x,y}\eta$, which means that $\nabla_{x,y}\eta = O(1/N)$. The difference of (3.109) written for $\alpha+1$ and for α gives

$$\nabla_{x,y}(p_{\alpha+1} - p_\alpha) = -g\rho_{\alpha+1}\nabla_{x,y} z_{\alpha+1} + g\rho_\alpha \nabla_{x,y} z_\alpha.$$

We use relation (3.90) to express p_α and $p_{\alpha+1}$ and we assume that the layers are equally distributed, so that we get

$$gh_{\alpha+1/2}\nabla_{x,y}\left(\frac{\rho_{\alpha+1}+\rho_\alpha}{2}\right) = g\rho_{\alpha+1}\nabla_{x,y} z_{\alpha+1/2} - g\rho_\alpha \nabla_{x,y} z_{\alpha+1/2}, \tag{3.110}$$

where $h_{\alpha+1/2} = h_\alpha = h_{\alpha+1}$. Finally, we divide equation (3.110) by $h_{\alpha+1/2}$ and we define $\frac{\partial \rho}{\partial z}\Big|_{z_{\alpha+1/2}} = \frac{\rho_{\alpha+1}-\rho_\alpha}{h_{\alpha+1/2}}$ to get the result. □

Proposition 8. *For $\nabla_{x,y} z_b$ small enough and for equally distributed layers, the stable equilibria of system (3.87)-(3.89) verify*

$$\partial_z \rho|_{z_\alpha} < 0, \quad \alpha = 1, \dots, N.$$

Proof of prop. 8. Let us define a perturbation around a static equilibrium

$$\begin{aligned} u_\alpha &= u'_\alpha, & w_\alpha &= w'_\alpha \\ \rho_\alpha &= R_\alpha + \rho'_\alpha, & \nabla_{x,y}\eta &= 0 \end{aligned}$$

The superscript $'$ denotes a first-order term. R_α is constant in space and time. The perturbation of the free surface is neglected. As a consequence, the space derivatives of h_α and z_α are zero for all α. The equations (3.87)-(3.89) are linearized around the static equilibrium. The Boussinesq approximation is performed for the sake of simplicity.

$$G'_{\alpha+1/2} = \sum_{j=1}^{\alpha} (\nabla_{x,y} \cdot (h_j u'_j)), \quad \alpha = 1, \dots, N, \tag{3.111}$$

$$\partial_t \rho'_\alpha = \frac{\rho_{\alpha+1} - \rho_\alpha}{h_\alpha} G'_{\alpha+1/2} - \frac{\rho_{\alpha-1} - \rho_\alpha}{h_\alpha} G'_{\alpha-1/2}, \quad \alpha = 1, \dots, N, \tag{3.112}$$

$$\rho_0 \partial_t u'_\alpha + \nabla_{x,y} p_\alpha = 0, \quad \alpha = 1, \dots, N. \tag{3.113}$$

In (3.112), the terms $\frac{\rho_{\alpha+1}-\rho_\alpha}{h_\alpha}$ and $\frac{\rho_{\alpha-1}-\rho_\alpha}{h_\alpha}$ are interpreted as $\frac{1}{2}\partial_z \rho|_{z_\alpha}$ and $-\frac{1}{2}\partial_z \rho|_{z_\alpha}$ respectively, so that we get

$$\partial_t \rho'_\alpha = K_\alpha \frac{G'_{\alpha+1/2} + G'_{\alpha-1/2}}{2}, \quad \alpha = 1, \dots, N, \tag{3.114}$$

where we have used the notation $K_\alpha = \partial_z \rho|_{z_\alpha}$. Next, $G_{\alpha+1/2}, G_{\alpha-1/2}$ are replaced by their expressions given by (3.112). In (3.113), the expression given by (3.90) for the pressure is substituted. We get

$$\partial_t \rho'_\alpha = K_\alpha \left(\sum_{j=1}^{\alpha} h_j \nabla_{x,y} \cdot u'_j - \frac{h_\alpha}{2} \nabla_{x,y} \cdot u'_\alpha \right), \quad \alpha = 1, \dots, N,$$

$$\rho_0 \partial_t u'_\alpha + g \left(\sum_{j=\alpha+1}^{N} h_j \nabla_{x,y} \rho'_j + \frac{h_\alpha}{2} \nabla_{x,y} \rho_\alpha \right) = 0, \quad \alpha = 1, \dots, N.$$

We describe the perturbations u'_α, ρ'_α as plane waves

$$u'_\alpha = \begin{pmatrix} u_{0,\alpha,x} \\ u_{0,\alpha,y} \end{pmatrix} e^{i(\Omega t - k_\alpha x - l_\alpha y)} \quad \rho'_\alpha = \rho_{0,\alpha} e^{i(\Omega t - k_\alpha x - l_\alpha y)}, \quad \alpha = 1, \dots, N,$$

with $\Omega, k_\alpha, l_\alpha$ positive real numbers. The linearized equations become

$$\omega \rho_{0,\alpha} + K_\alpha \frac{H_0}{N} \left(\sum_{j=1}^{\alpha} u_{0,j}(k_\alpha + l_\alpha) - \frac{u_{0,\alpha}}{2}(k_\alpha + l_\alpha) \right) = 0, \quad \alpha = 1, \dots, N,$$

$$\rho_0 \Omega u_{0,\alpha} - g \frac{H_0}{N} \left(\sum_{j=\alpha+1}^{N} \rho_{0,j}(k_\alpha + l_\alpha) + \frac{\rho_{0,\alpha}}{2}(k_\alpha + l_\alpha) \right) = 0, \quad \alpha = 1, \dots, N.$$

Let the two vectors $u_0 = (u_{0,1}, \dots, u_{0,N})^T$ and $\rho_0 = (\rho_{0,1}, \dots, \rho_{0,N})^T$. The previous system is rewritten as

$$\begin{aligned} \Omega I_N \rho_0 + K T_+^- u_0 = 0, \\ T_-^+ \rho_0 + \Omega I_N u_0 = 0, \end{aligned}$$

where I_N is the identity matrix of size N, T_-^+ is an upper triangular matrix with negative coefficients and KT_+^- is a lower triangular matrix, where the coefficients of line α have the sign of K_α. Then, necessarily, for the system to admit a solution, the coefficients $K_\alpha, \quad \alpha = 1, \dots, N$ must be negative. □

3.3.2 The layer-averaged Navier-Stokes-Fourier system

The hydrostatic approximation of the Navier-Stokes-Fourier system (3.25)-(3.27) obtained in paragraph 3.2.3 reads

$$\nabla \cdot \mathbf{U} = -\frac{\rho'(T)}{\rho^2 c_p} \left(\nabla \cdot (\lambda \nabla T) + \mu |\nabla_{x,y} \mathbf{u}|^2 + \mu \left| \frac{\partial \mathbf{u}}{\partial z} \right|^2 \right), \tag{3.115}$$

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{U}) = 0, \tag{3.116}$$

$$\frac{\partial (\rho \mathbf{u})}{\partial t} + \nabla_{x,y} \cdot (\rho \mathbf{u} \otimes \mathbf{u}) + \frac{\partial (\rho \mathbf{u} w)}{\partial z} + \nabla_{x,y} \int_z^\eta \rho g dz = \mu \Delta_{x,y} \mathbf{u} + \mu \frac{\partial}{\partial z} \left(\frac{\partial \mathbf{u}}{\partial z} \right), \tag{3.117}$$

with $T = T(\rho)$ a given function and where the notation $|\nabla_{x,y} \mathbf{u}|^2$ means $|\nabla_{x,y} \mathbf{u}|^2 = (\nabla_{x,y} \mathbf{u}) : (\nabla_{x,y} \mathbf{u})^T$. For the sake of simplicity, we have used the Stokes hypothesis, i.e. the second viscosity ζ is neglected, so that the viscosity terms in the momentum equation are written as $\mu \Delta \mathbf{u}$. The equation for the internal energy is

$$\frac{\partial}{\partial t}(\rho e) + \nabla \cdot (\rho \mathbf{U} e) = -p \nabla \cdot \mathbf{U} + \nabla \cdot (\lambda \nabla T) + \mu |\nabla_{x,y} \mathbf{u}|^2 + \mu \left| \frac{\partial \mathbf{u}}{\partial z} \right|^2,$$

Following the same strategy as in paragraph 3.3.1, we derive a layer-averaged version of the hydrostatic Navier-Stokes-Fourier system (3.115)-(3.117). A simplified formulation of the rheology terms is used, see [6].

Proposition 9. *Using the space $\mathbb{P}_{0,h}^{N,t}$ defined by (3.85) and the decomposition (3.86), the Galerkin approximation of the hydrostatic Navier-Stokes-Fourier system (3.25)-(3.27) completed with (3.10),(3.12),(3.11),(3.13) leads to the system*

$$\frac{\partial h}{\partial t}+\sum_{\alpha=1}^{N}\nabla_{x,y}\cdot(h_\alpha\mathbf{u}_\alpha)=-\sum_{\alpha=1}^{N}\frac{\rho'(T_\alpha)}{\rho_\alpha^2 c_p}(\mathcal{S}_{T,\alpha}-\mathcal{S}_{\mu,\alpha}), \tag{3.118}$$

$$\frac{\partial \rho_\alpha h_\alpha}{\partial t}+\nabla_{x,y}\cdot(\rho_\alpha h_\alpha\mathbf{u}_\alpha)=\rho_{\alpha+1/2}G_{\alpha+1/2}-\rho_{\alpha-1/2}G_{\alpha-1/2},\qquad \alpha=1,...,N, \tag{3.119}$$

$$\begin{aligned}\frac{\partial \rho_\alpha h_\alpha\mathbf{u}_\alpha}{\partial t}+\nabla_{x,y}\cdot(\rho_\alpha h_\alpha\mathbf{u}_\alpha\otimes\mathbf{u}_\alpha)+\nabla_{x,y}\big(h_\alpha p_\alpha\big)&=p_{\alpha+1/2}\nabla_{x,y}z_{\alpha+1/2}-p_{\alpha-1/2}\nabla_{x,y}z_{\alpha-1/2}\\&+\mathbf{u}_{\alpha+1/2}\rho_{\alpha+1/2}G_{\alpha+1/2}-\mathbf{u}_{\alpha-1/2}\rho_{\alpha-1/2}G_{\alpha-1/2}+\nabla_{x,y}\cdot(\mu h_\alpha\nabla_{x,y}\mathbf{u}_\alpha)\\&+\Gamma_{\alpha+1/2}(\mathbf{u}_{\alpha+1}-\mathbf{u}_\alpha)-\Gamma_{\alpha-1/2}(\mathbf{u}_\alpha-\mathbf{u}_{\alpha-1})-\kappa_\alpha\mathbf{u}_\alpha,\qquad \alpha=1,...,N,\end{aligned} \tag{3.120}$$

with

$$G_{\alpha+1/2}=-\sum_{j=1}^{N}\Big(\sum_{p=1}^{\alpha}l_p-\mathbb{1}_{j\leqslant\alpha}\Big)\nabla_{x,y}\cdot(h_j\mathbf{u}_j)+\sum_{j=1}^{\alpha}\frac{\rho'(T_j)}{\rho_j^2 c_p}(S_{T,j}-S_{\mu,j}), \tag{3.121}$$

$$\kappa_\alpha=\begin{cases}\kappa & \text{if}\quad \alpha=1\\ 0 & \text{if}\quad \alpha\neq 1\end{cases}, \tag{3.122}$$

$$\mathcal{S}_{T,\alpha}=\left(\lambda\nabla_{x,y}\cdot(h_\alpha\nabla_{x,y}T_\alpha)+2\lambda_{\alpha+1/2}\frac{T_{\alpha+1}-T_\alpha}{h_{\alpha+1}+h_\alpha}-2\lambda_{\alpha-1/2}\frac{T_\alpha-T_{\alpha-1}}{h_\alpha+h_{\alpha-1}}\right) \tag{3.123}$$

$$\lambda_{\alpha+1/2}=\lambda\quad \text{for}\quad \alpha=1,\dots,N-1$$

For $\alpha=0$, $2\lambda_{\alpha+1/2}\frac{T_{\alpha+1}-T_\alpha}{h_{\alpha+1}+h_\alpha}=FT_b^0$ if the Neumann boundary condition (3.71) is chosen, or $h_0=h_1$, $T_0=T_b^0$ if the Dirichlet boundary condition (3.72) is chosen. Likewise, for $\alpha=N$, $2\lambda_{\alpha+1/2}\frac{T_{\alpha+1}-T_\alpha}{h_{\alpha+1}+h_\alpha}=FT_s^0$ with the boundary condition (3.73), or $h_{N+1}=h_N$, $T_{N+1}=T_s^0$ with the boundary condition (3.74). The terms $\mathcal{S}_{\mu,\alpha}$, $\Gamma_{\alpha+1/2}$, $\mu_{\alpha+1/2}$ are respectively defined as

$$\mathcal{S}_{\mu,\alpha}=-h_\alpha\mu|\nabla_{x,y}\mathbf{u}_\alpha|^2-\Gamma_{\alpha+1/2}\frac{|\mathbf{u}_{\alpha+1}-\mathbf{u}_\alpha|^2}{2}-\Gamma_{\alpha-1/2}\frac{|\mathbf{u}_\alpha-\mathbf{u}_{\alpha-1}|^2}{2}-\kappa_\alpha|\mathbf{u}_\alpha|^2$$

$$\Gamma_{\alpha+1/2}=\frac{2\mu_{\alpha+1/2}}{h_{\alpha+1}+h_\alpha},$$

$$\mu_{\alpha+1/2}=\begin{cases}0 & \text{if}\quad \alpha=0\\ \mu & \text{if}\quad \alpha=1,\dots,N-1\\ 0 & \text{if}\quad \alpha=N.\end{cases}$$

The term $|\nabla_{x,y}\mathbf{u}_\alpha|^2$ *actually denotes*

$$\begin{aligned}|\nabla_{x,y}\mathbf{u}_\alpha|^2 &= (\nabla_{x,y}\mathbf{u}_\alpha) : (\nabla_{x,y}\mathbf{u}_\alpha)^T \\ &= \left(\frac{\partial u_\alpha}{\partial x}\right)^2 + \left(\frac{\partial u_\alpha}{\partial y}\right)^2 + \left(\frac{\partial v_\alpha}{\partial x}\right)^2 + \left(\frac{\partial v_\alpha}{\partial y}\right)^2.\end{aligned}$$

The temperature and viscosity terms have been simplified. In particular, the terms $\mu(\nabla_{x,y}u)|_{\alpha+1/2}\nabla_{x,y}z_{\alpha+1/2}, \mu(\nabla_{x,y}u)|_{\alpha-1/2}\nabla_{x,y}z_{\alpha-1/2}$ have been neglected, which is reasonable because the problems of interest are much vaster in the horizontal direction than in the vertical direction. Providing a detailed treatment of these terms is out of the scope of the present work. Notably, it would lead to very complicated terms in the fully discretized equations. For an exact integration of the viscosity terms, see [38], and for a simplified rheology, see [6]. The term $\mathcal{S}_{\mu,\alpha}$ is exactly the dissipative term that is obtained when the quantity $\mathbf{u}_\alpha.(\nabla_{x,y}\cdot(\mu h_\alpha\nabla_{x,y}\mathbf{u}_\alpha) + \Gamma_{\alpha+1/2}(\mathbf{u}_{\alpha+1}-\mathbf{u}_\alpha) - \Gamma_{\alpha-1/2}(\mathbf{u}_\alpha-\mathbf{u}_{\alpha-1}) - \kappa_\alpha\mathbf{u}_\alpha)$ is reformulated as a conservative term plus a dissipative term.

Proof of prop 9. The "Euler part" of the Navier-Stokes system is integrated as in the proof of proposition 4. Here, we deal only with the viscosity terms and the fact that $\nabla\cdot\mathbf{U}$ is no longer equal to 0. For the integration of the viscosity term in the momentum equation, we refer to [6]. The temperature diffusion term is integrated in the same manner. Integrating the equation on the divergence gives

$$\frac{\partial h_\alpha}{\partial t} + \nabla_{x,y}\cdot(h_\alpha\mathbf{u}_\alpha) = G_{\alpha+1/2} - G_{\alpha-1/2} - \frac{\rho'(T_\alpha)}{\rho_\alpha^2 c_p}(\mathcal{S}_{T,\alpha}-\mathcal{S}_{\mu,\alpha}). \tag{3.124}$$

The sum for $j=1,\dots,\alpha$ of the relations (3.124) with the boundary conditions (3.100) gives the expression (3.121) for $G_{\alpha+1/2}$.

□

Proposition 10. *The system* (3.118)-(3.120) *completed with the equation*

$$\begin{aligned}\frac{\partial}{\partial t}(\rho_\alpha h_\alpha e_\alpha) + \nabla_{x,y}\cdot(\rho_\alpha h_\alpha\mathbf{u}_\alpha e_\alpha) &= \rho_{\alpha+1/2}e_{\alpha+1/2}G_{\alpha+1/2} - \rho_{\alpha-1/2}e_{\alpha-1/2}G_{\alpha-1/2} \\ &\quad + p_\alpha\frac{\rho'(T_\alpha)}{\rho_\alpha^2 c_p}(\mathcal{S}_{T,\alpha}-\mathcal{S}_{\mu,\alpha}) + \mathcal{S}_{T,\alpha} - \mathcal{S}_{\mu,\alpha}\end{aligned} \tag{3.125}$$

admits, for smooth solutions, the energy balance

$$\begin{aligned}
&\frac{\partial}{\partial t}E_\alpha + \nabla_{x,y}\cdot\left(\mathbf{u}_\alpha\left(E_\alpha + h_\alpha p_\alpha - \mu h_\alpha \nabla_{x,y}\mathbf{u}_\alpha\right)\right)\\
&+\Gamma_{\alpha+1/2}\frac{|\mathbf{u}_{\alpha+\mathbf{1}}|^{\mathbf{2}} - |\mathbf{u}_\alpha|^{\mathbf{2}}}{2} - \Gamma_{\alpha-1/2}\frac{|\mathbf{u}_\alpha|^{\mathbf{2}} - |\mathbf{u}_{\alpha-\mathbf{1}}|^{\mathbf{2}}}{2}\\
&\quad= \left(\rho_{\alpha+1/2}\frac{|\mathbf{u}_{\alpha+1/2}|^2}{2} + g\rho_{\alpha+1/2}z_{\alpha+1/2}\right)G_{\alpha+1/2} + p_{\alpha+1/2}\big(G_{\alpha+1/2} - \frac{\partial z_{\alpha+1/2}}{\partial t}\big)\\
&\quad\quad- \left(\rho_{\alpha-1/2}\frac{|\mathbf{u}_{\alpha-1/2}|^2}{2} + g\rho_{\alpha-1/2}z_{\alpha-1/2}\right)G_{\alpha-1/2} - p_{\alpha-1/2}\big(G_{\alpha-1/2} - \frac{\partial z_{\alpha-1/2}}{\partial t}\big)\\
&\quad\quad-\frac{1}{2}\left(\rho_{\alpha+1/2}(\mathbf{u}_{\alpha+1/2} - \mathbf{u}_\alpha)^2 + gh_\alpha(\rho_{\alpha+1/2} - \rho_\alpha)\right)G_{\alpha+1/2}\\
&\quad\quad+\frac{1}{2}\left(\rho_{\alpha-1/2}(\mathbf{u}_{\alpha-1/2} - \mathbf{u}_\alpha)^2 - gh_\alpha(\rho_{\alpha-1/2} - \rho_\alpha)\right)G_{\alpha-1/2} + \mathcal{S}_{T,\alpha},
\end{aligned} \tag{3.126}$$

with

$$E_\alpha = \rho_\alpha\frac{h_\alpha|\mathbf{u}_\alpha|^2}{2} + \frac{\rho_\alpha g}{2}\left(z^2_{\alpha+1/2} - z^2_{\alpha-1/2}\right) + e_\alpha. \tag{3.127}$$

Note that in (3.126), we use the notation

$$\mathbf{u}_\alpha\nabla_{x,y}\mathbf{u}_\alpha = \begin{pmatrix} u_\alpha\frac{\partial u_\alpha}{\partial x} + v_\alpha\frac{\partial v_\alpha}{\partial x}\\ u_\alpha\frac{\partial u_\alpha}{\partial y} + v_\alpha\frac{\partial v_\alpha}{\partial y}\end{pmatrix}.$$

Remark 14. The energy e_α plays a role analogous to that of e in the continuous model in section 3.2. While in the Euler model it was enough to work with the kinetic and potential energies because $de/dt = 0$, here, an internal energy is needed in order to obtain an energy balance. The term $\mathcal{S}_{T,\alpha}$ is a heat flux, so its sign is unknown. The sum of $\mathcal{S}_{T,\alpha}$ over the layers gives

$$\sum_{\alpha=1}^{N}\mathcal{S}_{T,\alpha} = \lambda\sum_{\alpha=1}^{N}\nabla_{x,y}\cdot(h_\alpha\nabla_{x,y}T_\alpha) - \nabla T|_s.\mathbf{n}_s + \nabla T|_b.\mathbf{n}_b.$$

Proof of prop. 1. The proof is very similar to the proof of proposition 5. The kinetic energy contribution is the same as before, plus a contribution from the viscosity term. The quantity $I_{u,\alpha}$ described in (3.101) becomes

$$\begin{aligned}
I_{u,\alpha} = \frac{\partial}{\partial t}\left(\frac{\rho_\alpha h_\alpha u_\alpha^2}{2}\right) + \frac{\partial}{\partial x}\left(u_\alpha\left(\frac{\rho_\alpha h_\alpha u_\alpha^2}{2} - \mu h_\alpha\frac{\partial u_\alpha}{\partial x}\right)\right) + \frac{\partial}{\partial y}\left(v_\alpha\frac{\rho_\alpha h_\alpha u_\alpha^2}{2} - \mu h_\alpha u_\alpha\frac{\partial u_\alpha}{\partial y}\right)\\
- \rho_{\alpha+1/2}\frac{u^2_{\alpha+1/2}}{2}G_{\alpha+1/2} + \rho_{\alpha-1/2}\frac{u^2_{\alpha-1/2}}{2}G_{\alpha-1/2}\\
+ \rho_{\alpha+1/2}\frac{(u_{\alpha+1/2} - u_\alpha)^2}{2}G_{\alpha+1/2} - \rho_{\alpha-1/2}\frac{(u_{\alpha-1/2} - u_\alpha)^2}{2}G_{\alpha-1/2} + \mathcal{S}_{\mu,x,\alpha}
\end{aligned}$$

For the contribution of the pressure terms, the beginning of the proof is the same, but there is a difference when substituting $\nabla_{x,y}\cdot(h_\alpha \mathbf{u}_\alpha)$ in the sum of the pressure terms. Instead of eq. (3.102), we now obtain

$$I_{p,u,\alpha}+I_{p,v,\alpha}=p_\alpha\left(G_{\alpha+1/2}-G_{\alpha-1/2}-\frac{\partial h_\alpha}{\partial t}-\frac{\rho'(T_\alpha)}{\rho_\alpha^2 c_p}(\mathcal{S}_{T,\alpha}-\mathcal{S}_{\mu,\alpha})\right)-g\rho_\alpha h_\alpha \mathbf{u}_\alpha.\nabla_{x,y} z_\alpha,$$

so that instead of (3.103), we get

$$\begin{aligned}I_{p,u,\alpha}+I_{p,v,\alpha}&=p_{\alpha+1/2}\left(G_{\alpha+1/2}-\frac{\partial z_{\alpha+1/2}}{\partial t}\right)-p_{\alpha+1/2}\left(G_{\alpha-1/2}-\frac{\partial z_{\alpha-1/2}}{\partial t}\right)\\&+g\rho_\alpha\frac{h_\alpha}{2}\left(G_{\alpha+1/2}-G_{\alpha-1/2}\right)-g\rho_\alpha h_\alpha \mathbf{u}_\alpha.\nabla_{x,y} z_\alpha-g\rho_\alpha h_\alpha\frac{\partial z_\alpha}{\partial t}-p_\alpha\frac{\rho'(T_\alpha)}{\rho_\alpha^2 c_p}(\mathcal{S}_{T,\alpha}-\mathcal{S}_{\mu,\alpha}),\end{aligned}$$

The sum of (3.105) with (3.104) and (3.125) gives the final result. □

Remark 15. The layer-averaged Navier-Stokes system obtained in Prop. 9 has the form

$$\frac{\partial U}{\partial t}+\nabla_{x,y}\cdot F(U)=\mathcal{S}_p(U,z_b)+S_e(U,\partial_t U,\partial_x U)+S_{v,f}(U), \tag{3.128}$$

where the vector of unknowns is

$$U=(h,\rho_1 h_1,\dots,\rho_N h_N,q_{x,1},\dots,q_{x,N},q_{y,1},\dots,q_{y,N})^T,$$

with $q_{x,\alpha}=\rho_\alpha h_\alpha u_\alpha$, $q_{y,\alpha}=\rho_\alpha h_\alpha v_\alpha$. We denote by $F(U)=(F_x(U),F_y(U))^T$ the fluxes of the conservative part and by

$$\mathcal{S}_p(U,z_b)=\left(0,\dots,p_{3/2}\frac{\partial z_{3/2}}{\partial x}-p_{1/2}\frac{\partial z_{1/2}}{\partial x},\dots,p_{3/2}\frac{\partial z_{3/2}}{\partial y}-p_{1/2}\frac{\partial z_{1/2}}{\partial y},\dots\right)^T,$$

the non-conservative part of the pressure terms. The source terms are $S_e(U,\partial_t U,\partial_x U)$ and $S_{v,f}(U)$, representing respectively the mass and momentum exchanges and the viscous and friction effects. A numerical scheme for the simulation of the layer-averaged Navier-Stokes system is proposed in the companion paper [31]; it relies on the form (3.128).

3.4 Conclusion

In this paper we have derived an incompressible and hydrostatic model for variable density flows. This model is obtained by performing the incompressible limit of the compressible Navier-Stokes equations and incorporating a correction of order ε so as to obtain the correct energy balance. The resulting model does not rely on the Boussinesq approximation. It is mass-conservative; expansion or contraction can be observed as a result of the variation of a tracer concentration. A layer-averaged model is then proposed. The layer boundaries do not correspond to isopycnal surfaces and mass exchanges between

the layers are allowed. The equilibria of the layer-averaged Euler-Fourier model (in which the diffusion and viscosity effects are neglected) are found to be those of the classic Euler system. For smooth solutions, the layer-averaged model verifies an energy balance.

In [31], a numerical scheme is proposed and analyzed and the behaviour of the model is illustrated by means of several test cases.

Acknowledgments

The authors acknowledge the Inria Project Lab "Algae in Silico" for its financial support. This research is also supported by the ERC SLIDEQUAKES ERC-CG-2013-PE10-617472.

4

Numerical scheme and validation

Outline of the current chapter

The contents of this chapter will be submitted under the form of an article along with M.-O. Bristeau, F. Bouchut, A. Mangeney, J.Sainte-Marie and F. Souillé under the title "The Navier-Stokes system with temperature and salinity for free surface flows - Part II:

Numerical scheme and validation".

Abstract

In this paper, we propose a numerical scheme for the layer-averaged Euler-Fourier and Navier-Stokes-Fourier systems presented in part I [30]. These systems model free surface flows with density variations. We show that the finite volume scheme presented is well balanced with regards to the steady state of the lake at rest and preserves the positivity of the water height. A maximum principle on the density is also proved as well as a discrete entropy inequality in the case of the Euler-Fourier system. Some numerical validations are finally shown with comparisons to 3D analytical solutions and experiments.

Keywords: Navier-Stokes equations, free surface flows, variable density flows, layer-averaged formulation, finite volume scheme

4.1 Introduction

In this paper we present a numerical scheme for the 3D incompressible Navier-Stokes-Fourier system with free surface, as well as numerical test cases. This model describes variable density flows with free surface, the density variations coming from differences in temperature and/or salinity. The model is presented in the companion paper [30], in which a layer-averaged formulation is also given. The layer-averaged formulation suppresses the need for moving meshes [47], [49]. It allows to perform 3D simulations with a 2D fixed mesh.

Variable density flows are frequently studied by oceanographers. Different systems of coordinates exist, among which terrain-following coordinates and isopycnal coordinates. For a discussion of the advantages and disadvantages of the various coordinates frequently used in ocean models, the reader is referred to [71] and [129]. The layer-averaged model presented here is not a terrain-following coordinate model. Though the layer thicknesses are defined as fractions of the total water height, height coordinates are used. The model also differs from isopycnal coordinate models because in the layer-averaged formulation, the layers exchange mass between themselves, which means that the internal layer boundaries are actually not physical.

For the Euler part of the Navier-Stokes-Fourier system, a finite-volume formalism is adopted. The hydrostatic reconstruction technique is used [16]. Therefore, the topography is accurately represented and the scheme is well-balanced. Yet the discretization of the nonconservative pressure terms demands special care. For the viscosity terms, we use finite elements as in [6].

In [21], a similar model was studied and simulated. The scheme presented in [21] was a 2D $(x-z)$ scheme and relied on a kinetic interpretation. In the present work we present a fully 3D scheme which is more flexible in a certain sense, because the kinetic flux is only one of the possible choices for the numerical flux. With any flux consistent with the semi-discrete in time Euler system, the resulting scheme is well-balanced and preserves the nonnegativity of the water depth. A maximum principle on the density is satisfied. In order to prove an in-cell entropy inequality, we adopt a kinetic flux, already

used in the context of the Shallow Water equations in [109], [21]. The entropy inequality is satisfied for a constant topography and includes third-order rest terms. Moreover, the unknowns in [21] were not the same, which resulted in a complicated numerical scheme - nonlinear systems were solved at each step and a Newton fixed-point method was used. The present scheme is simpler and does not involve nonlinear systems. Finally, the present scheme is more stable than the scheme in [21], the CFL condition of which could actually degenerate and give a time step equal to zero. With the proposed scheme, the computational cost of the simulation of a non-Boussinesq flow is not greater than that of the simulation of a Boussinesq flow.

The proposed numerical scheme is validated on three test cases. The first test is a convergence test towards an analytical solution [39] for the Euler-Fourier system. In the second test, a lock exchange simulation is performed and the results are compared with experimental data available from the literature [3]. Finally, in two diffusion cases, the differences between the Navier-Stokes-Fourier and Boussinesq models are evidenced.

The paper is organized as follows. In section 4.2, the layer-averaged Navier-Stokes-Fourier and Euler-Fourier models introduced in [30] are recalled. A numerical scheme for the layer-averaged Euler-Fourier model is presented in section 4.3, its properties are studied. An extension of this scheme for the layer-averaged Navier-Stokes-Fourier model is presented in section 4.4. The numerical test cases are presented in section 4.5.

4.2 The layer-averaged models

We briefly recall here the features of the multilayer models studied here and presented in [30]. The multilayer Navier-Stokes-Fourier model is a layer-averaged version of the incompressible, hydrostatic Navier-Stokes-Fourier system

$$\nabla \cdot \mathbf{U} = -\frac{\rho'(T)}{\rho^2 c_p}(\nabla \cdot (\lambda \nabla T) + \sigma : D(U)), \tag{4.1}$$

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{U}) = 0, \tag{4.2}$$

$$\frac{\partial \rho \mathbf{u}}{\partial t} + \nabla \cdot (\rho \mathbf{u} \otimes \mathbf{u}) + \frac{\partial \rho \mathbf{u} w}{\partial z} + \nabla_{x,y} \int_z^\eta \rho g dz = \nabla_{x,y} \cdot \sigma + \frac{\partial}{\partial z}\left(\mu \frac{\partial \mathbf{u}}{\partial z}\right), \tag{4.3}$$

where $\mathbf{U}(t,x,y,z) = (u,v,w)^T$ is the velocity, $\mathbf{u} = (u,v)^T$ is the horizontal velocity vector and ρ is the density. The notation ∇ denotes $\nabla = \left(\frac{\partial}{\partial x}, (\frac{\partial}{\partial y}, (\frac{\partial}{\partial z}\right)^T$, $\nabla_{x,y}$ corresponds to the projection of ∇ on the horizontal plane i.e. $\nabla_{x,y} = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}\right)^T$. The quantity σ is the deviatoric part of the stress tensor. The fluid is assumed to be Newtonian, therefore σ is defined by

$$\sigma = -\zeta \nabla_{x,y} \cdot \mathbf{u} \mathbb{1} + 2\mu D_{x,y}(\mathbf{u}),$$

where μ is the viscosity coefficient and ζ is the second viscosity. The symmetric gradient of the velocity is $D_{x,y}(\mathbf{u}) = (\nabla_{x,y}\mathbf{u} + (\nabla_{x,y}\mathbf{u})^T)/2$. The temperature T is linked to the

density ρ via the equation of state $T = T(\rho)$. The heat conductivity is denoted by λ and the specific heat capacity at constant pressure by c_p. The energy balance for model (4.1)-(4.3) is

$$\frac{\partial}{\partial t}\left(\rho\frac{|\tilde{\mathbf{U}}|^2}{2}+\rho e\right)+\nabla\cdot\left(\mathbf{U}\left(\rho\frac{|\tilde{\mathbf{U}}|^2}{2}+\int_z^\eta \rho g dz+\rho e-\tilde{\sigma}\right)\right)=\nabla\cdot(\lambda\nabla T)+\rho\mathbf{g}.\mathbf{U},$$

with $\tilde{\mathbf{U}}=(u,v,0)^T$ and

$$\tilde{\sigma}=\begin{pmatrix}\sigma & & \partial_z u\\ & \sigma & \partial_z v\\ 0 & 0 & 0\end{pmatrix}.$$

We consider a free surface flow, therefore we assume

$$z_b(x,y)\leqslant z\leqslant \eta(t,x,y):=h(t,x,y)+z_b(x,y),$$

with $z_b(x,y)$ the bottom elevation and $h(t,x,y)$ the water depth, see figure 4.1.

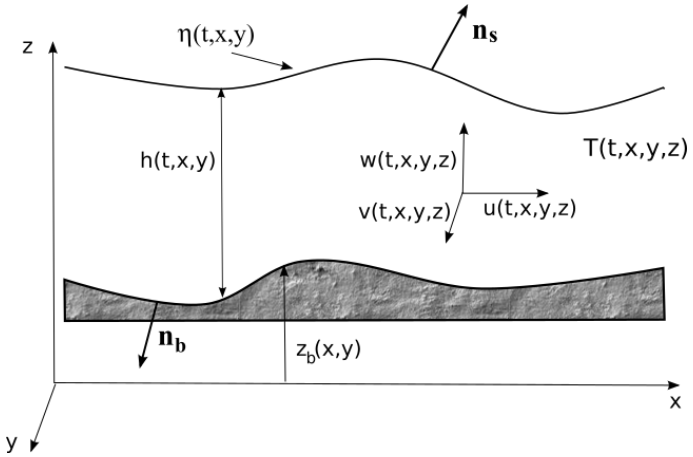


Figure 4.1 – Flow domain with water height $h(t,x,y)$, free surface $\eta(t,x,y)$ and bottom $z_b(x,y)$.

Let $\mathbf{n}_b$ and $\mathbf{n}_s$ be the unit outward normals at the bottom and at the free surface respectively defined by

$$\mathbf{n}_b=\frac{1}{\sqrt{1+|\nabla_{x,y}z_b|^2}}\begin{pmatrix}\nabla_{x,y}z_b\\ -1\end{pmatrix},\quad\text{and}\quad \mathbf{n}_s=\frac{1}{\sqrt{1+|\nabla_{x,y}\eta|^2}}\begin{pmatrix}-\nabla_{x,y}\eta\\ 1\end{pmatrix}.$$

On the bottom we prescribe an impermeability condition

$$\mathbf{U}.\mathbf{n}_b=0, \tag{4.4}$$

and a friction condition given e.g. by a Navier law

$$((-p\mathbb{1} + \sigma) \cdot \mathbf{n}_b) \cdot \mathbf{t}_i = -\kappa \mathbf{U} \cdot \mathbf{t}_i, \quad i = 1, 2 \tag{4.5}$$

with κ a Navier coefficient and $(\mathbf{t}_i, i = 1, 2)$ two tangential vectors. On the free surface, the kinematic boundary condition

$$\frac{\partial \eta}{\partial t} + \mathbf{u}(t, x, y, \eta) \cdot \nabla_{x,y} \eta - w(t, x, y, \eta) = 0, \tag{4.6}$$

is satisfied, along with the no stress condition

$$(-p\mathbb{1} + \sigma) \cdot \mathbf{n}_s = 0. \tag{4.7}$$

On solid walls, we prescribe a slip condition

$$\mathbf{U} \cdot \mathbf{n} = 0, \tag{4.8}$$

coupled with an homogeneous Neumann boundary condition

$$\frac{\partial \mathbf{u}}{\partial \mathbf{n}} = 0,$$

$\mathbf{n}$ being the outward normal to the considered wall. Boundary conditions for the temperature also have to be considered, we can choose either Neumann or Dirichlet conditions namely at the bottom

$$\lambda \nabla T \cdot \mathbf{n}_b = FT_b^0, \tag{4.9}$$

or

$$T_b = T_b^0 \tag{4.10}$$

and at the free surface

$$\lambda \nabla T \cdot \mathbf{n}_s = FT_s^0 \tag{4.11}$$

or

$$T_s = T_s^0 \tag{4.12}$$

where FT_b^0, FT_s^0 are two given temperature fluxes and T_b^0, T_s^0 are two given temperatures. Since $\rho = \rho(T)$, the boundary conditions for ρ naturally ensue from the boundary conditions for T.

The system is completed with some initial conditions

$$h(0, x, y) = h^0(x, y), \quad \rho(0, x, y) = \rho^0(x, y), \quad \mathbf{U}(0, x, y, z) = \mathbf{U}^0(x, y, z).$$

The system (4.1)-(4.3) was derived from the compressible Navier-Stokes-Fourier system in [30]. More specifically, the derivation consisted in performing the incompressible limit. This model respects the second principle of thermodynamics (non-decreasing entropy).

4.2.1 The multilayer Navier-Stokes-Fourier model

We consider a discretization of the fluid domain by layers, see Figure 4.2. In what follows,

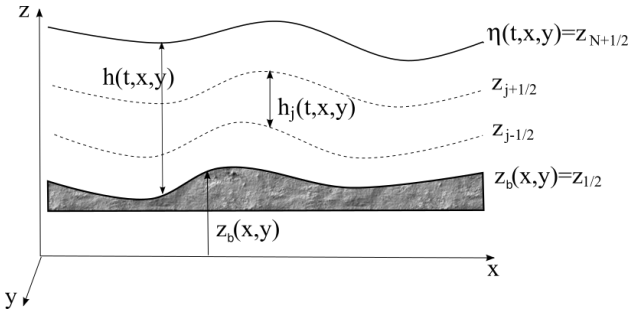


Figure 4.2 – Notations for the layerwise discretization.

h is the total water height and h_α is the thickness of the layer α. The layer α contains the points of coordinates (x, y, z) with $z \in L_\alpha(t, x, y) = (z_{\alpha-1/2}, z_{\alpha+1/2})$ and $\{z_{\alpha+1/2}\}_{\alpha=1,\dots,N}$ is defined by

$$\begin{cases} z_{\alpha+1/2}(t,x,y) = z_b(x,y) + \sum_{j=1}^{\alpha} h_j(t,x,y), \quad \alpha \in [0,\dots,N], \\ h_\alpha(t,x,y) = z_{\alpha+1/2}(t,x,y) - z_{\alpha-1/2}(t,x,y) = l_\alpha h(t,x,y), \end{cases}$$

and $\sum_{\alpha=1}^{N} l_\alpha = 1$.

The layer-averaged Navier-Stokes-Fourier system introduced in [30] reads

$$\frac{\partial h}{\partial t} + \sum_{\alpha=1}^{N} \nabla_{x,y} \cdot (h_\alpha \mathbf{u}_\alpha) = -\sum_{\alpha=1}^{N} \frac{\rho'(T_\alpha)}{\rho_\alpha^2 c_p} (\mathcal{S}_{T,\alpha} - \mathcal{S}_{\mu,\alpha}), \tag{4.13}$$

$$\frac{\partial \rho_\alpha h_\alpha}{\partial t} + \nabla_{x,y} \cdot (\rho_\alpha h_\alpha \mathbf{u}_\alpha) = \rho_{\alpha+1/2} G_{\alpha+1/2} - \rho_{\alpha-1/2} G_{\alpha-1/2}, \qquad \alpha = 1, ..., N, \tag{4.14}$$

$$\begin{aligned} \frac{\partial \rho_\alpha h_\alpha \mathbf{u}_\alpha}{\partial t} + \nabla_{x,y} \cdot (\rho_\alpha h_\alpha \mathbf{u}_\alpha \otimes \mathbf{u}_\alpha) + \nabla_{x,y}\big(h_\alpha p_\alpha\big) &= p_{\alpha+1/2} \nabla_{x,y} z_{\alpha+1/2} - p_{\alpha-1/2} \nabla_{x,y} z_{\alpha-1/2} \\ &+ \mathbf{u}_{\alpha+1/2} \rho_{\alpha+1/2} G_{\alpha+1/2} - \mathbf{u}_{\alpha-1/2} \rho_{\alpha-1/2} G_{\alpha-1/2} + \nabla_{x,y} \cdot (\mu h_\alpha \nabla_{x,y} \mathbf{u}_\alpha) \\ &+ \Gamma_{\alpha+1/2}(\mathbf{u}_{\alpha+1} - \mathbf{u}_\alpha) - \Gamma_{\alpha-1/2}(\mathbf{u}_\alpha - \mathbf{u}_{\alpha-1}) - \kappa_\alpha \mathbf{u}_\alpha, \qquad \alpha = 1, ..., N, \end{aligned} \tag{4.15}$$

with

$$G_{\alpha+1/2} = -\sum_{j=1}^{N} \Big(\sum_{p=1}^{\alpha} l_p - \mathbb{1}_{j \leqslant \alpha} \Big) \nabla_{x,y} \cdot (h_j \mathbf{u}_j) + \sum_{j=1}^{\alpha} \frac{\rho'(T_j)}{\rho_j^2 c_p} (S_{T,j} - S_{\mu,j}), \tag{4.16}$$

$$\kappa_\alpha = \begin{cases} \kappa & \text{if} \quad \alpha = 1 \\ 0 & \text{if} \quad \alpha \neq 1 \end{cases},$$

$$\mathcal{S}_{T,\alpha} = \left(\lambda \nabla_{x,y} \cdot (h_\alpha \nabla_{x,y} T_\alpha) + 2\lambda_{\alpha+1/2} \frac{T_{\alpha+1} - T_\alpha}{h_{\alpha+1} + h_\alpha} - 2\lambda_{\alpha-1/2} \frac{T_\alpha - T_{\alpha-1}}{h_\alpha + h_{\alpha-1}} \right) \tag{4.17}$$

$$\lambda_{\alpha+1/2} = \lambda \quad \text{for} \quad \alpha = 1, \dots, N-1,$$
$$T_\alpha = T(\rho_\alpha).$$

For $\alpha = 0$, $2\lambda_{\alpha+1/2}\frac{T_{\alpha+1}-T_\alpha}{h_{\alpha+1}+h_\alpha} = FT_b^0$ if the Neumann boundary condition (4.9) is chosen, or $h_0 = h_1$, $T_0 = T_b^0$ if the Dirichlet boundary condition (4.10) is chosen. Likewise, for $\alpha = N$, $2\lambda_{\alpha+1/2}\frac{T_{\alpha+1}-T_\alpha}{h_{\alpha+1}+h_\alpha} = FT_s^0$ with the boundary condition (4.11), or $h_{N+1} = h_N$, $T_{N+1} = T_s^0$ with the boundary condition (4.12). The dissipation term due to the viscous effects is

$$\mathcal{S}_{\mu,\alpha} = -h_\alpha \mu |\nabla_{x,y}\mathbf{u}_\alpha|^2 - \Gamma_{\alpha+1/2}\frac{|\mathbf{u}_{\alpha+1}-\mathbf{u}_\alpha|^2}{2} - \Gamma_{\alpha-1/2}\frac{|\mathbf{u}_\alpha-\mathbf{u}_{\alpha-1}|^2}{2} - \kappa_\alpha|\mathbf{u}_\alpha|^2 \quad (4.18)$$

$$\Gamma_{\alpha+1/2} = \frac{2\mu_{\alpha+1/2}}{h_{\alpha+1}+h_\alpha}, \quad (4.19)$$

$$\mu_{\alpha+1/2} = \begin{cases} 0 & \text{if} \quad \alpha = 0 \\ \mu & \text{if} \quad \alpha = 1, \dots, N-1 \\ 0 & \text{if} \quad \alpha = N. \end{cases} \quad (4.20)$$

The velocities $\mathbf{u}_{\alpha+1/2}$ and the densities $\rho_{\alpha+1/2}$ at the interfaces are defined by

$$v_{\alpha+1/2} = \begin{cases} v_\alpha & \text{if} \quad G_{\alpha+1/2} \leqslant 0 \\ v_{\alpha+1} & \text{if} \quad G_{\alpha+1/2} > 0 \end{cases} \quad (4.21)$$

for $v = \mathbf{u}, \rho$.
The presure terms p_α, $p_{\alpha+1/2}$ are given by

$$p_\alpha = g\left(\frac{\rho_\alpha h_\alpha}{2} + \sum_{j=\alpha+1}^{N}\rho_j h_j\right) \quad \text{and} \quad p_{\alpha+1/2} = g\sum_{j=\alpha+1}^{N}\rho_j h_j. \quad (4.22)$$

The pressure is hydrostatic. The terms $G_{\alpha+1/2}$ represent the mass exchanges between the layers. For the sake of simplicity, we have used the Stokes hypothesis, i.e. the second viscosity ζ has been neglected. Since the right-hand side of the total height conservation equation (4.13) is nonzero, we expect to observe dilatation and contraction due to the temperature diffusion and to the viscosity. We recall here the following result, obtained in [30].

Proposition 1. *The system* (4.13)-(4.15) *completed with the equation*

$$\frac{\partial}{\partial t}(\rho_\alpha h_\alpha e_\alpha) + \nabla_{x,y}\cdot(\rho_\alpha h_\alpha \mathbf{u}_\alpha e_\alpha) = \rho_{\alpha+1/2}e_{\alpha+1/2}G_{\alpha+1/2} - \rho_{\alpha-1/2}e_{\alpha-1/2}G_{\alpha-1/2} + p_\alpha\frac{\rho'(T_\alpha)}{\rho_\alpha^2 c_p}(\mathcal{S}_{T,\alpha} - \mathcal{S}_{\mu,\alpha}) + \mathcal{S}_{T,\alpha} - \mathcal{S}_{\mu,\alpha}$$

admits, for smooth solutions, the energy balance

$$
\begin{aligned}
&\frac{\partial}{\partial t}E_\alpha + \nabla_{x,y}\cdot\left(\mathbf{u}_\alpha\left(E_\alpha + h_\alpha p_\alpha - \mu h_\alpha \nabla_{x,y}\mathbf{u}_\alpha\right)\right)\\
&+\Gamma_{\alpha+1/2}\frac{|\mathbf{u}_{\alpha+1}|^2 - |\mathbf{u}_\alpha|^2}{2} - \Gamma_{\alpha-1/2}\frac{|\mathbf{u}_\alpha|^2 - |\mathbf{u}_{\alpha-1}|^2}{2}\\
&\qquad = \left(\rho_{\alpha+1/2}\frac{\mathbf{u}^2_{\alpha+1/2}}{2} + g\rho_{\alpha+1/2}z_{\alpha+1/2}\right)G_{\alpha+1/2} + p_{\alpha+1/2}\big(G_{\alpha+1/2} - \frac{\partial z_{\alpha+1/2}}{\partial t}\big)\\
&\qquad\quad - \left(\rho_{\alpha-1/2}\frac{\mathbf{u}^2_{\alpha-1/2}}{2} + g\rho_{\alpha-1/2}z_{\alpha-1/2}\right)G_{\alpha-1/2} - p_{\alpha-1/2}\big(G_{\alpha-1/2} - \frac{\partial z_{\alpha-1/2}}{\partial t}\big)\\
&\qquad\quad -\frac{1}{2}\left(\rho_{\alpha+1/2}(\mathbf{u}_{\alpha+1/2} - \mathbf{u}_\alpha)^2 + gh_\alpha(\rho_{\alpha+1/2} - \rho_\alpha)\right)G_{\alpha+1/2}\\
&\qquad\quad +\frac{1}{2}\left(\rho_{\alpha-1/2}(\mathbf{u}_{\alpha-1/2} - \mathbf{u}_\alpha)^2 - gh_\alpha(\rho_{\alpha-1/2} - \rho_\alpha)\right)G_{\alpha-1/2} + \mathcal{S}_{T,\alpha},
\end{aligned}
\tag{4.23}
$$

with

$$
E_\alpha = \rho_\alpha\frac{h_\alpha|\mathbf{u}_\alpha|^2}{2} + \frac{\rho_\alpha g h_\alpha z_\alpha}{2} + e_\alpha. \tag{4.24}
$$

Note that in (4.23), we use the notation

$$
\mathbf{u}_\alpha\nabla_{x,y}\mathbf{u}_\alpha = \begin{pmatrix} u\frac{\partial u}{\partial x} + v\frac{\partial v}{\partial x} \\ u\frac{\partial u}{\partial y} + v\frac{\partial v}{\partial y}\end{pmatrix}.
$$

The sum of Eqs. (4.23) for $\alpha = 1, \ldots, N$ gives

$$
\begin{aligned}
&\frac{\partial}{\partial t}\sum_{\alpha=1}^{N}E_\alpha + \sum_{\alpha=1}^{N}\nabla_{x,y}\cdot\mathbf{u}_\alpha\left(E_\alpha + h_\alpha p_\alpha\right)\\
&\qquad = \quad -\sum_{\alpha=1}^{N}\rho_{\alpha+1/2}\frac{|\mathbf{u}_{\alpha+1} - \mathbf{u}_\alpha|^2}{2}|G_{\alpha+1/2}| + \sum_{\alpha=1}^{N}\mathcal{S}_{T,\alpha}\\
&\qquad\quad -\frac{g}{2}\sum_{\alpha=1}^{N}\left(h_\alpha(\rho_{\alpha+1/2} - \rho_\alpha) + h_{\alpha+1}(\rho_{\alpha+1/2} - \rho_{\alpha+1})\right)G_{\alpha+1/2}.
\end{aligned}
$$

The sum of $\mathcal{S}_{T,\alpha}$ over the layers gives

$$
\sum_{\alpha=1}^{N}\mathcal{S}_{T,\alpha} = \lambda\sum_{\alpha=1}^{N}\nabla_{x,y}\cdot(h_\alpha\nabla_{x,y}T_\alpha) - \nabla T|_s\cdot\mathbf{n}_s + \nabla T|_b\cdot\mathbf{n}_b.
$$

As explained in [30], the terms on the last line of the right-hand side are third-order terms.

4.2.2 The layer-averaged Euler-Fourier system

What we refer to hereafter as the layer-averaged Euler-Fourier system is the system (4.13)-(4.15) without viscosity and without diffusion terms, i.e.

$$\frac{\partial h}{\partial t} + \sum_{\alpha=1}^{N} \nabla_{x,y} \cdot (h_\alpha \mathbf{u}_\alpha) = 0, \tag{4.25}$$

$$\frac{\partial \rho_\alpha h_\alpha}{\partial t} + \nabla_{x,y} \cdot (\rho_\alpha h_\alpha \mathbf{u}_\alpha) = \rho_{\alpha+1/2} G_{\alpha+1/2} - \rho_{\alpha-1/2} G_{\alpha-1/2}, \qquad \alpha = 1, ..., N, \tag{4.26}$$

$$\begin{aligned}\frac{\partial \rho_\alpha h_\alpha \mathbf{u}_\alpha}{\partial t} + \nabla_{x,y} \cdot (\rho_\alpha h_\alpha \mathbf{u}_\alpha \otimes \mathbf{u}_\alpha) + \nabla_{x,y}\big(h_\alpha p_\alpha\big) &= p_{\alpha+1/2} \nabla_{x,y} z_{\alpha+1/2} - p_{\alpha-1/2} \nabla_{x,y} z_{\alpha-1/2} \\ &+ \mathbf{u}_{\alpha+1/2} \rho_{\alpha+1/2} G_{\alpha+1/2} - \mathbf{u}_{\alpha-1/2} \rho_{\alpha-1/2} G_{\alpha-1/2}, \qquad \alpha = 1, ..., N.\end{aligned} \tag{4.27}$$

The quantity $G_{\alpha+1/2}$ (resp. $G_{\alpha-1/2}$) corresponds to mass exchange accross the interface $z_{\alpha+1/2}$ (resp. $z_{\alpha-1/2}$) and $G_{\alpha+1/2}$ is defined by

$$G_{\alpha+1/2} = \sum_{j=1}^{\alpha} \left(\frac{\partial h_j}{\partial t} + \nabla_{x,y} \cdot (h_j \mathbf{u}_j) \right) = - \sum_{j=1}^{N} \Big(\sum_{p=1}^{\alpha} l_p - \mathbb{1}_{j \leqslant \alpha} \Big) \nabla_{x,y} \cdot (h_j \mathbf{u}_j), \tag{4.28}$$

for $\alpha = 1, \ldots, N$. The energy balance verified by the system (4.25)-(4.27) is given in [30]. It is very similar to the balance (4.23), obviously without the viscosity and diffusion terms. In the balance for the Euler-Fourier system, the internal energy e_α does not intervene. It is actually equal to 0 because there is no volume variation.

4.3 Numerical scheme for the layer-averaged Euler-Fourier system

In this section, a numerical scheme for the layer-averaged Euler-Fourier system is designed and analyzed. It extends the work done by some of the authors in [6, 21]. Before specifying the scheme for the Euler-Fourier system, a common strategy for the time discretization of the Navier-Stokes-Fourier and Euler-Fourier systems is presented. The discretization of the diffusion terms does not present any additional difficulty, however including these terms considerably lengthens the equations. This is why it seems preferable to explain the numerical scheme for the Euler-Fourier system first. The advantages of the numerical scheme are the following

- it gives a 3D approximation of the Navier-Stokes-Fourier system, while only 2D situations were considered in [21]
- it can be implemented with any flux that is consistent with the homogeneous Saint-Venant system; the kinetic flux is used only for the discrete entropy property stated for a constant topography

- the scheme is endowed with strong stability properties (well-balanced, positivity of the water depth)

- convergence curves towards ad 3D non-stationary analytical solution with wet-dry interfaces were obtained, see paragraph 4.5.1.

4.3.1 Strategy for the time discretization

The system (4.13)-(4.15) has the form

$$\frac{\partial \mathbf{U}}{\partial t} + \nabla_{x,y} \cdot F(\mathbf{U}) = \mathcal{S}_p(\mathbf{U}, z_b) + S_e(\mathbf{U}, \partial_t \mathbf{U}, \partial_x \mathbf{U}) + S_{v,f}(\mathbf{U}), \tag{4.29}$$

where the vector of unknowns is

$$\mathbf{U} = (h, \rho_1 h_1, \dots, \rho_N h_N, q_{x,1}, \dots, q_{x,N}, q_{y,1}, \dots, q_{y,N})^T,$$

with $q_{x,\alpha} = \rho_\alpha h_\alpha u_\alpha$, $q_{y,\alpha} = \rho_\alpha h_\alpha v_\alpha$. We denote by $F(\mathbf{U}) = (F_x(\mathbf{U}), F_y(\mathbf{U}))^T$ the fluxes of the conservative part and by

$$\mathcal{S}_p(\mathbf{U}, z_b) = \left(0, \dots, p_{3/2}\frac{\partial z_{3/2}}{\partial x} - p_{1/2}\frac{\partial z_{1/2}}{\partial x}, \dots, p_{3/2}\frac{\partial z_{3/2}}{\partial y} - p_{1/2}\frac{\partial z_{1/2}}{\partial y}, \dots\right)^T,$$

the non-conservative part of the pressure terms. The source terms are $S_e(\mathbf{U}, \partial_t \mathbf{U}, \partial_x \mathbf{U})$ and $S_{v,f}(\mathbf{U})$, representing respectively the mass and momentum exchanges and the viscous and friction effects. Notice that, as a consequence of the layer-averaged discretization, the system (4.29) is made of only 2d (x, y) partial differential equations with source terms. Hence, the spacial approximation of the considered PDEs is performed on a 2d planar mesh. We consider discrete times t^n with $t^{n+1} = t^n + \Delta t^n$. For the time discretisation of the layer-averaged Navier-Stokes system (4.29) we adopt the following scheme

$$\mathbf{U}^{n+1} = \mathbf{U} - \Delta t^n \left(\nabla_{x,y} \cdot F(\mathbf{U}) - \mathcal{S}_p(\mathbf{U}, z_b)\right) + \Delta t^n S_e^{n+1} + \Delta t^n S_{v,f}^{n+l}, \tag{4.30}$$

where the integer $l = 0, 1/2, 1$ will be precised below. In (4.30) and wherever there is no ambiguity the superscript ^{n} has been omitted.

4.3.2 Semi-discrete (in time) scheme

From now on and until the end of this section, the system considered is the Euler-Fourier system. Similarly to [6], the semi-discrete in time scheme (4.30) with $S_{v,f}^{n+l} = 0$ reads

$$
\begin{aligned}
h_\alpha^{n+1/2} &= h_\alpha - \Delta t^n \nabla_{x,y}(h_\alpha \mathbf{u}_\alpha) && (4.31)\\
(\rho_\alpha h_\alpha)^{n+1/2} &= \rho_\alpha h_\alpha - \Delta t^n \nabla_{x,y}\cdot(\rho_\alpha h_\alpha \mathbf{u}_\alpha) && (4.32)\\
(\rho_\alpha h_\alpha \mathbf{u}_\alpha)^{n+1/2} &= \rho_\alpha h_\alpha \mathbf{u}_\alpha - \Delta t^n \Big(\nabla_{x,y}\cdot(\rho_\alpha h_\alpha \mathbf{u}_\alpha \otimes \mathbf{u}_\alpha) + \nabla_{x,y}\big(h_\alpha p_\alpha\big) \\
&\quad - p_\alpha \nabla_{x,y} h_\alpha + \rho_\alpha g h_\alpha \nabla_{x,y} z_\alpha \Big), && (4.33)\\
h^{n+1} &= h^{n+1/2} = \sum_{\alpha=1}^{N} h_\alpha^{n+1/2} = h - \Delta t^n \sum_{\alpha=1}^{N} \nabla_{x,y}\cdot(h_\alpha \mathbf{u}_\alpha), && (4.34)\\
(\rho_\alpha h_\alpha)^{n+1} &= (\rho_\alpha h_\alpha)^{n+1/2} + \Delta t^n \Big(\rho_{\alpha+1/2}^{n+1} G_{\alpha+1/2} - \rho_{\alpha-1/2}^{n+1} G_{\alpha-1/2} \Big), && (4.35)\\
(\rho_\alpha h_\alpha \mathbf{u}_\alpha)^{n+1} &= (\rho_\alpha h_\alpha \mathbf{u}_\alpha)^{n+1/2} + \Delta t^n \Big(\mathbf{u}_{\alpha+1/2}^{n+1} \rho_{\alpha+1/2}^{n+1} G_{\alpha+1/2} - \mathbf{u}_{\alpha-1/2}^{n+1} \rho_{\alpha-1/2}^{n+1} G_{\alpha-1/2} \Big), && (4.36)\\
& && (4.37)
\end{aligned}
$$

with

$$
G_{\alpha+1/2} = -\sum_{j=1}^{N} \Big(\sum_{p=1}^{\alpha} l_p - \mathbb{1}_{j\leqslant\alpha} \Big) \nabla_{x,y}\cdot(h_j \mathbf{u}_j). \tag{4.38}
$$

Since the relations

$$
p_{\alpha\pm1/2} = p_\alpha \mp \frac{\rho_\alpha g h_\alpha}{2}, \quad z_\alpha = \frac{z_{\alpha+1/2} + z_{\alpha-1/2}}{2}
$$

hold, in (4.33) we often use the identity

$$
p_{\alpha+1/2}\nabla_{x,y} z_{\alpha+1/2} - p_{\alpha-1/2}\nabla_{x,y} z_{\alpha-1/2} = p_\alpha \nabla_{x,y} h_\alpha - \rho_\alpha g h_\alpha \nabla_{x,y} z_\alpha. \tag{4.39}
$$

Following (4.30), Eqs. (4.31)-(4.34) also reads

$$
\mathbf{U}^{n+1/2} = \mathbf{U} - \Delta t^n \left(\nabla_{x,y}\cdot F(\mathbf{U}) - \mathcal{S}_p(\mathbf{U}, z_b) \right), \tag{4.40}
$$

and Eqs. (4.35)-(4.36) can be reformulated under the form

$$
\mathbf{U}^{n+1} = \mathbf{U}^{n+1/2} + \Delta t^n S_e^{n+1/2}. \tag{4.41}
$$

4.3.3 Finite volume formalism for the Euler part

In this paragraph, we propose a space discretization for the model (4.34)-(4.36) completed with (4.38). We first recall the general formalism of finite volumes on unstructured meshes. Let Ω denote the computational domain with boundary Γ, which we assume is polygonal. Let T_h be a triangulation of Ω for which the vertices are denoted by P_i with S_i the set of interior nodes and G_i the set of boundary nodes. The dual cells C_i are obtained by joining the centers of mass of the triangles surrounding each vertex P_i. We

use the following notations (see Fig. 4.3):

- K_i, set of subscripts of nodes P_j surrounding P_i,
- $|C_i|$, area of C_i,
- Γ_{ij}, boundary edge between the cells C_i and C_j,
- L_{ij}, length of Γ_{ij},
- $\mathbf{n}_{ij}$, unit normal to Γ_{ij}, outward to C_i ($\mathbf{n}_{ji} = -\mathbf{n}_{ij}$).

If P_i is a node belonging to the boundary Γ, we join the centers of mass of the triangles adjacent to the boundary to the middle of the edge belonging to Γ (see Fig. 4.3) and we denote

- Γ_i, the two edges of C_i belonging to Γ,
- L_i, length of Γ_i (for sake of simplicity we assume in the following that $L_i = 0$ if P_i does not belong to Γ),
- $\mathbf{n}_i$, the unit outward normal defined by averaging the two adjacent normals.

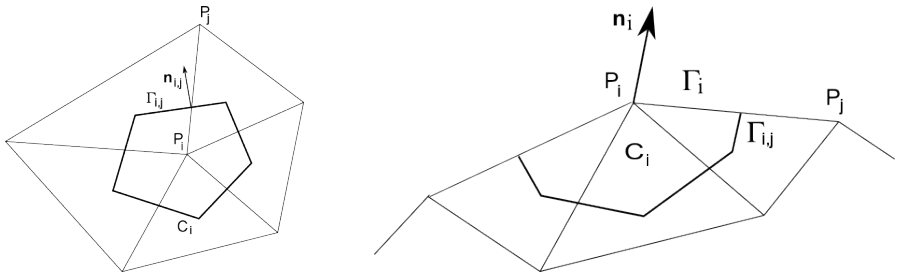


Figure 4.3 – *(a)* Dual cell C_i and *(b)* Boundary cell C_i.

We define the piecewise constant functions $\mathbf{U}^n(x, y)$ on cells C_i corresponding to time t^n and $z_b(x, y)$ as

$$\mathbf{U}^n(x, y) = \mathbf{U}_i^n, \quad z_b(x, y) = z_i, \quad \text{for } (x, y) \in C_i,$$

with $\mathbf{U}_i^n = (h_i^n, \rho_{1,i}^n h_{1,i}^n, \ldots, \rho_{N,i}^n h_{N,i}^n, q_{x,1,i}^n, \ldots, q_{x,N,i}^n, q_{y,1,i}^n, \ldots, q_{y,N,i}^n)^T$ i.e.

$$\mathbf{U}_i^n \approx \frac{1}{|C_i|} \int_{C_i} \mathbf{U}(t^n, x, y) dxdy, \quad z_i \approx \frac{1}{|C_i|} \int_{C_i} z_b(x, y) dxdy.$$

We will also use the notation

$$\mathbf{U}_{\alpha,i}^n \approx \frac{1}{|C_i|} \int_{C_i} \mathbf{U}_\alpha(t^n, x, y) dxdy,$$

with $\mathbf{U}_\alpha$ defined by

$$\mathbf{U}_\alpha = (h_\alpha, \rho_\alpha h_\alpha, \rho_\alpha h_\alpha u_\alpha, \rho_\alpha h_\alpha v_\alpha)^T. \tag{4.42}$$

Eigenvalues for the Euler part

Without the exchange terms, the system (4.25)-(4.27) reads

$$\frac{\partial h_\alpha}{\partial t} + \nabla_{x,y} \cdot (h_\alpha \mathbf{u}_\alpha) = 0 = 0, \tag{4.43}$$

$$\frac{\partial \rho_\alpha h_\alpha}{\partial t} + \nabla_{x,y} \cdot (\rho_\alpha h_\alpha \mathbf{u}_\alpha) = 0, \tag{4.44}$$

$$\frac{\partial \rho_\alpha h_\alpha \mathbf{u}_\alpha}{\partial t} + \nabla_{x,y} \cdot (\rho_\alpha h_\alpha \mathbf{u}_\alpha \otimes \mathbf{u}_\alpha) + p_\alpha \nabla_{x,y} h_\alpha - \rho_\alpha g h_\alpha \nabla_{x,y} z_\alpha = 0, \tag{4.45}$$

with

$$p_\alpha \nabla_{x,y} h_\alpha - \rho_\alpha g h_\alpha \nabla_{x,y} z_\alpha = g h_\alpha \left(\sum_{j=\alpha+1}^{N} \nabla_{x,y}(\rho_j h_j) + \frac{1}{2} \nabla_{x,y}(\rho_\alpha h_\alpha) \right) + \frac{\rho_\alpha g h_\alpha}{2} \nabla_{x,y} h_\alpha + \rho_\alpha g h_\alpha \left(\sum_{j=1}^{\alpha-1} \nabla_{x,y} h_j + \nabla_{x,y} z_b \right).$$

This system is the continuous version of system (4.31)-(4.33). We rewrite the system (4.43)-(4.45) under the form

$$\frac{\partial h}{\partial t} + \sum_{j=1}^{N} (l_j h \nabla_{x,y} \cdot \mathbf{u}_j) + \sum_{j=1}^{N} (l_j \mathbf{u}_j) \cdot \nabla_{x,y} h = 0, \tag{4.46}$$

$$\frac{\partial \rho_\alpha}{\partial t} + \mathbf{u}_\alpha . \nabla_{x,y} \rho_\alpha = 0, \tag{4.47}$$

$$\frac{\partial \mathbf{u}_\alpha}{\partial t} + (\mathbf{u}_\alpha \cdot \nabla_{x,y}) \mathbf{u}_\alpha + \frac{1}{\rho_\alpha h_\alpha} (p_\alpha \nabla_{x,y} h_\alpha - \rho_\alpha g h_\alpha \nabla_{x,y} z_\alpha) = 0, \tag{4.48}$$

and the quasilinear form of the system (4.46)-(4.48) writes

$$\frac{\partial \tilde{\mathbf{U}}}{\partial t} + A(\tilde{\mathbf{U}}) \nabla_{x,y} \tilde{\mathbf{U}} = s_b(\tilde{\mathbf{U}}), \tag{4.49}$$

with

$$\tilde{\mathbf{U}} = (h, \mathbf{u}_1, \dots, \mathbf{u}_N, \rho_1, \dots, \rho_N)^T,$$

and

$$A(\tilde{\mathbf{U}}) = \begin{pmatrix} A_1(\tilde{\mathbf{U}}) & A_2(\tilde{\mathbf{U}}) \\ A_3(\tilde{\mathbf{U}}) & A_4(\tilde{\mathbf{U}}) \end{pmatrix},$$

$$A_1(\tilde{\mathbf{U}}) = \begin{pmatrix} \sum_{j=1}^{N} l_j u_j & l_1 h & \dots & \dots & \dots & l_N h \\ \tilde{p}_1 & u_1 & 0 & \dots & \dots & 0 \\ \tilde{p}_2 & 0 & u_2 & 0 & \dots & 0 \\ \vdots & 0 & \ddots & u_j & \ddots & 0 \\ \vdots & 0 & \ddots & 0 & \ddots & 0 \\ \tilde{p}_N & 0 & 0 & \dots & 0 & u_N \end{pmatrix},$$

$$A_2(\tilde{\mathbf{U}}) = \begin{pmatrix} 0 & \dots & \dots & 0 \\ \frac{gh_1^2}{2} & 0 & \ddots & 0 \\ gh_2h_1 & \frac{gh_2^2}{2} & 0 & 0 \\ \vdots & 0 & \frac{gh_j^2}{2} & 0 \\ gh_Nh_1 & 0 & 0 & \frac{gh_N^2}{2} \end{pmatrix},$$

$$A_3(\tilde{\mathbf{U}}) = \mathbf{0}_N, \qquad A_4(\tilde{\mathbf{U}}) = \mathrm{diag}(u_j),$$

$$\tilde{p}_j = \frac{g}{\rho_j}\left(\rho_j l_j + \rho_j \sum_{i=1}^{j-1} l_i + \sum_{i=1}^{j-1} \rho_i l_i\right).$$

For the sake of simplicity, the expression of the matrix $A(\tilde{\mathbf{U}})$ given below corresponds to the 1D case i.e. for $v_i = 0$, $i = 1, \dots, N$. Notice that $A_2(\tilde{\mathbf{U}})$ and $A_3(\tilde{\mathbf{U}})^T$ are rectangular matrices with $N+1$ rows and N columns.

The following proposition holds, consisting in a version of the Cauchy's interlace theorem [79] in the case of non symmetric matrix.

Proposition 2. *The system (4.49) is strictly hyperbolic for $h > 0$ and the eigenvalues of $A(\tilde{\mathbf{U}})$ belong to the interval $(\lambda_{min}, \lambda_{max})$ with*

$$\lambda_{min} = \min_j\{u_j, v_j\} - \frac{\max\{\rho_j\}}{\min\{\rho_j\}}\sqrt{gh},$$
$$\lambda_{max} = \max\{u_j, v_j\} + \frac{\max\{\rho_j\}}{\min\{\rho_j\}}\sqrt{gh}.$$

Proof of prop. 2. Since $A(\tilde{\mathbf{U}})$ is a block-matrix, its eigenvalues consist in the eigenvalues of $A_1(\tilde{\mathbf{U}})$ completed with the set $\{u_i\}_{i=1}^N$. Writing the characteristic polynomial of $A_1(\tilde{\mathbf{U}})$ under the form (development e.g. along the first row)

$$P_{A_1} = \Pi_{i=1}^N(\lambda - u_i)\left(\lambda - \sum_{j=1}^{N} l_j u_j - \sum_{i=1}^{N} \frac{l_i h \tilde{p}_i}{\lambda - u_i}\right),$$

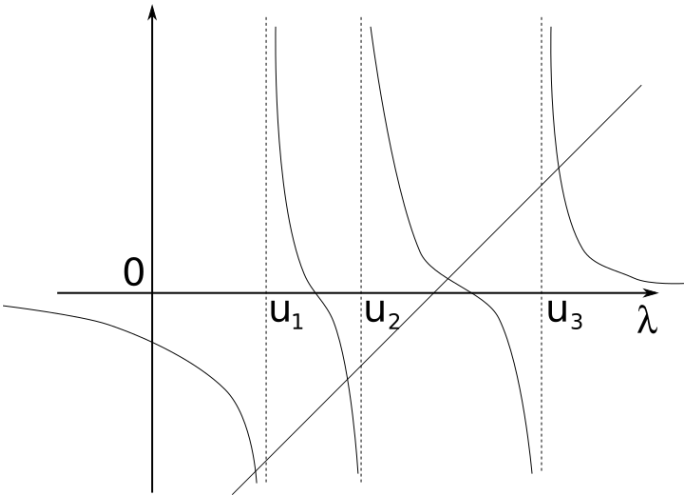


Figure 4.4 – The two functions $\lambda \mapsto Q_{A_1}(\lambda)$ and $\lambda \mapsto \lambda - \sum_{j=1}^{N} l_j u_j$, each intersection of the two curves is an eigenvalue of $A_1(\tilde{\mathbf{U}})$.

the eigenvalues of $A_1(\tilde{\mathbf{U}})$ satisfy

$$\lambda - \sum_{j=1}^{N} l_j u_j = Q_{A_1}(\lambda),$$

with $Q_{A_1}(\lambda) = \sum_{i=1}^{N} \frac{l_i h \tilde{p}_i}{\lambda - u_i}$. For $N = 3$, the functions $\lambda \mapsto Q_{A_1}(\lambda)$ and $\lambda \mapsto \lambda - \sum_{j=1}^{N} l_j u_j$ are depicted over Fig. 4.4 and it is easy to see that the four eigenvalues λ_i exists with the interlacing

$$\lambda_1 < u_1 \leqslant \lambda_2 \leqslant \ldots \leqslant u_3 < \lambda_4.$$

Moreover we have

$$Q_{A_1}(\lambda_{max}) = \sum_{i=1}^{N} \frac{l_i h \tilde{p}_i}{\lambda_{max} - u_i} \leqslant \frac{\min\{\rho_j\}}{\max\{\rho_j\}} \sum_{i=1}^{N} \frac{l_i h \tilde{p}_i}{\sqrt{gh}} \leqslant \frac{\max\{\rho_j\}}{\min\{\rho_j\}} \sqrt{gh} \leqslant \lambda_{max} - \sum_{j=1}^{N} l_j u_j,$$

and likewise

$$Q_{A_1}(\lambda_{min}) = \sum_{i=1}^{N} \frac{l_i h \tilde{p}_i}{\lambda_{min} - u_i} \geqslant -\frac{\min\{\rho_j\}}{\max\{\rho_j\}} \sum_{i=1}^{N} \frac{l_i h \tilde{p}_i}{\sqrt{gh}} \geqslant -\frac{\max\{\rho_j\}}{\min\{\rho_j\}} \sqrt{gh} \geqslant \lambda_{min} - \sum_{j=1}^{N} l_j u_j,$$

therefore the eigenvalues $\{\lambda_i\}_{i=1}^{N}$ of $A_1(\tilde{\mathbf{U}})$ satisfy

$$\lambda_{min} \leqslant \lambda_i \leqslant \lambda_{max}, \qquad i = 1, \ldots N,$$

proving the result.

□

The horizontal fluxes and the pressure terms

A finite volume scheme for solving the system (4.31)-(4.34) is a formula of the form

$$\mathbf{U}_i^{n+1/2} = \mathbf{U}_i - \sum_{j\in K_i} \sigma_{i,j}\mathcal{F}_{i,j} - \sigma_i \mathcal{F}_{e,i} + \sum_{j\in K_i} \sigma_{i,j}\mathcal{S}_p(\mathbf{U}_i, \mathbf{U}_j, z_{b,i}, z_{b,j}), \tag{4.50}$$

where using the notations of (4.30)

$$\sum_{j\in K_i} L_{i,j}\mathcal{F}_{i,j} \approx \int_{C_i} \nabla_{x,y} \cdot F(\mathbf{U}) dxdy, \tag{4.51}$$

with

$$\sigma_{i,j} = \frac{\Delta t^n L_{i,j}}{|C_i|}, \quad \sigma_i = \frac{\Delta t^n L_i}{|C_i|}.$$

Here we consider first-order explicit schemes where

$$\mathcal{F}_{i,j} = \begin{pmatrix} F_{i,j}^{h} \\ F_{i,j}^{\rho_1 h_1} \\ \vdots \\ F_{i,j}^{\rho_N h_N} \\ F_{i,j}^{\mathbf{u}_1} \\ \vdots \\ F_{i,j}^{\mathbf{u}_N} \end{pmatrix}. \tag{4.52}$$

and

$$F_{i,j}^{h} = \sum_{\alpha=1}^{N} F_{i,j}^{h_\alpha}, \tag{4.53}$$

and for the boundary nodes

$$\mathcal{F}_{i,e} = \begin{pmatrix} F_{i,e}^{h} \\ F_{i,e}^{\rho_1 h_1} \\ \vdots \\ F_{i,e}^{\rho_N h_N} \\ F_{i,e}^{\mathbf{u}_1} \\ \vdots \\ F_{i,e}^{\mathbf{u}_N} \end{pmatrix}. \tag{4.54}$$

The fluxes $F_{i,j}^{h_\alpha}$, $F_{i,j}^{\rho_\alpha h_\alpha}$, $F_{i,j}^{\mathbf{u}_\alpha}$ appearing in expressions (4.52),(4.53), (4.54) are numerical fluxes such that

$$F_{i,j}^{m_\alpha} = F^{m_\alpha}(\mathbf{U}_{\alpha,i}, \mathbf{U}_{\alpha,j}, \mathbf{n}_{i,j}),$$

with $m = h, \rho h, \mathbf{u}$, $\alpha = 1, \dots, N$.

Relation (4.50) tells how to compute the values $\mathbf{U}_i^{n+1/2}$ knowing $\mathbf{U}_i$ and discretized values $z_{b,i}$ of the topography. Following (4.51), the term $\mathcal{F}_{i,j}$ in (4.50) denotes an interpolation of the normal component of the flux $F(\mathbf{U}).\mathbf{n}_{i,j}$ along the edge $C_{i,j}$. The functions $F(\mathbf{U}_i, \mathbf{U}_j, \mathbf{n}_{i,j}) \in \mathbb{R}^{2N+1}$ are the numerical fluxes, see [32].

Until now, the expression for the numerical fluxes is not detailed since any numerical fluxes (Rusanov, HLL,...) can be used [32]. In paragraph 4.3.4 we define $\mathcal{F}(\mathbf{U}_i, \mathbf{n}_{i,j})$ using kinetic fluxes and we prove a discrete entropy inequality for the system. The computation of the value $\mathbf{U}_{i,e}$, which denotes a value outside C_i (see Fig. 4.3-*(b)*), defined such that the boundary conditions are satisfied, and the definition of the boundary flux $F(\mathbf{U}_i, \mathbf{U}_{e,i}, \mathbf{n_i})$ are described in [6]. Notice that we assume a flat topography on the boundaries i.e. $z_{b,i} = z_{b,i,e}$.

For the discretization of the pressure source term $\mathcal{S}_p(\mathbf{U}, z_b)$, we adopt a strategy defined below.

The hydrostatic reconstruction technique

The hydrostatic reconstruction scheme (HR scheme for short) for the Saint-Venant system has been introduced in [16] in the 1d case and described in 2d for unstructured meshes in [18]. The HR in the context of the kinetic description for the Saint-Venant system has been studied in [17].

In order to take into account the topography variations and to preserve relevant equilibria, the HR leads to a modified version of (4.50) under the form

$$U_i^{n+1/2} = U_i - \sum_{j \in K_i} \sigma_{i,j} \mathcal{F}_{i,j}^* - \sigma_i \mathcal{F}_{i,e} + \sum_{j \in K_i} \sigma_{i,j} \mathcal{S}_{p,i,j}^*, \tag{4.55}$$

where

$$\begin{aligned} \mathcal{F}_{i,j}^* &= F(U_{i,j}^*, U_{j,i}^*, \mathbf{n}_{i,j}), & (4.56) \\ \mathcal{S}_{p,i,j}^* &= S_p(U_i, U_{i,j}^*, z_{b,i}, z_{b,j}, \mathbf{n}_{i,j}) \\ &= \begin{pmatrix} 0 \\ \tilde{p}_{1,i,j}^*(h_{1,i,j} - h_{1,i})\mathbf{n}_{i,j} - g\rho_{1,i}\tilde{h}_{1,i,j}^*(z_{1,i,j} - z_{1,i})\mathbf{n}_{i,j} \\ \vdots \\ \tilde{p}_{\alpha,i,j}^*(h_{\alpha,i,j} - h_{\alpha,i})\mathbf{n}_{i,j} - g\rho_{\alpha,i}\tilde{h}_{\alpha,i,j}^*(z_{\alpha,i,j} - z_{\alpha,i})\mathbf{n}_{i,j} \\ \vdots \end{pmatrix} & (4.57) \end{aligned}$$

with

$$\begin{array}{l} z^*_{b,i,j} = \max(z_{b,i}, z_{b,j}), \quad h^*_{i,j} = \max(h_i + z_{b,i} - z^*_{b,i,j}, 0), \\ U^*_{i,j} = (h^*_{i,j}, \rho_{1,i} l_1 h^*_{i,j}, \ldots, \rho_{N,i} l_N h^*_{i,j}, \rho_{1,i} l_1 h^*_{i,j} u_{1,i}, \ldots, \rho_{N,i} l_N h^*_{i,j} u_{N,i}, \ldots)^T, \\ z^*_{\alpha,i,j} = z^*_{b,i,j} + \left(\frac{l_\alpha}{2} + \sum_{j=1}^{\alpha-1} l_j\right) h^*_{i,j}, \\ z_{\alpha,i,j} = \frac{z^*_{\alpha,i,j} + z^*_{\alpha,j,i}}{2}, \\ h_{\alpha,i,j} = \frac{h^*_{\alpha,i,j} + h^*_{\alpha,j,i}}{2}, \\ \tilde{h}^*_{\alpha,i,j} = \frac{h_{\alpha,i} + h^*_{\alpha,i,j}}{2}, \\ \tilde{p}^*_{\alpha,i,j} = \frac{p_{\alpha,i} + p^*_{\alpha,i,j}}{2}, \end{array} \tag{4.58}$$

and

$$z_\alpha = z_b + \left(\frac{l_\alpha}{2} + \sum_{j=1}^{\alpha-1} l_j\right) h, \qquad p_\alpha = \frac{\rho_\alpha g h_\alpha}{2} + \sum_{j=\alpha+1}^{N} \rho_j g h_j.$$

Throughout this work, the * refers to the HR technique.

Remark 1. Since the quantity $\mathcal{S}(\mathbf{U}, z_b)$ appearing in (4.29) contains non conservative terms, its integration over the cell C_i is not straightforward and we have used the result proposed by Bouchut [32, Proposition 5.3] to obtain the expression (4.57).

The vertical exchange terms

We give the fully discrete expression of the step for the vertical exchanges, described by equations (4.35)-(4.36). The step for the vertical exchanges consists in

$$U_i^{n+1} = U_i^{n+1/2} + \Delta t^n \mathcal{G}_i^{n+1}, \tag{4.59}$$

with

$$\mathcal{G}_i^{n+1} = \begin{pmatrix} 0 \\ \rho^{n+1}_{3/2,i} G_{3/2,i} \\ \rho^{n+1}_{5/2,i} G_{5/2,i} - \rho^{n+1}_{3/2,i} G_{3/2,i} \\ \vdots \\ \rho^{n+1}_{N-1/2,i} G_{N-1/2,i} - \rho^{n+1}_{N-3/2,i} G_{N-3/2,i} \\ -\rho^{n+1}_{N-1/2,i} G_{N-1/2,i} \\ u^{n+1}_{3/2,i} \rho^{n+1}_{3/2,i} G_{3/2,i} \\ u^{n+1}_{5/2,i} \rho^{n+1}_{5/2,i} G_{5/2,i} - u^{n+1}_{3/2,i} \rho^{n+1}_{3/2,i} G_{3/2,i} \\ \vdots \\ u^{n+1}_{N-1/2,i} \rho^{n+1}_{N-1/2,i} G_{N-1/2,i} - u^{n+1}_{N-3/2,i} \rho^{n+1}_{N-3/2,i} G_{N-3/2,i} \\ -u^{n+1}_{N-1/2,i} \rho^{n+1}_{N-1/2,i} G_{N-1/2,i} \end{pmatrix}.$$

This system and its numerical resolution are studied afterwards, in section 4.3.3.

The variable update

Hence the sum of relations (4.55) and (4.59) gives

$$U_i^{n+1} = U_i - \dots, \tag{4.60}$$

The space discretization of the system (4.34)-(4.36) and its numerical resolution allow to determine the quantities

$$h_i^{n+1}, \rho_{\alpha,i}^{n+1}, (\rho_\alpha h_\alpha \mathbf{u}_\alpha)_i^{n+1},$$

for any $i \in I$ knowing the quantities $\{h_j, (\rho_\alpha h_\alpha)_j, (\rho_\alpha h_\alpha \mathbf{u}_\alpha)_j\}_{j\in I}$.

Thus, it is possible to recover $\rho_{\alpha,i}^{n+1}$ and $\mathbf{u}_{\alpha,i}^{n+1}$ from

$$\rho_{\alpha,i}^{n+1} = \frac{(\rho_\alpha h_\alpha)_i^{n+1}}{l_\alpha h_i^{n+1}}, \qquad \mathbf{u}_{\alpha,i}^{n+1} = \frac{(\rho_\alpha h_\alpha \mathbf{u}_\alpha)_i^{n+1}}{(\rho_\alpha h_\alpha)_i^{n+1}}.$$

Concerning the temperature $T_{\alpha,i}^{n+1}$, we use the formula

$$T_{\alpha,i}^{n+1} = \rho^{-1}(\rho_{\alpha,i}^{n+1}).$$

Properties of the numerical scheme

Proposition 3. *Consider a consistent numerical flux $\mathcal{F}$ for the homogeneous problem (4.31)-(4.34) i.e. Eq. (4.40) with $\mathcal{S}_p(\mathbf{U}, z_b) = 0$) that preserves the non-negativity of the water depth h_i under the corresponding CFL condition, then the finite volume scheme (4.55)-(4.59)*

(i) preserves the non-negativity of the water depth,

(ii) preserves the steady state of a lake at rest,

(iii) is consistent with the system (4.25)-(4.27),(4.28).

Proof. *(i)* For $\alpha = 1, \dots, N$, $h_\alpha^{n+1/2}$ is obtained using the fully discrete version of (4.31), that is to say,

$$h_{\alpha,i}^{n+1/2} = h_{\alpha,i} - \sum_{j\in K_i} \sigma_{i,j} \mathcal{F}_{i,j}^{h_\alpha}$$

Since the flux $\mathcal{F}$ preserves the positivity for the shallow water equations, then $h_{\alpha,i}^{n+1/2}$ is positive. As a sum of positive terms, h^{n+1} is positive. This proves *(i)*.

(ii) In the constant density case and with a non-flat topography, *(ii)* is proved in [21]. In the variable density case with flat topography, the proof is trivial because at equilibrium, the density is constant in each of the layers. In the general case, the continuous

equations describing the static equilibrium are

$$\begin{aligned} u_\alpha &= 0, \quad \alpha = 1, \dots, N, \\ \nabla_{x,y}\eta &= 0, \\ \nabla_{x,y}(h_\alpha p_\alpha) &= p_{\alpha+1/2}\nabla_{x,y} z_{\alpha+1/2} - p_{\alpha-1/2}\nabla_{x,y} z_{\alpha-1/2}, \quad \alpha = 1, \dots, N. \end{aligned} \tag{4.61}$$

While the first two equations are easy to write at the discrete level, the third one is not. There is no simple discrete formulation of (4.61). Yet, for

$$\mathbf{u}_{\alpha,i} = 0, \quad \forall \alpha, \forall i, \qquad \text{and} \qquad \eta_i = \eta_{eq}, \quad \forall i \tag{4.62}$$

with η_{eq} a constant, due to the use of the HR technique, the proposed numerical scheme verifies

$$\begin{aligned} h_{\alpha,i}^{n+1/2} &= h_{\alpha,i}, \\ (\rho h)_{\alpha,i}^{n+1/2} &= (\rho h)_{\alpha,i}, \\ (\rho h)_{\alpha,i}^{n+1} &= (\rho h)_{\alpha,i}^{n+1/2}, \\ (\rho h \mathbf{u})_{\alpha,i}^{n+1} &= (\rho h \mathbf{u})_{\alpha,i}^{n+1/2}, \end{aligned}$$

for all α and all i. Therefore, starting from a situation described by (4.62) and a discrete version of (4.61), the discrete equilibrium is preserved. Starting from a situation near the equilibrium, the system will evolve towards equilibrium. We have empirically checked that a system initially far from equilibrium will reach an equilibrium, see [21].

(*iii*) Since the flux $\mathcal{F}$ is consistent with the homogeneous ($\mathcal{S}_p(U, z_b) = 0$)), Eq. (4.57) is a consistent discretization of the non-conservative pressure terms, and (4.59) is a consistent discretization of the vertical exchange terms, then (*iii*) is true.

□

Proposition 4. *Consider a consistent numerical flux $\mathcal{F}$ for the homogeneous problem (4.31)-(4.34) i.e. Eq. (4.40) with $\mathcal{S}_p(\mathbf{U}, z_b) = 0$ that preserves nonnegativity of $h_i(t)$ and such that*

$$F_{i,j}^{\rho_\alpha h_\alpha} = \rho_{\alpha,i,j} F_{i,j}^{h_\alpha}, \tag{4.63}$$

with

$$\rho_{\alpha,i,j} = \left\{ \begin{array}{ll} \rho_{\alpha,i} & \text{if } F_{i,j}^{h_\alpha} \geqslant 0 \\ \rho_{\alpha,j} & \text{if } F_{i,j}^{h_\alpha} \leqslant 0 \end{array} \right\} \tag{4.64}$$

then the numerical scheme (4.55)-(4.59) satisfies a maximum principle on the density i.e. for any α, i one has

$$\rho_{\alpha,i}^{n+1} \leqslant \max_\alpha \{\rho_{\alpha,i}, \rho_{\alpha,j}\}, \qquad \forall j \in K_i.$$

Remark 2. The formula (4.63) was initially proposed in [89] see also [32].

Proof of prop 4. Let us first deal with the horizontal exchanges. Due to the choice of $F_{i,j}^{\rho_\alpha h_\alpha}$, the numerical discretization of (4.32) can be decomposed into

$$(\rho_\alpha h_\alpha)_i^{n+1/2} = \rho_{\alpha,i}\left(h_{\alpha,i}^n - \sum_{j\in K_i}\sigma_{i,j}|F_{i,j}^{h_\alpha}|_+\right) - \sum_{j\in K_i}\sigma_{i,j}\rho_{\alpha,j}|F_{i,j}^{h_\alpha}|_-$$

The right hand side of this expression is positive. Indeed, $h_{\alpha,i}^n - \sum_{j\in K_i}\sigma_{i,j}|F_{i,j}^{h_\alpha}|_+$ is positive due to the CFL condition and $-\sum_{j\in K_i}\sigma_{i,j}\rho_{\alpha,j}|F_{i,j}^{h_\alpha}|_-$ is positive because $|F_{i,j}^{h_\alpha}|_-$ is negative. The right hand side is factorized by $\max\{\rho_{\alpha,i},\rho_{\alpha,j}\}$

$$(\rho_\alpha h_\alpha)_i^{n+1/2} \leqslant \max\{\rho_{\alpha,i},\rho_{\alpha,j}\}\left(h_{\alpha,i}^n - \sum_{j\in K_i}\sigma_{i,j}|F_{i,j}^{h_\alpha}|_+ - \sum_{j\in K_i}\sigma_{i,j}|F_{i,j}^{h_\alpha}|_-\right).$$

Therefore, dividing by $h_{\alpha,i}^{n+1/2}$ positive, we obtain that

$$\rho_{\alpha,i}^{n+1/2} \leqslant \max\{\rho_{\alpha,i},\rho_{\alpha,j}\}, \qquad \forall j\in K_i. \tag{4.65}$$

We deal next with the vertical exchanges. We subtract the equation

$$h_{\alpha,i}^{n+1} = h_{\alpha,i}^{n+1/2} + \Delta t(G_{\alpha+1/2,i} - G_{\alpha-1/2,i})$$

multiplied by $\rho_{\alpha,i}^{n+1}$ from equation (4.35). It comes

$$\rho_{\alpha,i}^{n+1}h_{\alpha,i}^{n+1/2} = \rho_{\alpha,i}^{n+1/2}h_{\alpha,i}^{n+1/2} + \Delta t((\rho_{\alpha+1/2,i}^{n+1} - \rho_{\alpha,i}^{n+1})G_{\alpha+1/2,i} - (\rho_{\alpha-1/2,i}^{n+1} - \rho_{\alpha,i}^{n+1})G_{\alpha-1/2,i}).$$

This relation can be rewritten as

$$(H_{N,i}^{n+1/2} + \Delta t G_{N,i})\rho_i^{n+1} = (\rho_i h_i)^{n+1/2}$$

where $H_{N,i}^{n+1/2}$ is a diagonal matrix of size N with coefficients $H_{N,i_j\,1\leqslant j\leqslant N} = h_j^{n+1/2}$ and the matrix $G_{N,i}$ is given by

$$G_{N,i} = \begin{pmatrix}
\frac{|G_{3/2,i}|_+}{h_{1,i}^{n+1/2}} & -\frac{|G_{3/2,i}|_+}{h_{1,i}^{n+1/2}} & 0 & 0 & \dots & 0 \\
\frac{|G_{3/2,i}|_-}{h_{1,i}^{n+1/2}} & \ddots & \ddots & 0 & \dots & 0 \\
0 & \ddots & \ddots & \ddots & 0 & 0 \\
\vdots & 0 & \frac{|G_{\alpha-1/2,i}|_-}{h_{\alpha,i}^{n+1/2}} & \frac{|G_{\alpha+1/2,i}|_+}{h_{\alpha,i}^{n+1/2}} - \frac{|G_{\alpha-1/2,i}|_-}{h_{\alpha,i}^{n+1/2}} & -\frac{|G_{\alpha+1/2,i}|_+}{h_{\alpha,i}^{n+1/2}} & 0 \\
\vdots & \ddots & 0 & \ddots & \ddots & -\frac{|G_{N-1/2,i}|_+}{h_{N,i}^{n+1/2}} \\
0 & \dots & 0 & 0 & \frac{|G_{N-1/2,i}|_-}{h_{N,i}^{n+1/2}} & -\frac{|G_{N-1/2,i}|_-}{h_{N,i}^{n+1/2}}
\end{pmatrix}$$

If $h_{\alpha,i}^{n+1/2} = 0$ for all α, then we trivially have $\rho_{\alpha,i}^{n+1} = 0$ for all α. Let us now assume that there exists a layer α such that $h_{\alpha,i}^{n+1} > 0$. Then the matrix $H_{N,i}^{n+1/2} + \Delta t G_{N,i}$ is a strictly diagonally dominant matrix. Therefore, it is invertible and the entries of its inverse are all nonnegative - see the proof made in [23]. The matrix $H_{N,i}^{n+1/2} + \Delta t G_{N,i}$ is an M-matrix. Let $\mathbf{1}$ be the vector the entries of which are all equal to 1. We notice that

$$(H_{N,i}^{n+1/2} + \Delta t G_{N,i})\mathbf{1} = h_i^{n+1/2},$$

so we also have $\mathbf{1} = (H_{N,i}^{n+1/2} + \Delta t G_{N,i})^{-1} h_i^{n+1/2}$. Let $(\rho_i h_i)^{n+1/2}$ be the vector the entries of which are $(\rho_i h_i)_\alpha^{n+1/2} = \rho_{\alpha,i}^{n+1/2} h_{\alpha,i}^{n+1/2}$, $\alpha = 1, \dots, N$. Then

$$||(H_{N,i}^{n+1/2} + \Delta t G_{N,i})^{-1} (\rho_i h_i)^{n+1/2}||_\infty \leqslant ||\rho_i^{n+1/2}||_\infty ||(H_{N,i}^{n+1/2} + \Delta t G_{N,i})^{-1} h_i^{n+1/2}||_\infty,$$

which is exactly

$$||\rho_i^{n+1}||_\infty \leqslant ||\rho_i^{n+1/2}||_\infty. \tag{4.66}$$

To conclude, relationship (4.65) is applied to $\rho_{\alpha,i}^{n+1/2}$ for all α. Combining with (4.66) gives the maximum principle on the density. □

Remark 3. Let us present a more accurate result for the maximum principle on the vertical exchanges. Let α, α_0 and α_1 such that $1 \leqslant \alpha_0 < \alpha < \alpha_1 \leqslant N$ and

$$G_{\alpha_j+1/2,i} < 0, \quad G_{\alpha_j-1/2,i} > 0 \qquad \text{for } j \in \{0, 1\}.$$

The coefficients of the lines α_0 and α_1 of matrix $G_{N,i}$ are respectively $(G_{N,i})_{\alpha_0,j} = \delta_{\alpha_0,j}$ and $(G_{N,i})_{\alpha_1,j} = \delta_{\alpha_1,j}$ with $\delta_{k,l}$ the Kronecker symbol, which means that $\rho_{\alpha_0,i}^{n+1} = \rho_{\alpha_0,i}^{n+1/2}$ and $\rho_{\alpha_1,i}^{n+1} = \rho_{\alpha_1,i}^{n+1/2}$. The system can be solved using forward elimination and backward substitution. Denoting by $\rho_{k-l,i}^{n+1/2}$ the vector $(\rho_{k,i}^{n+1/2}, \rho_{k+1,i}^{n+1/2}, \dots, \rho_{l,i}^{n+1/2})^T$, we have the following results

$$||\rho_{1-\alpha_0,i}^{n+1}||_\infty \leqslant ||\rho_{1-\alpha_0,i}^{n+1/2}||_\infty, \quad ||\rho_{\alpha_0-\alpha_1,i}^{n+1}||_\infty \leqslant ||\rho_{\alpha_0-\alpha_1,i}^{n+1/2}||_\infty, \quad ||\rho_{\alpha_0-N,i}^{n+1}||_\infty \leqslant ||\rho_{\alpha_0-N,i}^{n+1/2}||_\infty.$$

The layers receiving no mass from the layers above and below them separate groups of layers which exchange mass between themselves.

Remark 4. For the following semi-implicit scheme for the vertical exchanges, the maximum principle on the density of proposition 4 is also verified:

$$\begin{aligned}(\rho_{\alpha,i} h_{\alpha,i})^{n+1} = (\rho_{\alpha,i} h_{\alpha,i})^{n+1/2} &+ \frac{\Delta t}{2}(\rho_{\alpha+1/2,i}^{n+1} G_{\alpha+1/2,i} - \rho_{\alpha-1/2,i}^{n+1} G_{\alpha-1/2,i}) \\ &+ \frac{\Delta t}{2}(\rho_{\alpha+1/2,i}^{n+1/2} G_{\alpha+1/2,i} - \rho_{\alpha-1/2,i}^{n+1/2} G_{\alpha-1/2,i}).\end{aligned}$$

The more accurate maximum principle presented in remark 3 is verified as well. However, in the rest of the paper, we work only with the fully implicit scheme for the vertical

exchanges. More specifically, proposition 5 is stated only for the fully implicit scheme.

Remark 5. We have presented a first order in space discretization of the system. In practice, we apply a formally second order extension in space and time presented in [18] and [6]. More specifically, the modified Heun scheme which is used is presented in [6]. For the obtained numerical scheme we are able to prove the consistence, the well-balancing and the non-negativity of the water depth. But a discrete entropy inequality such as the one in proposition 5 has not yet been obtained. The proof of proposition 5 cannot be adapted for the second order scheme; a different strategy would be needed.

Remark 6. At each time step, to advance

- from h_i^n to $h_i^{n+1/2}$
- from $(\rho_{\alpha,i}h_{\alpha,i})^n$ to $(\rho_{\alpha,i}h_{\alpha,i})^{n+1/2}$

convex combinations are used, which gives the scheme a certain stability. Then, to advance from $(\rho_{\alpha,i}h_{\alpha,i})^{n+1/2}$ to $(\rho_{\alpha,i}h_{\alpha,i})^{n+1}$, the fact that the matrix of the system (matrix $H_{N,i}^{n+1/2} + \Delta t G_{N,i}$, defined in the proof of proposition 4) is an M-matrix gives stability to the computation.

4.3.4 Kinetic fluxes

Whereas the proposed numerical scheme can be adapted to any finite volume solver for the classical Saint-Venant system, in Section 4.5, the numerical simulations are performed using a kinetic solver and hence the numerical fluxes in (4.52) in the kinetic context now are specified.

To define the numerical fluxes, we introduce the functions χ_0, M_α

$$\chi_0(z_1, z_2) = \frac{1}{4\pi}\mathbb{1}_{z_1^2+z_2^2\leqslant 4},$$
$$M_\alpha = M(U_\alpha, \xi, \gamma) = \frac{h_\alpha}{c_\alpha^2}\chi_0\left(\frac{\xi - u_\alpha}{c_\alpha}, \frac{\gamma - v_\alpha}{c_\alpha}\right),$$

with $c_\alpha = \sqrt{p_\alpha/\rho_\alpha}$, U_α defined by (4.42) and where $(\xi, \gamma) \in \mathbb{R}^2$. We also define the quantity M_α^ρ by

$$M_\alpha^\rho = \rho_\alpha M_\alpha.$$

The quantity M_α^ρ satisfies the following moment relations

$$\int_{\mathbb{R}^2} \begin{pmatrix} 1 \\ \xi \end{pmatrix} \rho_\alpha M(U_\alpha, \xi, \gamma)\, d\xi d\gamma = \begin{pmatrix} \rho_\alpha h_\alpha \\ \rho_\alpha h_\alpha u_\alpha \\ \rho_\alpha h_\alpha v_\alpha \end{pmatrix}, \tag{4.67}$$

$$\int_{\mathbb{R}^2} \begin{pmatrix} \xi^2 \\ \xi\gamma \\ \gamma^2 \end{pmatrix} \rho_\alpha M(U_\alpha, \xi, \gamma)\, d\xi d\gamma = \begin{pmatrix} \rho_\alpha h_\alpha u_\alpha^2 + h_\alpha p_\alpha \\ \rho_\alpha h_\alpha u_\alpha v_\alpha \\ \rho_\alpha h_\alpha v_\alpha^2 + h_\alpha p_\alpha \end{pmatrix}. \tag{4.68}$$

Hence in the context of the kinetic description, the fluxes appearing in (4.52) have the expressions

$$F_{i,j}^{h_\alpha} = \int_{\mathbb{R}^2} M_{\alpha,i,j}\zeta_{i,j}d\xi d\gamma, \quad F_{i,j}^{\rho_\alpha h_\alpha} = \rho_{\alpha,i,j}F_{i,j}^{h_\alpha}, \quad F_{i,j}^{\mathbf{u}_\alpha} = \rho_{\alpha,i,j}\int_{\mathbb{R}^2}\begin{pmatrix}\xi\\ \gamma\end{pmatrix} M_{\alpha,i,j}\zeta_{i,j}d\xi d\gamma \tag{4.69}$$

with

$$M_{\alpha,i,j} = M^*_{\alpha,i,j}\mathbb{1}_{\zeta_{i,j}\geqslant 0} + M^*_{\alpha,j,i}\mathbb{1}_{\zeta_{i,j}\leqslant 0}, \tag{4.70}$$

where $M^*_{\alpha,i,j} = M(U^*_{\alpha,i,j},\xi,\gamma)$, $U^*_{\alpha,i,j} = (l_\alpha h^*_{i,j}, l_\alpha h^*_{i,j}u_{\alpha,i}, l_\alpha h^*_{i,j}v_{\alpha,i})^T$. The * refers to the HR technique, see (4.58). The density $\rho_{\alpha,i,j}$ is defined by (4.64) and

$$\zeta_{i,j} = \begin{pmatrix}\xi\\ \gamma\end{pmatrix}.\mathbf{n}_{i,j}.$$

We give here some details about Proposition 3 in the case of the kinetic flux. We consider the equation

$$f_{\alpha,i}^{n+1/2-} = M_{\alpha,i} - \sum_{j\in K_i}\sigma_{i,j}\zeta_{i,j}\mathbb{1}_{\zeta_{i,j}\geqslant 0}M^*_{\alpha,i,j} - \sum_{j\in K_i}\sigma_{i,j}M^*_{\alpha,j,i}\zeta_{i,j}\mathbb{1}_{\zeta_{i,j}\leqslant 0} \tag{4.71}$$

Using (4.67), we see that integrating equation (4.71) for $\alpha = 1,\dots,N$ with respect to ξ,γ gives the HR scheme for (4.31).

Let us now give the CFL condition for the kinetic flux. There exists a velocity $v_m \geqslant 0$ such that for all α, i

$$|\xi| \geqslant v_m \text{ or } |\gamma| \geqslant v_m \Rightarrow M(U_{\alpha,i},\xi,\gamma) = 0.$$

This means that $|u_{\alpha,i}| + |v_{\alpha,i}| + \sqrt{2gh_i} \leqslant v_m$. A CFL condition strictly less than one is considered

$$\tilde{\sigma}_i v_m \leqslant \beta < \frac{1}{2} \quad \text{for all } i,$$

where $\tilde{\sigma}_i = \Delta t^n \Sigma_{j\in K_i} L_{i,j}/|C_i|$, and β is a given constant. Under this CFL condition, the kinetic function $f_{\alpha,i}^{n+1/2-}$ remains non-negative, i.e.

$$f_{\alpha,i}^{n+1/2-} \geqslant 0, \quad \forall(\xi,\gamma)\in\mathbb{R}^2, \forall i, \forall\alpha.$$

The proof can be found in [6]. Therefore, the water depth $h^{n+1/2}$ is non-negative. Note that the CFL condition does not depend on the vertical exchange terms because they are treated implicitly.

4.3.5 Discrete entropy inequality

In this paragraph, a discrete entropy inequality is proved in the case of a flat topography. The crucial point of the numerical scheme is the treatment of the pressure source term

$\mathcal{S}_{p,\alpha}$ written under the form (4.39). Indeed the other terms appearing in the numerical scheme are either conservative – and hence easily incorporated in the numerical fluxes – or similar to terms appearing in the constant density case, see [6]. The term $\mathcal{S}_{p,\alpha}$ is an extension of the topography term for the Saint-Venant system but in a far more complex setting. Proposition 5 is interesting since until now, the properties satisfied by the numerical scheme detailed in paragraph 4.3.3 concern Eqs. (4.31)-(4.38) except Eq. (4.33). Hence, for the momentum equation (4.33), only the equilibrium at rest is proved.

In the context of the kinetic description, the relation between the mass and momentum fluxes is simple (see (4.69)) and it is possible to slightly modify the definitions (4.57) in order to obtain an in cell discrete entropy inequality. The authors do not claim that only the kinetic fluxes allow to obtain such a result but, as in [17], it is not clear whether the result holds with other numerical fluxes in particular the definition (4.73) is related to the kinetic description and not easily available for other finite volume solver (Rusanov, HLL,...).

We consider in this paragraph a discrete form of (4.39) that is slightly different from (4.57) and defined by

$$\mathcal{S}_{p,\alpha,i,j} = p_{\alpha,i}(\hat{h}_{\alpha,i,j} - h_{\alpha,i})\mathbf{n}_{i,j} - g\rho_{\alpha,i}h_{\alpha,i}(z_{\alpha,i,j} - z_{\alpha,i})\mathbf{n}_{i,j}, \tag{4.72}$$

with

$$\begin{aligned}
z_{\alpha+1/2,i} &= z_{b,i} + \sum_{l=1}^{\alpha} h_{l,i}, \quad \text{with } z_{b,i} = z_{b,j} = cst, \ \forall j, \\
z_{\alpha,i} &= \frac{z_{\alpha+1/2,i} + z_{\alpha-1/2,i}}{2}, \\
z_{\alpha,i,j} &= \frac{z_{\alpha,i} + z_{\alpha,j}}{2}, \\
\hat{h}_{\alpha,i,j} &= \int_{\mathbb{R}^2} (M_{\alpha,i}\mathbb{1}_{\zeta_{i,j}\leqslant 0} + M_{\alpha,j}\mathbb{1}_{\zeta_{i,j}\geqslant 0})d\xi d\gamma, = \int_{\mathbb{R}^2} M_{\alpha,i,j}d\xi d\gamma \qquad (4.73) \\
p_{\alpha,i} &= \frac{\rho_{\alpha,i}}{2} \frac{\int_{\mathbb{R}^2} \left| \begin{pmatrix} \xi \\ \gamma \end{pmatrix} - \mathbf{u}_{\alpha,i} \right|^2 M_{\alpha,i}d\xi d\gamma}{\int_{\mathbb{R}^2} M_{\alpha,i}d\xi d\gamma}. \qquad (4.74)
\end{aligned}$$

Since we consider a flat topography, the notations associated with the HR do not appear in the previous definitions. The discretization (4.72) is very similar to the discretization (4.57) written for a flat topography, but a bit of upwinding with respect to the advection is included in (4.72). Note that this kind of upwinding could help to reduce the error term due to the topography in [17]. For a non-flat topography, the HR induces upwinding with respect to the topography, not with respect to the advection.

The main interest of the following proposition is to justify the discretization (4.72) but for three reasons it is a partial result:

- it only concerns flat topography situations whereas it is well known that the nu-

merical treatment of topography source terms is a very difficult issue,

- the extension of the expression (4.72) to the situation of a non flat topography is not natural since in the simple case of a single layer with a constant density i.e. the classical Saint-Venant system, the definition (4.72) does not exactly match with previous works of some of the authors [17] (whereas definition (4.57) exactly reduces to the scheme studied in [17] in the Saint-Venant case),

- the numerical tests carried out with the two possible discretizations of $\mathcal{S}_{p,\alpha,i,j}$, namely (4.57) and (4.72) give very similar results especially similar convergence curves, see paragraph 4.5.1.

For these reasons and even if the result proposed in the following proposition is an interesting stability property, the numerical simulations presented in Section 4.4 have been obtained using the discretization (4.57).

Proposition 5. *When considering a flat bottom, the scheme (4.55),(4.59) with the fluxes defined by (4.69) satisfies an in cell fully discrete entropy inequality having the form*

$$
\begin{aligned}
&E_{\alpha,i}^{n+1} - E_{\alpha,i} + \sum_{j\in K_i} \sigma_{i,j} \int_{\mathbb{R}^2} \left(g z_{\alpha,i,j} + \frac{\xi^2+\gamma^2}{2} - \frac{1}{2}\left| \begin{pmatrix} \xi \\ \gamma \end{pmatrix} - \tilde{\mathbf{u}}_{\alpha,i,j} \right|^2 \right) \rho_{\alpha,i,j} M_{\alpha,i,j} \zeta_{i,j} d\xi d\gamma \\
&-\Delta t^n \rho_{\alpha+1/2,i}^{n+1} \left(\frac{|\mathbf{u}_{\alpha+1/2,i}^{n+1}|^2}{2} + g z_{\alpha+1/2,i} \right) G_{\alpha+1/2,i} + \Delta t^n \left(\rho_{\alpha-1/2,i}^{n+1} \frac{|\mathbf{u}_{\alpha-1/2,i}^{n+1}|^2}{2} + g z_{\alpha-1/2,i} \right) G_{\alpha-1/2,i} \\
&- p_{\alpha+1/2,i}^{n*} \left(\Delta t^n G_{\alpha+1/2,i} - z_{\alpha+1/2,i}^{n+1} + z_{\alpha+1/2,i} \right) + p_{\alpha-1/2,i}^{n*} \left(\Delta t^n G_{\alpha-1/2,i} - z_{\alpha-1/2,i}^{n+1} + z_{\alpha-1/2,i} \right) \\
&= d_{\alpha,i} + e_{\alpha,i} + f_{\alpha,i}
\end{aligned}
$$

where $E_{\alpha,i}$ is the discrete energy

$$
E_{\alpha,i} = \rho_{\alpha,i} h_{\alpha,i} \frac{|\mathbf{u}_{\alpha,i}|^2}{2} + \rho_{\alpha,i} g h_{\alpha,i} z_{\alpha,i},
$$

and where $d_{\alpha,i}$ is a sum of non-positive and hence dissipative terms whereas $e_{\alpha,i}$ (resp. $f_{\alpha,i}$) contains errors terms of magnitude $\mathcal{O}(diam(C_i)^3)$ (resp. $\mathcal{O}(\Delta t^n)^2$). The expressions

of $d_{\alpha,i}$ and $e_{\alpha,i}$ are given by

$$\begin{aligned} d_{\alpha,i} =& -\sum_{j\in K_i}\sigma_{i,j}\frac{|\tilde{\mathbf{u}}_{\alpha,i,j}-\mathbf{u}_{\alpha,i}|^2}{2}\int_{\mathbb{R}^2}\rho_{\alpha,i}M_{\alpha,i}|\zeta_{i,j}|d\xi d\gamma \\ &+\sum_{j\in K_i}\sigma_{i,j}\frac{|\mathbf{u}_{\alpha,i}-\mathbf{u}_{\alpha,j}|^2}{4}\int_{\{\zeta_{i,j}\leqslant 0\}}\rho_{\alpha,j}M_{\alpha,j}\zeta_{i,j}d\xi d\gamma \end{aligned} \tag{4.75}$$

$$\begin{aligned} e_{\alpha,i} =& \sum_{j\in K_i}\sigma_{i,j}g\int_{\mathbb{R}^2}\left(\rho_{\alpha,i}h_{\alpha,i}\mathbf{u}_{\alpha,i}-\rho_{\alpha,i,j}M_{\alpha,i,j}\begin{pmatrix}\xi\\ \gamma\end{pmatrix}\right)(z_{\alpha,i,j}-z_{\alpha,i})\cdot\mathbf{n}_{i,j}d\xi d\gamma \\ &-\sum_{j\in K_i}\sigma_{i,j}p_{\alpha,i}\int_{\mathbb{R}^2}(M_{\alpha,i,j}-M_{\alpha,i})(\tilde{\mathbf{u}}_{\alpha,i,j}-\mathbf{u}_{\alpha,i})\cdot\mathbf{n}_{i,j}d\xi d\gamma \\ &-\sum_{j\in K_i}\sigma_{i,j}(\tilde{\mathbf{u}}_{\alpha,i,j}-\mathbf{u}_{\alpha,i})\cdot\int_{\{\zeta_{i,j}\leqslant 0\}}\left[\left(\begin{pmatrix}\xi\\ \gamma\end{pmatrix}-\mathbf{u}_{\alpha,i}\right)\rho_{\alpha,i}M_{\alpha,i}-\left(\begin{pmatrix}\xi\\ \gamma\end{pmatrix}-\mathbf{u}_{\alpha,j}\right)\rho_{\alpha,j}M_{\alpha,j}\right]\zeta_{i,j}d\xi d\gamma \\ &-\sum_{j\in K_i}\sigma_{i,j}\frac{|2\tilde{\mathbf{u}}_{\alpha,i,j}-\mathbf{u}_{\alpha,i}-\mathbf{u}_{\alpha,j}|^2}{4}\int_{\{\zeta_{i,j}\leqslant 0\}}\rho_{\alpha,j}M_{\alpha,j}\zeta_{i,j}d\xi d\gamma. \end{aligned} \tag{4.76}$$

The quantity $f_{\alpha,i}$ is defined by

$$\begin{aligned} f_{\alpha,i} =& \Delta t^n\frac{gh_{\alpha,i}}{2}\left((\rho_{\alpha,i}^{n+1}-\rho_{\alpha+1/2,i}^{n+1})G_{\alpha+1/2,i}+(\rho_{\alpha,i}^{n+1}-\rho_{\alpha-1/2,i}^{n+1})G_{\alpha-1/2,i}\right) \\ &+g\left((\rho_{\alpha,i}h_{\alpha,i})^{n+1}-\rho_{\alpha,i}^{n+1}h_{\alpha,i}\right)(z_{\alpha,i}^{n+1}-z_{\alpha,i})+\frac{(\rho_{\alpha,i}h_{\alpha,i})^{n+1}}{2}|\mathbf{u}_{\alpha,i}^{n+1}-\mathbf{u}_{\alpha,i}|^2 \\ &-\Delta t^n\rho_{\alpha+1/2,i}^{n+1}\frac{\mathbf{u}_{\alpha,i}-\mathbf{u}_{\alpha,i}^{n+1}}{2}.(\mathbf{u}_{\alpha,i}^{n+1}-2\mathbf{u}_{\alpha+1/2,i}^{n+1}+\mathbf{u}_{\alpha,i})G_{\alpha+1/2,i} \\ &+\Delta t^n\rho_{\alpha-1/2,i}^{n+1}\frac{\mathbf{u}_{\alpha,i}-\mathbf{u}_{\alpha,i}^{n+1}}{2}.(\mathbf{u}_{\alpha,i}^{n+1}-2\mathbf{u}_{\alpha-1/2,i}^{n+1}+\mathbf{u}_{\alpha,i})G_{\alpha-1/2,i}, \end{aligned}$$

where the first line is the discrete version of the term appearing in the last line of the continuous energy balance given in Eq. (4.23) (see also [30]) and the other lines come from the explicit time scheme and vanish in the semi-discrete (in space) case. The velocity $\tilde{\mathbf{u}}_{\alpha,i,j}$ is defined by

$$\tilde{\mathbf{u}}_{\alpha,i,j}=\frac{\displaystyle\int_{\mathbb{R}^2}M_{\alpha,i,j}\begin{pmatrix}\xi\\ \gamma\end{pmatrix}d\xi d\gamma}{\displaystyle\int_{\mathbb{R}^2}M_{\alpha,i,j}d\xi d\gamma}, \tag{4.77}$$

where $M_{\alpha,i,j}$ is defined by (4.70) and the pressure $p_{\alpha\pm1/2,i}^{n*}$ is defined by

$$p_{\alpha\pm1/2,i}^{n*}=p_{\alpha,i}\mp\rho_{\alpha,i}^{n+1}g\frac{h_{\alpha,i}}{2}.$$

Remark 7. Since we use an explicit time scheme it is natural to have error terms $f_{\alpha,i}$ of order $\mathcal{O}(\Delta t^n)^2$. Concerning the error terms $e_{\alpha,i}$ due to the space discretization, we point out that they are of order $\mathcal{O}(diam(C_i)^3)$ i.e. smaller than residuals with second terms.

Proof of prop. 5. The proof of this proposition is long but only contains simple computations. The authors have not found a simpler presentation.

Starting from the set of discrete equations (4.60), the energy balance for the cell i at the layer α is obtained by performing the sum of the two following quantities:

- the mass conservation equation over the layer α in (4.60) multiplied by $gz_{\alpha,i} - |\mathbf{u}_{\alpha,i}|^2/2$
- the momentum equation in (4.60) over the layer α multiplied by $\mathbf{u}_{\alpha,i}$.

In other words, the energy balance comes from a rewriting of the quantity

$$\begin{aligned}\mathcal{E}_{\alpha,i} \quad := \quad & \big(gz_{\alpha,i} - \frac{|\mathbf{u}_{\alpha,i}|^2}{2}\big)\Big((\rho_{\alpha,i}h_{\alpha,i})^{n+1} - \rho_{\alpha,i}h_{\alpha,i} + \sum_{j\in K_i}\sigma_{i,j}\rho_{\alpha,i,j}F_{i,j}^{h_\alpha} \\ & +\Delta t^n\Big(\rho^{n+1}_{\alpha+1/2,i}G_{\alpha+1/2,i} - \rho^{n+1}_{\alpha-1/2,i}G_{\alpha-1/2,i}\Big)\Big) \\ & +\mathbf{u}_{\alpha,i}.\Bigg(\Big((\rho_{\alpha,i}h_{\alpha,i}\mathbf{u}_{\alpha,i})^{n+1} - \rho_{\alpha,i}h_{\alpha,i}\mathbf{u}_{\alpha,i} + \sum_{j\in K_i}\sigma_{i,j}(F_{i,j}^{h_\alpha\mathbf{u}_{\alpha,i}} - \mathcal{S}_{p,\alpha,i,j}) \\ & -\Delta t^n\Big(\rho^{n+1}_{\alpha+1/2,i}\mathbf{u}^{n+1}_{\alpha+1/2,i}G_{\alpha+1/2,i} - \rho^{n+1}_{\alpha-1/2,i}\mathbf{u}^{n+1}_{\alpha-1/2,i}G_{\alpha-1/2,i}\Big)\Bigg)\,. \end{aligned} \tag{4.78}$$

Since the manipulations necessary to obtain the result are, to some extent tedious, we proceed as follows: first we consider the terms involving time derivatives then those involving the horizontal fluxes and finally, we consider the vertical exchange terms.

The discrete time derivatives The terms appearing in (4.78) reads

$$\mathcal{E}^t_{\alpha,i} := \big(gz_{\alpha,i} - \frac{|\mathbf{u}_{\alpha,i}|^2}{2}\big)\Big((\rho_{\alpha,i}h_{\alpha,i})^{n+1} - \rho_{\alpha,i}h_{\alpha,i}\Big) + \mathbf{u}_{\alpha,i}.\left((\rho_{\alpha,i}h_{\alpha,i}\mathbf{u}_{\alpha,i})^{n+1} - \rho_{\alpha,i}h_{\alpha,i}\mathbf{u}_{\alpha,i}\right),$$

and can be rewritten under the form

$$\begin{aligned}\mathcal{E}^t_{\alpha,i} = (\rho_{\alpha,i}h_{\alpha,i})^{n+1}\frac{|\mathbf{u}^{n+1}_{\alpha,i}|^2}{2} - (\rho_{\alpha,i}h_{\alpha,i})\frac{|\mathbf{u}_{\alpha,i}|^2}{2} + (\rho_{\alpha,i}gh_{\alpha,i}z_{\alpha,i})^{n+1} - (\rho_{\alpha,i}gh_{\alpha,i}z_{\alpha,i}) \\ - (\rho_{\alpha,i}gh_{\alpha,i})^{n+1}(z^{n+1}_{\alpha,i} - z_{\alpha,i}) - \frac{(\rho_{\alpha,i}h_{\alpha,i})^{n+1}}{2}|\mathbf{u}^{n+1}_{\alpha,i} - \mathbf{u}_{\alpha,i}|^2. \end{aligned} \tag{4.79}$$

The last term in (4.79) is classical in the context of explicit in time schemes and give rise to a non-negative term in the energy balance.

The horizontal fluxes The quantities we are now considering are

$$\mathcal{E}^{xy}_{\alpha,i,j} := \big(gz_{\alpha,i} - \frac{|\mathbf{u}_{\alpha,i}|^2}{2}\big)\rho_{\alpha,i,j}\int_{\mathbb{R}^2} M_{\alpha,i,j}\zeta_{i,j}d\xi d\gamma \\ + \mathbf{u}_{\alpha,i}.\left(\rho_{\alpha,i,j}\int_{\mathbb{R}^2}\begin{pmatrix}\xi\\ \gamma\end{pmatrix} M_{\alpha,i,j}\zeta_{i,j}d\xi d\gamma - \mathcal{S}_{p,\alpha,i,j}\right).$$

And using (4.72) we rewrite $\mathcal{E}^{xy}_{\alpha,i,j}$ under the form

$$\begin{aligned}\mathcal{E}^{xy}_{\alpha,i,j} \quad := \quad & \int_{\mathbb{R}^2}\left(gz_{\alpha,i,j} + \frac{\xi^2+\gamma^2}{2} - \frac{1}{2}\left|\begin{pmatrix}\xi\\ \gamma\end{pmatrix} - \tilde{\mathbf{u}}_{\alpha,i,j}\right|^2\right)\rho_{\alpha,i,j}M_{\alpha,i,j}\zeta_{i,j}d\xi d\gamma \\ & + \int_{\mathbb{R}^2}\frac{1}{2}\left(\left|\begin{pmatrix}\xi\\ \gamma\end{pmatrix} - \tilde{\mathbf{u}}_{\alpha,i,j}\right|^2 - \left|\begin{pmatrix}\xi\\ \gamma\end{pmatrix} - \mathbf{u}_{\alpha,i}\right|^2\right)\rho_{\alpha,i,j}M_{\alpha,i,j}\zeta_{i,j}d\xi d\gamma \\ & -p_{\alpha,i}(\hat{h}_{\alpha,i,j} - h_{\alpha,i})\mathbf{u}_{\alpha,i}.\mathbf{n}_{i,j} \\ & +g\int_{\mathbb{R}^2}\rho_{\alpha,i,j}M_{\alpha,i,j}\big(z_{\alpha,i,j} - z_{\alpha,i}\big)\big(\mathbf{u}_{\alpha,i}.\mathbf{n}_{i,j} - \zeta_{i,j}\big)d\xi d\gamma \qquad (4.80) \\ & +g\int_{\mathbb{R}^2}(\rho_{\alpha,i}h_{\alpha,i} - \rho_{\alpha,i,j}(z_{\alpha,i,j} - z_{\alpha,i})\mathbf{u}_{\alpha,i}\cdot\mathbf{n}_{i,j}d\xi d\gamma. \qquad (4.81)\end{aligned}$$

The last two lines of the previous equation reduce to

$$-\sum_{j\in K_i}\sigma_{i,j}p_{\alpha,i}\int_{\mathbb{R}^2}(M_{\alpha,i,j} - M_{\alpha,i})(\tilde{\mathbf{u}}_{\alpha,i,j} - \mathbf{u}_{\alpha,i})\cdot\mathbf{n}_{i,j}d\xi d\gamma.$$

Now, using the definition (4.77), we rewrite the second line of (4.81), denoted $\mathcal{E}^{2,xy}_{\alpha,i,j}$,

under the form

$$\begin{aligned}
\mathcal{E}^{2,xy}_{\alpha,i,j} &= -\frac{1}{2}(\tilde{\mathbf{u}}_{\alpha,i,j}-\mathbf{u}_{\alpha,i}).\mathbf{n}_{i,j}\int_{\mathbb{R}^2}\left|\begin{pmatrix}\xi\\ \gamma\end{pmatrix}-\mathbf{u}_{\alpha,i}\right|^2\rho_{\alpha,i}M_{\alpha,i}d\xi d\gamma\\
&\quad -\int_{\mathbb{R}^2}\frac{1}{2}\left|\begin{pmatrix}\xi\\ \gamma\end{pmatrix}-\mathbf{u}_{\alpha,i}\right|^2\rho_{\alpha,i,j}M_{\alpha,i,j}\zeta_{i,j}d\xi d\gamma\\
&\quad +\int_{\mathbb{R}^2}\frac{1}{2}\left|\begin{pmatrix}\xi\\ \gamma\end{pmatrix}-\tilde{\mathbf{u}}_{\alpha,i,j}\right|^2\rho_{\alpha,i,j}M_{\alpha,i,j}\zeta_{i,j}d\xi d\gamma\\
&\quad +\frac{1}{2}(\tilde{\mathbf{u}}_{\alpha,i,j}-\mathbf{u}_{\alpha,i}).\mathbf{n}_{i,j}\int_{\mathbb{R}^2}\left|\begin{pmatrix}\xi\\ \gamma\end{pmatrix}-\mathbf{u}_{\alpha,i}\right|^2\rho_{\alpha,i}M_{\alpha,i}d\xi d\gamma\\
&= -p_{\alpha,i}\int_{\mathbb{R}^2}M_{\alpha,i}(\tilde{\mathbf{u}}_{\alpha,i,j}-\mathbf{u}_{\alpha,i}).\mathbf{n}_{i,j}d\xi d\gamma\\
&\quad -(\tilde{\mathbf{u}}_{\alpha,i,j}-\mathbf{u}_{\alpha,i}).\int_{\mathbb{R}^2}\left(\begin{pmatrix}\xi\\ \gamma\end{pmatrix}-\frac{\tilde{\mathbf{u}}_{\alpha,i,j}+\mathbf{u}_{\alpha,i}}{2}\right)\rho_{\alpha,i,j}M_{\alpha,i,j}\zeta_{i,j}d\xi d\gamma\\
&\quad +(\tilde{\mathbf{u}}_{\alpha,i,j}-\mathbf{u}_{\alpha,i}).\mathbf{n}_{i,j}\int_{\mathbb{R}^2}\frac{1}{2}\left|\begin{pmatrix}\xi\\ \gamma\end{pmatrix}-\mathbf{u}_{\alpha,i}\right|^2\rho_{\alpha,i}M_{\alpha,i}d\xi d\gamma, \qquad (4.82)
\end{aligned}$$

where the definition (4.74) has been used. Let $\mathbf{n}_{i,j}=(n_{1,i,j},n_{2,i,j})^T$, because the Gibbs equilibrium $M_{\alpha,i}$ is an even function of the variables $\xi-u_{\alpha,i}$ and $\gamma-v_{\alpha,i}$, we get

$$\begin{aligned}
\frac{1}{2}\int_{\mathbb{R}^2}\left|\begin{pmatrix}\xi\\ \gamma\end{pmatrix}-\mathbf{u}_{\alpha,i}\right|^2\rho_{\alpha,i}M_{\alpha,i}d\xi d\gamma n_{1,i,j} &= \int_{\mathbb{R}^2}\frac{(\xi-u_{\alpha,i})^2+(\gamma-v_{\alpha,i})^2}{2}\rho_{\alpha,i}M_{\alpha,i}d\xi d\gamma n_{1,i,j}\\
&= \int_{\mathbb{R}^2}(\xi-u_{\alpha,i})^2\rho_{\alpha,i}M_{\alpha,i}d\xi d\gamma n_{1,i,j} = \int_{\mathbb{R}^2}(\xi-u_{\alpha,i})\xi n_{1,i,j}\rho_{\alpha,i}M_{\alpha,i}d\xi d\gamma\\
&= \int_{\mathbb{R}^2}(\xi-u_{\alpha,i})\zeta_{i,j}\rho_{\alpha,i}M_{\alpha,i}d\xi d\gamma, \qquad (4.83)
\end{aligned}$$

where we have used that since the function $(\xi-u_{\alpha,i})M_{\alpha,i}$ is even then

$$\int_{\mathbb{R}^2}(\xi-u_{\alpha,i})\gamma n_{2,i,j}\rho_{\alpha,i}M_{\alpha,i}d\xi d\gamma=0.$$

Likewise, we obtain

$$\frac{1}{2}\int_{\mathbb{R}^2}\left|\begin{pmatrix}\xi\\ \gamma\end{pmatrix}-\mathbf{u}_{\alpha,i}\right|^2\rho_{\alpha,i}M_{\alpha,i}d\xi d\gamma n_{2,i,j}=\int_{\mathbb{R}^2}(\gamma-v_{\alpha,i})\zeta_{i,j}\rho_{\alpha,i}M_{\alpha,i}d\xi d\gamma. \qquad (4.84)$$

And therefore, using Eqs. (4.83),(4.84), we can rewrite the last line of relation (4.82) under the form

$$(\tilde{\mathbf{u}}_{\alpha,i,j}-\mathbf{u}_{\alpha,i}).\int_{\mathbb{R}^2}\left(\begin{pmatrix}\xi\\ \gamma\end{pmatrix}-\mathbf{u}_{\alpha,i}\right)\rho_{\alpha,i}M_{\alpha,i}\zeta_{i,j}d\xi d\gamma,$$

leading to the following expression for $\mathcal{E}^{2,xy}_{\alpha,i,j}$

$$\begin{aligned}\mathcal{E}^{2,xy}_{\alpha,i,j} &= -p_{\alpha,i}\int_{\mathbb{R}^2} M_{\alpha,i}(\tilde{\mathbf{u}}_{\alpha,i,j}-\mathbf{u}_{\alpha,i}).\mathbf{n}_{i,j}d\xi d\gamma \\ &\quad -(\tilde{\mathbf{u}}_{\alpha,i,j}-\mathbf{u}_{\alpha,i}).\int_{\mathbb{R}^2}\left(\begin{pmatrix}\xi\\ \gamma\end{pmatrix}-\frac{\tilde{\mathbf{u}}_{\alpha,i,j}+\mathbf{u}_{\alpha,i}}{2}\right)\rho_{\alpha,i,j}M_{\alpha,i,j}\zeta_{i,j}d\xi d\gamma \\ &\quad +(\tilde{\mathbf{u}}_{\alpha,i,j}-\mathbf{u}_{\alpha,i}).\int_{\mathbb{R}^2}\left(\begin{pmatrix}\xi\\ \gamma\end{pmatrix}-\mathbf{u}_{\alpha,i}\right)\rho_{\alpha,i}M_{\alpha,i}\zeta_{i,j}d\xi d\gamma.\end{aligned}$$

We rewrite the second line of $\mathcal{E}^{2,xy}_{\alpha,i,j}$ under the form

$$\frac{|\tilde{\mathbf{u}}_{\alpha,i,j}-\mathbf{u}_{\alpha,i}|^2}{2}\int_{\mathbb{R}^2}\rho_{\alpha,i,j}M_{\alpha,i,j}\zeta_{i,j}d\xi d\gamma-(\tilde{\mathbf{u}}_{\alpha,i,j}-\mathbf{u}_{\alpha,i}).\int_{\mathbb{R}^2}\left(\begin{pmatrix}\xi\\ \gamma\end{pmatrix}-\mathbf{u}_{\alpha,i}\right)\rho_{\alpha,i,j}M_{\alpha,i,j}\zeta_{i,j}d\xi d\gamma,$$

or equivalently

$$\begin{aligned}&\frac{|\tilde{\mathbf{u}}_{\alpha,i,j}-\mathbf{u}_{\alpha,i}|^2}{2}\int_{\mathbb{R}^2}\rho_{\alpha,i,j}M_{\alpha,i,j}\zeta_{i,j}d\xi d\gamma-(\tilde{\mathbf{u}}_{\alpha,i,j}-\mathbf{u}_{\alpha,i}).\int_{\mathbb{R}^2}\left(\begin{pmatrix}\xi\\ \gamma\end{pmatrix}-\mathbf{u}_{\alpha,i,j}\right)\rho_{\alpha,i,j}M_{\alpha,i,j}\zeta_{i,j}d\xi d\gamma \\ &\qquad\qquad +(\tilde{\mathbf{u}}_{\alpha,i,j}-\mathbf{u}_{\alpha,i}).\int_{\mathbb{R}^2}(\mathbf{u}_{\alpha,i}-\mathbf{u}_{\alpha,i,j})\rho_{\alpha,i,j}M_{\alpha,i,j}\zeta_{i,j}d\xi d\gamma.\end{aligned}$$

Using the previous identities and considering the cases $\zeta_{i,j}\geqslant 0$ and $\zeta_{i,j}\leqslant 0$, simple computations give the following expression for $\mathcal{E}^{2,xy}_{\alpha,i,j}$

$$\begin{aligned}\mathcal{E}^{2,xy}_{\alpha,i,j} &= -p_{\alpha,i}\int_{\mathbb{R}^2} M_{\alpha,i}(\tilde{\mathbf{u}}_{\alpha,i,j}-\mathbf{u}_{\alpha,i}).\mathbf{n}_{i,j}d\xi d\gamma \\ &\quad +\frac{|\tilde{\mathbf{u}}_{\alpha,i,j}-\mathbf{u}_{\alpha,i}|^2}{2}\int_{\{\zeta_{i,j}\geqslant 0\}}\rho_{\alpha,i}M_{\alpha,i}\zeta_{i,j}d\xi d\gamma \\ &\quad +(\tilde{\mathbf{u}}_{\alpha,i,j}-\mathbf{u}_{\alpha,i}).\int_{\{\zeta_{i,j}\leqslant 0\}}\left[\left(\begin{pmatrix}\xi\\ \gamma\end{pmatrix}-\mathbf{u}_{\alpha,i}\right)\rho_{\alpha,i}M_{\alpha,i}-\left(\begin{pmatrix}\xi\\ \gamma\end{pmatrix}-\mathbf{u}_{\alpha,j}\right)\rho_{\alpha,j}M_{\alpha,j}\right]\zeta_{i,j}d\xi d\gamma \\ &\quad +\frac{1}{2}(\tilde{\mathbf{u}}_{\alpha,i,j}-\mathbf{u}_{\alpha,i}).\int_{\{\zeta_{i,j}\leqslant 0\}}\big(\tilde{\mathbf{u}}_{\alpha,i,j}-\mathbf{u}_{\alpha,i}-2(\mathbf{u}_{\alpha,j}-\mathbf{u}_{\alpha,i})\big)\rho_{\alpha,j}M_{\alpha,j}\zeta_{i,j}d\xi d\gamma,\end{aligned}$$

and the last line of the previous expression can be written under the form

$$\begin{aligned}&-\frac{1}{2}|\tilde{\mathbf{u}}_{\alpha,i,j}-\mathbf{u}_{\alpha,i}|^2\int_{\{\zeta_{i,j}\leqslant 0\}}\rho_{\alpha,j}M_{\alpha,j}\zeta_{i,j}d\xi d\gamma \\ &\qquad\qquad +(\tilde{\mathbf{u}}_{\alpha,i,j}-\mathbf{u}_{\alpha,i}).(\tilde{\mathbf{u}}_{\alpha,i,j}-\mathbf{u}_{\alpha,j})\int_{\{\zeta_{i,j}\leqslant 0\}}\rho_{\alpha,j}M_{\alpha,j}\zeta_{i,j}d\xi d\gamma,\end{aligned}$$

or equivalently

$$-\frac{1}{2}|\tilde{\mathbf{u}}_{\alpha,i,j}-\mathbf{u}_{\alpha,i}|^2\int_{\{\zeta_{i,j}\leqslant 0\}}\rho_{\alpha,j}M_{\alpha,j}\zeta_{i,j}d\xi d\gamma$$
$$-\frac{|\mathbf{u}_{\alpha,i}-\mathbf{u}_{\alpha,j}|^2}{4}\int_{\{\zeta_{i,j}\leqslant 0\}}\rho_{\alpha,j}M_{\alpha,j}\zeta_{i,j}d\xi d\gamma+\frac{|2\tilde{\mathbf{u}}_{\alpha,i,j}-\mathbf{u}_{\alpha,i}-\mathbf{u}_{\alpha,j}|^2}{4}\int_{\{\zeta_{i,j}\leqslant 0\}}\rho_{\alpha,j}M_{\alpha,j}\zeta_{i,j}d\xi d\gamma,$$

where the identity $ab=(a+b)^2/4-(a-b)^2/4$ has been used. Hence, the final expression for $\mathcal{E}^{2,xy}_{\alpha,i,j}$ is given by

$$\begin{aligned}\mathcal{E}^{2,xy}_{\alpha,i,j} &= -p_{\alpha,i}\int_{\mathbb{R}^2}M_{\alpha,i}(\tilde{\mathbf{u}}_{\alpha,i,j}-\mathbf{u}_{\alpha,i}).\mathbf{n}_{i,j}d\xi d\gamma\\
&+\frac{|\tilde{\mathbf{u}}_{\alpha,i,j}-\mathbf{u}_{\alpha,i}|^2}{2}\int_{\mathbb{R}^2}\rho_{\alpha,i}M_{\alpha,i}|\zeta_{i,j}|d\xi d\gamma-\frac{|\mathbf{u}_{\alpha,i}-\mathbf{u}_{\alpha,j}|^2}{4}\int_{\{\zeta_{i,j}\leqslant 0\}}\rho_{\alpha,j}M_{\alpha,j}\zeta_{i,j}d\xi d\gamma\\
&+(\tilde{\mathbf{u}}_{\alpha,i,j}-\mathbf{u}_{\alpha,i}).\int_{\{\zeta_{i,j}\leqslant 0\}}\left[\left(\begin{pmatrix}\xi\\ \gamma\end{pmatrix}-\mathbf{u}_{\alpha,i}\right)\rho_{\alpha,i}M_{\alpha,i}-\left(\begin{pmatrix}\xi\\ \gamma\end{pmatrix}-\mathbf{u}_{\alpha,j}\right)\rho_{\alpha,j}M_{\alpha,j}\right]\zeta_{i,j}d\xi d\gamma\\
&+\frac{|2\tilde{\mathbf{u}}_{\alpha,i,j}-\mathbf{u}_{\alpha,i}-\mathbf{u}_{\alpha,j}|^2}{4}\int_{\{\zeta_{i,j}\leqslant 0\}}\rho_{\alpha,j}M_{\alpha,j}\zeta_{i,j}d\xi d\gamma,\end{aligned}$$

where in the previous expression, the second line is nonnegative and the last two lines are third order terms (i.e. of order $\mathcal{O}(\mathrm{d}iam(C_i)^3)$) when considering Lipschitz continuous solutions.

Then for the third line of (4.81), performing simple manipulations we have

$$\begin{aligned}\mathcal{P}^{xy}_{\alpha,i,j} &= -p_{\alpha,i}(\hat{h}_{\alpha,i,j}-h_{\alpha,i})\mathbf{u}_{\alpha,i}.\mathbf{n}_{i,j}\\
&= -p_{\alpha,i}\int_{\mathbb{R}^2}(M_{\alpha,i,j}-M_{\alpha,i})d\xi d\gamma\mathbf{u}_{\alpha,i}.\mathbf{n}_{i,j}\\
&= -p_{\alpha,i}\int_{\mathbb{R}^2}(M_{\alpha,i,j}\tilde{\mathbf{u}}_{\alpha,i,j}.\mathbf{n}_{i,j}-M_{\alpha,i}\mathbf{u}_{\alpha,i}.\mathbf{n}_{i,j})d\xi d\gamma\\
&\quad +p_{\alpha,i}\int_{\mathbb{R}^2}M_{\alpha,i,j}(\tilde{\mathbf{u}}_{\alpha,i,j}-\mathbf{u}_{\alpha,i}).\mathbf{n}_{i,j}d\xi d\gamma\\
&= -p_{\alpha,i}\int_{\mathbb{R}^2}(M_{\alpha,i,j}\zeta_{i,j}-M_{\alpha,i}\mathbf{u}_{\alpha,i}.\mathbf{n}_{i,j})d\xi d\gamma\\
&\quad +p_{\alpha,i}\int_{\mathbb{R}^2}M_{\alpha,i,j}(\tilde{\mathbf{u}}_{\alpha,i,j}-\mathbf{u}_{\alpha,i}).\mathbf{n}_{i,j}d\xi d\gamma\\
&= -p_{\alpha,i}\int_{\mathbb{R}^2}(M_{\alpha,i,j}\zeta_{i,j}-M_{\alpha,i}\mathbf{u}_{\alpha,i}.\mathbf{n}_{i,j})d\xi d\gamma\\
&\quad +p_{\alpha,i}\int_{\mathbb{R}^2}M_{\alpha,i}(\tilde{\mathbf{u}}_{\alpha,i,j}-\mathbf{u}_{\alpha,i}).\mathbf{n}_{i,j}d\xi d\gamma\\
&\quad +p_{\alpha,i}\int_{\mathbb{R}^2}(M_{\alpha,i,j}-M_{\alpha,i})(\tilde{\mathbf{u}}_{\alpha,i,j}-\mathbf{u}_{\alpha,i}).\mathbf{n}_{i,j}d\xi d\gamma\end{aligned}\tag{4.85}$$

where the definitions (4.77),(4.73) have been used. Notice that for the second term in the first line of (4.85), we have

$$\sum_{j\in K_i} p_{\alpha,i}\int_{\mathbb{R}^2} M_{\alpha,i}\mathbf{u}_{\alpha,i}.\mathbf{n}_{i,j}d\xi d\gamma = p_{\alpha,i}\int_{\mathbb{R}^2} M_{\alpha,i}\mathbf{u}_{\alpha,i}d\xi d\gamma.\sum_{j\in K_i}\mathbf{n}_{i,j} = 0,$$

and using the discrete form of the continuity equation, for the first term in the first line of (4.85) we get

$$-\sum_{j\in K_i}\sigma_{i,j}p_{\alpha,i}\int_{\mathbb{R}^2} M_{\alpha,i,j}\zeta_{i,j}d\xi d\gamma = p_{\alpha,i}\big(h_{\alpha,i}^{n+1} - h_{\alpha,i} - \Delta t^n(G_{\alpha+1/2,i} - G_{\alpha-1/2,i})\big).$$

The vertical exchange terms It remains to examine the contribution of the vertical exchange terms over the energy balance, namely in Eq. (4.78) the quantity

$$\begin{aligned}\mathcal{V}_{\alpha,i} \;:=\; & \Delta t^n\big(gz_{\alpha,i} - \frac{|\mathbf{u}_{\alpha,i}|^2}{2}\big)\Big(\rho^{n+1}_{\alpha+1/2,i}G_{\alpha+1/2,i} - \rho^{n+1}_{\alpha-1/2,i}G_{\alpha-1/2,i}\Big)\\ & +\Delta t^n\mathbf{u}_{\alpha,i}.\Big(\rho^{n+1}_{\alpha+1/2,i}\mathbf{u}^{n+1}_{\alpha+1/2,i}G_{\alpha+1/2,i} - \rho^{n+1}_{\alpha-1/2,i}\mathbf{u}^{n+1}_{\alpha-1/2,i}G_{\alpha-1/2,i}\Big).\end{aligned}$$

And we write

$$\begin{aligned}\mathcal{V}_{\alpha,i} \;=\; & \Delta t^n\rho^{n+1}_{\alpha+1/2,i}\frac{|\mathbf{u}^{n+1}_{\alpha+1/2,i}|^2}{2}G_{\alpha+1/2,i} - \Delta t^n\rho^{n+1}_{\alpha-1/2,i}\frac{|\mathbf{u}^{n+1}_{\alpha-1/2,i}|^2}{2}G_{\alpha-1/2,i}\\ & +\Delta t^n g\Big(\rho^{n+1}_{\alpha+1/2,i}z_{\alpha+1/2,i}G_{\alpha+1/2,i} - \rho^{n+1}_{\alpha-1/2,i}z_{\alpha-1/2,i}G_{\alpha-1/2,i}\Big)\\ & -\Delta t^n\rho^{n+1}_{\alpha+1/2,i}\left(\frac{|\mathbf{u}^{n+1}_{\alpha+1/2,i} - \mathbf{u}^{n+1}_{\alpha,i}|^2}{2} + \frac{\mathbf{u}_{\alpha,i} - \mathbf{u}^{n+1}_{\alpha,i}}{2}.(\mathbf{u}^{n+1}_{\alpha,i} - 2\mathbf{u}^{n+1}_{\alpha+1/2,i} + \mathbf{u}_{\alpha,i})\right)G_{\alpha+1/2,i}\\ & +\Delta t^n\rho^{n+1}_{\alpha-1/2,i}\left(\frac{|\mathbf{u}^{n+1}_{\alpha-1/2,i} - \mathbf{u}^{n+1}_{\alpha,i}|^2}{2} + \frac{\mathbf{u}_{\alpha,i} - \mathbf{u}^{n+1}_{\alpha,i}}{2}.(\mathbf{u}^{n+1}_{\alpha,i} - 2\mathbf{u}^{n+1}_{\alpha-1/2,i} + \mathbf{u}_{\alpha,i})\right)G_{\alpha-1/2,i}\\ & -\Delta t^n g\frac{h_{\alpha,i}}{2}\Big(\rho^{n+1}_{\alpha+1/2,i}G_{\alpha+1/2,i} + \rho^{n+1}_{\alpha-1/2,i}G_{\alpha-1/2,i}\Big).\end{aligned} \tag{4.86}$$

Remark 8. When $G_{\alpha+1/2,i}\leqslant 0$, then the third line of (4.86) is nonnegative namely,

$$-\Delta t^n\rho^{n+1}_{\alpha+1/2,i}\frac{|\mathbf{u}_{\alpha,i} - \mathbf{u}^{n+1}_{\alpha,i}|^2}{2}G_{\alpha+1/2,i} = \mathcal{O}\big((\Delta t^n)^3\big),$$

whereas for $G_{\alpha+1/2,i} > 0$

$$\begin{aligned}\frac{|\mathbf{u}^{n+1}_{\alpha+1/2,i} - \mathbf{u}^{n+1}_{\alpha,i}|^2}{2} - \frac{\mathbf{u}_{\alpha,i} - \mathbf{u}^{n+1}_{\alpha,i}}{2}.(\mathbf{u}^{n+1}_{\alpha,i} - 2\mathbf{u}^{n+1}_{\alpha+1/2,i} + \mathbf{u}_{\alpha,i})\\ = \frac{\mathbf{u}^{n+1}_{\alpha+1,i} - \mathbf{u}^{n+1}_{\alpha,i}}{2}(\mathbf{u}^{n+1}_{\alpha+1,i} - 2\mathbf{u}_{\alpha,i} + \mathbf{u}^{n+1}_{\alpha,i}) = \mathcal{O}\big((\Delta t^n)^2\big).\end{aligned}$$

Hence, when it is not a dissipative term, the third line of (4.86) is a $\mathcal{O}\big((\Delta t^n)^3\big)$ term. The same result holds for the fourth line of (4.86).

All the contributions

Now, summarizing the computations carried out in the previous paragraphs, we sum all the contributions, namely $\mathcal{E}^t_{\alpha,i}$, $\mathcal{E}^{xy}_{\alpha,i,j}$ and $\mathcal{V}_{\alpha,i}$ leading to a new expression for $\mathcal{E}_{\alpha,i}$ under the form

$$E^{n+1}_{\alpha,i} - E_{\alpha,i} + \sum_{j\in K_i} \sigma_{i,j} \int_{\mathbb{R}^2} \left(g z_{\alpha,i,j} + \frac{\xi^2+\gamma^2}{2} - \frac{1}{2}\left| \begin{pmatrix} \xi \\ \gamma \end{pmatrix} - \tilde{\mathbf{u}}_{\alpha,i,j} \right|^2 \right) \rho_{\alpha,i,j} M_{\alpha,i,j} \zeta_{i,j} d\xi d\gamma$$
$$-\Delta t^n \rho^{n+1}_{\alpha+1/2,i} \left(\frac{|\mathbf{u}^{n+1}_{\alpha+1/2,i}|^2}{2} + g z_{\alpha+1/2,i} \right) G_{\alpha+1/2,i} + \Delta t^n \rho^{n+1}_{\alpha-1/2,i} \left(\frac{|\mathbf{u}^{n+1}_{\alpha-1/2,i}|^2}{2} + g z_{\alpha-1/2,i} \right) G_{\alpha-1/2,i}$$
$$= d_{\alpha,i} + e_{\alpha,i} + F_{\alpha,i}$$

where $d_{\alpha,i}$ are non-positive and hence dissipative terms whereas $e_{\alpha,i}$ are errors terms. $d_{\alpha,i}$ and $e_{\alpha,i}$ are given by (4.75) and (4.76) respectively. The quantity $F_{\alpha,i}$ is defined by

$$\begin{aligned}
F_{\alpha,i} &= (\rho_{\alpha,i} g h_{\alpha,i})^{n+1} (z^{n+1}_{\alpha,i} - z_{\alpha,i}) - p_{\alpha,i}\big(h^{n+1}_{\alpha,i} - h_{\alpha,i} - \Delta t^n (G_{\alpha+1/2,i} - G_{\alpha-1/2,i})\big) \\
&\quad -\Delta t^n \rho^{n+1}_{\alpha+1/2,i} g \frac{h_{\alpha,i}}{2} G_{\alpha+1/2,i} - \Delta t^n \rho^{n+1}_{\alpha-1/2,i} g \frac{h_{\alpha,i}}{2} G_{\alpha-1/2,i} + \frac{(\rho_{\alpha,i} h_{\alpha,i})^{n+1}}{2} |\mathbf{u}^{n+1}_{\alpha,i} - \mathbf{u}_{\alpha,i}|^2 \\
&\quad -\Delta t^n \rho^{n+1}_{\alpha+1/2,i} \left(\frac{|\mathbf{u}^{n+1}_{\alpha+1/2,i} - \mathbf{u}^{n+1}_{\alpha,i}|^2}{2} + \frac{\mathbf{u}_{\alpha,i} - \mathbf{u}^{n+1}_{\alpha,i}}{2} . (\mathbf{u}^{n+1}_{\alpha,i} - 2\mathbf{u}^{n+1}_{\alpha+1/2,i} + \mathbf{u}_{\alpha,i}) \right) G_{\alpha+1/2,i} \\
&\quad +\Delta t^n \rho^{n+1}_{\alpha-1/2,i} \left(\frac{|\mathbf{u}^{n+1}_{\alpha-1/2,i} - \mathbf{u}^{n+1}_{\alpha,i}|^2}{2} + \frac{\mathbf{u}_{\alpha,i} - \mathbf{u}^{n+1}_{\alpha,i}}{2} . (\mathbf{u}^{n+1}_{\alpha,i} - 2\mathbf{u}^{n+1}_{\alpha-1/2,i} + \mathbf{u}_{\alpha,i}) \right) G_{\alpha-1/2,i},
\end{aligned}$$

and the first three lines of $F_{\alpha,i}$ can be rewritten under the form

$$\begin{aligned}
F^1_{\alpha,i} &= (\rho_{\alpha,i} g h_{\alpha,i})^{n+1} (z^{n+1}_{\alpha,i} - z_{\alpha,i}) - p_{\alpha,i}\big(h^{n+1}_{\alpha,i} - h_{\alpha,i}\big) + \frac{(\rho_{\alpha,i} h_{\alpha,i})^{n+1}}{2} |\mathbf{u}^{n+1}_{\alpha,i} - \mathbf{u}_{\alpha,i}|^2 \\
&\quad +\Delta t^n \Big(p^{n*}_{\alpha+1/2,i} G_{\alpha+1/2,i} - p^{n*}_{\alpha-1/2,i} G_{\alpha-1/2,i} \Big) \\
&\quad +\Delta t^n \frac{g h_{\alpha,i}}{2} \Big((\rho^{n+1}_{\alpha,i} - \rho^{n+1}_{\alpha+1/2,i}) G_{\alpha+1/2,i} + (\rho^{n+1}_{\alpha,i} - \rho^{n+1}_{\alpha-1/2,i}) G_{\alpha-1/2,i} \Big) . \qquad (4.87)
\end{aligned}$$

Now, we rewrite the first two terms of the first line of (4.87) under the form

$$\begin{aligned}
F^2_{\alpha,i} &= \frac{(\rho_{\alpha,i} g h_{\alpha,i})^{n+1}}{2} (z^{n+1}_{\alpha+1/2,i} + z^{n+1}_{\alpha-1/2,i} - z_{\alpha+1/2,i} - z_{\alpha-1/2,i}) \\
&\quad -p_{\alpha,i}\big(z^{n+1}_{\alpha+1/2,i} - z^{n+1}_{\alpha-1/2,i} - z_{\alpha+1/2,i} + z_{\alpha-1/2,i}\big) \\
&= -p^{n*}_{\alpha+1/2,i}(z^{n+1}_{\alpha+1/2,i} - z_{\alpha+1/2,i}) + p^{n*}_{\alpha-1/2,i}(z^{n+1}_{\alpha-1/2,i} - z_{\alpha-1/2,i}) \\
&\quad +g \Big((\rho_{\alpha,i} h_{\alpha,i})^{n+1} - \rho^{n+1}_{\alpha,i} h_{\alpha,i} \Big) (z^{n+1}_{\alpha,i} - z_{\alpha,i}).
\end{aligned}$$

Hence, we have for $F_{\alpha,i}$

$$
\begin{aligned}
F_{\alpha,i} \;=\;& p^{n*}_{\alpha+1/2,i}\big(\Delta t^n G_{\alpha+1/2,i} - z^{n+1}_{\alpha+1/2,i} + z_{\alpha+1/2,i}\big) \\
& - p^{n*}_{\alpha-1/2,i}\big(\Delta t^n G_{\alpha-1/2,i} - z^{n+1}_{\alpha-1/2,i} + z_{\alpha-1/2,i}\big)\Big) \\
& - \Delta t^n \rho^{n+1}_{\alpha+1/2,i}\left(\frac{|\mathbf{u}^{n+1}_{\alpha+1/2,i} - \mathbf{u}^{n+1}_{\alpha,i}|^2}{2} + \frac{\mathbf{u}_{\alpha,i} - \mathbf{u}^{n+1}_{\alpha,i}}{2}.(\mathbf{u}^{n+1}_{\alpha,i} - 2\mathbf{u}^{n+1}_{\alpha+1/2,i} + \mathbf{u}_{\alpha,i})\right) G_{\alpha+1/2,i} \\
& + \Delta t^n \rho^{n+1}_{\alpha-1/2,i}\left(\frac{|\mathbf{u}^{n+1}_{\alpha-1/2,i} - \mathbf{u}^{n+1}_{\alpha,i}|^2}{2} + \frac{\mathbf{u}_{\alpha,i} - \mathbf{u}^{n+1}_{\alpha,i}}{2}.(\mathbf{u}^{n+1}_{\alpha,i} - 2\mathbf{u}^{n+1}_{\alpha-1/2,i} + \mathbf{u}_{\alpha,i})\right) G_{\alpha-1/2,i} \\
& + \Delta t^n \frac{g h_{\alpha,i}}{2}\Big((\rho^{n+1}_{\alpha,i} - \rho^{n+1}_{\alpha+1/2,i})G_{\alpha+1/2,i} + (\rho^{n+1}_{\alpha,i} - \rho^{n+1}_{\alpha-1/2,i})G_{\alpha-1/2,i}\Big) \\
& + g\Big((\rho_{\alpha,i}h_{\alpha,i})^{n+1} - \rho^{n+1}_{\alpha,i}h_{\alpha,i}\Big)(z^{n+1}_{\alpha,i} - z_{\alpha,i}) + \frac{(\rho_{\alpha,i}h_{\alpha,i})^{n+1}}{2}|\mathbf{u}^{n+1}_{\alpha,i} - \mathbf{u}_{\alpha,i}|^2,
\end{aligned}
$$

the first two lines being conservatives terms and the four following ones being error terms corresponding to $f_{\alpha,i}$. The fifth line is the discrete version of the term appearing in the last line of the continuous energy balance given in Eq. (4.23) (see also [30]) and the last line of the previous relation comes from the time scheme and is of order $\mathcal{O}\big((\Delta t^n)^2\big)$.

In order to conclude the proof, it remains to prove that all the quantities appearing in $e_{\alpha,i}$ are third order terms i.e. of magnitude $\mathcal{O}(\mathrm{d}iam(C_i)^3)$. The terms in each of the sums in $e_{\alpha,i}$ are obviously second-order terms. Since the sum is made on the faces, gradients of second-order terms appear. These gradients are indeed third-order terms. □

If one wishes to introduce even more upwinding in the discretization of $S_{p,\alpha}$ by using the discretization

$$
\mathcal{S}_{p,\alpha,i,j} = p_{\alpha,i}(\hat{h}_{\alpha,i,j} - h_{\alpha,i})\mathbf{n}_{i,j} - g\rho_{\alpha,i,j}\hat{h}_{\alpha,i,j}(z_{\alpha,i,j} - z_{\alpha,i})\mathbf{n}_{i,j},
$$

Proposition 5 still holds but the expression of the rest term $e_{\alpha,i}$ becomes

$$
\begin{aligned}
e_{\alpha,i} \;=\;& \sum_{j\in K_i} \sigma_{i,j} g \int_{\mathbb{R}^2} \rho_{\alpha,i,j} M_{\alpha,i,j}\big(z_{\alpha,i,j} - z_{\alpha,i}\big)\big(\mathbf{u}_{\alpha,i} - \mathbf{u}_{\alpha,i,j}\big)\cdot\mathbf{n}_{i,j}\, d\xi d\gamma \\
& + \sum_{j\in K_i} \sigma_{i,j} g \int_{\mathbb{R}^2} \rho_{\alpha,i,j} M_{\alpha,i,j}\big(z_{\alpha,i,j} - z_{\alpha,i}\big)\big(\mathbf{u}_{\alpha,i,j}\cdot\mathbf{n}_{i,j} - \zeta_{i,j}\big) d\xi d\gamma \\
& - \sum_{j\in K_i} \sigma_{i,j} p_{\alpha,i} \int_{\mathbb{R}^2} (M_{\alpha,i,j} - M_{\alpha,i})(\tilde{\mathbf{u}}_{\alpha,i,j} - \mathbf{u}_{\alpha,i})\cdot\mathbf{n}_{i,j}\, d\xi d\gamma \\
& - \sum_{j\in K_i} \sigma_{i,j}(\tilde{\mathbf{u}}_{\alpha,i,j} - \mathbf{u}_{\alpha,i})\cdot\int_{\{\zeta_{i,j}\leqslant 0\}} \left[\left(\begin{pmatrix}\xi\\ \gamma\end{pmatrix} - \mathbf{u}_{\alpha,i}\right)\rho_{\alpha,i}M_{\alpha,i} - \left(\begin{pmatrix}\xi\\ \gamma\end{pmatrix} - \mathbf{u}_{\alpha,j}\right)\rho_{\alpha,j}M_{\alpha,j}\right]\zeta_{i,j}\, d\xi d\gamma \\
& - \sum_{j\in K_i} \sigma_{i,j}\frac{|2\tilde{\mathbf{u}}_{\alpha,i,j} - \mathbf{u}_{\alpha,i} - \mathbf{u}_{\alpha,j}|^2}{4}\int_{\{\zeta_{i,j}\leqslant 0\}} \rho_{\alpha,j}M_{\alpha,j}\zeta_{i,j}\, d\xi d\gamma.
\end{aligned}
$$

This new expression of $e_{\alpha,i}$ also contains only third-order error terms, thus the result is not deteriorated.

4.4 Numerical scheme for the layer-averaged Navier-Stokes-Fourier system

The discretization of the full layer-averaged Navier-Stokes-Fourier is presented. The main difficulties have already been tackled in section 4.3.

4.4.1 Semi-discrete (in time) scheme

The semi-discrete in time scheme (4.30) yields the following system

$$h^{n+1} = h^{n+1/2} = h - \Delta t^n \sum_{\alpha=1}^{N} \nabla_{x,y} \cdot (h_\alpha \mathbf{u}_\alpha) - \sum_{\alpha=1}^{N} \frac{\rho'(T_\alpha)}{\rho_\alpha^2 c_p} (S_{T,\alpha} - S_{\mu,\alpha}), \tag{4.88}$$

$$(\rho_\alpha h_\alpha)^{n+1/2} = \rho_\alpha h_\alpha - \Delta t^n \nabla_{x,y} \cdot (\rho_\alpha h_\alpha \mathbf{u}_\alpha) \tag{4.89}$$

$$\begin{aligned}(\rho_\alpha h_\alpha \mathbf{u}_\alpha)^{n+1/2} = {} & \rho_\alpha h_\alpha \mathbf{u}_\alpha - \Delta t^n \Big(\nabla_{x,y} \cdot (\rho_\alpha h_\alpha \mathbf{u}_\alpha \otimes \mathbf{u}_\alpha) + \nabla_{x,y} \big(h_\alpha p_\alpha \big) \\ & - p_\alpha \nabla_{x,y} h_\alpha + \rho_\alpha g h_\alpha \nabla_{x,y} z_\alpha \Big),\end{aligned} \tag{4.90}$$

$$(\rho_\alpha h_\alpha)^{n+1} = (\rho_\alpha h_\alpha)^{n+1/2} - \Delta t^n \Big(\rho^{n+1}_{\alpha+1/2} G_{\alpha+1/2} - \rho^{n+1}_{\alpha-1/2} G_{\alpha-1/2} \Big), \tag{4.91}$$

$$\begin{aligned}(\rho_\alpha h_\alpha \mathbf{u}_\alpha)^{n+1} = {} & (\rho_\alpha h_\alpha \mathbf{u}_\alpha)^{n+1/2} - \Delta t^n \Big(\mathbf{u}^{n+1}_{\alpha+1/2} \rho^{n+1}_{\alpha+1/2} G_{\alpha+1/2} - \mathbf{u}^{n+1}_{\alpha-1/2} \rho^{n+1}_{\alpha-1/2} G_{\alpha-1/2} \\ & + \nabla_{x,y} \cdot (\mu h^{n+1}_\alpha \nabla_{x,y} \mathbf{u}^{n+1}_\alpha) + \Gamma^{n+1}_{\alpha+1/2} (\mathbf{u}^{n+1}_{\alpha+1} - \mathbf{u}^{n+1}_\alpha) \\ & - \Gamma^{n+1}_{\alpha-1/2} (\mathbf{u}^{n+1}_\alpha - \mathbf{u}^{n+1}_{\alpha-1}) - \kappa_\alpha \mathbf{u}_\alpha \Big),\end{aligned} \tag{4.92}$$

with

$$G_{\alpha+1/2} = -\sum_{j=1}^{N} \Big(\sum_{p=1}^{\alpha} l_p - \mathbb{1}_{j \leqslant \alpha} \Big) \nabla_{x,y} \cdot (h_j \mathbf{u}_j) + \sum_{j=1}^{\alpha} \frac{\rho'(T_j)}{\rho_j^2 c_p} (S_{T,j} - S_{\mu,j}). \tag{4.93}$$

Note that the definition of the mass exchange terms $G_{\alpha+1/2}$ is different from the definition of the mass exchange terms for the Euler-Fourier system given in (4.38).

4.4.2 Spatial discretization of the diffusion terms

The Euler part of the system is discretized in space as in section 4.3.3. We present here only the discretization of the diffusion terms. For the discretization of the viscosity terms in the momentum equation, we refer to [6]. A classical $\mathbb{P}_1$ finite element type approximation with mass-lumping is used. Let us define the number of cells N_x as well

as the vector of unknowns in layer α

$$\mathbf{U}_\alpha = (h_{\alpha,1},\dots,h_{\alpha,N_x},(\rho_\alpha h_\alpha)_1,\dots,(\rho_\alpha h_\alpha)_{Nx}, \\ (\rho_\alpha h_\alpha u_\alpha)_1,\dots,(\rho_\alpha h_\alpha u_\alpha)_{Nx},(\rho_\alpha h_\alpha v_\alpha)_1,\dots,(\rho_\alpha h_\alpha v_\alpha)_{Nx})^T$$

and the vector containing the temperatures in layer α

$$\mathbf{T}_\alpha = (T_{\alpha,1},\dots,T_{\alpha,N_x})^T.$$

The discretization of $\nabla_{x,y}\cdot(\mu h_\alpha^{n+1}\nabla_{x,y}\mathbf{u}_\alpha^{n+1})+\Gamma_{\alpha+1/2}^{n+1}(\mathbf{u}_{\alpha+1}^{n+1}-\mathbf{u}_\alpha^{n+1})-\Gamma_{\alpha-1/2}^{n+1}(\mathbf{u}_\alpha^{n+1}-\mathbf{u}_{\alpha-1}^{n+1})$ reads

$$-\Delta t^n \mathcal{K}_{\mu,\alpha}\mathbf{U}_\alpha + \Delta t^n \mathcal{M}_{\mu,\alpha+1/2}(\mathbf{U}_{\alpha+1}-\mathbf{U}_\alpha) - \Delta t^n \mathcal{M}_{\mu,\alpha+1/2}(\mathbf{U}_\alpha - \mathbf{U}_{\alpha-1}),$$

with the $4N_x \times 4N_x$ block matrices

$$\mathcal{K}_{\mu,\alpha} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \mathcal{K}'_{\mu,\alpha} & 0 \\ 0 & 0 & 0 & \mathcal{K}'_{\mu,\alpha} \end{pmatrix}, \quad \mathcal{M}_{\mu,\alpha+1/2} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \mathcal{M}'_{\mu,\alpha+1/2} & 0 \\ 0 & 0 & 0 & \mathcal{M}'_{\mu,\alpha+1/2} \end{pmatrix}$$

where the non-zero coefficients are given by

$$(\mathcal{K}'_{\mu,\alpha})_{j,i} = \frac{3}{A_j}\frac{\mu}{(\rho_\alpha h_\alpha)_i}\int_\Omega h_\alpha \nabla_{x,y}\varphi_i \cdot \nabla_{x,y}\varphi_j dxdy,$$
$$(\mathcal{M}'_{\mu,\alpha+1/2})_{j,i} = \frac{\mu}{(\rho_\alpha h_\alpha)_i}\frac{\delta_{i,j}}{h_{\alpha+1,i}+h_{\alpha,i}}.$$

The φ_i are the basis functions and $\delta_{i,j}$ is the Kronecker symbol. The area A_j is the area of the support of the test function φ_j.
The discretization of the terms $S_{T,\alpha}$ due to temperature diffusion is similar. To discretize $\frac{\rho'(T_\alpha)}{\rho_\alpha^2 c_p}S_{T,\alpha}$, we propose the simplification

$$\frac{\rho'(T_{\alpha,i})}{\rho_{\alpha,i}^2 c_p}(S_{T,\alpha})_i.$$

The term $S_{\mu,\alpha}$ ensures that the energy of the layer-averaged system is consistent, see [30]. Indeed, the following identity holds

$$\mathbf{u}_\alpha \cdot \nabla_{x,y}\cdot(\mu h_\alpha \nabla_{x,y}\mathbf{u}_\alpha) = \nabla_{x,y}\cdot(\mu h_\alpha \mathbf{u}_\alpha \nabla_{x,y}\mathbf{u}_\alpha) \\ + \Gamma_{\alpha+1/2}\frac{|\mathbf{u}_{\alpha+1}|^2-|\mathbf{u}_\alpha|^2}{2} - \Gamma_{\alpha-1/2}\frac{|\mathbf{u}_\alpha|^2-|\mathbf{u}_{\alpha-1}|^2}{2} + S_{\mu,\alpha}.$$

It is important to ensure the same consistency at the discrete level, i.e. an analogous identity should be verified at the discrete level. This gives guidelines for the discretization

of $S_{\mu,\alpha}$. A similar problem is studied in [67] in the case of the compressible Navier-Stokes equations. However the numerical analysis of the scheme with the dissipation and diffusion terms has not been done yet. It will be performed in a future work.

4.5 Numerical validation

In this section, we confront the proposed numerical scheme to three test cases. In the first one we present convergence results in the case of 3d analytical solution for the Euler-Fourier system. The second test deals with the simulation of the lock exchange phenomenon and the comparison with experimental results available in the literature. The third test consists in two simple diffusion cases for which we present a validation with an analytical solution and comparisons between the Navier-Stokes-Fourier and Boussinesq models.

4.5.1 Analytic solution

In [39], the authors have proposed an analytic solution for the Euler-Fourier system i. e. the system (4.1)-(4.3) with $\lambda = 0, \mu = 0$. The following proposition holds, its proof is detailed in [39].

Proposition 6. *For any nonnegative function $s \mapsto \rho(s)$ and for some $(a, \alpha, \eta, h_0) \in \mathbb{R}^3 \times \mathbb{R}_+$, let us consider the functions h, u, v, w, p, ϕ defined for $(x, y) \in [-L/2, L/2]^2$, $t \geqslant t_0$ by*

$$
\begin{aligned}
h(t,x,y) &= \max\left\{0, h_0 - \alpha \frac{(x-\eta\cos(\omega t))^2 + (y-\eta\sin(\omega t))^2}{2}\right\},\\
u(t,x,y,z) &= -\eta\omega\sin(\omega t),\\
v(t,x,y,z) &= \eta\omega\cos(\omega t),\\
w(t,x,y,z) &= -\alpha\eta\omega\left(x\sin(\omega t) - y\cos(\omega t)\right),\\
p(t,x,y,z) &= p^a(t) + \int_z^{h+z_b} \rho(T(t,x,y,z_1))dz_1,\\
T(t,x,y,z) &= a\left(h + z_b - z\right),
\end{aligned}
$$

with $\omega = \sqrt{\alpha g}$ and with a bottom topography defined by $z_b(x,y) = \frac{\alpha}{2}(x^2+y^2)$, then h, u, v, w, p, ϕ as defined previously satisfy the 3D hydrostatic Euler system with variable density (system (4.1)-(4.3) with $\lambda = 0, \mu = 0$) completed with the kinematic boundary conditions (4.4), (4.6).

For the numerical validation the parameters are set to $\eta = 0.1$, $h_0 = 0.1$, $a = 10$, $\alpha = 1$ and $L = 4$ and we consider a simplified equation of state given by $\rho(T) = \rho_0 + \beta T$ with $\rho_0 = 1000$ and $\beta = 10$. The free surface is plotted at different times in Figure 4.5 highlighting the planar motion of the fluid in the bowl. On Figure 4.6, the density in the slice plane (x, y=0, z) after a period for a mesh with 31316 triangles and 30 layers is plotted.

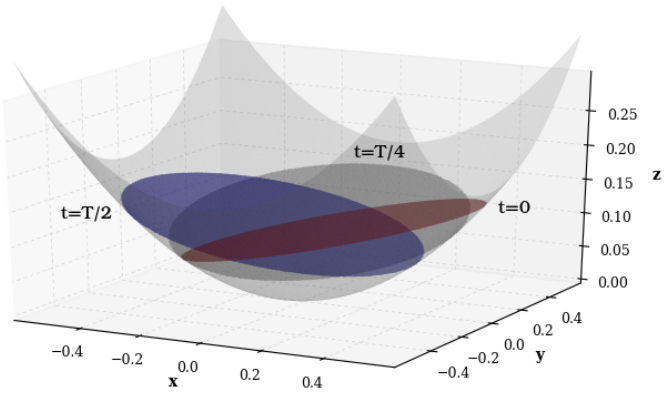


Figure 4.5 – Analytical solution of prop. 6, 3D planar surface in a parabolic bowl: free surface at $t = 0$ (red), $t = \tau/4$ (dark grey), $t = \tau/2$ (blue), with the period τ defined by $\tau = 2\pi/\omega$.

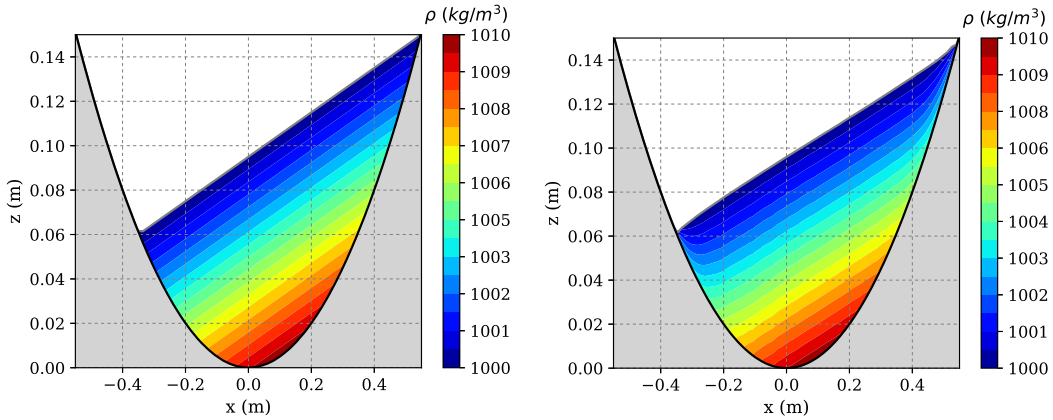


Figure 4.6 – Numerical result of the parabolic bowl with variable density. Free surface and density contour in the slice plane (x, y=0, z) at initial time (left) and at time $\tau = 2\pi/\omega$ with first order scheme (right).

The convergence towards the analytic solution is assessed by plotting the logarithm of the cumulative error (in L^2-norm) at time $\tau = 2\pi/\omega$ versus the space discretization (i.e. $\log(l_0/l_i)$ where l_0 is the average edge length of the mesh i and l_0 the average edge length of the coarsest mesh) for several unstructured meshes with 934, 2194, 4020, 6408, 9066, 12674 triangles. More precisely, the error in L^2-norm is computed by summing on all the nodes of the mesh at each time step, for each layer. Then, the cumulative error is obtained by summing the errors at each time step, normalized by the time step.

Figures 4.7 and 4.8 show the cumulative errors obtained with a constant number of layers equal to 10 (the mesh is refined in the horizontal direction but not in the vertical direction). The analytic solution being non-stationary, errors in time and space

accumulate over time and the theoretical rate of convergence is thus hard to obtain.

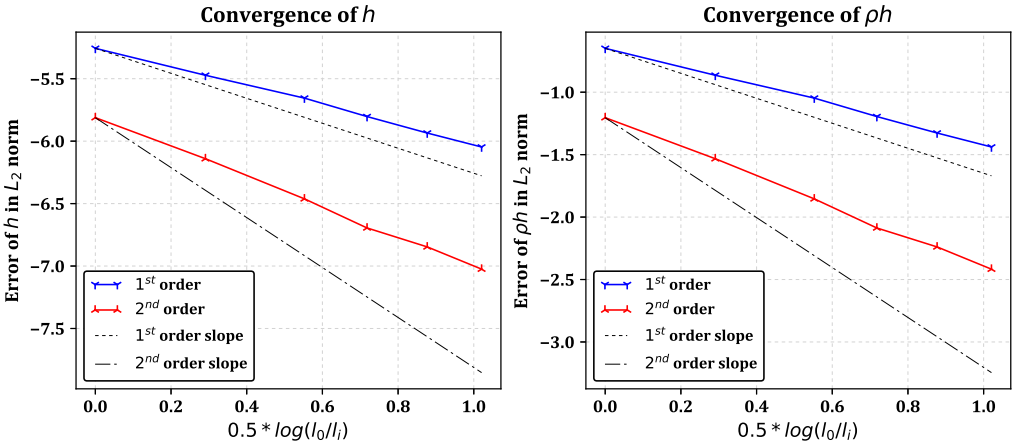


Figure 4.7 – Convergence of h and ρh in L^2-norm towards the analytical solution, constant number of layers.

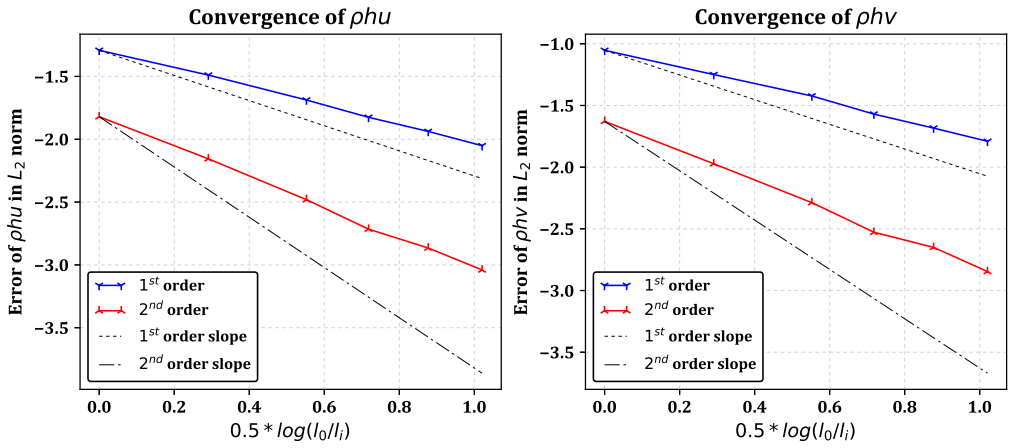


Figure 4.8 – Convergence of ρhu and ρhv in L^2-norm towards the analytical solution, constant number of layers.

Figures 4.9 and 4.10 show the cumulative error in L^2 norm for meshes with 934, 2194, 4020, 6408, 9066, 12674 triangles and 10, 12, 14, 16, 18, 20 layers respectively. Increasing the number of layers while refining the horizontal mesh is a reasonable idea because by doing so, the proportions of the 3D wedge cells are preserved. A super-convergence phenomenon can be observed for ρh when the number of layers is increased as the mesh is refined. A rate of convergence higher than the theoretical rate is also observed for ρhu and ρhv. The faster one refines the mesh in the vertical direction, the higher the convergence rates obtained. The results shown on figures 4.9 and 4.10 prove the stability of the numerical scheme for the Euler-Fourier system.

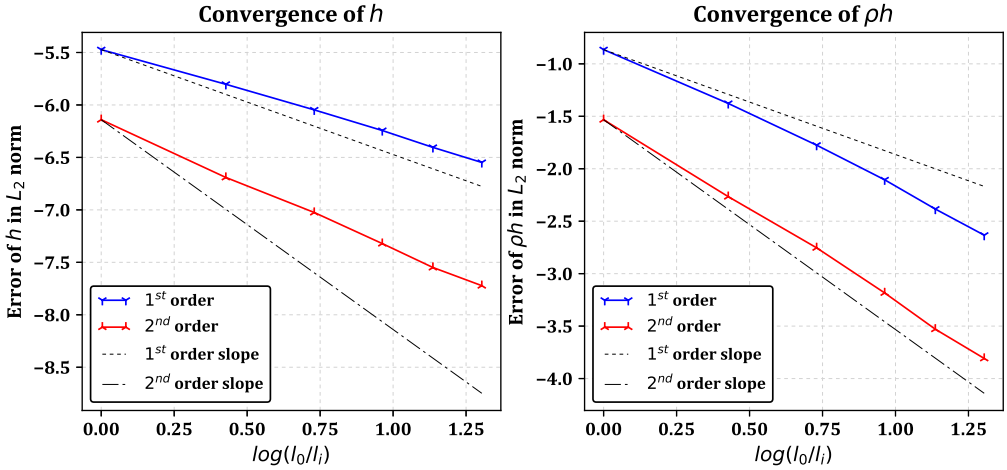


Figure 4.9 – Convergence of h and ρh in L^2-norm towards the analytical solution, increasing number of layers.

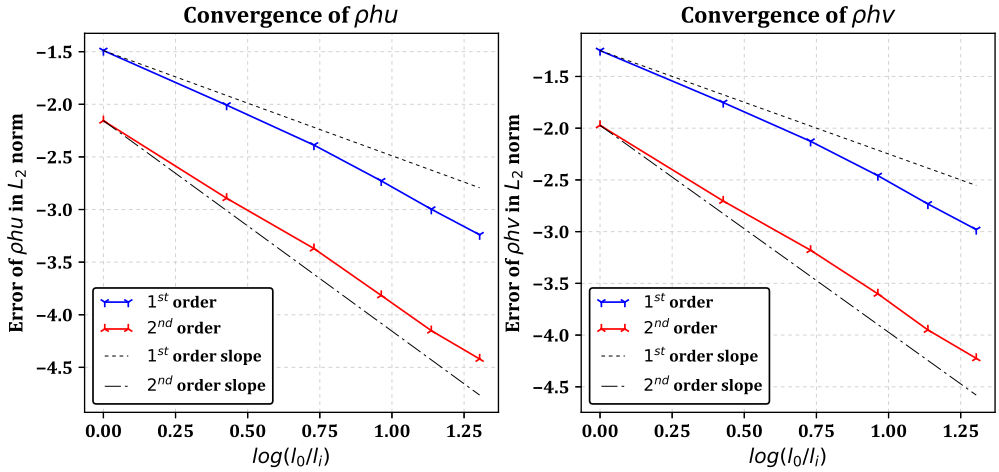


Figure 4.10 – Convergence of ρhu and ρhv in L^2-norm towards the analytical solution, increasing number of layers.

To summarize the results on the convergence of the schemes and to overcome the difficulty of interpreting the super-convergence cases, we propose another way of plotting the results in the case of multi-layer models. The error is plotted as a function of the horizontal space step l_i and of the vertical proportion $l_p = 1/N$ at the same time. Figure 4.11 shows the horizontal convergence rates of the first- and second-order schemes, as well as the first order vertical convergence. Note that the vertical discretization proportion does not correspond to the vertical space step because the water depth varies ($h_\alpha = l_p h$).

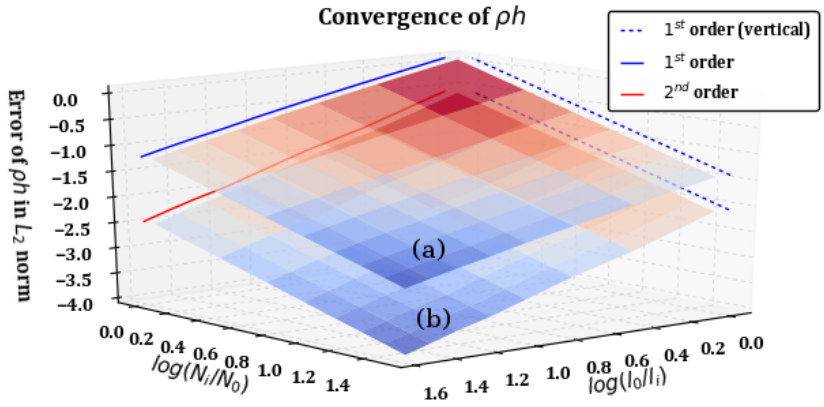


Figure 4.11 – Error of ρh in L^2-norm as a function of vertical and horizontal discretization for first order (a) and second order (b) numerical schemes.

4.5.2 Lock exchange

Gravity currents triggered by lock-exchanges are encountered in many applications and their numerical simulation is a challenge. In this section we show the ability of our numerical scheme to properly simulate the propagation of lock-exchange induced density currents. The computed front position is compared to the experiments carried out by Adduce & al. [3]. The results presented were obtained with the Navier-Stokes-Fourier model where the dissipation terms $S_{\mu,\alpha}$ due to the viscosity were neglected.

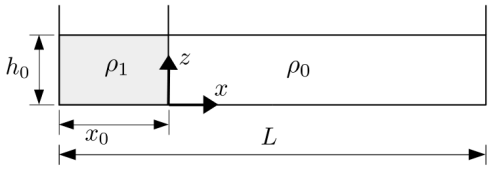


Figure 4.12 – Fluid domain of the lock-exchange test case

Initially, fluids of different densities, $\rho_1 = 1090kg.m^{-3}$ and $\rho_0 = 1000kg.m^{-3}$ respectively, are at rest and separated by a wall located at $x_0 = 0.3m$. When the vertical barrier is removed, the denser fluid flows under the lighter one due to the difference in the hydrostatic pressure. The initial water height is $h_0 = 0.3m$ and the length of the domain is $L = 3m$. The reduced gravity is defined by $g^* = g\,(1-\gamma)$ where $\gamma = \rho_0/\rho_1$ is the density ratio. We also define the buoyancy velocity as $u_b = \sqrt{g^* h_0}$. The Navier-Stokes equations are usually made dimensionless using the Grashof number defined by

$$Gr = \left(\frac{u_b h_0}{\nu}\right)^2, \tag{4.94}$$

with ν the kinematic viscosity. Simulations have been carried out with a Grashof number of $Gr = 2.53 \times 10^8$ on different meshes. The evolution of the density and the position of

the front are presented in figures 4.13 and 4.14 respectively. The initial opening of the gate has been delayed ($1s$) to take into account the time of opening in the experiment. The figure 4.14 shows the convergence of the numerical scheme and we observe a good matching between the numerical simulation and the experimental data of Adduce & Al. [3].

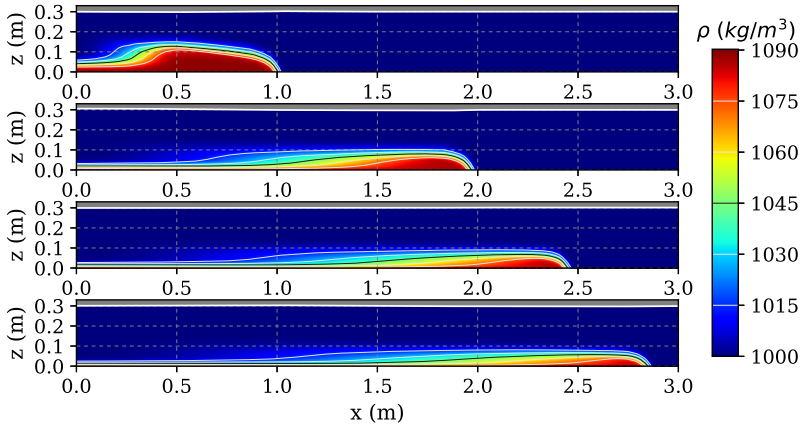


Figure 4.13 – Computed density with the most refined mesh in the slice plane $(x, y = 0, z)$ with $Gr = 2.53 \times 10^8$ at times $t = 3$, 7, 9 and $11s$ (from top to bottom).

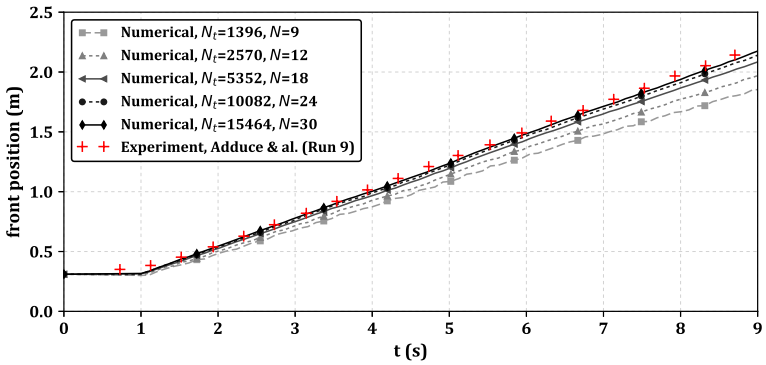


Figure 4.14 – Front position as a function of time for different meshes with comparison to Adduce & al. [3] experimental results (where N_t is the number of triangles and N is the number of layers).

4.5.3 Diffusion

In this section we consider simple diffusion cases in which a basin is initially at rest i.e. $\mathbf{U} = (0, 0, 0)^T$ and the free surface and bottom are flat and equal to $h(t_0, x, y) = h_0$ and $z_b(t_0, x, y) = 0$ respectively. In the basin, the temperature is initially distributed such that $T(t_0, x, y, z) = T_0(z)$, where $T_0 = T_0(z)$ is a given function.

First we consider the system

$$\nabla \cdot \mathbf{U} = 0, \tag{4.95}$$

$$\rho_0 c_p \frac{\partial T}{\partial t} + \nabla \cdot (T\mathbf{U}) = \nabla \cdot (\lambda \nabla T), \tag{4.96}$$

$$\rho_0 \left(\frac{\partial \mathbf{u}}{\partial t} + \nabla_{x,y} \cdot (\mathbf{u} \otimes \mathbf{u}) + \frac{\partial (\mathbf{u} w)}{\partial z} \right) + \nabla_{x,y} \int_z^\eta \rho g dz = 0, \tag{4.97}$$

i.e. where the Boussinesq assumption is made - ρ_0 is a constant. Starting from the initial conditions described above, it is easy to see that the velocity and density for $t \geqslant t_0$ are given by $\mathbf{U} = (0,0,0)^T$ and $\rho = \rho(T)$, where $T = T(t,z)$ is governed by the heat equation:

$$\begin{cases} \dfrac{\partial T}{\partial t} = \mathcal{D}_0 \dfrac{\partial^2 T}{\partial z^2}, \\ T(t_0, z) = T_0(z). \end{cases} \tag{4.98}$$

The coefficient $\mathcal{D}_0$ is the diffusivity defined by $\mathcal{D}_0 = \frac{\lambda_0}{\rho_0 c_p}$.

Now considering the system (4.1)-(4.3) with $\mu = 0$ i.e. without the Boussinesq assumption and starting also from the initial conditions described above, it is easy to see that the velocity and temperature for $t \geqslant t_0$ is given by $\mathbf{U} = (0,0,\overline{w})^T$ and $T = T(\rho)$, where $\overline{w}$ is defined by:

$$\bar{w} = -\frac{\lambda_0}{c_p} \int_{z_b}^z \frac{\rho'}{\rho^2} \frac{\partial^2 T}{\partial z^2} dz,$$

where $\rho' = \frac{\partial \rho}{\partial T}$ is deduced from the equation of state and $\rho = \rho(t,z)$ is governed by the equation:

$$\begin{cases} \dfrac{\partial \rho}{\partial t} + \overline{w} \dfrac{\partial \rho}{\partial z} = \dfrac{\lambda_0}{c_p} \dfrac{\rho'}{\rho} \dfrac{\partial^2 T}{\partial z^2}, \\ \rho(t_0, z) = \rho_0(z). \end{cases} \tag{4.99}$$

The velocity $\overline{w}$ is the result of the fluid dilatation. When $\overline{w} = 0$ the two models are almost identical, the only difference being the diffusivity $\mathcal{D}$. It is either a constant in the Boussinesq model or defined by $\mathcal{D} = \frac{\lambda_0}{\rho c_p}$ in the case of the Navier-Stokes-Fourier model. In the following examples we show that for common water equations of state, $\overline{w}$ is sufficiently small to have no noticeable effect on T and ρ. However, this term is essential in order to obtain rigorous mass conservation.

Analytical solution

In this test, the initial temperature T_0 (and density ρ_0) is constant in the domain and a Dirichlet boundary condition is applied at the bottom with a temperature of $T_b = 0$ (cf. figure 4.15). At the free surface we impose a homogeneous Neumann boundary condition ($\phi_T|_\eta = 0$, where ϕ_T is the thermal flux).

Dimensionless parameters are defined such that $\tilde{z} = z/h_0$, $\tilde{t} = \frac{\mathcal{D}_0 t}{h_0^2}$ and $T = T_b +$

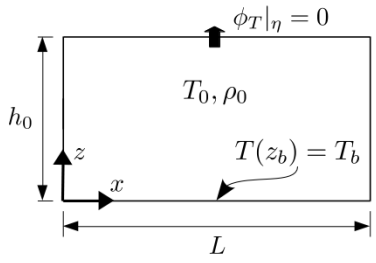


Figure 4.15 – Fluid domain of the diffusion test case with Dirichlet boundary condition at the bottom.

$(T_0 - T_b)\tilde{T}$. From equation (4.98) we obtain the dimensionless heat equation:

$$\frac{\partial \tilde{T}}{\partial \tilde{t}} = \frac{\partial^2 \tilde{T}}{\partial \tilde{z}^2} \tag{4.100}$$

From equation (4.99) we obtained a slightly different dimensionless heat equation

$$\frac{\partial \tilde{T}}{\partial \tilde{t}} - \frac{\rho(T_0 - T_b)}{h_0} \frac{\partial \tilde{T}}{\partial \tilde{z}} \int_{z_b}^{z} \frac{\rho'}{\rho^2} \frac{\partial \tilde{T}}{\partial \tilde{z}} dz = \frac{\partial^2 \tilde{T}}{\partial \tilde{z}^2}. \tag{4.101}$$

We compare the computed temperature to the analytical solution (4.102) with the parameters $\tilde{T}_0 = 1$, $\tilde{T}_b = 0$ (i.e. $T = \tilde{T}$) and a simplified equation of state given by: $T(\rho) = \frac{\rho_0 - \rho}{\beta}$ with $\rho_0 = 1000$ and $\beta = 10$. The analytical solution of the equation (4.100) in this case is given by the error function

$$\tilde{T}_{an}(\tilde{z}, \tilde{t}) = \frac{2}{\sqrt{\pi}} \int_0^{\frac{\tilde{z}}{2\sqrt{\tilde{t}}}} exp(-\xi^2) d\xi. \tag{4.102}$$

We obtain a good matching between the numerical results and the analytical solution both with the Navier-Stokes-Fourier system (with $\mu = 0$) and the Boussinesq system (4.95)-(4.97) which validates the numerical treatment of the diffusion. The results are plotted on Figure 4.16 in the case of the Navier-Stokes-Fourier model.

Note that the mass is strictly conserved in the Navier-Stokes-Fourier model, even though the density varies. Dilatation effects induce a variation of the water height such that the integral of the density over the total volume of fluid stays constant over time. This is not the case with the Boussinesq model, where the volume is not affected by temperature variation and mass conservation is violated (cf. Figure 4.18). However, there is no noticeable difference between the analytical solution and the Navier-Stokes-Fourier numerical solution even though $\rho' = -\beta$ has been chosen higher than its real physical value (in the case of water $0.03 < \rho' < 0.13$).

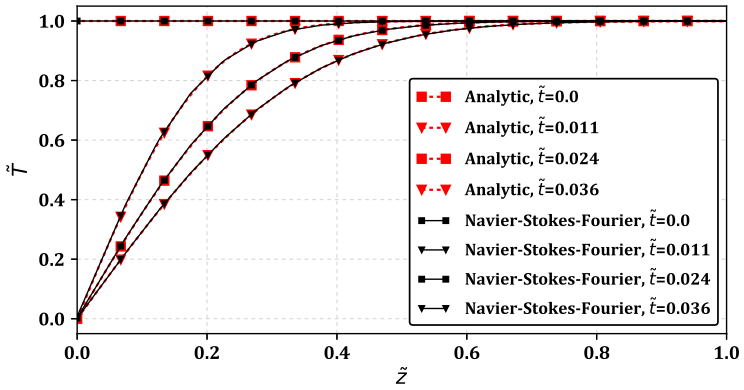


Figure 4.16 – Dimensionless temperature $\tilde{T}$ as a function of z/h_0 at different times and comparison between analytical solution and numerical simulation with a number of layers equal to $N = 20$.

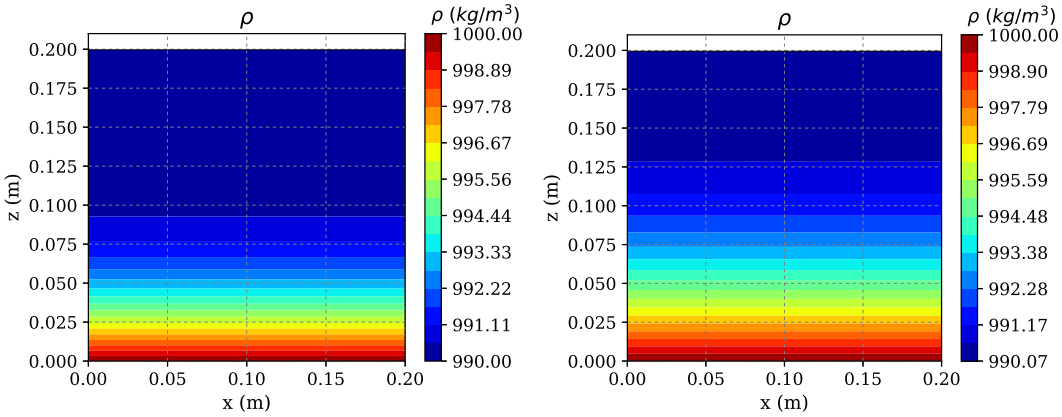


Figure 4.17 – Evolution of the density in the slice plane (x, y=0, z) with the Navier-Stokes-Fourier model at time $\tilde{t} = 0.03$ (left) and $\tilde{t} = 0.06$ (right).

Thermal equilibrium

In this second test a well stratified fluid is considered. No exterior forcing is applied and zero thermal flux boundary conditions are considered so that only the internal diffusion due to gradients of temperature affects the evolution of the density in the fluid (cf. Figure 4.19). Given a non null thermal conductivity, the density converges towards a stationary and uniform solution.

For the numerical simulation the parameters have been set to $h_0 = 2m$, $\rho_0 = 995.52kg.m^{-3}$ and $\rho_1 = 999.76kg.m^{-3}$. We now use a more realistic water equation of state defined by $T(\rho) = 4 + \sqrt{\frac{\rho_0 - \rho}{\beta \rho_0}}$ with $\beta = 6.63 \times 10^{-6}$ and $\rho_0 = 1000kg.m^{-3}$. This gives the following initial temperatures $T_0 = 30°C$ and $T_1 = 10°C$. We define the dimensionless z coordinate and time as $\tilde{z} = z/h_0$ and $\tilde{t} = \frac{a_m t}{h_0^2}$, where a_m is the initial

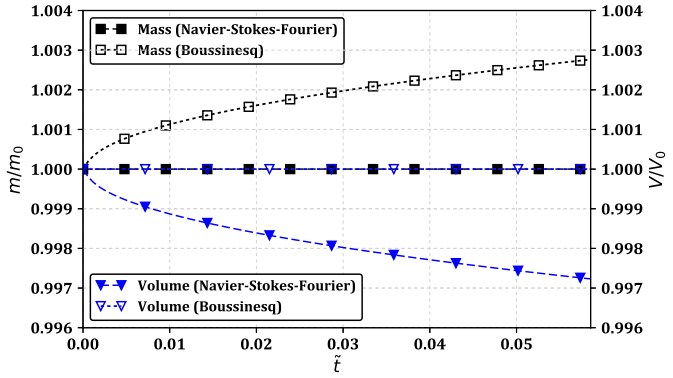


Figure 4.18 – Evolution of the mass ratio (m/m_0) and volume ratio (V/V_0) for the Navier-Stokes-Fourier and Boussinesq models.

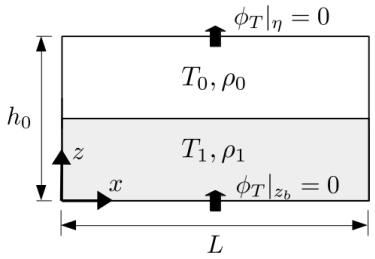


Figure 4.19 – Fluid domain for the diffusion test case

mean diffusivity defined by $a_m = \frac{\lambda}{\rho_m c_p}$ with $\rho_m = (\rho_0 + \rho_1)/2$.

The evolution of density and temperature obtained with both Navier-Stokes-Fourier and Boussinesq is plotted on Figure 4.20. The temperature converges rigorously towards an equilibrium temperature of $T_{eq} = T_m = (T_0 + T_1)/2 = 20°C$ in the case of the Boussinesq model. For the Navier-Stokes-Fourier model, the equilibrium temperature is shifted below T_m due to dilatation effects and is equal to $T_{eq} = 19.977°C$. It is not yet clear why the equilibrium temperature is lower than T_m. Note that for both models the density does not converge towards ρ_m as the equation of state is not linear.

4.6 Conclusion

In this work, we have proposed and analyzed a finite-volume scheme to solve the Navier-Stokes-Fourier equations, which describe free-surface variable density flows. With any flux consistent with the semi-discrete in time Euler system, the proposed scheme is well-balanced and preserves the non-negativity of the water depth. In the case of a kinetic flux, a discrete entropy balance is proved for a flat topography. The numerical scheme is validated. The confrontation is made with results obtained with the simulation of

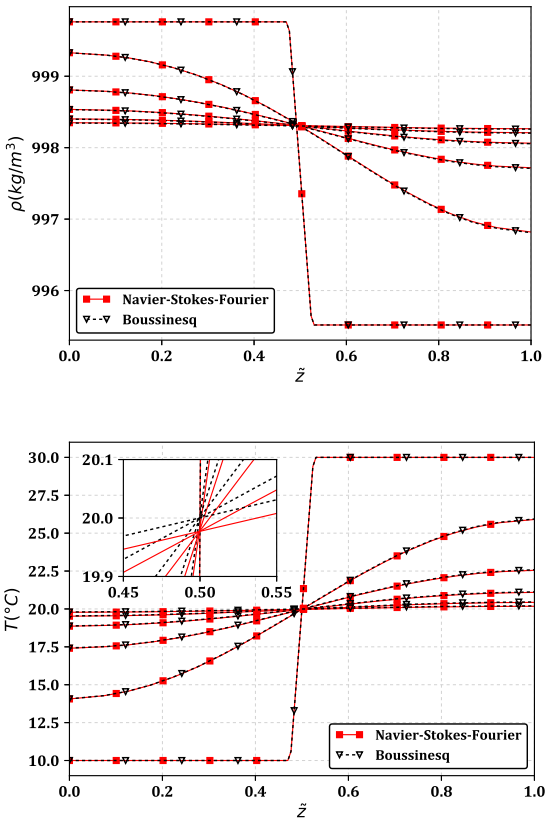


Figure 4.20 – Density (top) and temperature (bottom) against z/h_0 with the Navier-Stokes-Fourier and Boussinesq models at times $\tilde{t} = 0$, 0.07, 0.16, 0.24, 0.33 and 0.42 with $N = 20$

the Boussinesq system. Notably, in a simple thermal diffusion case, the equilibrium temperature is not the same for the two systems.

A discrete entropy balance for a non-flat topography has yet to be obtained. Another challenge is the design and analysis of a numerical scheme for a system with a non-Newtonian rheology. With a non-Newtonian rheology, we expect to be able to simulate complex interactions between the viscous effect and the temperature fluxes. Finally, simulations of the Navier-Stokes-Fourier system could be performed to investigate the propagation of internal waves in a stratified ocean.

Acknowledgments

The authors acknowledge the Inria Project Lab "Algae in Silico" for its financial support. This research is also supported by the ERC SLIDEQUAKES ERC-CG-2013-PE10-617472.

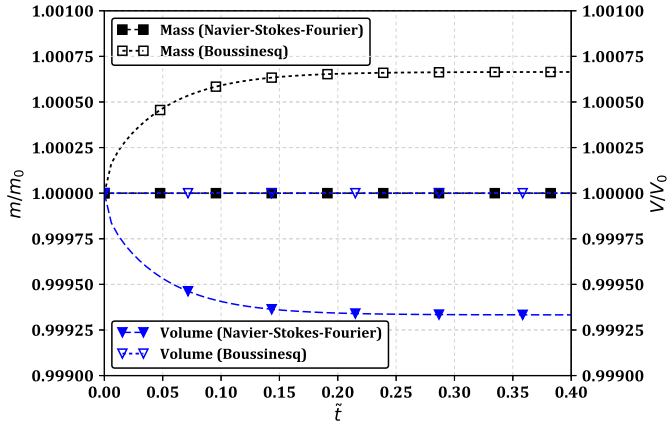


Figure 4.21 – Evolution of the mass ratio (m/m_0) and volume ratio (V/V_0) for the Navier-Stokes-Fourier and the Boussinesq models.

List of main symbols in Chapters 3 and 4

Symbol	Description
x, y	Coordinates in the horizontal plane
z	Coordinate in the vertical direction
t	Time
g	Gravitational acceleration
η	Free surface elevation
z_b	Bottom elevation
h	Water depth
ρ	Water density
$\mathbf{U}$	Velocity vector
$\mathbf{u}$	Horizontal velocity vector
w	Vertical velocity
p	Pressure
Σ	Stress tensor
σ	Viscosity tensor
e	Internal specific energy
H	Specific enthalpy
s	Specific entropy
T	Temperature
S	Salinity
μ	Viscosity
ζ	Second viscosity
κ	Navier friction coefficient
p^a	Atmospheric pressure
λ	Heat conductivity
Q_T	Heat flux (in the absence of salt)
k^T	Molecular diffusivity of temperature
k^S	Molecular diffusivity of salt
c_p	Heat capacity at constant pressure
μ_S	Chemical potential of seawater
$\mathbf{F^T}$	Heat flux (in the presence of salt)
$\mathbf{F^S}$	Salt flux
ε	Scaling parameter
p_{ref}	Constant reference pressure
p_0	Rescaled variable part of the pressure

Symbol	Description
e_0	Rescaled specific internal energy
e_0^{eq}	Rescaled specific internal energy constrained by $p_0(\rho, T) = 0$
T^{eq}	Temperature constrained by $p_0(\rho, T) = 0$
s_0	Rescaled specific entropy
λ_0	Rescaled heat conductivity
c_{p0}	Rescaled heat capacity at constant pressure
α	Layer index
l_α	Height fraction of layer α
$z_{\alpha+1/2}$	Interface between layers α and $\alpha + 1$
L_α	Fluid domain delimitated by $z_{\alpha-1/2}$ and $z_{\alpha+1/2}$
$G_{\alpha+1/2}$	Mass exchange term at interface $z_{\alpha+1/2}$
E_α	Mechanical energy in layer α
$\mathcal{S}_{T,\alpha}$	Vertically averaged heat flux in layer α
$\mathcal{S}_{\mu,\alpha}$	Vertically averaged viscous dissipation terms in layer α

The symbols used for the description of the numerical schemes are presented on pages 126 and 130.

Bibliography

[1] R. Abgrall and S. Karni. "Two-layer shallow water system: a relaxation approach". In: *SIAM J. Sci. Comput.* 31 (2009), pp. 1603–1627.

[2] C. Acary-Robert, E. D. Fernández-Nieto, G. Narbona-Reina, and P. Vigneaux. "A Well-balanced Finite Volume-Augmented Lagrangian Method for an Integrated Herschel-Bulkley Model". In: *Journal of Scientific Computing* 53 (2012), pp. 608–641.

[3] C. Adduce, G. Sciortino, and S. Proietti. "Gravity Currents Produced by Lock Exchanges: Experiments and Simulations with a Two-Layer Shallow-Water Model with Entrainment". In: *Journal of Hydraulic Engineering* 138 (2012), pp. 111–121.

[4] T. Alazard. "Low Mach number limit of the full Navier-Stokes equations". In: *Archive for Rational Mechanics and Analysis* 180.1 (2006), pp. 1–73.

[5] N. Alibaud, P. Azerad, and D. Isèbe. "A non-local non-monotone conservation law for dune morphodynamics". In: *Differential and Integral Equations* 23.1-2 (2010), pp. 155–158.

[6] S. Allgeyer et al. "Numerical approximation of the 3d hydrostatic Navier-Stokes system with free surface". working paper or preprint. 2017.

[7] B. Alvarez-Samaniego and P. Azerad. "Existence of travelling-wave solutions and local well-posedness of the Fowler equation". In: *Discrete and Continuous Dynamical Systems, series B* 12.04 (2009), pp. 671–692.

[8] C. Ancey. "Stochastic modeling in sediment dynamics: Exner equation for planar bed incipient bed load transport conditions". In: *Journal of Geophysical Research* 115 (2010).

[9] B. Andreotti, Y. Forterre, and O. Pouliquen. *Granular Media. Between Fluid and Solid.* Cambridge University Press, 2013.

[10] A. Arakawa and V. R. Lamb. "A Potential Enstrophy and Energy Conserving Scheme for the Shallow Water Equations". In: *Monthly Weather Review* 109 (1980), pp. 18–36.

[11] A. Arakawa and V. R. Lamb. “Computational design of the basic dynamical processes of the UCLA general circulation model”. In: *Methods in Computational Physics: Advances in Research and Applications* 17 (1977), pp. 173–265.

[12] K. Ashida and M. Michiue. “Studies on bed-load transport rate in open channel flows”. In: *Proceedings of the International Association for Hydraulic Research International Symposium on River Mechanics*. Asian Institute of Technology, Bangkok. 1973, pp. 407–417.

[13] F. Auclair et al. “A non-hydrostatic non-Boussinesq algorithm for free-surface ocean modelling”. In: *Ocean Modelling* (2018).

[14] E. Audusse. “A multilayer Saint-Venant model : Derivation and numerical validation”. In: *Discrete Contin. Dyn. Syst. Ser. B* 5.2 (2005), pp. 189–214.

[15] E. Audusse, F. Bouchut, M.-O. Bristeau, R. Klein, and B. Perthame. “A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows”. In: *SIAM J. Sci. Comput.* 25.6 (2004), pp. 2050–2065.

[16] E. Audusse, F. Bouchut, M.-O. Bristeau, R. Klein, and B. Perthame. “A Fast and Stable Well-Balanced Scheme with Hydrostatic Reconstruction for Shallow Water Flows”. In: *SIAM J. Sci. Comput.* 25.6 (2004), pp. 2050–2065.

[17] E. Audusse, F. Bouchut, M.-O. Bristeau, and J. Sainte-Marie. “Kinetic entropy inequality and hydrostatic reconstruction scheme for the Saint-Venant system”. In: *Mathematics of Computation* 85.302 (2016), pp. 2815–2837.

[18] E. Audusse and M.-O. Bristeau. “A well-balanced positivity preserving second-order scheme for Shallow Water flows on unstructured meshes”. In: *J. Comput. Phys.* 206.1 (2005), pp. 311–333.

[19] E. Audusse and M.-O. Bristeau. “Finite-volume solvers for a multilayer Saint-Venant system”. In: *Int. J. Appl. Math. Comput. Sci.* 17.3 (2007), pp. 311–319.

[20] E. Audusse, M.-O. Bristeau, and A. Decoene. “Numerical simulations of 3d free surface flows by a multilayer Saint-Venant model”. In: *Internat. J. Numer. Methods Fluids* 56.3 (2008), pp. 331–350.

[21] 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 solution”. In: *Journal of Computational Physics* 230 (2011), pp. 3453–3478.

[22] 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”. In: *ESAIM: M2AN* 45 (2011), pp. 169–200.

[23] E. Audusse, M.-O. Bristeau, and J. Sainte-Marie. “Kinetic entropy for layer-averaged hydrostatic Navier-Stokes equations”. *submitted*. 2017.

[24] E. Audusse, C. Chalons, and P. Ung. “A simple three-wave approximate Riemann solver for the Saint-Venant-Exner equations”. In: *Numerical Methods in Fluids* 87.10 (2018), pp. 508–528.

[25] E. Audusse et al. “Sediment transport modelling: relaxation schemes for Saint-Venant - Exner and three layer models”. In: *ESAIM: Proceedings and Surveys* 38 (2012), pp. 78–98.

[26] O. Aydin. “Effects of viscous dissipation on the heat transfer in a forced pipe flow. Part 2: Thermally developing flow”. In: *Energy Conversion and Management* 46 (2005), pp. 3091–3102.

[27] O. Aydin. “Effects of viscous dissipation on the heat transfer in forced pipe flow. Part 1: both hydrodynamically and thermally fully developed flow”. In: *Energy Conversion and Management* 46 (2005), pp. 757–769.

[28] F. Benkhaldoun, S. Sahmim, and M. Seaid. “Solution of the sediment transport equation using a finite volume method based on sign matrix”. In: *SIAM J. Sci. Comput.* 31.4 (2009), pp. 2866–2889.

[29] C. Berthon, M. Bessemoulin-Chatard, and H. Mathis. “Numerical convergence rate for a diffusive limit of hyperbolic systems: p-system with damping”. In: *SMAI Journal of Computational Mathematics* (2016).

[30] L. Boittin, F. Bouchut, M.-O. Bristeau, A. Mangeney, J. Sainte-Marie, and F. Souille. “The incompressible Navier-Stokes-Fourier system with free surface, Part I: Model & layer-averaged formulation”. In: (2018).

[31] L. Boittin, F. Bouchut, M.-O. Bristeau, A. Mangeney, J. Sainte-Marie, and F. Souille. “The incompressible Navier-Stokes-Fourier system with free surface, Part II: Numercial scheme & validation”. In: (2018).

[32] F. Bouchut. *Nonlinear Stability of Finite Volume Methods for Hyperbolic Conservation Laws*. Birkhaüser Verlag, 2004.

[33] F. Bouchut and T. M. de Luna. “An entropy-satisfying scheme for two-layer Shallow-Water equations with uncoupled treatment”. In: *ESAIM: M2AN* 42 (2008), pp. 683–698.

[34] F. Bouchut and V. Zeitlin. “A robust well-balanced scheme for multi-layer shallow water equations”. In: *Discrete Contin. Dyn. Syst. Ser. B* 13 (2010), pp. 739–758.

[35] A. Bouharguane and B. Mohammadi. “Minimization principles for the evolution of a soft sea bed interacting with a shallow sea”. In: *International Journal of Computational Fluid Dynamics* 26.3 (2012).

[36] J. V. Boussinesq. *Théorie analytique de la chaleur mise en harmonie avec la thermodynamique et avec la théorie mécanique de la lumière*. Vol. 2. Paris: Gathier-Villars, 1903.

[37] Y. Brenier. “Homogeneous hydrostatic flows with convex velocity profiles”. In: *Nonlinearity* 12.3 (1999), pp. 495–512.

[38] M.-O. Bristeau, B. D. Martino, C. Guichard, and J. Sainte-Marie. “Layer-averaged Euler and Navier-Stokes equations”. In: *Communications in Mathematical Sciences* 15.5 (2017), pp. 1221–1246.

[39] M.-O. Bristeau, B. D. Martino, A. Mangeney, J. Sainte-Marie, and F. Souillé. "Various analytical solutions for the incompressible Euler and Navier-Stokes systems with free surface". preprint. 2018.

[40] M.-J. Castro, J. Macías, and C. Parés. "A Q-scheme for a class of systems of coupled conservation laws with source term. Application to a two-layer 1-D shallow water system". In: *M2AN Math. Model. Numer. Anal.* 35.1 (2001), pp. 107–127.

[41] M. Castro, J. García-Rodríguez, J. González-Vida, J. Macías, C. Parés, and M. Vázquez-Cendón. "Numerical simulation of two-layer shallow water flows through channels with irregular geometry". In: *J. Comput. Phys.* 195.1 (2004), pp. 202–235.

[42] V. Casulli. "Semi-implicit Finite Difference Methods for the Two-Dimensional Shallow Water Equations". In: *Journal of Computational Physics* 86 (1990), pp. 56–74.

[43] F. Charru. "Selection of the ripple length on a granular bed sheared by a liquid flow". In: *Physics of Fluids* 18 (2006).

[44] F. Charru and E. J. Hinch. "Ripple formation on a particle bed sheared by a viscous liquid. Part 1. Steady flow". In: *Journal of Fluid Mechanics* 550 (2006), pp. 111–121.

[45] S. Cordier, M. H. Le, and T. M. de Luna. "Bedload transport in shallow water models: Why splitting (may) fail, how hyperbolicity (can) help". In: *Advances in Water Resources* 34 (2011), pp. 980–989.

[46] J. M. K. Dake and D. R. F. Harleman. "Thermal stratification in lakes: Analytical and laboratory studies". In: *Water Resources Research* 5.2 (), pp. 484–495. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/WR005i002p00484.

[47] A. Decoene and J.-F. Gerbeau. "Sigma transformation and ALE formulation for three-dimensional free surface flows". In: *International Journal for Numerical Methods in Fluids* 59 (2009), pp. 357–386.

[48] M. J. C. Díaz, E. D. Fernández-Nieto, and A. M. Ferreiro. "Sediment transport models in Shallow Water equations and numerical approach by high order finite volume methods". In: *Computers and Fluids* 37.3 (2008), pp. 299–316.

[49] J. Donea, A. Huerta, J.-P. Ponthot, and A. Rodríguez-Ferran. "Encyclopedia of Computational Mechanics". In: ed. by E. Stein, R. de Borst, and T. J. R. Hughes. Vol. 1. John Wiley and Sons Ltd., 2004. Chap. 14, Arbitrary Lagrangian–Eulerian Methods.

[50] H. A. Einstein. *The Bed-Load Function for Sediment Transportation in Open Channel Flows*. Technical Report 1026. United States Department of Agriculture, 1950.

[51] F. Engelund and E. Hansen. *Investigations of flow in alluvial streams.* Vol. 35. Acta polytechnica Scandinavica/Civil engineering and building construction series. Danish Academy of Technical Sciences, 1966.

[52] F. Engelund and O. Skovgaard. "On the origin of meandering and braiding in alluvial streams". In: *Journal of Fluid Mechanics* 57.289-302 (1973).

[53] C. Escauriaza, C. Paola, and V. R. Voller. "Gravel-Bed Rivers: Processes and Disasters". In: ed. by D. Tsutsumi and J. Laronne. John Wiley and Sons Ltd., 2017. Chap. 1.

[54] F. M. Exner. "Über die Wechselwirkung zwischen Wasser und Geschiebe in Flüssen". In: *Akademie der Wissenschaften in Wien. Mathematisch-naturwissenschaftliche Klasse* 134.2a (1925), pp. 165–204.

[55] R. Eymard and P. H. Gunawan. "Staggered scheme for the Exner-shallow water equations". In: *Computational Geosciences* 19.6 (2015), pp. 1197–1206.

[56] E. Feireisl and A. Novotný. "The Low Mach Number Limit for the Full Navier–Stokes–Fourier System". In: *Archive for Rational Mechanics and Analysis* 186 (2007), pp. 77–107.

[57] E. Fernández-Nieto, E. Koné, and T. Chacón Rebollo. "A multilayer method for the hydrostatic Navier-Stokes equations: a particular weak solution". In: *Journal of Scientific Computing* 60.2 (2014), pp. 408–437.

[58] E. Fernández-Nieto, T. M. de Luna, G. Narbona-Reina, and J. D. Zabsonré. "Formal deduction of the Saint-Venant-Exner model including arbitrarily sloping sediment beds and associated energy". In: *Mathematical Modelling and Numerical Analysis* (2016).

[59] S. Ferrari and F. Saleri. "A new two-dimensional Shallow Water model including pressure effects and slow varying bottom topography". In: *M2AN Math. Model. Numer. Anal.* 38.2 (2004), pp. 211–234.

[60] F. Filbet. "A finite volume scheme for the Patlak–Keller–Segel chemotaxis model". In: *Numerische Mathematik* 104.4 (2006), pp. 457–488.

[61] A. C. Fowler. "Geomorphological fluid mechanics". In: ed. by A. Provenzale, A. Provenzale, and N. Balmforth. 211. Springer-Verlag, Berlin, 2001. Chap. Dunes and drumlins, pp. 430–454.

[62] J. Fredsøe. "On the development of dunes in erodible channel". In: *Journal of Fluid Mechanics* 64 (1974), pp. 1–16.

[63] J.-F. Gerbeau and B. Perthame. "Derivation of viscous Saint-Venant System for Laminar Shallow Water". In: *Discrete and Continuous Dynamical Systems, series B* 1.1 (2001), pp. 89–102.

[64] B. Gomez and M. Church. "An Assessment of Bed Load Sediment Transport Formulae for Gravel Bed Rivers". In: *Water Resources Research* 25.6 (1989), pp. 1161–1186.

[65] T. Goudon and M. Parisot. "Finite Volume schemes on unstructured grids for non-local models: Application to the simulation of heat transport in plasmas". In: *Journal of Computational Physics* 231 (2012), pp. 8188–8208.

[66] T. Goudon and M. Parisot. "On the Spitzer-Härm regime and nonlocal approximations: Modeling, analysis and numerical simulations". In: *Multiscale Modeling & Simulation* 9.2 (2011), pp. 568–600.

[67] D. Grapsas, R. Herbin, W. Kheriji, and J.-C. Latché. "An unconditionally stable staggered pressure correction scheme for the compressible Navier-Stokes equations". In: *SMAI Journal of Computational Mathematics* 2 (2016), pp. 51–97.

[68] A. J. Grass. *Sediment transport by waves and currents.* Tech. rep. FL29. SERC London Cent. Mar. Technol., 1981.

[69] R. J. Greatbatch, Y. Lu, and Y. Cai. "Relaxing the Boussinesq Approximation in Ocean Circulation Models". In: *Journal of Atmospheric and Oceanic Technology* 18.11 (2001), pp. 1911–1923.

[70] E. Grenier. "On the derivation of homogeneous hydrostatic equations". In: *ESAIM: M2AN* 33.5 (1999), pp. 965–970.

[71] S. M. Griffies et al. "Developments in ocean climate modelling". In: *Ocean Modelling* 2.3 (2000), pp. 123 –192.

[72] A. F. Gulbransen, V. L. Hauge, and K.-A. Lie. "A Multiscale Mixed Finite Element Method for Vuggy and Naturally Fractured Reservoirs". In: *SPE Journal* 15.2 (2010).

[73] H. P. Gunawan. "Numerical simulation of shallow water equations and related models". PhD thesis. Université Paris-Est, 2015.

[74] F. Harlow and A. Amsden. *Fluid Dynamics.* Tech. rep. LA-4700. Los Alamos National Laboratory, 1971.

[75] R. Herbin, W. Kheriji, and J.-C. Latché. "On some implicit and semi-implicit staggered schemes for the shallow water and Euler equations". In: *ESAIM: Mathematical Modelling and Numerical Analysis* 48 (2014), pp. 1807–1857.

[76] B. R. Hodges, J. Imberger, A. Saggio, and K. B. Winters. "Modeling basin-scale internal waves in a stratified lake". In: *Limnology and Oceanography* 45.7 (2000), pp. 1603–1620.

[77] R. X. Huang, X. Jin, and X. Zhang. "An Oceanic General Circulation Model in Pressure Coordinates". In: *Advances in Atmospheric Sciences* 18.1 (2001).

[78] A. Hudson and P. K. Sweby. "Formulations for Numerically Approximating Hyperbolic Systems Governing Sediment Transport". In: *Journal of Scientific Computing* 19.1-3 (2003), pp. 225–252.

[79] S. Hwang. "Cauchy's Interlace Theorem for Eigenvalues of Hermitian Matrices". In: *The American Mathematical Monthly* 111.2 (2004), pp. 157–159.

[80] IOC, SCOR, and IAPSO. *The international thermodynamic equation of seawater – 2010: Calculation and use of thermodynamic properties*. Manuals and Guides 56. http://unesdoc.unesco.org/images/0018/001881/188170e.pdf. Intergovernmental Oceanographic Commission, 2010.

[81] F. James, P.-Y. Lagrée, H.-M. Le, and M. Legrand. "Towards a new friction model for shallow water equations through an interactive viscous layer". https://hal.archives-ouvertes.fr/hal-01341563v2.

[82] D. J. Jerolmack and D. Mohrig. "A unified model for subaqueous bed form dynamics". In: *Water Resources Research* 41 (2005).

[83] E. F. Keller and L. A. Segel. "Traveling Bands of Chemotactic Bacteria: A Theoretical Analysis". In: *Journal of Theoretical Biology* 30 (1971), pp. 235–248.

[84] J. F. Kennedy. "The mechanics of dunes and antidunes in erodible bed channels". In: *Journal of Fluid Mechanics* 16 (1963), pp. 521–544.

[85] S. Klainerman and A. Majda. "Compressible and Incompressible Fluids". In: *Communications on Pure and Applied Mathematics* 35 (1982), pp. 629–651.

[86] S. Klainerman and A. Majda. "Singular Limits of Quasilinear Hyperbolic Systems with Large Parameters and the Incompressible Limit of Compressible Fluids". In: *Communications on Pure and Applied Mathematics* 34 (1981), pp. 481–524.

[87] P.-Y. Lagrée. "A triple deck model of ripple formation and evolution". In: *Physics of Fluids* 15.8 (2003).

[88] E. Lajeunesse, L. Malverti, and F. Charru. "Bed load transport in turbulent flow at the grain scale: Experiments and modeling". In: *Journal of Geophysical Research* 115 (2010).

[89] B. Larrouturou. "How to preserve the mass fractions positivity when computing compressible multi-component flows". In: *J. Comp. Phys.* 95.1 (1991), pp. 59 –84.

[90] C. Lin, J. Han, C. Bennett, and R. L. Parsons. "Case History Analysis of Bridge Failures due to Scour". In: *International Symposium of Climatic Effects on Pavement and Geotechnical Infrastructure*. A.S.C.E. 2013.

[91] R. Lopez, D. Vericat, and R. J. Batalla. "Evaluation of bed load transport formulae in a large regulated gravel bed river: The lower Ebro (NE Iberian Peninsula)". In: *Journal of Hydrology* 510 (2014), pp. 164–181.

[92] Y. Lu. "Including Non-Boussinesq Effects in Boussinesq Ocean Circulation Models". In: *Journal Of Physical Oceanography* 31 (2000).

[93] C. Lusso. "Modélisation numérique des écoulements gravitaires viscoplastiques avec transition fluide/solide". PhD thesis. Université Paris-Est, 2013.

[94] D. A. Lyn and M. Altinakar. "St. Venant-Exner Equations for Near-Critical and Transcritical Flows". In: *Journal of Hydraulic Engineering* 128.579-587 (2002).

[95] D. K. Lysne. "Movement of sand in tunnels". In: *Proc. A.S.C.E.* 95 (1969), pp. 1835–1846.

[96] R. Manning. "On the flow of water in open channels and pipes". In: *Transactions of the Institution of Civil Engeers of Ireland* XX (1891), pp. 161–207.

[97] F. Marche. "Derivation of a new two-dimensional viscous shallow water model with varying topography, bottom friction and capillary effects". In: *European Journal of Mechanic /B* 26 (2007), pp. 49–63.

[98] N. Masmoudi and T. Wong. "On the Hs Theory of Hydrostatic Euler Equations". In: *Archive for Rational Mechanics and Analysis* 204.1 (2012), pp. 231–271.

[99] T. J. McDougall, R.-J. Greatbatch, and Y. Lu. "On Conservation Equations in Oceanography: How Accurate Are Boussinesq Ocean Models?" In: *Journal of Physical Oceanography* 32.5 (2002), pp. 1574–1584.

[100] G. L. Mellor and T. Ezer. "Sea level variations induced by heating and cooling: An evaluation of the Boussinesq approximation in ocean models ". In: *Journal Of Geophysical Research* 100.C10 (1995), pp. 20,565–20,577.

[101] S. Menoni and C. Margottini, eds. *Inside Risk: A Strategy for Sustainable Risk Mitigation*. Springer, 2011.

[102] A. F. Messiter. "Boundary-layer flow near the trailing edge of a flat plate". In: *SIAM J. Appl. Math.* 18.1 (1970), pp. 241–257.

[103] H. E. Reports, ed. *Formulas for Bed-Load Transport*. International Association for Hydraulic Structures Research. 1948.

[104] N. Münchener Rückversicherungs-Gesellschaft Geo Risks Research. *Loss events worldwide 1980-2014: 10 costliest floods ordered by insured losses*. Tech. rep. Munich RE, 2015.

[105] V. Y. Neiland. "Propagation of perturbation upstream with interaction between a hypersonic flow and a boundary layer". In: *Mekhanika Zhidkosti i Gaza* 4.53-57 (1970).

[106] T. Pähtz, J. F. Kok, E. J. R. Parteli, and H. J. Herrmann. "Flux saturation length of sediment transport". In: *Physical Review Letters* 111 (2013).

[107] S. Paolucci. *On the filtering of sound from the Navier-Stokes equations*. Technical Report 82-8257. Sandia National Laboratories, 1982.

[108] C. S. Patlak. "Random walk with persistence and external bias". In: *The bulletin of mathematical biophysics* 15.3 (1953), pp. 311–338.

[109] B. Perthame and C. Simeoni. "A kinetic scheme for the Saint-Venant system with a source term". In: *Calcolo* 38.4 (2001), pp. 201–231.

[110] A. Radice et al. "Management of flood hazard via hydro-morphological river modelling. The case of the Mallero in Italian Alps". In: *Journal of Flood Risk Management* 6 (2013), pp. 197–209.

[111] S. Rahmstorf. "Thermohaline circulation: The current climate". In: *Nature* 421.699 (2003).

[112] A. Recking. "A comparison between flume and field bed load transport data and consequences for surface-based bed load transport prediction". In: *Water Resources Research* 46 (2010).

[113] L. van Rijn. "Sediment transport - Part I: bed load - Part II: suspended load". In: *Journal of Hydraulic Division* 110 (1984).

[114] G. Rosatti, L. Bonaventura, A. Deponti, and G. Garegnani. "An accurate and efficient semi-implicit method for section-averaged free-surface flow modelling". In: *International Journal for Numerical Methods in Fluids* 65 (2011), pp. 448–473.

[115] A. J.-C. 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". In: *Comptes Rendus de l'Académie des Sciences* 73.147-154 (1871).

[116] J. Sainte-Marie. "Vertically averaged models for the free surface Euler system. Derivation and kinetic interpretation". In: *Math. Models Methods Appl. Sci. (M3AS)* 21.3 (2011), pp. 459–490.

[117] G. P. Schurtz, P. D. Nicolaï, and M. Busquet. "A nonlocal electron conduction model for multidimensional radiation hydrodynamics codes". In: *Physics of Plasmas* 7.10 (2000), pp. 4238–4249.

[118] A. Shields. *Anwendung der Änlichkeitsmechanik und der Turbulenzforschung auf die Geschiebebewegung*. Mitteilung der Preußischen Versuchanstalt für Wasserbau 26. Berlin: Preußische Versuchanstalt für Wasserbau, 1936.

[119] *Software CROCO*. https://www.croco-ocean.org.

[120] *Software Freshkiss3D*. http://freshkiss3d.gforge.inria.fr.

[121] *Software HEC-RAS*. http://www.hec.usace.army.mil/software/hec-ras/.

[122] *Software ICON-ESM*. https://www.mpimet.mpg.de/en/science/models/icon-esm/.

[123] *Software MIKE HYDRO River*. https://www.mikepoweredbydhi.com/products/mike-hydro-river.

[124] *Software MITgcm*. http://mitgcm.org.

[125] *Software NEMO*. https://www.nemo-ocean.eu.

[126] *Software POM*. http://www.ccpo.odu.edu/POMWEB/.

[127] *Software TELEMAC-MASCARET*. http://opentelemac.org.

[128] SOGREAH. *Modélisation hydrodynamique de l'Etang de Berre et des milieux annexes*. Tech. rep. GIPREB (Gestion intégrée, prospective et restauration de l'étang de Berre), 2009.

[129] Y. T. Song and T. Y. Hou. "Parametric vertical coordinate formulation for multiscale, Boussinesq, and non-Boussinesq ocean modeling". In: *Ocean Modelling* 11 (2006), pp. 298–332.

[130] N. Soontiens, M. Stastna, and M. L. Waite. "Trapped internal waves over topography: Non-Boussinesq effects, symmetry breaking and downstream recovery jumps". In: *Physics of Fluids* 25.6 (2013), p. 066602.

[131] B. Spinewine. "Two-layer flow behaviour and the effects of granular dilatancy in dam-break induced sheet-flow". PhD thesis. Université catholique de Louvain, 2005.

[132] K. Stewartson and P. G. Williams. "Self-induced separation". In: *Proceedings of the Royal Society A* 312 (1969), pp. 181–206.

[133] P. A. Tassi, S. Rhebergen, C. A. Vionnet, and O. Bokhove. "A discontinuous Galerkin finite element model for river bed evolution under shallow flows". In: *Computer Methods in Applied Mechanics and Engineering* 197.33-40 (2008), pp. 2930–2947.

[134] UNESCO. *Tenth report of the joint panel on oceanographic tables and standards.* Tech. rep. UNESCO Technical Papers in Marine Science, 1981.

[135] P. Ung. "Simulation numérique du transport sédimentaire : aspects déterministes et stochastiques". PhD thesis. Université d'Orléans, 2016.

[136] S. Vater. "A New Projection Method for the Zero Froude Number Shallow Water Equations". PhD thesis. Free University Berlin, 2004.

[137] J. D. Zabsonré, C. Lucas, and E. D. Fernández-Nieto. "An energetically consistent viscous sedimentation model". In: *Mathematical Models and Methods in Applied Sciences* 19.3 (2009), pp. 477–499.

[138] Y. Zech, S. Soares-Frazão, B. Spinewine, C. Savary, and L. Goutière. "Inertia effects in bed-load transport models". In: *Canadian Journal of Civil Engineering* 36.10 (2009), pp. 1587–1597.

Modeling, analysis and simulation of two geophysical flows
Sediment transport and variable density flows

Abstract

The present thesis deals with the modeling and numerical simulation of complex geophysical flows. Two processes are studied: sediment transport, and variable density flows. For both flows, the approach is the same. In each case, a reduced vertically-averaged model is derived from the 3D Navier-Stokes equations by making a specific asymptotic analysis. The models verify stability properties. Attention is paid to preserving these properties at the discrete level, in particular the entropy stability. The behavior of both models is illustrated numerically.
Concerning the sediment transport model, the sediment layer is first studied alone. Then, a coupled sediment-water model is presented and simulated. The influence of a viscosity term in the model for the sediment layer is investigated. Due to this viscosity term, the sediment flux is non-local. A transport threshold is added to the model. The water layer is modeled by the Shallow Water equations. Adding some non-locality to the model allows to simulate dune growth and propagation.
In the variable density flow model, the density is a function of one or several tracers such as temperature and salinity. The model derivation consists in removing the dependence of the density on the pressure. A layer-averaged formulation of the model is proposed, which is subsequently used to propose a numerical discretization. The numerical simulations emphasize the differences between this model and a model relying on the classical Boussinesq approximation.

Keywords: geophysical flows, sediment transport, non-local flux, variable density flow, multilayer model, numerical simulation

---

Modélisation, analyse et simulation de deux écoulements géophysiques
Transport sédimentaire et écoulement à densité variable

Résumé

Cette thèse traite de la modélisation et de la simulation numérique d'écoulements géophysiques complexes. Deux types d'écoulements sont étudiés, le transport sédimentaire par charriage et les écoulements à densité variable. La démarche suivie est la même pour les deux phénomènes. Dans chaque cas, un modèle réduit, moyenné suivant la verticale est dérivé à partir des équations de Navier-Stokes 3D en suivant une certaine asymptotique. Les modèles possèdent des propriétés de stabilité. Ces propriétés sont ensuite préservées au niveau discret, en particulier l'inégalité d'entropie.
En ce qui concerne le transport sédimentaire, la couche de sédiments est d'abord traitée seule, puis un modèle couplé pour les sédiments et l'eau est présenté et simulé. L'influence d'un terme de viscosité est étudiée. La présence du terme de viscosité rend le flux sédimentaire non-local. Un seuil pour le transport est introduit dans le modèle. L'eau est modélisée par les équations Shallow Water. L'ajout d'effets non-locaux permet de simuler la croissance et la propagation d'une dune.
Dans le modèle pour les écoulements à densité variable, la densité varie en fonction d'un ou plusieurs traceurs tels que la température et la salinité. La dérivation consiste à enlever la dépendance en pression dans la loi d'état du fluide. Une formulation moyennée suivant la verticale est proposée ; cette formulation est par la suite utilisée pour proposer une discrétisation. Les simulations font ressortir les différences entre le modèle étudié et un modèle classique reposant sur l'approximation de Boussinesq.

Mots clés : écoulements géophysiques, transport sédimentaire, flux non-local, écoulement à densité variable, modèle multicouches, simulation numérique

---

Inria Paris
2 rue Simone Iff – 75012 Paris – France