Low Froude regime and implicit kinetic schemes for the Saint-Venant system

Mathieu Rigal

Sorbonne Université

École doctorale de Sciences Mathématiques de Paris Centre

Thèse

pour obtenir le titre de

Docteur en Sciences
de Sorbonne Université

Discipline: Mathématiques Appliquées

présentée et soutenue par

Mathieu Rigal

Régime bas Froude et schémas cinétiques implicites pour les équations de Saint-Venant

---

Thèse dirigée par **Jacques Sainte-Marie**
et encadrée par **Nina Aguillon** et **Nathalie Ayi**

soutenue le 14 Novembre 2022 devant le jury composé de :

M.	**Jacques SAINTE-MARIE**	Directeur de thèse
Mme.	**Claire CHAINAIS-HILLAIRET**	Rapporteure
Mme.	**Gladys NARBONA-REINA**	Rapporteure
M.	**Benoît PERTHAME**	Examinateur
Mme.	**Anaïs CRESTETTO**	Examinatrice
M.	**Nicolas SEGUIN**	Président
Mme.	**Nina AGUILLON**	Co-encadrante de thèse
Mme.	**Nathalie AYI**	Co-encadrante de thèse

Projet soutenu par la Région Ile de France

Remerciements

L'aboutissement de cette thèse doit beaucoup aux nombreux soutiens reçus durant ces trois années. Je souhaiterais ici prendre le temps d'exprimer ma gratitude à toutes les personnes qui, de près ou de loin, ont pris part à cette expérience.

Ces remerciements vont tout d'abord à mes encadrantes, Nina et Nathalie, qui m'ont proposé ce sujet de thèse et m'ont accompagné tout au long de ce parcours. Depuis mon arrivée au laboratoire, elles ont su rester à mon écoute, et aussi bien leurs conseils que leur pédagogie et leur gentillesse m'ont été d'une aide précieuse pour avancer et faire face aux difficultés de la recherche. Cela a été une chance et un véritable plaisir de travailler avec elles.

Un grand merci aussi à Jacques qui, au fil des réunions, a pu me guider dans cette thèse grâce à ses nombreuses suggestions. Malgré son emploi du temps chargé, il est resté très disponible et je lui en suis reconnaissant. Sa patience, sa bienveillance et les discussions instructives que nous avons eues ont été autant de ressources qui m'ont permis de mener à bien ce projet et de progresser.

J'aimerais bien sûr remercier Claire Chainais-Hillairet, Gladys Narbona-Reina, Benoît Perthame, Anaïs Crestetto ainsi que Nicolas Seguin pour avoir consacré de leur temps en acceptant de participer au jury de thèse.

Je n'oublie pas non plus Laurent Boudin et Jean-Yves Chemin, aux côtés de qui j'ai travaillé dans le cadre du monitorat. Grâce à eux, j'ai été mis dans de bonnes conditions et cela a été une période très formatrice pour moi. Je les en remercie.

Mille mercis également à Julien Guieu pour son aide sur le plan administratif, rythmée par les ordres de mission, commandes et autres tickets, toujours dans la bonne humeur.

Je dédie enfin ces remerciements à mes collègues de bureau pour tous les bons moments passés ensemble. A Apolline, qui m'a fait découvrir les bandes dessinées de Manu Larcenet. A Chourouk, et son enthousiasme (partagé !) pour la musique de Chopin. A Matthias, arrivé en même temps que moi à l'Inria Paris il y a plus de trois ans et demi déjà. A Frédérique et les B.A.-BA d'autre fois. Mais aussi à Nelly, Juliette, Norbert, Emma, Jesús, Nicolas, Eugenio, Victor, Suraj, Edouard, Van Thanh, Haibo, Siwar, Fabien, et tous les autres. Je leur souhaite à tous bonne continuation, et bonne chance à celles et ceux qui soutiendront dans un avenir proche.

Low Froude regime and implicit kinetic schemes for the Saint-Venant system

Abstract : In this thesis we study implicit time discretizations for the Saint-Venant system. First we consider the issue of the low Froude regime in the two dimensional case. The ability to seamlessly transition towards this limiting regime poses two main issues, namely the computational cost of handling fast scales and the correct description of the asymptotic dynamic. The former is traditionally dealt with the use of implicit-explicit time integrators, whereas the latter requires the numerical error to be uniformly bounded with respect to the scale parameter. Especially, it is important for nearly incompressible states to satisfy some form of stability. This motivates the refinement of an existing criterion allowing to predict whether a scheme is accurate at low Froude numbers, which we validate through numerical examples. Furthermore the proposed semi-implicit schemes are based on a wave splitting enabling the well balanced property.

Then we focus on kinetic schemes for the one dimensional Saint-Venant system. In the case of a flat bathymetry, we obtain a fully implicit scheme preserving the water height positivity and admitting a discrete entropy inequality without any restriction on the time step. A simplified version of this scheme allows to explicitly rewrite the update at the macroscopic level. In order to account for varying bottoms, we examine an iterative strategy making use of the hydrostatic reconstruction. This approach requires a CFL condition to converge, in exchange of what we obtain a positive update with a discrete entropy inequality that always dissipates the energy of the system. This is an improvement over the fully explicit version of the scheme, which can sometimes increase the energy. We perform numerical tests to assess the efficiency and qualitative aspects of the proposed schemes.

Key words : Saint-Venant system, implicit-explicit methods, kinetic schemes, low Froude regime, asymptotic preserving methods, finite volumes, entropy inequality.

Contents

Chapter 1

General introduction

1.1 Introduction

This work is devoted to the study of the Saint-Venant equations — also known as the Shallow Water system, mainly under the prism of numerical analysis. More specifically we aim to design, analyze and implement numerical methods for approximating the solutions of the Saint-Venant system both in one and two spatial dimensions. The principal aspect of this work revolves around assessing the interest of using an implicit in time type of strategy, together with ensuring good properties at the discrete level that mirror the features of the continuous model.

1.1.1 Motivation

To begin with, we would like to motivate this goal and, to do so, we start by the more general picture of *geophysical flows*. With upwards of seventy percent of the Earth covered in water, this is a broad subject that, if well understood, can help us predict and solve numerous issues of great interest. Water management is probably among the most vital ones, as it implies the control of water quality and its availability. The term quality makes for instance echo to the degree of salinity near an estuary or the evolution of pollutants in the vicinity of an industrial area or a sewage treatment plant. It is critical especially for running water or swimming places. The availability is important for households but also agriculture (irrigation), energy production (dams, cooling of nuclear plants) and even leisure (swimming, kayaking, fishing). Another point of interest is the forecasting of natural disasters and the mitigation of their consequences. Natural disasters can include among other things floodings, tsunamis, tidal waves or dam break. Finally we want to mention the understanding of ocean hydrodynamics, which is relevant for various reasons. First, it is deeply coupled with climate change through factors such as the rise of sea level and its impacts on the coasts, changes in chemical constitution, as well as hurricane formation due to a greater heat transfer into the atmosphere. Secondly, the understanding of ocean hydrodynamics can allow us to better take advantage of its resources (marine energy, sea food) while preserving the ecosystems.

The common denominator in the above examples is that we are faced with a hydrodynamical evolution problem. The free surface Navier-Stokes system stands as one of the most precise models to describe this physical phenomena, but it also has a high level of complexity, which is problematic for two reasons. Firstly, it makes it difficult to study the mathematical properties of the model itself. Secondly, it is a real challenge to derive numerical methods that yield qualitatively good approximations at a reasonable computational cost. A simplification can be achieved under some scaling hypotheses and by averaging the incompressible Navier-Stokes system over the vertical. Doing so we reduce the dimension of the problem by one, and obtain a hierarchy of depth averaged free surface flow models. Among them, the Saint-Venant system constitutes one of the simplest yet accurate nonlinear model. Historically, it was introduced by de Saint Venant [61] in 1871. More than a century later, its rigorous mathematical derivation was proposed by Gerbeau and Perthame [33] for the one dimensional case, and by Marche [52] for the two dimensional one.

As a nonlinear system of hyperbolic conservation laws, the Saint-Venant equations are

classically treated with finite volumes schemes. Such methods are naturally conservative, since the cell values are updated through an exchange of fluxes with their neighbors. In order to produce good results, the two main ingredients are accuracy and stability. An accurate scheme is able to approximate the solution with a small error for a given resolution in time and space, whereas a stable scheme is able to yield qualitatively good approximations by keeping them in some domain of physical validity. Here we purposely use the term stability in a broad sense, which encompasses methods decreasing some energy, avoiding the apparition of spurious oscillations, keeping the water height positive or preserving steady states. One particular family of steady states regroups the lake at rest stationary states, which correspond to motionless flows with a flat free surface. It is important to take into consideration these stationary flows as most flows are a small perturbation of a lake at rest. Another fundamental aspect is that we are interested in solutions of the Saint-Venant equations that satisfy an entropy inequality. Accuracy and stability can enter in competition with each other, as for instance a scheme dissipating the energy will tend to be more diffusive and have a greater error. Yet, one must not overlook the stability as we believe it is essential to have exploitable results. To some degree, stability can be improved by using implicit (or semi-implicit) time integrators. In this thesis we shall see to what extent this statement is true, and how it can be taken advantage of.

1.1.2 Structure of the document

In a first part, we focus on the two dimensional Saint-Venant system in the regime of low Froude numbers. The Froude number is a dimensionless scale parameter obtained by taking the ratio between the material velocity of fluid particles and the celerity of propagation of surface gravity waves. Hence this regime coincides to the case where the particles travel much slower than the surface waves and which, in real life, can correspond to coastal flows, lakes or rivers. Usual finite volumes schemes with explicit time integration struggle to approximate this type of flows in an efficient and accurate way. One issue is that the stability of such methods requires to enforce a CFL condition making the time step proportional to the Froude number. Therefore when the latter tends to zero, the time step has to vanish which renders the scheme unusable. The other issue is that numerical solutions can become inaccurate when the Froude number decreases. To avoid this, one must be careful to ensure the error of consistency to be bounded uniformly in the scale parameter. To overcome these obstacles, we study an asymptotic preserving semi-implicit approach based on a wave splitting. The implicited part is linear in order to get a reasonable computational cost. The advantage of this method is that the CFL constraint is no more dependent of the Froude number, and we can use large time steps. We also justify that a centered discretization of the surface waves linear operator prevents any loss of accuracy as the Froude number goes to zero. Furthermore, lakes at rest are preserved thanks to the choice of splitting.

The second part of this work is dedicated to the analysis of an implicit kinetic scheme in the one dimensional case. The kinetic framework is attractive, as we replace the nonlinear system by a linear scalar equation which is much easier to discretize and study. It is also a practical way to design numerical schemes preserving the water height and

admitting a discrete entropy inequality, two properties that are satisfied at the continuous level by the solutions of the Saint-Venant system. Especially, negative water heights should be avoided at all cost as it has no physical meaning, and makes the system non hyperbolic. In the past years, kinetic schemes have been successful in achieving these two properties in absence of source term [56][57]. In order to deal with a varying bottom, a kinetic interpretation of the hydrostatic reconstruction has been proposed recently in [10] and was used in an explicit framework. This allows, in the context of a varying bathymetry, to obtain once again the positivity of the water height. However it was shown that this scheme can sometimes increase the energy of the system, which is due to the error induced by the explicit time discretization. This motivates the use of an implicit strategy that we investigate. In the end we are able to obtain a discrete entropy inequality which always dissipates the energy. Furthermore in absence of bathymetry we have an implicit kinetic scheme whose update can be rewritten explicitly at the macroscopic level. On the other hand the resolution over a varying bathymetry involves an iterative process.

1.2 The Saint-Venant system in one dimension

1.2.1 Properties of the model

The Saint-Venant system models a fluid flow in a geometry delimited below by a fixed bathymetry and above by a moving free surface (air-fluid interface). More precisely, it describes the conservation of the water height and the balance of the discharge — also called momentum. In the one dimensional case, we denote these scalar quantities by $h(t,x)$ and $q(t,x)$. After introducing the bathymetry profile $z(x)$ and the gravitational acceleration constant g, the Saint-Venant system reads

$$\begin{cases} \dfrac{\partial h}{\partial t} + \dfrac{\partial q}{\partial x} = 0 \\ \dfrac{\partial q}{\partial t} + \dfrac{\partial}{\partial x}\Big(\dfrac{q^2}{h} + \dfrac{g}{2}h^2\Big) = -ghz' \end{cases} . \tag{1.1}$$

We remark that the discharge becomes a conservative variable when the bathymetry is flat. It can be related to the water height with the equality $q = hu$ where $u(t,x)$ is the horizontal velocity of the fluid averaged over the vertical. It is also usual to consider the notation $\zeta(t,x) := h + z$ for the free surface. These quantities of interest are illustrated in Figure 1.1.

The water height should remain positive at all times, and we look for solutions valued in some convex set $\mathcal{U}$ satisfying

$$\mathcal{U} \subset \mathbb{R}_+ \times \mathbb{R} .$$

We have the convenient vector notation of (1.1)

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

with $F(U) = (hu, hu^2 + gh^2/2)^T$ and $S(U,z) = (0, -gh\partial_x z)^T$. The flux Jacobian reads

$$DF(U) = \begin{pmatrix} 0 & 1 \\ gh - u^2 & 2u \end{pmatrix} ,$$

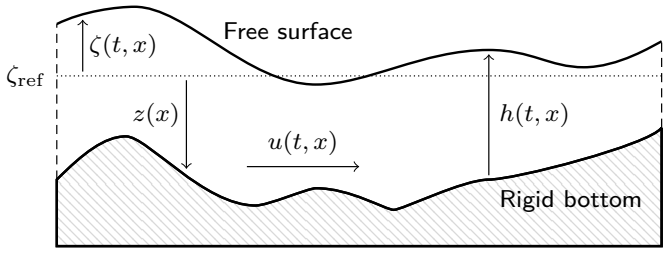

Figure 1.1: Quantities of interest in the one dimensional case. The reference free surface elevation ζ_{ref} is chosen such that it coincides with the lake at rest by convention.

and the real eigenvalues are $\lambda_\pm(U) = u \pm \sqrt{gh}$ with the celerity $\sqrt{gh}$ related to the surface waves, and as a consequence problem (1.1) is an hyperbolic system of nonlinear conservation laws with source term. It is well known that solutions of the Saint-Venant system can exhibit discontinuities even when the initial data is taken smooth, thus we have to consider them in the weak sense. Because weak solutions are not unique, System (1.1) is ill posed and we need a criterion to choose which solution is the physical one. The classical method [64] in the homogeneous case consists in adding a parabolic viscous perturbation of the form

$$\partial_t U + \partial_x F(U) = \nu \partial^2_{xx} U \ , \quad \nu > 0 \ ,$$

admitting a unique and smooth solution U_ν for a given initial data. The criterion is then to select the weak solution as the limit of the perturbed solution when $\nu \to 0$. Equivalently (see Lefloch [50], Chapter 1, Section 3 for the scalar case), this amounts to ask for all the entropy-entropy flux pairs (η, G_η) with η convex on $\mathcal{U}$ to satisfy the entropy inequality

$$\partial_t \eta(U) + \partial_x G_\eta(U) \leq 0 \ . \tag{1.3}$$

For smooth solutions, the inequality becomes an equality. The entropy flux G_η associated to η has to verify $\nabla G_\eta(U) = DF(U)^T \nabla \eta(U)$ so that in the smooth case (1.3) is obtained upon multiplication of (3.2) by $\nabla \eta(U)^T$ from the left. One such pair is given by the energy of the system and the associated flux

$$E(U) = \frac{hu^2}{2} + \frac{gh^2}{2} + ghz, \quad G(U) = \Big(E + \frac{gh^2}{2}\Big)u \ . \tag{1.4}$$

We call entropy solution of the Saint-Venant system any solution of (1.1) verifying the entropy inequality (1.3) for $\eta = E$. Since the work of Lions, Perthame and Souganidis [59], the existence of entropy solutions is ensured when the bathymetry is flat. The proof uses a kinetic formulation [51][55], meaning that all entropies are accounted for, not only the energy E. When only using a single entropy, we get the less precise kinetic representation which can nevertheless be used to derive numerical schemes with nice properties such as positivity and discrete entropy inequality. Such schemes are investigated in the third chapter.

Finally we have to mention the steady states of the one dimensional Saint-Venant system. They are obtained by looking for solutions that are constant in time. From the water height conservation, we deduce that steady states are characterized by a discharge q constant in space. Together with the momentum equation this yields Bernoulli's principle when the data is smooth

$$\begin{cases} \partial_x hu = 0 \\ \partial_x\Big(hu^2 + \dfrac{g}{2}h^2\Big) = -ghz' \end{cases} \qquad \Longrightarrow \qquad \partial_x\Big(\frac{u^2}{2} + g(h+z)\Big) = 0 \ .$$

A particular type of steady flows is obtained by setting $u = 0$, implying also $q = 0$. This coincides to stationary states or motionless flows. In this case, the momentum equation leads to a balance between the pressure variation and the source term, and we deduce that in wet areas the free surface is flat

$$\partial_x\Big(\frac{g}{2}h^2\Big) = -ghz' \qquad \Longrightarrow \qquad h + z \equiv \text{Cst} \ .$$

Because of that, a stationary state is often called lake at rest or hydrostatic equilibrium. In real life, most flows can be seen as a perturbation around a lake at rest, therefore is especially important to preserve these at the discrete level. Such methods are called *well balanced*, and will be achieved in two different ways throughout this thesis. In Chapter 2 we will split the flux into a convection operator and a surface wave operator according to [14][34]. The surface wave operator is obtained by linearizing the Saint-Venant system around a lake at rest, and this is what makes it easy to propose well balanced discretizations. In the third chapter we will use the *hydrostatic reconstruction* introduced in [6].

1.2.2 Derivation, domain of validity and extensions

We finish this section by briefly recalling the arguments used to derive the d dimensional Saint-Venant system, where d is either one or two. Departing from the incompressible $d+1$ dimensional Navier-Stokes system, the main idea is to reduce the dimension by averaging along the vertical direction, and then perform an asymptotic analysis using the hydrostatic and shallow water assumptions [33][48][52]. These assumptions are important as they shape the domain of validity of the Saint-Venant system. They are enumerated in the below.

- The flow is incompressible in dimension $d+1$;
- The free surface can be described by a single valued function ζ;
- The water depth is small compared to the characteristic wavelength;
- The velocity profile varies slowly in the vertical direction;
- The non hydrostatic part of the pressure is neglected;

We comment on these assumptions, as they point to the strong suit and limitations of the Saint-Venant model. Despite the flow being assummed incompressible in dimension $d+1$, the Saint-Venant system shares the same structure as the compressible isentropic Euler system. In fact, the compressibility of the Saint-Venant system should be understood in the sense that the free surface is able to evolve throughout time. This means that the model does not feature sound waves, but it can represent surface gravity waves. One advantage of depth averaged models over more complex ones such as the Navier-Stokes system is that the geometry of the fluid domain is naturally accounted for, and we don't need to keep track of the free surface. On the other hand, the non folding assumption prevents the model to describe the wave breaking phenomena responsible for the formation of wave rolls near the shoreline for instance.

Since it is a vertically averaged model, the Saint-Venant system better handles flows with a slowly varying velocity profile over the vertical axis. On the contrary, a situation not well represented is given by a strong change of horizontal velocity depending on depth, which can be induced in real life by wind friction or a gradient of temperature. For instance, a circular flow in the vertical plane correspond to a change of sign of the velocity, which is not seen by the model. When the vertical description of the flow is relevant such as in high sea, multi-layer models [7] can be used instead as they offer a good compromise between low complexity and accuracy. Another assumption restricting the domain of validity of the shallow water equations stems from the fact that the non hydrostatic pressure is neglected. Only keeping the hydrostatic part of the pressure means that the latter is exclusively related to the effect of gravity (i.e. the weight exerted by the mass of fluid above). A consequence is that dispersive effects are not featured in the model. The Serre-Green Naghdi model should then be preferred.

Despite the previous limitations, the Saint-Venant model yields accurate results in many situations of interest such as rivers or coastal flows. In addition, it admits several possible extensions to account for various additional physical phenomena. These range from the friction at the bottom required to represent granular flows, the friction induced by the wind at the free surface, the Coriolis force which is relevant in ocean dynamics, to the transport of a density of salt or pollutants. Variations in time of the bathymetry profile can also be considered with the Saint-Venant–Exner model. In this model it is a sediment layer standing on top of the fixed bottom that is allowed to evolve, see [31] for a derivation from the 3D Navier-Stokes system.

1.3 Finite volumes in the one dimensional case

Finite volumes are well suited to approximate the solutions of hyperbolic systems of conservation laws with a source term such as the Saint-Venant equations. Their main advatages lie in their ability to conserve the quantities of interest, to deal with discontinuous weak solutions, and to treat complex geometries by the mean of unstructured meshes. In this thesis we will consider both the one and two dimentional Saint-Venant systems. However in the two dimensional case we will limit to cartesian meshes that will enable the analysis of a given scheme through its modified PDE. For this reason we restrict this presentation to 1D finite volumes, as the methods for the two dimensional

case are obtained from one dimensional numerical fluxes projected along the principal directions.

1.3.1 Integral formulation and Riemann problem

The concept of finite volumes is closely tied with the integral formulation of conservation laws. Considering a strong solution U of the Saint-Venant system (1.1) with flat bathymetry, we obtain the equality

$$\int_a^b U(t_2,x)\,\mathrm{d}x - \int_a^b U(t_1,x)\,\mathrm{d}x = \int_{t_1}^{t_2} F(U(\tau,a))\,\mathrm{d}\tau - \int_{t_1}^{t_2} F(U(\tau,b))\,\mathrm{d}\tau \tag{1.5}$$

by integrating (3.2) over the control volume $[t_1,t_2]\times[a,b]\subset\mathbb{R}_+\times\mathbb{R}$ and by applying Green's formula. This equality means that the variation of the quantities of interest contained in $[a,b]$ between times t_1 and t_2 is entirely determined by the fluxes at the boundaries a and b. Let us introduce the spatial discretization made of cells $C_i = [x_{i-1/2}, x_{i+1/2}]$ for $i\in\mathbb{Z}$, where $x_{i\pm1/2}$ represents the position of the left and right interfaces. Now if we apply the previous formula to cell C_i between times 0 and Δt, we get that

$$U_i(\Delta t) - U_i(0) = \frac{1}{|C_i|}\left(\int_0^{\Delta t} F(U(\tau,x_{i-1/2}))\,\mathrm{d}\tau - \int_0^{\Delta t} F(U(\tau,x_{i+1/2}))\,\mathrm{d}\tau\right), \tag{1.6}$$

where $U_i(t)$ is the average of the solution at time t over the cell C_i. No approximation has been made to obtain (1.6) which is exact. All the difficulty is then to evaluate the remaining integrals of the interfacial fluxes, which for the Saint-Venant system this is in general not possible. In practice we will replace the solution $U_i(0)$ by a cellwise constant approximation $\overline{U}_i^0$. If we then zoom on the neighborhood of an interface $x_{i+1/2}$, at time $t=0$ the data is made of two constant states $\overline{U}_i^0$ and $\overline{U}_{i+1}^0$ separated by a discontinuity. Performing the translation $y = x - x_{i+1/2}$, this gives us a Riemann problem of the form

$$\begin{cases} \partial_t\widetilde{U} + \partial_y F(\widetilde{U}) = 0 \\ \widetilde{U}(0,y) = \mathbb{1}_{y<0}\widetilde{U}_L + \mathbb{1}_{y>0}\widetilde{U}_R \end{cases}. \tag{1.7}$$

It is well known that solutions $\widetilde{U}$ of (1.7) are self slimilar, meaning that they verify $\widetilde{U}(t,y) = \widetilde{U}(y/t)$. More precisely they are made of up to two waves, each one either a shock or a rarefaction. We recall that a shock is a traveling discontinuity dissipating the entropy, and a rarefaction is a C^1 fan profile preserving the entropy. In general, the left and right states $\widetilde{U}_L$ and $\widetilde{U}_R$ are connected to an intermediate state $\widetilde{U}_I$, in which case we have two waves. Such an example is given in Figure 1.2, where the left-going wave is a shock connecting $\widetilde{U}_L$ to $\widetilde{U}_I$, and where the right-going wave is a rarefaction connecting $\widetilde{U}_I$ to $\widetilde{U}_R$. It is also possible that $\widetilde{U}_L$ and $\widetilde{U}_R$ are on the same Hugoniot locus, that is to say that they can be connected directly by one wave without the need for an intermediate state.

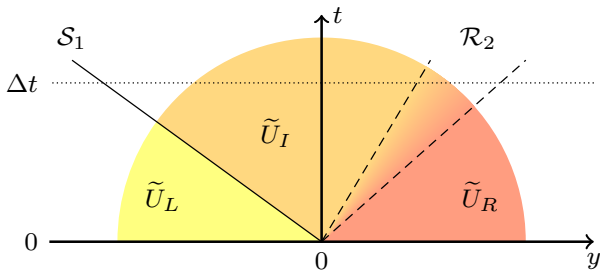

Figure 1.2: Example of solution to the Riemann problem featuring a 1-shock and a 2-rarefaction.

We briefly recall how to construct entropy solutions for the Riemann problem (1.7). Two states U_L, U_I are connected together through a 1-wave if

$$u_I = \mathcal{W}_1(h_I, h_L, u_L) \overset{\text{def}}{=} \begin{cases} u_L + 2\sqrt{gh_L} - 2\sqrt{gh_I} & \text{if } h_I \leq h_L \text{ (rarefaction)} \\ u_L - \sqrt{\frac{g}{2}\Big(\frac{h_L}{h_I} - \frac{h_I}{h_L}\Big)(h_L - h_I)} & \text{if } h_I > h_L \text{ (shock)} \end{cases},$$

and two states U_I, U_R are connected through a 2-wave if

$$u_I = \mathcal{W}_2(h_I, h_R, u_R) \overset{\text{def}}{=} \begin{cases} u_R - 2\sqrt{gh_R} + 2\sqrt{gh_I} & \text{if } h_I \leq h_R \text{ (rarefaction)} \\ u_R + \sqrt{\frac{g}{2}\Big(\frac{h_R}{h_I} - \frac{h_I}{h_R}\Big)(h_R - h_I)} & \text{if } h_I > h_R \text{ (shock)} \end{cases}.$$

To establish these conditions, we make use of the fact that the 1-Riemann invariant $u - 2\sqrt{gh}$ (resp. the 2-Riemann invariant $u + 2\sqrt{gh}$) remains constant through a 2-rarefaction (resp. a 1-rarefaction), and that the speed σ of a shock has to satisfy the so called Rankine-Hugoniot jump relation. Then the intermediate water height $\widetilde{h}_I$ of the solution is found as the root of

$$h \in \mathbb{R}_+ \longmapsto \mathcal{W}_1(h, \widetilde{h}_L, \widetilde{u}_L) - \mathcal{W}_2(h, \widetilde{h}_R, \widetilde{u}_R)\ , \tag{1.8}$$

and the intermediate velocity $\widetilde{u}_I$ is given by

$$\widetilde{u}_I = \mathcal{W}_1(\widetilde{h}_I, \widetilde{h}_L, \widetilde{u}_L) = \mathcal{W}_2(\widetilde{h}_I, \widetilde{h}_R, \widetilde{u}_R)\ .$$

Finally, we deduce the position of the shocks and the profile of the rarefactions that appear in the solution. We refer to [44], Chapter 7 for more details on the complete procedure.

1.3.2 The HLL approximate Riemann solver

In 1959, Godunov [60] proposed to approximate Equality (1.6) by substituting the interfacial flux $F(U(t, x_{i+1/2}))$ with $F(\widetilde{U}(0; \overline{U}_i^0, \overline{U}_{i+1}^0))$, where $\widetilde{U}(y/t; \overline{U}_i^0, \overline{U}_{i+1}^0)$ is the self similar solution of the Riemann problem (1.7) given by the initial condition $\mathbb{1}_{y<0}\overline{U}_i^0 +$

$\mathbb{1}_{y>0}\overline{U}^0_{i+1}$. However solving this Riemann problem is rather expensive, as one has to find the root of the nonlinear function (1.8).

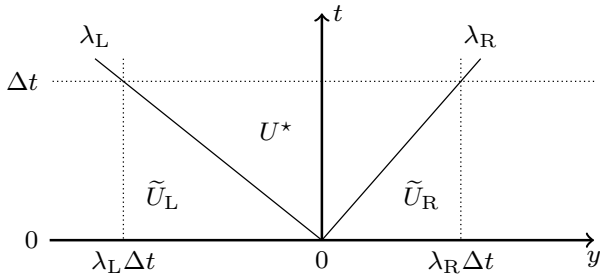

Figure 1.3: Approximate Riemann solver with two discontinuities.

Alternatively we can consider an approximate Riemann solver. The simplification proposed by Harten, Lax and Van Leer in [40] consists to introduce two fictitious waves traveling at speeds $\lambda_L < \lambda_R$ as illustrated by Figure 1.3. Instead of solving exactly the Riemann problem, we try to connect the exterior states $\widetilde{U}_L, \widetilde{U}_R$ to the average $U^\star$ of the solution between these characteristics. If we assume that the fan produced by the fictitious waves encompasses all the waves arising in the exact solution of problem (1.7), then $U^\star$ can be computed explicitly using the integral formulation (1.5) applied to the control volume $[0, \Delta t] \times [\lambda_L \Delta t, \lambda_R \Delta t]$

$$\int_{\lambda_L \Delta t}^{\lambda_R \Delta t} \widetilde{U}\Big(\frac{s}{\Delta t}; \widetilde{U}_L, \widetilde{U}_R\Big)\, \mathrm{d}s = \int_{\lambda_L \Delta t}^{\lambda_R \Delta t} (\mathbb{1}_{y<0}\widetilde{U}_L + \mathbb{1}_{y>0}\widetilde{U}_R)\, \mathrm{d}y + \Delta t F(\widetilde{U}_L) - \Delta t F(\widetilde{U}_R)\ ,$$

which leads to the so called consistency condition

$$U^\star = \frac{1}{\lambda_R - \lambda_L}\Big(\lambda_R \widetilde{U}_R - \lambda_L \widetilde{U}_L + F(\widetilde{U}_L) - F(\widetilde{U}_R)\Big)\ . \tag{1.9}$$

Typically we will estimate the velocities λ_L, λ_R from the eigenvalues of the jacobians $DF(\widetilde{U}_L)$ and $DF(\widetilde{U}_R)$. We can then introduce the piecewise constant function

$$U^{\mathrm{HLL}}(y/t; \widetilde{U}_L, \widetilde{U}_R) = \begin{cases} \widetilde{U}_L & \text{if } y/t \le \lambda_L \\ U^\star & \text{if } \lambda_L < y/t < \lambda_R \\ \widetilde{U}_R & \text{if } y/t \ge \lambda_R \end{cases}\ . \tag{1.10}$$

as a substitute of the exact solution $\widetilde{U}(y/t; \widetilde{U}_L, \widetilde{U}_R)$, and define the HLL flux F^{HLL} as

$$\begin{aligned} F^{\mathrm{HLL}}(\widetilde{U}_L, \widetilde{U}_R) &= \frac{1}{\Delta t}\int_0^{\lambda_R \Delta t} U^{\mathrm{HLL}}\Big(\frac{s}{\Delta t}; \widetilde{U}_L, \widetilde{U}_R\Big)\, \mathrm{d}s - \lambda_R \widetilde{U}_R + F(\widetilde{U}_R) \\ &= -\frac{1}{\Delta t}\int_{\lambda_L \Delta t}^{0} U^{\mathrm{HLL}}\Big(\frac{s}{\Delta t}; \widetilde{U}_L, \widetilde{U}_R\Big)\, \mathrm{d}s + \lambda_L \widetilde{U}_L + F(\widetilde{U}_R)\ . \end{aligned} \tag{1.11}$$

The second equality holds true thanks to the consistency condition (1.9). Note that definition (1.11) can be seen as an analogy with Godunov's flux which satisfies

$$F(\widetilde{U}(0; \widetilde{U}_L, \widetilde{U}_R)) = \frac{1}{\Delta t}\int_0^{\lambda_R \Delta t} \widetilde{U}\Big(\frac{s}{\Delta t}; \widetilde{U}_L, \widetilde{U}_R\Big)\, \mathrm{d}s - \lambda_R \widetilde{U}_R + F(\widetilde{U}_R)$$

$$= -\frac{1}{\Delta t}\int_{\lambda_L \Delta t}^{0} \widetilde{U}\Big(\frac{s}{\Delta t}; \widetilde{U}_L, \widetilde{U}_R\Big)\,\mathrm{d}s + \lambda_L \widetilde{U}_L + F(\widetilde{U}_R)$$

through the integral formulation applied to the control volumes $[0, \Delta t] \times [\lambda_L \Delta t, 0]$ and $[0, \Delta t] \times [0, \lambda_R \Delta t]$. Plugging the definition (1.10) of U^{HLL} in (1.11) we finally get

$$F^{\mathrm{HLL}}(\widetilde{U}_L, \widetilde{U}_R) = \begin{cases} F(\widetilde{U}_R) & \text{if } \lambda_L, \lambda_R \leq 0 \\ \dfrac{\lambda_R F(\widetilde{U}_L) - \lambda_L F(\widetilde{U}_R) + \lambda_L \lambda_R (\widetilde{U}_R - \widetilde{U}_L)}{\lambda_R - \lambda_L} & \text{if } \lambda_L < 0 < \lambda_R \\ F(\widetilde{U}_L) & \text{if } \lambda_L, \lambda_R \geq 0 \end{cases} .$$

We remark that the case where velocities λ_L, λ_R have the same sign is treated by an upwinding. Interestingly, if we symmetrize the fictitious waves by setting $\lambda_R = -\lambda_L$ we obtain the Rusanov flux. In practice, the HLL flux introduces less diffusion than the Rusanov flux and is thus more accurate. An extension of the HLL flux called HLLC was proposed by Toro in [66], with the purpose of accounting for a contact discontinuity as a third wave. Such a modification is not required for the one dimensional Saint-Venant system, as we only have two waves and none of them is a contact discontinuity. However in the two dimensional case the transverse velocity gives a contact discontinuity which makes the HLLC an interesting choice of discretization.

1.3.3 Time integration

The Saint-Venant system is a first order evolution problem involving a time derivative. To discretize this term, one needs to use some time integrator whose choice will of course impact the properties of the overall scheme. We start by illustrating the effect it can have on the stability by considering a toy problem. This will help to better understand what kind of issues occur in the richer dynamics offered by the nonlinear Saint-Venant system, and what benefits come with using implicit time integration. We consider the solutions in $L^2([0,1])$ of the following initial value problem with periodic boundary conditions

$$\frac{\partial u}{\partial t} + a\,\frac{\partial u}{\partial x} = 0 \ , \quad u(0,x) = u^0(x) \ , \quad u(t,0) = u(t,1) \ . \tag{1.12}$$

This corresponds to the scalar linear transport equation whose speed $a > 0$ is taken constant. Multiplying Equation (1.12) by u, we obtain that the energy $E(u) = u^2/2$ also satisfies the same transport equation. This implies that the total energy $\int_{[0,1]} E(u)(t,x)\,\mathrm{d}x$ is a constant function of time, especially it is non increasing. We would like a numerical approximation to satisfy this stability property at the discrete level. Let us first introduce a θ-scheme combined to the forward Euler method in time, where the solution at time Δt is approximated in cell j by u_j^1 defined as

$$\frac{u_j^1 - u_j^0}{\Delta t} + \frac{a}{\Delta x}\Big[\theta(u_j^0 - u_{j-1}^0) + (1-\theta)(u_{j+1}^0 - u_j^0)\Big] = 0 \ . \tag{1.13}$$

This scheme is parameterized by $\theta \in [0,1]$, which can be interpreted as taking a convex combination between the upwind and downwind discretization of the spatial derivative.

By analogy with the continuous energy, we can write the update induced on the discrete energy $E(u_j^1)$ by multiplying (1.13) with u_j^0. Remarking that for any $(a,b) \in \mathbb{R}^2$ there holds

$$a\,(b-a) = E(b) - E(a) - \frac{1}{2}(b-a)^2 \ ,$$

we obtain the following equation

$$\begin{aligned}\frac{E(u_j^1) - E(u_j^0)}{\Delta t} + \frac{a}{\Delta x}\Big[\theta(E(u_j^0) - E(u_{j-1}^0)) + (1-\theta)(E(u_{j+1}^0) - E(u_j^0))\Big] = \\ \frac{1}{2\Delta t}(u_j^1 - u_j^0)^2 - \frac{a}{2\Delta x}\Big[\theta(u_j^0 - u_{j-1}^0)^2 - (1-\theta)(u_{j+1}^0 - u_j^0)^2\Big] \ .\end{aligned}$$

Of the three terms appearing in the right hand side, only the second one, in factor of θ, has a negative sign no matter the value of $\theta \in [0,1]$. We say that it corresponds to the dissipation related to the upwinding. The third term is positive but can be made zero by taking $\theta = 1$, which is the upwind scheme. On the other hand we cannot get rid of the first term resulting from the time discretization. Since its sign is always positive, and because it is not always dominated by the upwind dissipation, the overall scheme (1.13) will increase the total energy in some occurrences. In other words, under no condition is the explicit scheme (1.13) ensured to dissipate the total energy.

Now consider the implicit version of the previous scheme, meaning that we use a backward Euler strategy consisting to evaluate the fluxes at time Δt

$$\frac{\widetilde{u}_j^1 - \widetilde{u}_j^0}{\Delta t} + \frac{a}{\Delta x}\Big[\theta(\widetilde{u}_j^1 - \widetilde{u}_{j-1}^1) + (1-\theta)(\widetilde{u}_{j+1}^1 - \widetilde{u}_j^1)\Big] = 0 \ . \tag{1.14}$$

Multiplying the scheme (1.14) by $\widetilde{u}_j^1$ and performing the same operations as before, the discrete energy is shown to satisfy

$$\begin{aligned}\frac{E(\widetilde{u}_j^1) - E(\widetilde{u}_j^0)}{\Delta t} + \frac{a}{\Delta x}\Big[\theta(E(\widetilde{u}_j^1) - E(\widetilde{u}_{j-1}^1)) + (1-\theta)(E(\widetilde{u}_{j+1}^1) - E(\widetilde{u}_j^1))\Big] = \\ -\frac{1}{2\Delta t}(\widetilde{u}_j^1 - \widetilde{u}_j^0)^2 - \frac{a}{2\Delta x}\Big[\theta(\widetilde{u}_j^1 - \widetilde{u}_{j-1}^1)^2 - (1-\theta)(\widetilde{u}_{j+1}^1 - \widetilde{u}_j^1)^2\Big] \ .\end{aligned} \tag{1.15}$$

Now we see that all the terms on the right hand side can be made negative when $\theta = 1$. Especially, no constraint on the time step is required. We loose this property if $\theta < 1$.

The use of implicit time integrator can make the upwind scheme ($\theta = 1$) unconditionally stable in the sense that it is granted to always dissipate the total energy. An analogy of this fact is encountered in the third chapter where we work with kinetic entropies to be introduced later.

In Chapter 2 we will among other things focus on a linear wave equation representing the propagation of surface waves in the two dimensional case. Implicit time stepping methods will again prove advantageous, and we illustrate why through a simplified example. We are interested in the solutions from $L^2(\mathbb{R})$ of

$$\begin{cases} \partial_t u + \partial_x v = 0 \\ \partial_t v + a\partial_x u = 0 \end{cases} \ . \tag{1.16}$$

In this case the total energy is given by $E_{\rm tot}(u) = a\,\|u\|_{L^2}^2 + \|v\|_{L^2}^2$ and remains constant in time. In fact we have using (1.16)

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|u\|_{L^2}^2 = \int_{\mathbb{R}} u\,\partial_t u\,\mathrm{d}x = -\int_{\mathbb{R}} u\,\partial_x v\,\mathrm{d}x = \int_{\mathbb{R}} v\,\partial_x u\,\mathrm{d}x = -\frac{1}{a}\int_{\mathbb{R}} v\,\partial_t v\,\mathrm{d}x = -\frac{1}{2a}\|v\|_{L^2}^2\ .$$

Now consider the explicit scheme with centered fluxes

$$\frac{u_j^1 - u_j^0}{\Delta t} + \frac{v_{j+1}^0 - v_{j-1}^0}{2\Delta x} = 0\ , \quad \frac{v_j^1 - v_j^0}{\Delta t} + a\,\frac{u_{j+1}^0 - u_{j-1}^0}{2\Delta x} = 0\ , \tag{1.17}$$

where we assume that sequences $(u_j^0)_{j\in\mathbb{Z}}$ and $(v_j^0)_{j\in\mathbb{Z}}$ are in $\ell^2(\mathbb{Z};\mathbb{R})$. Multiplying the first equality from (1.17) by au_j^0 and the second equality by v_j^0, and then summing both over $j\in\mathbb{Z}$ we find

$$\begin{aligned}\frac{a}{2\Delta t}\Big(\sum_{j\in\mathbb{Z}}(u_j^1)^2 - (u_j^0)^2 - (u_j^1 - u_j^0)^2\Big) + \frac{1}{2\Delta t}\Big(\sum_{j\in\mathbb{Z}}(v_j^1)^2 - (v_j^0)^2 - (v_j^1 - v_j^0)^2\Big) = \\ - \frac{a}{2\Delta x}\sum_{j\in\mathbb{Z}} u_j^0(v_{j+1}^0 - v_{j-1}^0) - \frac{a}{2\Delta x}\sum_{j\in\mathbb{Z}} v_j^0(u_{j+1}^0 - u_{j-1}^0)\ .\end{aligned} \tag{1.18}$$

The two sums on the right hand side cancel each other by performing a change of index

$$\sum u_j^0(v_{j+1}^0 - v_{j-1}^0) = \sum u_j^0\,v_{j+1}^0 - \sum u_j^0\,v_{j-1}^0 = \sum u_{j-1}^0\,v_j^0 - \sum u_{j+1}^0\,v_j^0\ . \tag{1.19}$$

We deduce from (1.18) and (1.19) that the discrete total energy associated with (1.17) satisfies

$$\frac{1}{\Delta t}\sum\Big(\Big[a(u_j^1)^2 + (v_j^1)^2\Big] - \Big[a(u_j^0)^2 + (v_j^0)^2\Big]\Big) = \frac{1}{\Delta t}\sum\Big[a(u_j^1 - u_j^0)^2 + (v_j^1 - v_j^0)^2\Big]\ .$$

The right hand side is always positive, thus the total energy increases. In order to stabilize this explicit scheme, one would need to add some numerical viscosity together with a CFL condition. On the other hand, if we consider the implicit version of (1.17) given by

$$\frac{\widetilde{u}_j^1 - \widetilde{u}_j^0}{\Delta t} + \frac{\widetilde{v}_{j+1}^1 - \widetilde{v}_{j-1}^1}{2\Delta x} = 0\ , \quad \frac{\widetilde{v}_j^1 - \widetilde{v}_j^0}{\Delta t} + a\,\frac{\widetilde{u}_{j+1}^1 - \widetilde{u}_{j-1}^1}{2\Delta x} = 0\ , \tag{1.20}$$

then the corresponding total energy satisfies

$$\frac{1}{\Delta t}\sum\Big(\Big[a(\widetilde{u}_j^1)^2 + (\widetilde{v}_j^1)^2\Big] - \Big[a(\widetilde{u}_j^0)^2 + (\widetilde{v}_j^0)^2\Big]\Big) = -\frac{1}{\Delta t}\sum\Big[a(\widetilde{u}_j^1 - \widetilde{u}_j^0)^2 + (\widetilde{v}_j^1 - \widetilde{v}_j^0)^2\Big]\ .$$

The last equality is obtained similarly as in the explicit case, only multiplying equalities from (1.20) by $a\widetilde{u}_j^1$ and $\widetilde{v}_j^1$ respectively.

When approximating the one dimensional wave equation (1.16) *by an explicit approach, one needs to add numerical viscosity and constrain the time step by a CFL condition in order to dissipate the total energy. On the contrary, the implicit scheme* (1.20) *dissipates the energy without requiring any additional viscosity nor CFL condition. In the second chapter we will see that having a numerical viscosity in the context of low Froude numbers results in poorly accurate results. An implicit time stepping method rids us from the need of having such a term, and much improved results will ensue. This is yet again an advantage of using implicit methods.*

1.4 State of the art and contributions

1.4.1 Low Froude accurate implicit-explicit schemes

In Chapter 2 we will consider the two dimensional case, where the quantities of interest depend on the spatial coordinates $(x, y) \in \mathbb{R}^2$ and where the scalar discharge q is replaced by the vector $Q(t, x, y) = (q, r)^T(t, x, y)$ of $\mathbb{R}^2$. Similarly to the one dimensional case, we introduce the horizontal velocity $V(t, x, y) \in \mathbb{R}^2$ so that we are able to write $Q = hV$. The corresponding system is given by

$$\begin{cases} \dfrac{\partial h}{\partial t} + \dfrac{\partial q}{\partial x} + \dfrac{\partial r}{\partial y} = 0 \\ \dfrac{\partial q}{\partial t} + \dfrac{\partial}{\partial x}\Big(\dfrac{q^2}{h} + \dfrac{g}{2}h^2\Big) + \dfrac{\partial}{\partial y}\Big(\dfrac{qr}{h}\Big) = -gh\dfrac{\partial z}{\partial x} \\ \dfrac{\partial r}{\partial t} + \dfrac{\partial}{\partial x}\Big(\dfrac{qr}{h}\Big) + \dfrac{\partial}{\partial y}\Big(\dfrac{r^2}{h} + \dfrac{g}{2}h^2\Big) = -gh\dfrac{\partial z}{\partial y} \end{cases} . \tag{1.21}$$

A dimensionless version of the 2D Saint-Venant system (3.78) is obtained by substituting the gravity constant g with the inverse of the quadratic characteristic Froude number Fr. If we denote by $\widetilde{U} = (\widetilde{h}, \widetilde{q}, \widetilde{r})^T \in \mathbb{R}^3$ the dimensionless physical quantities of order unity, then the dimensionless Saint-Venant system admits the following vector form

$$\partial_t \widetilde{U} + \nabla \cdot F(\widetilde{U}) = S(\widetilde{U}, \widetilde{z}) \ , \tag{1.22}$$

In what follows we will drop the tildes, and detail the dimensionless flux and source term

$$F(U) = \begin{pmatrix} hV^T \\ hV \otimes V + h^2 \mathbf{I}_2/(2\mathsf{Fr}^2) \end{pmatrix} \in \mathbb{R}^{3\times 2}, \qquad S(U, z) = \begin{pmatrix} 0 \\ -h\nabla z/\mathsf{Fr}^2 \end{pmatrix} \in \mathbb{R}^3 \ . \tag{1.23}$$

The flux in direction $n \in \mathbb{S}^2$ is given by $F(U)n$, and its jacobian $DF(U; n)$ admits the eigenvalues $\lambda_\pm(U; n) = (V \cdot n) \pm \sqrt{h}/\mathsf{Fr}$ and $\lambda_0(U; n) = (V \cdot n)$. In the low Froude regime, the celerity of the surface gravity waves $\sqrt{h}/\mathsf{Fr}$ appearing in $\lambda_\pm(U; n)$ is several orders of magnitude greater than the material velocity of particles $\|V\|$. For instance in coastal flows and lakes we frequently encounter Froude numbers around 10^{-2}. For stability purposes, it is well known that explicit finite volumes schemes require a CFL condition where, up to a multiplicative constant independent of Fr, the ratio between the time resolution Δt and the spatial resolution δ has to be smaller than the inverse of the greatest wave velocity

$$\frac{\Delta t}{\delta} \leq \mathrm{CFL} \times \frac{\mathsf{Fr}}{\mathsf{Fr}\|V\| + \sqrt{h}} \ .$$

This implies that the time step has to scale as $\mathcal{O}(\mathsf{Fr})$, which renders explicit methods inefficient when the characteristic Froude number becomes small, and even unusable in the vanishing limit $\mathsf{Fr} \to 0$. Note that the low Froude regime is analogous to the low Mach regime for the Euler and Navier-Stokes systems, and the literature on the topic is quite vast. We give a glimpse of it in the following lines.

State of the art. A way to get rid of the over-restrictive CFL condition is to use implicit time integrators. We are more specifically interested in implicit-explicit Runge-Kutta methods (IMEX-RK), which give the possibility to just implicit a linear part of the equations for efficiency purposes, while offering arbitrary high orders of accuracy depending on which Butcher tables are used. A general study on IMEX-RK methods has been conducted by Boscarino et al. in [16][15], where questions related to the order conditions, stability properties and the loss of accuracy in the asymptotic regime have been answered. A total variation diminishing (TVD) MOOD-based strategy was proposed by Michel-Dansac and coauthors first in the case of second order Butcher tables [29], and then in the case of arbitrary high order Butcher tables [53]. An IMEX-RK scheme is based on a splitting of the equations in a part conveying the slow dynamics — typically a convective phenomena, and a part representing the fast dynamics. Usually the splitting leads to a wave equation for the fast dynamics. This is the case of the splitting proposed by Giraldo and Restelli [34] and later used by Bispen et al. [14] in the context of evolution Galerkin schemes (FVEG). This splitting consists to linearize the Saint-Venant system around the lake at rest to obtain a linear fast surface waves operator. In a similar fashion Hack et al. [39] propose to decompose the pressure of the isentropic Navier-Stokes equations in a slow nonlinear and a fast linear contributions. Likewise a splitting of the pressure for the full Euler equations is studied by Noelle et al. in [54]. The idea of decomposing the pressure by performing a multiscale expansion stems from the work of Klein [46].

We say that a scheme is asymptotically stable when it produces stable results while using time increments that are uniform in the scale parameter. Another concern is related to the ability of the scheme to remain accurate with the model in the asymptotic regime. For this to be true, we need the error of the scheme to remain uniformly bounded with respect to the scale parameter, which is referred to as asymptotic consistency. When both the asymptotic stability and asymptotic consistency are satisfied, we obtain the notion of asymptotic preserving property (AP) introduced by Jin [43][42], and which means that the scheme converges to a consistent discretization of the limiting equations. To know how to design asymptotically consistent methods, we must first understand the behavior of the continuous solutions in the asymptotic regime. In [45], Klainerman and Majda established in a rigorous manner that the compressible Euler equations converge to an incompressible model in the low Mach limit, assuming that the initial condition is a perturbed incompressible state and under further technical assumptions on the functional spaces. In agreement with this result, Schochet [63] obtained an estimate stating that when the initial data is close to an incompressible state within a distance controlled by the Mach number, then the solution remains so. This characterization of the asymptotic regime carries to the low Froude Saint-Venant system over a flat bathymetry, and we can formally recover the incompressible limit in the periodic case by performing an expansion in powers of Froude

$$\begin{aligned} h(t,x,y\,;\mathsf{Fr}) &= h_{(0)}(t,x,y) + \mathsf{Fr}\, h_{(1)}(t,x,y) + \mathsf{Fr}^2\, h_{(2)}(t,x,y) + \mathcal{O}(\mathsf{Fr}^3) \\ V(t,x,y\,;\mathsf{Fr}) &= V_{(0)}(t,x,y) + \mathsf{Fr}\, V_{(1)}(t,x,y) + \mathsf{Fr}^2\, V_{(2)}(t,x,y) + \mathcal{O}(\mathsf{Fr}^3) \end{aligned} . \tag{1.24}$$

We then inject this expansion in (1.22), identify the powers of Froude alike and obtain

that the solution $(h, V)^T$ belongs to the set of well prepared data

$$\left\{ \sum_{k\in\mathbb{N}} \mathsf{Fr}^k \begin{pmatrix} h_{(k)} \\ V_{(k)} \end{pmatrix} : \mathbb{T}^2 \to \mathbb{R}^3,\ \nabla(h_{(0)} + z) = \nabla h_{(1)} = 0,\ \nabla \cdot (h_{(0)} V_{(0)}) = 0 \right\}. \quad (1.25)$$

When the bathymetry is flat, this set implies the incompressibility of the leading order term in the sense that we have $\nabla h_{(0)} = 0$ and thus $\nabla \cdot (h_{(0)} V_{(0)}) = h_{(0)} \nabla \cdot V_{(0)} = 0$. It is important to restrict to initial conditions that are well prepared in order to avoid initial boundary layers.

A key ingredient to achieve asymptotic consistency is for the numerical approximation to mimic the behavior of continuous solutions. With that in mind, one of the conditions is that a discrete version of (1.25) should be preserved by the scheme. Thanks to the estimate from Schochet [63] we can be more specific, as we know that in the limit $\mathsf{Fr} \to 0$ a solution of (1.22) arising from a nearly incompressible initial data remains close to the initial condition advected by the flow, at least when the bottom is flat. An interesting idea due to Dellacherie [27][28] is to check whether the modified PDE of a given scheme satisfies a linearized version of this property. The modified PDE is a system of equations that incorporates the error of consistency of the scheme back into the original model. Therefore it better describes the behavior of the scheme by accounting for its diffusion or dispersion. In [8][9] Audusse et al. applied this criterion to first order Godunov schemes for the Saint-Venant system with Coriolis source. The defect of these methods is well characterized by the proposed criterion, and comes from a wrong scaling in the pressure that unnecessarily increases the diffusion. This issue was already pointed to by Guillard and Viozat [38], and one solution is to center the discretization of the pressure term. In [2][3] Arun an coauthors study the properties of a second order IMEX-RK scheme applied to the Euler isentropic system, and prove that the criterion from Dellacherie is satisfied when linearizing the nonlinear convection operator. We also point to Barsukow's thesis [11] where a similar criterion was studied in the context of the linear wave equation.

Although we will not focus on them, we mention relaxation methods which have recently received an extensive coverage. These methods allow to design cheap asymptotically consistent schemes by enriching the model with relaxation variables. The latter are used to construct an augmented Riemann solver that remains accurate in the asymptotic regime. Generally, the relaxation is designed such that the Riemann problem only develops linearly degenerated waves, and is simpler to solve. In [21], Bouchut et al. proposed an explicit relaxation scheme satisfying the asymptotic preserving property for the barotropic Euler equations. The scheme is based on the Suliciu relaxation system with two velocities. Usually an explicit scheme is not asymptotically stable, but in this case the CFL condition was made independent of the scale parameter at the expense of becoming parabolic (i.e. $\Delta t \leq \mathrm{CFL} \times \delta^2$). This scheme has been adapted to the full Euler equations in [22] and uses a semi-implicit time integrator. Relaxation techniques offer a favorable context for IMEX integrators, as one can choose to implicit the linear update of the relaxation variables, while the physical variables are treated with the explicit nonlinear Riemann solver. A slightly different relaxation strategy was proposed by Berthon et al. [12] together with a fully implicit time integrator. The difference with the previous technique is that an additional relaxation term for the pressure is used

to cure the excessive diffusion. An IMEX-RK version of this scheme was proposed by Klingenberg et al. in [47], and was then extended to the case with gravitational source term in [65]. Another possibility is to use a Lagrange-projection strategy to decouple the slow material waves form the fast acoustic ones and treat them with an IMEX solver. A Suliciu-type relaxation can then be applied to solve the acoustic waves accurately. We refer to the thesis of Girardin [35] and to the work of Chalons et al. [25] for more details on this last approach.

Contributions. We consider an IMEX-RK strategy for discretizing (1.22). The splitting we use is the one from Giraldo and Restelli [34], which is given by

$$\partial_t U + \nabla \cdot H(U,z) + [\nabla \cdot (F-H) - S](U,z) = 0 \ , \tag{1.26}$$

where $H(U,z) \in \mathbb{R}^{3\times 2}$ is a convective flux defined by

$$H(U,z) = \begin{pmatrix} 0 \\ hV \otimes V + \frac{1}{2\mathsf{Fr}^2}(h+z)^2 \, \mathbf{I}_2 \end{pmatrix} \ , \tag{1.27}$$

and where we introduce $L(U,z) := [\nabla \cdot (F-H) - S](U,z)$ a linear operator representing the propagation of surface waves

$$L(U,z) = \begin{pmatrix} 0 & 1 & 0 \\ -z/\mathsf{Fr}^2 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \frac{\partial U}{\partial x} + \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -z/\mathsf{Fr}^2 & 0 & 0 \end{pmatrix} \frac{\partial U}{\partial y} + \begin{pmatrix} 0 \\ (-z/\mathsf{Fr}^2)\partial_x z \\ (-z/\mathsf{Fr}^2)\partial_y z \end{pmatrix} \ . \tag{1.28}$$

We emphasize on the advantages offered by the decomposition (1.27)–(1.28) of the Saint-Venant flux, which we enumerate below.

- When replacing either H or L by zero in (1.26), the resulting system remains hyperbolic and admits a conservative writing for the water height and for the discharge when the bathymetry is flat;
- For any unit vector $n \in \mathbb{S}^2$, the convective flux $H(U,z)n$ has all its eigenvalues independent of Fr and is thus adapted for an explicit treatment;
- The operator L is linear. More specifically it coincides with the Saint-Venant system linearized around the lake at rest $h+z=0$, $V=(0,0)^T$. The linearity of L is important as it will be treated implicitly, and means that the computational cost of inverting the associated matrix will be smaller than for a nonlinear operator;

We focus on the periodic case and restrict to initial conditions belonging to the set of well prepared data (1.25), which are relevant when considering the low Froude regime. One of our goal is to predict whether a scheme is accurate at low Froude numbers, and explain why it is so. To this end we study the modified PDE of our scheme for an arbitrary IMEX-RK method when the convection part is neglected ($H \equiv 0$) and when the bathymetry is flat. This modified PDE will take the form

$$\partial_t U + LU = (R_{\Delta t} - R_\delta)U \ , \tag{1.29}$$

where $R_{\Delta t}$ and R_δ are respectively the error of consistency related to the time and spatial discretizations. We believe that it is reasonable to neglect H as it only accounts for the slow dynamics, and an eventual loss of accuracy will rather come from the wave operator L. We recall that the criterion proposed by Dellacherie (Theorem 2.2 in [28]) and used by Arun et al. in [2][3] consists to check whether the modified PDE (1.29) leaves the set of incompressible states

$$\mathcal{E} \overset{\text{def}}{=} \left\{ \begin{pmatrix} h \\ V \end{pmatrix} \in (L^2(\mathbb{T}^2))^3,\ \nabla h = 0,\ \nabla \cdot V = 0 \right\} \tag{1.30}$$

invariant. This is sufficient to ensure that a nearly incompressible data $(h, V)^T \in \mathcal{E} + \mathcal{O}(\mathsf{Fr})$ remains so throughout the time. We propose a refined version of this criterion by doing the following remarks

1. The wave equation $(\partial_t + L)U = 0$ admits $\mathcal{E}$ as a set of steady states. This means that rather than checking if the modified PDE (1.29) leaves $\mathcal{E}$ invariant, we should be more specific and verify if its solutions $U = (h, hV)^T$ satisfy

$$(h, V)^T(t = 0, \cdot) \in \mathcal{E} \implies U(t \geq 0, \cdot) = U(0, \cdot)\ . \tag{1.31}$$

   This argument is compatible with Proposition 2.1 from [28] since we set the convection flux H to zero.

2. In reality we will not consider flows that are exactly incompressible, but nearly incompressible, and we want the solutions U of the modified PDE (1.29) to satisfy

$$(h, V)^T(t = 0, \cdot) \in \mathcal{E} + \mathcal{O}(\mathsf{Fr}) \implies \|U(t \geq 0, \cdot) - U(0, \cdot)\|_{L^2} = \mathcal{O}(\mathsf{Fr})\ . \tag{1.32}$$

   This condition corresponds to the one encountered in Theorem 1.3 from [27]. Property (1.31) acts as a sufficient condition for (1.32), but is somewhat restrictive. We relax it with the following criterion

$$(h, V)^T(t = 0, \cdot) \in \mathcal{E} \implies \|U(t \geq 0, \cdot) - U(0, \cdot)\|_{L^2} = \mathcal{O}(\mathsf{Fr})\ . \tag{1.33}$$

We call (1.33) the *low Froude accuracy* criterion and show by a simple triangle inequality that it is still a sufficient condition to have (1.32). We then obtain the following statement

Proposition 1.4.1. *Consider a time semi-discretization scheme using a IMEX-RK method. Without any assumption on the Butcher tables, this scheme is low Froude accurate. This means that when neglecting the convective flux* (1.27)*, the modified PDE of the corresponding scheme keeps any data initially incompressible close to the initial condition. In addition the set of well prepared data* (1.25) *is left invariant by the update of the IMEX-RK scheme when the Butcher tables are* globally stiffly accurate *and when the implicit table is of type A or CK.*

The terminology related to Butcher tables will be reviewed later in Chapter 2. What enables us to obtain Proposition 1.4.1 is that in the modified PDE (1.29) we have $R_\delta = 0$ (no spatial discretization has been made) and we can show that

$$(h, V)^T \in \mathcal{E} \implies (h, hV)^T \in \ker L \cap \ker R_{\Delta t}\ .$$

Still in the time semi-discrete case, when neglecting the convective flux H we prove that the modified PDE (1.29) decreases the L^2 energy under some algebraic condition on the Butcher table. It is important to satisfy this condition in the perspective of having the asymptotic stability, and which allows the use of large time steps. As expected, we check that common implicit Butcher tables verify this condition, but not explicit ones. The better stability of implicit schemes in the setting of the wave equation was suggested through the toy problem (1.16), and so the generalization to the 2D case with arbitrary high order Butcher tables is not a surprise.

Proposition 1.4.1 is valuable, as it indicates that in the context of a fully discrete scheme, an eventual loss of accuracy cannot come from the IMEX-RK time discretization. We study various spatial discretizations for the surface waves operator (1.28). It appears that using a spatial discretization with an order of accuracy strictly larger than that of the time discretization prevents the loss of asymptotic consistency. Indeed a consequence of a higher order accuracy in space is that R_δ becomes negligible compared to $R_{\Delta t}$, and we can apply the same reasoning that led to Proposition 1.4.1. When using first order methods, this is compatible with the solution proposed in [8][9][28], which consists to center the pressure discretization. We verify this with the first order IMEX-Euler method, and compare a centered difference discretization with a Godunov scheme that adds numerical viscosity. In presence of viscosity, the method is first order in space and strongly degrades the approximation when the Froude number becomes small. On the other hand, in absence of viscosity the surface waves operator is discretized at second order, and the results seem insensible to the scale parameter (asymptotic consistency). These numerical results can be seen in the two middle plots from Figure 1.4. We are able to justify that the refined low Froude accuracy condition (1.33) is satisfied by the centered scheme but not by the Godunov scheme.

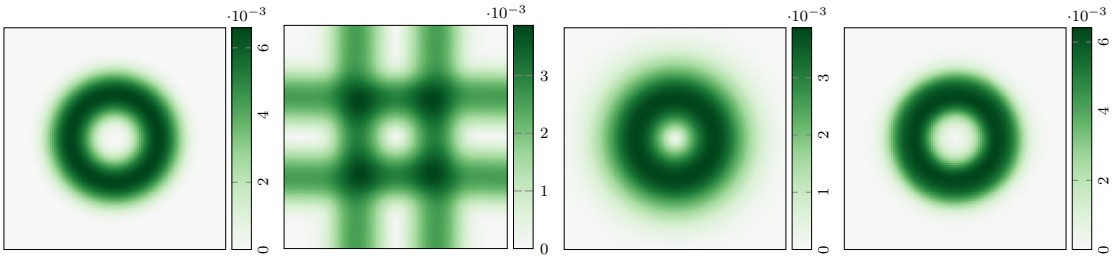

Figure 1.4: Approximated local Froude number for the Gresho vortex. From left to right: reference solution, IMEX-Euler with upwind flux, IMEX-Euler with centered flux, ARS-(2,2,2) with centered flux.

Interestingly, a centered discretization of the surface wave operator L combined to a second order IMEX-RK method produces accurate approximations, see the right plot in Figure 1.4. This outcome is not so obvious to predict since the spatial error operator R_δ can no more be neglected, and we verify that incompressible states do not belong to $\ker R_\delta$ in general. Yet we obtain the proposition below.

Proposition 1.4.2. *Consider a second order IMEX-RK method. When combined with a centered discretization of the surface waves operator* (1.28)*, we get a scheme that is*

low Froude accurate. Provided that the IMEX-RK method is globally stiffly accurate and that the implicit Butcher table is of type A or CK, a discrete version of the set of well prepared data (1.25) *is preserved during each update.*

This good result is explained by the fact that the error in time somehow dominates the error in space. To validate our approach, we also compare this scheme to discretizations of the surface gravity waves operator involving modified stencils used in [11][41]. These stencils offer the advantage to yield a modified PDE satisfying the more restrictive sufficient condition (1.31). Despite this we weren't able to detect any noticeable difference between these discretizations and the standard centered scheme, which is a good point for this latter. It is also an additional argument confirming the ability of the low Froude accuracy criterion to predict whether a scheme remains accurate or not in the asymptotic regime.

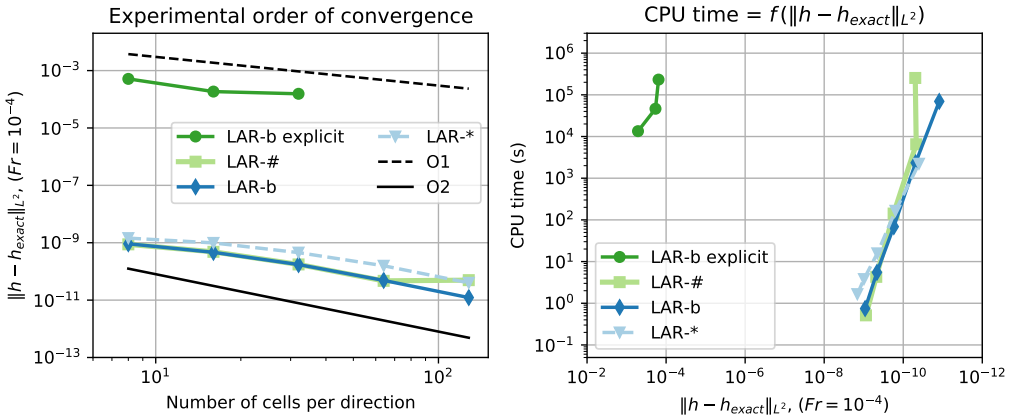

Figure 1.5: Error and efficiency curves for the Gresho testcase. LAR-$\star$ refers to the ARS-(2,2,2) scheme with a centered discretization of the surface gravity waves. The two other schemes LAR-$\flat$ and LAR-$\sharp$ use the same ARS-(2,2,2) double Butcher table, but with a modified second order stencils for the surface gravity waves.

All considered schemes preserve the lake at rest stationary states. Efficiency wise, the use of semi-implicit methods over explicit ones seems to become advantageous for Froude numbers smaller than 10^{-2}, at least for the Gresho testcase. The error and efficiency curves of various schemes have been compiled in Figure 1.5 for the case $\mathsf{Fr} = 10^{-4}$. We see that despite using a modified stencil referred to previously, the second order explicit scheme produces much less accurate results than its semi-implicit counterparts. In fact its error is more than six orders of magnitudes greater for an equivalent computational cost. This is partly explained by the fact that small times steps have to be used to achieve stability.

We finish by mentioning that implementation for the two dimensional case was done in Matlab. Great care has been brought to re-usability through the use of classes, such that previous functionalities can be accessed without re-implementing them in every scheme. For instance, new Butcher tables can be set by just specifying their coefficients,

and no additional line of code is needed. The program is fully vectorized for greater efficiency.

1.4.2 Implicit kinetic schemes and iterative methods

The Saint-Venant system is a macroscopic law that predicts the evolution of quantities of interest observable at our scale. Yet the underlying mechanism can be described more precisely at the mesoscopic scale by the mean of Boltzmann-type kinetic equations. The kinetic theory aims at describing a cloud of particles such as a gas or a fluid. To this end we introduce a positive function $f(t,x,\xi)$ representing the density of particles which, at time t, are located at the point x and travel with velocity $\xi \in \mathbb{R}$. We call f the distribution function, and when integrating it with respect to the velocity variable we recover macroscopic quantities of interest. An overview on kinetic theory can for instance be found in the book from Perthame [55].

Both the macroscopic Saint-Venant system and the kinetic description are part of the field of continuum mechanics, where the particles constitutive of the matter (here the fluid) fill the whole space. However in the kinetic setting one also accounts for the binary collisions occurring between particles of different velocities. The evolution of the distribution function is governed by the kinetic equation

$$\partial_t f + \xi \partial_x f = Q[f](t,x,\xi)/\varepsilon \ , \tag{1.34}$$

where $Q[f]$ is a collision operator accounting for the evolution of the population of particles, and $1/\varepsilon$ is the collision frequency. Several choices are possible for the collision term, and we consider one of the simplest given by the BGK operator (Bhatnagar, Gross and Krook)

$$Q_{\mathrm{BGK}}[f](t,x,\xi) \overset{\mathrm{def}}{=} M(U_f(t,x),\xi) - f(t,x,\xi) \ , \quad U_f = \int_{\mathbb{R}} \begin{pmatrix} 1 \\ \xi \end{pmatrix} f(t,x,\xi)\,\mathrm{d}\xi \ .$$

Over a given period of time, the number of collisions increases when ε decreases, and in the limit $\varepsilon \to 0$ the distribution of particles converges to some hydrodynamic equilibrium $M(U_f,\xi)$ that we refer to as the Gibbs equilibrium or maxwellian distribution. The Gibbs equilibrium plays a crucial role as it allows to link the kinetic equation description to the Saint-Venant system by recovering the macroscopic quantities of interest U and the Saint-Venant flux $F(U)$ when integrating. In fact we can for instance consider

$$M(U,\xi) = \frac{1}{g\pi}\sqrt{\left(2gh - (\xi-u)^2\right)_+} \quad \Longrightarrow \quad \int_{\mathbb{R}} \begin{pmatrix} 1 \\ \xi \\ \xi^2 \end{pmatrix} M(U,\xi)\,\mathrm{d}\xi = \begin{pmatrix} h \\ F(U) \end{pmatrix} \ . \tag{1.35}$$

In (1.35), when integrating $M(U,\xi)$ against ξ we get the first component of $F(U)$ which coincides with hu the second component of U. Then the Saint-Venant system (1.1) is obtained by integrating the kinetic representation given by

$$\partial_t M + \xi \partial_x M - g z' \partial_\xi M = Q(t,x,\xi) \ , \qquad \int_{\mathbb{R}} \begin{pmatrix} 1 \\ \xi \end{pmatrix} Q(t,x,\xi)\,\mathrm{d}\xi = 0 \ , \tag{1.36}$$

against the vector $(1, \xi)^T$. In the above $Q(t, x, \xi)$ has to be understood as the distribution obtained by taking the limit of $Q_{\mathrm{BGK}}[f]/\varepsilon$ when ε vanishes. More details about the relation between (1.1) and (1.36) can be found in [58]. The upside is that we replaced the initial nonlinear system of two equations by a linear scalar transport equation with collision term, at the price of having to compute an integral. An interesting idea is to propose a simple solver for (1.36), and then to integrate it over the velocities $\xi \in \mathbb{R}$ in order to obtain a numerical scheme for the macroscopic law. This is the approach followed by *kinetic schemes*, and we propose to combine them with implicit time integrators.

The kinetic schemes we consider are based on the fact that over flat bathymetries, the kinetic interpretation (1.36) is obtained by relaxing the kinetic equation (1.34) in the limit $\varepsilon \to 0$. In practice we perform a BGK splitting, meaning that instead of directly discretizing (1.34), we alternate between a collision step and a transport step written as

$$\begin{cases} \partial_t f = (M(U_f, \xi) - f)/\varepsilon \\ \partial_t f + \xi \partial_x f = 0 \end{cases} .$$

In the limit $\varepsilon \to 0$, the collision step projects the initial data onto the space of maxwellians, which can then be transported in the second step. This alternating process gives rise to the transport-projection method from [18]. In our case we discretize the transport step by an upwind implicit scheme

$$\frac{f_i^{1-} - M_i^0}{\Delta t} + \frac{\xi}{\Delta x}\Big(\mathbb{1}_{\xi<0}(f_{i+1}^{1-} - f_i^{1-}) + \mathbb{1}_{\xi>0}(f_i^{1-} - f_{i-1}^{1-})\Big) = 0 \ , \tag{1.37}$$

with $M_i^0 = M(U_i^0, \xi)$. As suggested in the example of the previous toy problem (1.12), having an upwind discretization of the flux is important even when using an implicit time stepping algorithm, as we can otherwise increase some energy. In order to write the update at the macroscopic level, we need to invert the system defined by (1.37) and compute the integrals against 1 and ξ. Hence over a mesh of N cells the updates $h^1 \in \mathbb{R}^N$ and $(hu)^1 \in \mathbb{R}^N$ take the form

$$h^1 = \int_{\mathbb{R}} (\mathbf{I} + \sigma \mathbf{L}(\xi))^{-1}[M^0 + \sigma B](\xi)\,\mathrm{d}\xi \ , \quad (hu)^1 = \int_{\mathbb{R}} \xi(\mathbf{I} + \sigma \mathbf{L}(\xi))^{-1}[M^0 + \sigma B](\xi)\,\mathrm{d}\xi \ , \tag{1.38}$$

with $\mathbf{I} + \sigma \mathbf{L}(\xi)$ a triangular matrix from $\mathbb{R}^{N\times N}$ and $B(\xi) \in \mathbb{R}^N$ a vector accounting for the boundary conditions. We discuss these terms more in detail in Chapter 3, and state the main result below.

Proposition 1.4.3. *The update* (1.38) *is conservative and consistent with the Saint-Venant system, and keeps the water height positive. Besides, it admits a discrete entropy inequality which always dissipates the energy when the coefficients of M^0 are half-disk maxwellians given by* (1.35)*. These properties hold without any condition on the time step.*

We will see that proving the existence of a discrete entropy inequality is akin to obtaining equality (1.15) for the discrete energy of the toy problem. The difference is that now the

proof is made harder by requiring the use of a kinetic entropy, to be defined in Chapter 3. Although we are able to obtain an analytical expression for the inverse of the mass matrix $\mathbf{I}+\sigma\mathbf{L}(\xi)$, in practice it is not possible to compute the integrals (1.38) when using the half-disk maxwellian (1.35). Substituting this reference maxwellian with

$$M(U,\xi) = \frac{h}{2\sqrt{3}\,c}\mathbb{1}_{|\xi-u|\leq\sqrt{3}\,c} \tag{1.39}$$

enables to explicitly compute formulas (1.38) while retaining the positivity of the scheme. It can be implemented numerically with a quadratic cost with respect to the number of cells. This is due to the fact that we lose the sparsity of the mass matrix when inverting it.

Still in the case without bathymetry, we consider the alternative offered by an iterative method to approximate (1.36). The iterative process is based on a Gauss-Jacobi decomposition and reads

$$f^0(\xi) = M^0\ , \qquad \left\{ \begin{array}{l} (1+\alpha)f^{k+1}(\xi) = (\alpha\mathbf{I}-\sigma\mathbf{L})M^k + M^0 + \sigma B[M^k] \\ M^{k+1} = M(\{U_f\}^{k+1},\xi) \end{array} \right. , \tag{1.40}$$

with k denoting the iteration index and $\alpha \geq 0$ a relaxation parameter. This time we don't need to invert any matrix, and the integrals of update (1.40) against 1 and ξ can be expressed explicitly when using the half-disk maxwellian (1.35). The drawback is that we will need a CFL condition on the time step to ensure both the convergence and the positivity. Moreover we get the following proposition giving the existence of a kinetic entropy inequality.

Proposition 1.4.4. *Assuming that it converges, the sequence $(f^k(\xi))_k \subset \mathbb{R}^N$ defined by* (1.40) *satisfies a kinetic entropy inequality that dissipates the energy from some rank.*

The iterative approach is also useful to treat varying bathymetries which we investigate next. The source term is treated through the hydrostatic reconstruction introduced in [6]. Without entering too much into the details, the kinetic interpretation of the hydrostatic reconstruction proposed in [10] introduces nonlinear terms in the kinetic scheme (1.37). Due to this we must use an iterative approach similar to (1.40) and given below.

$$\left\{ \begin{array}{l} (1+\alpha)f_i^{k+1} = M_i^0 + \alpha M_i^k - \sigma\xi(M_{i+1/2}^k - M_{i-1/2}^k) + \sigma(\xi-u_i^k)(M_{i+1/2-}^k - M_{i-1/2+}^k) \\ M_i^{k+1} = M\big(\{U_f\}_i^{k+1},\xi\big) \\ M_{i\pm1/2}^{k+1} = \mathbb{1}_{\xi<0}M\big(\{U_f\}_{i\pm1/2+}^{k+1},\xi\big) + \mathbb{1}_{\xi>0}M\big(\{U_f\}_{i\pm1/2-}^{k+1},\xi\big) \end{array} \right. \tag{1.41}$$

The macroscopic quantities $\{U_f\}_{i+1/2-}$ and $\{U_f\}_{i+1/2+}$ denote the reconstructed states in the left and right neighborhoods of interface $i+1/2$ following the hydrostatic reconstruction procedure to be recalled in Chapter 3. We are able to prove the positivity of this method as well as an estimate on the kinetic entropy. This estimate guarantees the dissipation of energy from some rank, which is an improvement over the explicit version of the scheme. In fact, Audusse et al. proved in [10] that the explicit kinetic scheme with hydrostatic

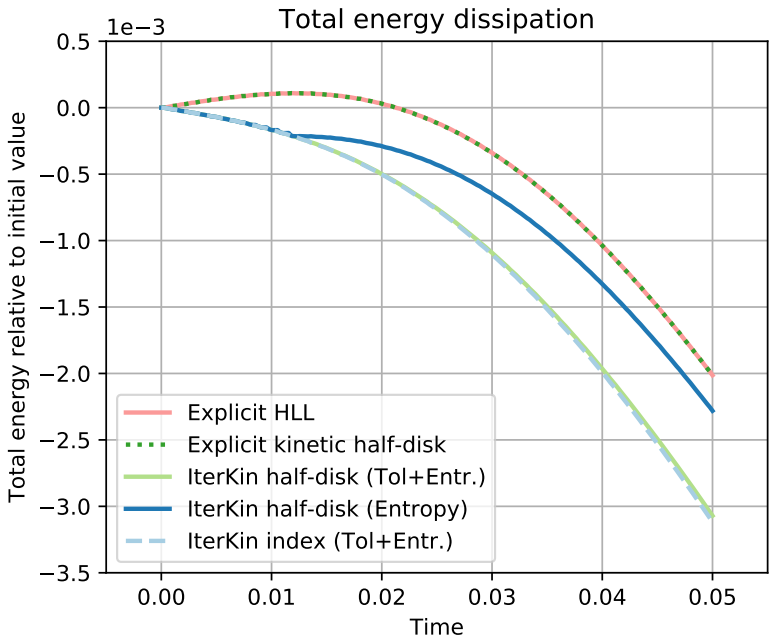

Figure 1.6: Dissipation of total energy is achieved by the iterative kinetic schemes, but not by the explicit method. The testcase is given by a varying bathymetry, and at initial time both the velocity and free surface elevation are constant. Periodic boundary conditions are used.

reconstruction admits a discrete entropy inequality with an error term that can sometimes increase the energy. The benefit of using iterative kinetic schemes is well illustrated by the testcase presented in Figure 1.6. To finish we demonstrate the convergence of our algorithm under some assumptions relative to the boundedness of the iterated solution h^k and $(hu)^k$. One of the assumption is that the water height should be bounded away from zero. We believe there is hope to get rid of this restriction, as the iterative process performed quite well on the parabolic bowl testcase featured in Figure 1.7. This testcase is characterized by an evolving wet/dry front over a non flat bottom, which is rather complex to resolve. Therefore in the near future we would like to investigate another way to obtain the convergence of the iterative scheme that includes the possibility of having dry areas.

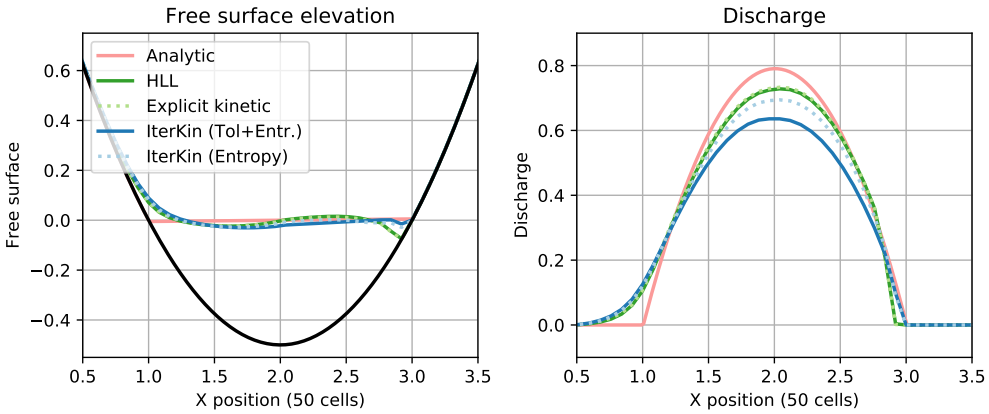

Figure 1.7: Comparing kinetic strategies on the parabolic bowl testcase.

Chapter 2

Low Froude accurate implicit-explicit schemes

2.1 Introduction

The goal of this article is to design an efficient accurate numerical scheme for the Saint-Venant system under low Froude regime in two spatial dimensions. A good comprehension of this model benefits a wide spectrum of applications, ranging from water management, interplay between ocean dynamics and climate change, erosion of the coastline, as well as forecasting natural disasters among other examples.

At low Froude numbers, the Saint-Venant system models flows featuring a slow convection of fluid particles and a fast propagation of surface gravity waves. The velocity attributed to these two phenomena vary drastically, inducing a multiscale behavior in time. For reference, in coastal flows the order of magnitude between these characteristic speeds is about 10^{-2}. Because of such a great disparity, usual Godunov methods with explicit time integration are inefficient to approximate low Froude flows, since the CFL condition required for stability dictates a prohibitively small time step. Indeed, the latter is restricted by the inverse of the maximum wave speed, which effectively means that the scheme tries to resolve the scale of faster dynamics. As a result, numerical computations are significantly slowed down, and the increased number of iterations implies a surge in numerical diffusion, regardless of the order. A solution is then to implicit the resolution of the fast dynamics to enable the use of larger time steps, and justifies the recent interest of implicit-explicit (IMEX) methods [3][14][15][22][25][34][39][53][65].

By analogy with the Euler system, as the Froude number vanishes the shallow water equations with flat bottom converge, at least for suitable initial conditions and periodic boundary conditions, towards an incompressible-like system in the sense that the water height can be seen as a density. A rigorous proof of this convergence result can be found in the work of Klainerman and Majda [45], and a compatible estimate is stated by Schochet [63]. Hence it is important to satisfy this property at the discrete level, and for that reason asymptotic preserving schemes (AP) seem to be judicious, as they converge towards a consistent discretization of the limiting system, see [42] for a review on the topic. On the contrary, a non AP scheme could have a bad behavior, with for instance the emergence of a wrong scaling in the approximated pressure [38], or the formation of spurious oscillations caused by the loss of near-incompressibility as studied by Dellacherie in [28]. Thus the AP property is of paramount importance in achieving accurate results at low Froude numbers, and a key ingredient in this regard is the ability of a scheme to preserve nearly incompressible states. Several methods satisfying this criterion have been proposed recently. In [28][9][8], the first order case is handled through the cancelling of the numerical diffusion associated to the discharge. However one has to be worried about the stability of the method when removing diffusive terms, and it might not work for the Saint-Venant system without any regularizing contribution such as the Coriolis force. Instead, the approach exploited in [41][11] consists in adding terms rather than removing diffusion, and seems a safer way to go. At second order, the scheme proposed in [3] requires no modification but the splitting used is nonlinear in the acoustic operator, and is very costly to solve implicitly. We also want to mention relaxation strategies which are suitable for designing AP schemes, see for instance [12][21][22][25][47][65].

When designing schemes for the Saint-Venant system, great care has to be brought to the preservation of the lake at rest steady state. In fact, many geophysical flows can

be seen as a perturbation of this state, which is given by a flat free surface and a zero discharge. Numerically, we need to ensure that the pressure variation is exactly balanced by the source term, and schemes entering this scope are said to be well balanced (WB). This has proven challenging and numerous works have been devoted to this issue, with for instance the hydrostatic reconstruction [6] and its kinetic interpretation [10]. In the case of IMEX methods, where it is not desirable to apply nonlinear reconstructions to the implicit update, the system has to be splitted appropriately so that lakes at rest are naturally accounted for.

In the light of the above, we strive to design a high order IMEX scheme offering a linear treatment of surface gravity waves, stable under a convection-driven CFL, well balanced, AP, and keeping the compressible component of the flow under control. To achieve this, the present document is organized as follows. The second section is dedicated to a recall of the main literature results concerning the limiting system, AP property and the stability of nearly-incompressible states. As we shall see, the latter will be easier to characterize for the linear surface waves system, and will give rise to the *low Froude accuracy* criterion. In the third section, a wave splitting is proposed and IMEX time semi-discretizations using *diagonally implicit Runge-Kutta* methods (IMEX-DIRK) are studied. We show that in this framework, schemes are always low Froude accurate and well balanced. Regarding the asymptotic consistency, it is obtained formally for a specific class of Butcher tableaux. When neglecting the convection operator, the use of scale-independent time steps is granted to lead to an L^2 stable semi-discretization of the survace waves under some algebraic condition on the corresponding Butcher table. In the next section, we take a look at fully discretized schemes. We illustrate the well known defect of a standard first order upwinded method, and check that the low Froude accuracy criterion is able to detect it. We also discuss a simple fix consisting in having an order of accuracy in space striclty greater to that of the time discretization. Then, we consider a second order scheme in time and space with centered discretization of the surface gravity waves. Thanks to the notion of low Froude accuracy introduced earlier, we justify the good results obtained with this method for small Froude numbers. Interestingly the centered discretization of the surface waves doesn't satisfy the criterion from Dellacherie (Theorem 2.2 in [28]), which to a certain extent is more restrictive than the low Froude accuracy condition. Finally, to illustrate its good behavior, we compare the standard centered discretization to modified stencils proposed in the litterature, and get similar results for scale independent time steps over the Gresho vortex testcase.

2.2 The low Froude singular limit

2.2.1 Dimensionless formulation of the Saint-Venant system

The Saint-Venant system, also known as the shallow water system, describes the conservation of the water height $h(t,x,y)$ and the balance of the discharge vector

$Q(t,x,y) = (q,r)^T(t,x,y)$ according to the following set of equations:

$$\begin{cases} \dfrac{\partial h}{\partial t} + \dfrac{\partial q}{\partial x} + \dfrac{\partial r}{\partial y} = 0 \\ \dfrac{\partial q}{\partial t} + \dfrac{\partial}{\partial x}\Big(\dfrac{q^2}{h} + \dfrac{g}{2}h^2\Big) + \dfrac{\partial}{\partial y}\Big(\dfrac{qr}{h}\Big) = -gh\dfrac{\partial z}{\partial x} \\ \dfrac{\partial r}{\partial t} + \dfrac{\partial}{\partial x}\Big(\dfrac{qr}{h}\Big) + \dfrac{\partial}{\partial y}\Big(\dfrac{r^2}{h} + \dfrac{g}{2}h^2\Big) = -gh\dfrac{\partial z}{\partial y} \end{cases} \tag{2.1}$$

where $z(x,y)$ is the fixed bathymetry, and g the gravitational acceleration constant. It is also usual to consider the free surface $\zeta(t,x,y) := h + z$ and the fluid velocity $V(t,x,y) = (v,w)^T := Q/h$. We detail the dimensionless formulation for the Saint-Venant system so as to exhibit the dominant terms in the low Froude regime. For this we introduce the dimensionless quantities:

$$\tilde{t} = \frac{t}{t^*}, \quad \tilde{x} = \frac{x}{l^*}, \quad \tilde{y} = \frac{y}{l^*}, \quad \tilde{h} = \frac{h}{h^*}, \quad \widetilde{V} = \frac{V}{u^*}, \quad \widetilde{Q} = \frac{Q}{h^*u^*}, \quad \tilde{z} = \frac{z}{h^*}$$

The star denotes a strictly positive characteristic constant chosen such that the dimensionless quantities are of order one. In this new set of coordinates, the equation (3.78) reads:

$$\begin{cases} \dfrac{h^*}{t^*}\dfrac{\partial \tilde{h}}{\partial \tilde{t}} + \dfrac{h^*u^*}{l^*}\left(\dfrac{\partial \tilde{q}}{\partial \tilde{x}} + \dfrac{\partial \tilde{r}}{\partial \tilde{y}}\right) = 0 \\ \dfrac{h^*u^*}{t^*}\dfrac{\partial \tilde{q}}{\partial \tilde{t}} + \dfrac{h^*(u^*)^2}{l^*}\left(\dfrac{\partial \tilde{h}\tilde{v}^2}{\partial \tilde{x}} + \dfrac{\partial \tilde{h}\tilde{v}\tilde{w}}{\partial \tilde{y}}\right) + \dfrac{g(h^*)^2}{2l^*}\dfrac{\partial \tilde{h}^2}{\partial \tilde{x}} = -\dfrac{g(h^*)^2}{l^*}\tilde{h}\dfrac{\partial \tilde{z}}{\partial \tilde{x}} \\ \dfrac{h^*u^*}{t^*}\dfrac{\partial \tilde{r}}{\partial \tilde{t}} + \dfrac{h^*(u^*)^2}{l^*}\left(\dfrac{\partial \tilde{h}\tilde{v}\tilde{w}}{\partial \tilde{x}} + \dfrac{\partial \tilde{h}\tilde{w}^2}{\partial \tilde{y}}\right) + \dfrac{g(h^*)^2}{2l^*}\dfrac{\partial \tilde{h}^2}{\partial \tilde{y}} = -\dfrac{g(h^*)^2}{l^*}\tilde{h}\dfrac{\partial \tilde{z}}{\partial \tilde{y}} \end{cases}$$

Dividing the first equality by h^*u^*/l^* and the momentum equations by $h^*(u^*)^2/l^*$, defining $\mathsf{Fr} = u^*/\sqrt{gh^*}$, $\mathsf{St} = l^*/(u^*t^*)$ and dropping the tildes for the sake of legibility we get:

$$\begin{cases} \mathsf{St}\,\dfrac{\partial h}{\partial t} + \dfrac{\partial q}{\partial x} + \dfrac{\partial r}{\partial y} = 0 \\ \mathsf{St}\,\dfrac{\partial q}{\partial t} + \left(\dfrac{\partial hv^2}{\partial x} + \dfrac{\partial hvw}{\partial y}\right) + \dfrac{1}{2\mathsf{Fr}^2}\dfrac{\partial h^2}{\partial x} = -\dfrac{h}{\mathsf{Fr}^2}\dfrac{\partial z}{\partial x} \\ \mathsf{St}\,\dfrac{\partial r}{\partial t} + \left(\dfrac{\partial hvw}{\partial x} + \dfrac{\partial hw^2}{\partial y}\right) + \dfrac{1}{2\mathsf{Fr}^2}\dfrac{\partial h^2}{\partial y} = -\dfrac{h}{\mathsf{Fr}^2}\dfrac{\partial z}{\partial y} \end{cases} \tag{$\mathcal{P}_{\mathsf{Fr}}$}$$

The characteristic quantities Fr and St correspond respectively to the Froude and the Strouhal numbers. The choice $\mathsf{St} = \mathcal{O}(\mathsf{Fr}^{-1})$ is attributed to an acoustic time scale, whereas $\mathsf{St} = \mathcal{O}(1)$ is rather associated with convective phenomena. Hence it is this second choice that we make. We will also adopt the more compact vector notation for system ($\mathcal{P}_{\mathsf{Fr}}$) as given below:

$$\frac{\partial U}{\partial t} + \nabla \cdot F(U) = S(U,z)$$

where the flux F and source term S are defined by:

$$F(U) = \begin{pmatrix} hV^T \\ hV \otimes V + h^2 \mathbf{I}_2/(2\mathsf{Fr}^2) \end{pmatrix} \in \mathbb{R}^{3\times 2}, \qquad S(U,z) = \begin{pmatrix} 0 \\ -h\nabla z/\mathsf{Fr}^2 \end{pmatrix} \in \mathbb{R}^3 \tag{2.2}$$

For any vector n belonging to the unit sphere $\mathbb{S}^2$, we denote $F(U;n) = F(U)n$ the directional flux, whose Jacobian with respect to the variables (h, Q) is given by:

$$DF(U;n) = \begin{pmatrix} 0 & n^T \\ (h\mathbf{I}_2/\mathsf{Fr}^2 - V\otimes V)n & V\otimes n + (V\cdot n)\mathbf{I}_2 \end{pmatrix} \in \mathbb{R}^{3\times 3} \tag{2.3}$$

The corresponding eigenvalues are given for $j \in \{-1, 0, 1\}$ by $\lambda_j(U;n) = V \cdot n + jc$, where $c = \sqrt{h}/\mathsf{Fr}$ is the velocity of surface waves. Especially, we see that the system is strictly hyperbolic in wet areas, where we have $h > 0$. Classical explicit finite volumes schemes will be stable under the usual CFL condition $\delta/\Delta t \geq 2\,\lambda_{\max}$ with $\lambda_{\max}$ being the maximum wave velocity arising from the discretized data on a cartesian mesh $\cup_{i,j} C_{i,j}$. For approximate Riemann solvers this velocity is estimated as an upper bound of the aforementioned eigenvalues, hence we need to have over every interface:

$$\frac{\delta}{\Delta t} \geq 2\,\|V\|\Big(1 + \frac{c}{\|V\|}\Big)$$

with δ the spatial mesh resolution and $\|V\|$ the greatest velocity along the interface normal in neighboring cells. This is problematic in the low Froude regime, corresponding to the case where $c\,\|V\| \gg 1$, meaning that gravity waves are propagating much faster than material waves. In this regime, the time step will be severely restricted by the fast dynamics, whereas interest might only concern convection-type/convective phenomena.

We dedicate the remainder of this section to present the asymptotic properties of the dimensionless system $(\mathcal{P}_{\mathsf{Fr}})$.

2.2.2 The limiting equations

In order to investigate the solutions of $(\mathcal{P}_{\mathsf{Fr}})$ in the limit $\mathsf{Fr} \to 0$, a standard practice [14][39][65] is to assume that they admit the following expansion with respect to the Froude number:

$$\begin{aligned} h(t,x,y\,;\mathsf{Fr}) &= h_{(0)}(t,x,y) + \mathsf{Fr}\,h_{(1)}(t,x,y) + \mathsf{Fr}^2\,h_{(2)}(t,x,y) + \mathcal{O}(\mathsf{Fr}^3) \\ V(t,x,y\,;\mathsf{Fr}) &= V_{(0)}(t,x,y) + \mathsf{Fr}\,V_{(1)}(t,x,y) + \mathsf{Fr}^2\,V_{(2)}(t,x,y) + \mathcal{O}(\mathsf{Fr}^3) \end{aligned} \tag{2.4}$$

It is important to note that the terms $h_{(j)}, V_{(j)}$ appearing above do only depend on time and space variables, but not on the Froude number. Inserting this in $(\mathcal{P}_{\mathsf{Fr}})$, we identify/extract the terms in factor of different Froude powers/magnitudes. Terms in Fr^{-2} give:

$$\frac{1}{2\mathsf{Fr}^2}\nabla(h_{(0)})^2 = -\frac{1}{\mathsf{Fr}^2}h_{(0)}\nabla z \implies h_{(0)}\nabla(h_{(0)} + z) = 0 \tag{2.5}$$

This means that the leading order free surface should remain flat in wet areas. Now considering the terms in Fr^0 in the water height conservation law we have:

$$\frac{\partial h_{(0)}}{\partial t} + \frac{\partial (h_{(0)} v_{(0)})}{\partial x} + \frac{\partial (h_{(0)} w_{(0)})}{\partial y} = 0 \tag{2.6}$$

From (2.5) we know that $h_{(0)} + z$ is constant in space, and so is $\partial_t (h_{(0)} + z) = \partial_t h_{(0)}$. Making use of this, we integrate the above equation over the spatial domain Ω and apply Gauss' divergence formula:

$$0 = |\Omega| \frac{\partial h_{(0)}}{\partial t} + \iint_\Omega \nabla \cdot (h_{(0)} V_{(0)}) \, \mathrm{d}x\mathrm{d}y = |\Omega| \frac{\partial h_{(0)}}{\partial t} + \int_{\partial\Omega} h_{(0)} V_{(0)} \cdot n_{\partial\Omega} \, \mathrm{d}\sigma$$

Assuming either that $\Omega = \mathbb{T}^2$, or that the non-penetration condition $V \cdot n_{\partial\Omega} = 0$ hold at the boundary, the last integral cancels out and thus $h_{(0)}$ is time independent. If $\Omega = \mathbb{R}^2$, the same conclusion holds assuming the sub-linear growth condition is satisfied by the discharge, that is to say $hV = o(|x|, |y|)$. As a consequence, Equation (2.6) becomes a divergence-free condition on the leading order discharge:

$$\nabla \cdot Q_{(0)} = \nabla \cdot (h_{(0)} V_{(0)}) = 0 \tag{2.7}$$

Next, we collect terms in Fr^{-1} from the momentum balance equation of $(\mathcal{P}_{\mathsf{Fr}})$, resulting in the relation:

$$\frac{1}{\mathsf{Fr}} \nabla \cdot (h_{(0)} h_{(1)} \mathbf{I}_2) = -\frac{1}{\mathsf{Fr}} h_{(1)} \nabla z \implies h_{(0)} \nabla h_{(1)} = 0 \tag{2.8}$$

Finally, terms in Fr^0 taken from the same equation form the relation:

$$\begin{aligned}
&\frac{\partial}{\partial t}(h_{(0)} V_{(0)}) + \nabla \cdot (h_{(0)} V_{(0)} \otimes V_{(0)} + [h_{(0)} h_{(2)} + (h_{(1)})^2/2] \mathbf{I}_2) = -h_{(2)} \nabla z \\
\implies &\frac{\partial}{\partial t}(h_{(0)} V_{(0)}) + h_{(0)} (V_{(0)} \cdot \nabla) V_{(0)} + h_{(0)} \nabla h_{(2)} = 0 \qquad (2.9) \\
\implies &\frac{\partial}{\partial t} V_{(0)} + (V_{(0)} \cdot \nabla) V_{(0)} + \nabla h_{(2)} = 0 \, . \qquad (2.10)
\end{aligned}$$

To get the second equality we have used that $\nabla \cdot (h_{(0)} V_{(0)} \otimes V_{(0)}) = h_{(0)} (V_{(0)} \cdot \nabla) V_{(0)}$ due to the divergence-free condition (2.7). The last equality is a consequence of the time independence of $h_{(0)}$ under adequate boundary conditions. The limiting system of $(\mathcal{P}_{\mathsf{Fr}})$ shall be noted $(\mathcal{P}_0)$ and is obtained, at least formally, by gathering the equations satisfied by the leading order terms $h_{(0)}, V_{(0)}$. In accordance with equations (2.5) and (2.7), we introduce the following set of functions:

$$\mathbb{W} \stackrel{\text{def}}{=} \{(\overline{h}, \overline{V}) : \mathbb{T}^2 \to \mathbb{R}^3, \ \nabla(\overline{h} + z) = 0, \ \nabla \cdot (\overline{h}\, \overline{V}) = 0\} \, . \tag{2.11}$$

Together with Equation (2.10), we get the following writing of the limiting system:

$$\left\{ \begin{array}{l} \forall t > 0, \ (\overline{h}(t, \cdot), \overline{V}(t, \cdot)) \in \mathbb{W} \\ \dfrac{\partial}{\partial t} \overline{V} + (\overline{V} \cdot \nabla) \overline{V} + \nabla \Pi = 0 \end{array} \right. \tag{$\mathcal{P}_0$}$$

The term Π in $(\mathcal{P}_0)$ coincides with the remaining term $h_{(2)}$ from (2.10) and can be seen as a pressure. At least when the initial condition (h^0, V^0) belongs to $\mathbb{W}$, we can relate this pressure to the leading order terms by taking the divergence of (2.9), leading to:

$$(\nabla h_{(0)} + h_{(0)}\nabla) \cdot [\nabla h_{(2)} + (V_{(0)} \cdot \nabla)V_{(0)}] = 0 \tag{2.12}$$

If additionally the bathymetry is flat, we get a Laplace equation on $h_{(2)}$. On the other hand, if the incompressibile constraint isn't initially fulfilled by the leading order terms, a boundary layer might appear for $0 < \mathsf{Fr} \ll 1$.

Remark 2.2.1. *Some important remarks are in order:*

1. *The condition* (2.8) *on the second order term $h_{(1)}$ doesn't appear in the limiting system $(\mathcal{P}_0)$. However for small but non zero Froude numbers, it is important to take into consideration as it implies that the fluctuations of the free surface elevation $\zeta = h + z$ are in $\mathcal{O}(\mathsf{Fr}^2)$, that is to say $\nabla\zeta = \mathcal{O}(\mathsf{Fr}^2)$. For a flat bathymetry, we also have that the spatial variations of the pressure $p(h) = h^2/2$ are in $\mathcal{O}(\mathsf{Fr}^2)$. This is true because $p(h) = h_{(0)}^2/2 + \mathsf{Fr}h_{(0)}h_{(1)} + \mathsf{Fr}^2(h_{(1)}^2/2 + h_{(0)}h_{(2)}) + \mathcal{O}(\mathsf{Fr}^3)$ and $\nabla h_{(0)} = \nabla h_{(1)} = 0$ in that case. For this reason, we will have to consider the so-called set of* well prepared data *defined below, whose denomination will become clearer in the next point:*

$$\mathbb{W}_{\mathrm{p}} = \left\{ \sum_{k\in\mathbb{N}} \mathsf{Fr}^k \begin{pmatrix} h_{(k)} \\ V_{(k)} \end{pmatrix} : \mathbb{T}^2 \to \mathbb{R}^3, \ \begin{pmatrix} h_{(0)} \\ V_{(0)} \end{pmatrix} \in \mathbb{W} \text{ and } \nabla h_{(1)} = 0 \right\} \tag{2.13}$$

2. *We have only provided a formal derivation of the limiting system $(\mathcal{P}_0)$, based on the assumption that expansion* (2.4) *exists. Its existence in the case without source term was justified rigorously by Klainerman and Majda. More precisions about the required functional spaces and convergence results can be found in their paper [45]. Without entering too much into the details, one of the requirements is for the initial pressure to fluctuate in $\mathcal{O}(\mathsf{Fr}^2)$. Thus if the initial condition $(h, V)(t = 0, \cdot)$ belongs to $\mathbb{W}_{\mathrm{p}}$, we will say that it is well prepared since we recover a similar constraint over the pressure.*

3. *In the case where the initial data is not well prepared, the solution isn't granted to admit expansion* (2.4) *anymore. In this situation a boundary layer appears, in which the solution transitions into the set of well prepared data, see for instance [67]. In this work we will only focus on well prepared data, where it makes sense to use large times steps.*

2.2.3 Asymptotic preserving property

Intuitively, for well prepared initial data the solutions of $(\mathcal{P}_{\mathsf{Fr}})$ are good approximations to the solutions of $(\mathcal{P}_0)$ when the Froude number is small, because we have $(h, V) = (h_{(0)}, V_{(0)}) + \mathcal{O}(\mathsf{Fr})$. With this in mind, we would like to construct numerical schemes complying with this asymptotic behavior. More specifically, if $\mathcal{P}_{\delta,\mathsf{Fr}}$ is a consistent discretization of $\mathcal{P}_{\mathsf{Fr}}$ associated to some time or spatial step/resolution δ, we would like

the limit scheme $\mathcal{P}_{\delta,0}$ to yield a consistent approximation of the continuous system $(\mathcal{P}_0)$. Keeping the consistency while taking the low Froude limit indicates that the consistency error should be uniform in Fr. Otherwise, the consistency error would blow up in the limit $\mathsf{Fr} \to 0$, δ fixed. In a similar fashion, we ask for the region of stability of the scheme to be scale independent. Such methods are called asymptotic preserving, see [42] for in depth explanations on this concept. The AP property can be summarized by the commuting diagram in Figure 2.1, and motivates the definition below.

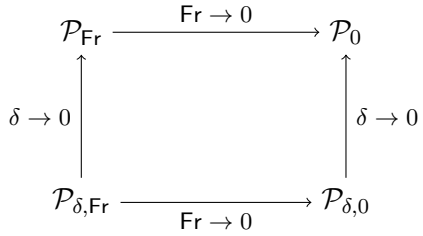

Figure 2.1: Asymptotic preserving diagram.

Definition 2.2.2 (Asymptotic preserving property)**.** *A numerical discretization $\mathcal{P}_{\delta,\mathsf{Fr}}$ of $(\mathcal{P}_{\mathsf{Fr}})$ is said to be* asymptotic preserving *if it satisfies the two following conditions:*

1. *It is* asymptotically consistent*, meaning that the limit discretization $\mathcal{P}_{\delta,0}$ results in a consistent discretization of the continuous limiting system $(\mathcal{P}_0)$.*

2. *It is* asymptotically stable*, meaning that the stability constraint on the time step has to be scale-independent.*

We have seen that for well prepared data, the leading order height and discharge satisfy the two static equations $\nabla(h_{(0)} + z) = 0$ and $\nabla \cdot (h_{(0)} V_{(0)}) = 0$. In the next section, we further characterize how close the solutions (h, V) of $(\mathcal{P}_{\mathsf{Fr}})$ are from the incompressible set of data $\mathbb{W}$.

2.2.4 Stability of nearly incompressible states

In this section, we recall important invariance and convergence results related to the set of well prepared data, due to Schochet [63] and Dellacherie [28]. These results have been obtained for the barotropic Euler equations without source term, for which the Saint-Venant system with flat bathymetry constitutes one particular case. Hence we will assume the absence of varying bottom thereafter. In this setting, note that the space $\mathbb{W}$ defined in (2.11) is now the set of space-constant water heights h and divergence-free velocities V. Seeing h as a density, $\mathbb{W}$ can thus be considered as the set of incompressible states.

Let us motivate the analysis by remarking that any initial condition $(h, V)(t = 0, \cdot) \in \mathbb{W}_{\mathrm{p}}$ parameterized by Fr admits a limit $(h_{(0)}, V_{(0)})$ in $\mathbb{W}$ as Fr vanishes. This means that for small Froude numbers, the parameterized initial condition will be close to $\mathbb{W}$. The relevant question is whether or not the state $(h, V)(t > 0, \cdot\,; \mathsf{Fr})$ remains close to $\mathbb{W}$ at

later times for a fixed small Froude number. To answer this question, we will decompose the solutions in an incompressible part and a compressible (or acoustic) one. We shall see that the compressible component is in some way controlled by the initial data, ensuring the invariance of nearly-incompressible states. For this we introduce the incompressible and acoustic L^2 energy spaces defined by:

$$\mathcal{E} \stackrel{\text{def}}{=} \left\{ \begin{pmatrix} h \\ V \end{pmatrix} \in (L^2(\mathbb{T}^2))^3,\ \nabla h = 0,\ \nabla \cdot V = 0 \right\}$$

$$\mathcal{E}^\perp \stackrel{\text{def}}{=} \left\{ \begin{pmatrix} h \\ V \end{pmatrix} \in (L^2(\mathbb{T}^2))^3,\ \iint_{\mathbb{T}^2} h \,\mathrm{d}x\mathrm{d}y = 0,\ \exists \phi \in H^1(\mathbb{T}^2), V = \nabla\phi \right\}.$$

Every function in $(L^2(\mathbb{T}^2))^3$ can be written uniquely as the sum of one element of $\mathcal{E}$ and one element of $\mathcal{E}^\perp$. This is known as the Helmoltz-Leray decomposition

$$(L^2(\mathbb{T}^2))^3 = \mathcal{E} \oplus \mathcal{E}^\perp .$$

Note also that $\mathcal{E}$ is the set of square integrable functions lying in $\mathbb{W}$. Because the spaces $\mathcal{E}$ and $\mathcal{E}^\perp$ are orthogonal for the usual scalar product on L^2, we can introduce the orthogonal projection, or Leray projector, defined by

$$\forall U \in (L^2(\mathbb{T}^2))^3,\ P_{\mathcal{E}} U \stackrel{\text{def}}{=} U_{\mathcal{E}} \quad \text{with} \quad U_{\mathcal{E}} \in \mathcal{E},\ \text{and}\ U - U_{\mathcal{E}} \in \mathcal{E}^\perp .$$

It is then interesting to recast system $(\mathcal{P}_{\mathsf{Fr}})$ in (h, V) coordinates as below:

$$\frac{\partial U}{\partial t} + \mathcal{K}(U) + \mathcal{G}(U) = 0, \tag{2.14}$$

$$U = \begin{pmatrix} h \\ V \end{pmatrix}, \quad \mathcal{K}(U) = (V \cdot \nabla)U, \quad \mathcal{G}(U) = \begin{pmatrix} h(\nabla \cdot V) \\ \mathsf{Fr}^{-2} \nabla h \end{pmatrix}$$

Operator $\mathcal{K}$ accounts for convective phenomena occurring over a time scale of order $\mathcal{O}(1)$, whereas $\mathcal{G}$ stands for the gravity waves propagating over a short time scale of order $\mathcal{O}(\mathsf{Fr})$. We are now able to recall the convergence result from Schochet [63]:

Theorem 2.2.3 (Schochet, 1994)**.** *Let $U = (h, V)^T$ and $U^* = (h^*, V^*)$ be respective solutions of*

$$\begin{cases} \partial_t U + \mathcal{K}(U) + \mathcal{G}(U) = 0 \\ U(t = 0, \cdot) = U^0(\cdot) \end{cases} \tag{2.15}$$

$$\begin{cases} \partial_t U^* + P_{\mathcal{E}} \mathcal{K}(U^*) = 0 \\ U^*(t = 0, \cdot) = P_{\mathcal{E}} U^0(\cdot) \end{cases} \tag{2.16}$$

on $\mathbb{T}^2$. Then $U^*(t \geq 0, \cdot) \in \mathcal{E}$, and

$$\begin{cases} \|h - h_{\mathcal{E}}\|_{L^2}(t = 0) = \mathcal{O}(\mathsf{Fr}^2) \\ \|V - V_{\mathcal{E}}\|_{L^2}(t = 0) = \mathcal{O}(\mathsf{Fr}) \end{cases} \Longrightarrow \begin{cases} \|h - h^*\|_{L^2}(t > 0) = \mathcal{O}(\mathsf{Fr}^2) \\ \|V - V^*\|_{L^2}(t > 0) = \mathcal{O}(\mathsf{Fr}) \end{cases}$$

In Theorem 2.2.3, the projected system (2.16) is equivalent to the incompressible Euler system with constant density, and thus coincides with problem $(\mathcal{P}_0)$. In fact we have

$U^*(t,x,y) = P_{\mathcal{E}}\big(U^0(x,y) - \int_0^t \mathcal{H}(U^*)(s,x,y)\,\mathrm{d}s\big)$ which belongs to $\mathcal{E}$ at all times. As a consequence we get $\mathcal{H}(U^*) = (0,(V\cdot\nabla)V)^T$ and $P_{\mathcal{E}}\mathcal{H}(U^*) = (0,\ (V\cdot\nabla)V+\psi)^T$ with ψ a vector field satisfying both $\nabla\cdot[(V\cdot\nabla)V+\psi] = 0$ and the existence of a scalar function ϕ such that $\psi = \nabla\phi$. Up to a constant, the valid choice for ϕ coincides with the pressure Π appearing in (2.12). Hence Theorem 2.2.3 ensures us that if the initial condition is close to an incompressible state, the solution U remains close to being incompressible at later times. However, the closeness is stated with respect to the solution U^* of the projected system which is not practical to manipulate, and we would like to have a simpler criterion. For that reason, in [28] Dellacherie studies the linearization of Equation (2.14) around the constant-in-space state $(\overline{h},\overline{V})$ with $\overline{h}\geq 0$:

$$\frac{\partial U}{\partial t} + KU + GU = 0, \qquad KU = (\overline{V}\cdot\nabla)U, \quad GU = \begin{pmatrix} \overline{h}\nabla\cdot V \\ \mathsf{Fr}^{-2}\nabla h \end{pmatrix} \tag{2.17}$$

Defining the incompressible and acoustic energies

$$E_{\mathcal{E}}(t) = \frac{1}{\mathsf{Fr}^2}\|h_{\mathcal{E}}\|_{L^2}^2 + \overline{h}\,\|V_{\mathcal{E}}\|_{L^2}^2\ , \qquad E_{\mathcal{E}^\perp}(t) = \frac{1}{\mathsf{Fr}^2}\|h-h_{\mathcal{E}}\|_{L^2}^2 + \overline{h}\,\|V-V_{\mathcal{E}}\|_{L^2}^2\ ,$$

the following result is given in [28]:

Theorem 2.2.4 (Dellacherie, 2010)**.** *Let $U=(h,V)^T$ and $U^*=(h^*,V^*)^T$ be respective solutions of*

$$\begin{cases} \partial_t U + KU + GU = 0 \\ U(t=0,\cdot) = U^0(\cdot) \end{cases} \tag{2.18}$$

$$\begin{cases} \partial_t U^* + KU^* = 0 \\ U^*(t=0,\cdot) = P_{\mathcal{E}}U^0(\cdot) \end{cases} \tag{2.19}$$

Then $P_{\mathcal{E}}U = U^*$, and the incompressible and acoustic energies remain constant in time:

$$\frac{\mathrm{d}}{\mathrm{d}t}E_{\mathcal{E}} = \frac{\mathrm{d}}{\mathrm{d}t}E_{\mathcal{E}^\perp} = 0\,.$$

As a consequence we recover the estimate

$$\begin{cases} \|h-h_{\mathcal{E}}\|_{L^2}(t=0) = \mathcal{O}(\mathsf{Fr}^2) \\ \|V-V_{\mathcal{E}}\|_{L^2}(t=0) = \mathcal{O}(\mathsf{Fr}) \end{cases} \Longrightarrow \begin{cases} \|h-h^*\|_{L^2}(t>0) = \mathcal{O}(\mathsf{Fr}^2) \\ \|V-V^*\|_{L^2}(t>0) = \mathcal{O}(\mathsf{Fr}) \end{cases}$$

since in this case $E_{\mathcal{E}^\perp}(t\geq 0) = E_{\mathcal{E}^\perp}(t=0) = \mathcal{O}(\mathsf{Fr}^2)$.

The projected system is now equivalent to a linear transport equation on the divergence-free velocity, combined with a constant-in-time water height. More precisely, the solution U^* of (2.19) is given by $U^0_{\mathcal{E}}\circ\gamma$ with the characteristic γ defined as

$$\gamma : (t,x,y)\in\mathbb{R}_+\times\mathbb{T}^2 \longmapsto (x - t\,\overline{V}_x, y - t\,\overline{V}_y)\ . \tag{2.20}$$

The key ingredient of Theorem 2.2.4 is that Equation (2.17) is $\mathcal{E}$- and $\mathcal{E}^\perp$-invariant, see [28] for a proof. This sheds some light on why basic schemes fail to capture the correct behavior by introducing unwanted oscillations. In fact, what might happen in the linear case (2.17) is that the numerical method fails to satisfy a discrete analogy of the

$\mathcal{E}$-invariance. This is especially problematic when this failure occurs in the discretization of the wave operator G, since it corresponds to the fast dynamics and will rapidly amplify the oscillations of the compressible component, eventually making it dominant. The ability of a scheme to preserve or not the space $\mathcal{E}$ will be assessed by its modified PDE [68], whose definition is recalled below.

Definition 2.2.5 (Modified PDE)**.** *A p^{th} order modified PDE associated to a scheme is an equation whose solutions are approximated by that scheme up to $\mathcal{O}(\delta^{p+2})$ terms, with δ the time and/or space resolution.*

According to this definition, a modified PDE better models the behavior of the scheme when compared to the original problem (2.17). However the latter usually differs from the former since modified PDEs incorporate some consistency error, either under the form of diffusive or dispersive terms. Because it is necessary to consider those new terms, Theorem 2.2.4 cannot be used directly, and we need the more general result from [28]:

Theorem 2.2.6 (Dellacherie, 2010)**.** *Let $\mathcal{F}$ denote a linear differential operator such that the following equation is well-posed on $L^\infty(\mathbb{R}_+; (L^2(\mathbb{T}^2))^3)$*

$$\partial_t U + \mathcal{F}U = 0 \ . \tag{2.21}$$

Let U and U^* be solutions of (2.21) with respective initial condition $U(t=0,\cdot) = U^0(\cdot)$ and $U^*(t=0,\cdot) = P_\mathcal{E} U^0(\cdot)$. Then the following holds:

1. *$\|U^0 - P_\mathcal{E} U^0\|_{L^2} = \mathcal{O}(\mathsf{Fr}) \Longrightarrow \|U - U^*\|_{L^2}(t \geq 0) = \mathcal{O}(\mathsf{Fr})$ where the time $t = \mathcal{O}(1)$ is bounded as $\mathsf{Fr} \to 0$. Since $\mathcal{E}$ is not left invariant, in general U^* does not belong to $\mathcal{E}$ and thus $U^* \neq P_\mathcal{E} U$.*

2. *Assume $\mathcal{F}$ is such that $(\partial_t + \mathcal{F})U = 0$ leaves $\mathcal{E}$ invariant. Then we can substitute U^* with $P_\mathcal{E} U$ in the point above.*

In Theorem 2.2.6, equation $\partial_t U + \mathcal{F}U = 0$ stands for the modified PDE. Analogously to the original system (2.17), the $\mathcal{E}$-invariance seems to be a key ingredient enabling the modified PDE to preserve nearly incompressible states. In fact this is a sufficient condition to get the relevant estimate implied in the second point of Theorem 2.2.6. On the other hand, for the first point to hold we only need the well-posedness of the modified PDE, implying that the dependence of the solution on the initial condition is smooth.

2.2.5 The near stationary condition

Theorem 2.2.6 has been used to detect and fix the inaccuracy observed with first order upwinded schemes. In [8][9][28][38] the proposed solution consists in removing the diffusion on the discharge component in order to meet the requirement of the sufficient condition found in Theorem 2.2.6, while hoping to retain the stability property of the scheme. In [11] a similar objective was achieved, although instead of removing diffusive terms, additional crossed partial derivatives were discretized, such that $\mathcal{E}$ is encompassed in the kernel of the resulting diffusion operator. This latter solution potentially causes

less issues regarding the stability in the case of explicit methods, and can be extended to higher order methods on cartesian meshes.

Although the mentioned modifications are successful in getting much improved results at low Froude numbers, we believe that the criterion from Theorem 2.2.6 could be refined and made more accurate in its ability to detect and explain an eventual loss of consistency in the asymptotic regime. In fact we can do the following remarks.

1. Theorem 2.2.6 ensures the solution of the modified PDE (2.21) remains close to some incompressible data, but without specifying which one. In this regard, Theorem 2.2.4 is stronger because it implies that solutions should remain close to an incompressible state determined from the initial condition. Indeed the incompressible state in question is given by $U^0_{\mathcal{E}} \circ \gamma$ the solution of (2.19), with γ defined in (2.20). In other words the linear PDE $(\partial_t + K + G)U = 0$ acts on the elements of $\mathcal{E}$ by translating them at speed $\overline{V}$. Therefore instead of asking the modified PDE to keep $\mathcal{E}$ invariant as in the second point of Theorem 2.2.6, we can be more specific and verify if its solutions U satisfy

$$U(t=0,\cdot) \in \mathcal{E} \implies U(t \geq 0, \cdot) = U(0, \gamma(t,\cdot)) \ . \tag{2.22}$$

   This way we are able to detect the situations where the solution deviates from $U^0 \circ \gamma$ even if it remains in $\mathcal{E}$.

2. In practice we will consider small Froude numbers that are strictly positive, and the initial condition need not be exactly incompressible. Hence we want to know if the modified PDE (2.21) satisfies

$$U(t=0,\cdot) \in \mathcal{E} + \mathcal{O}(\mathsf{Fr}) \implies \|U(t \geq 0, \cdot) - U_{\mathcal{E}}(0, \gamma(t,\cdot))\|_{L^2} = \mathcal{O}(\mathsf{Fr}) \ , \tag{2.23}$$

   which is an analogy of the implication found in Theorem 2.2.6. The property (2.22) constitutes a sufficient condition with respect to (2.23). However (2.22) is restrictive and can be relaxed through a small modification

$$U(t=0,\cdot) \in \mathcal{E} \implies \|U(t \geq 0, \cdot) - U(0, \gamma(t,\cdot))\|_{L^2} = \mathcal{O}(\mathsf{Fr}) \ . \tag{2.24}$$

   Thanks to the last condition (2.24) we will now be able to detect situations where initially incompressible solutions deviate *too quickly* from $U^0 \circ \gamma$ while remaining in $\mathcal{E} + \mathcal{O}(\mathsf{Fr})$. Such an example will be encountered in Section 2.4.3 with the first order Rusanov scheme.

We retain the condition (2.24) as a good indicator to predict whether a scheme can remain accurate in the low Froude regime. We have the following result which confirms that (2.24) is a sufficient condition for (2.23) to hold.

Theorem 2.2.7. *Let $\mathcal{F}$ be a linear differential operator such that* (2.21) *is well-posed. We assume that any solution U of* (2.21) *with initial condition U^0 satisfies the near stationary condition*

$$U^0 \in \mathcal{E} \Longrightarrow \|U - U^0 \circ \gamma\|_{L^2}(t \geq 0) = \mathcal{O}(\mathsf{Fr}) \ . \tag{2.25}$$

Then the following holds for any convective time τ bounded as $\mathsf{Fr} \to 0$

$$\|U^0 - P_{\mathcal{E}} U^0\|_{L^2} = \mathcal{O}(\mathsf{Fr}) \Longrightarrow \|U - U^0_{\mathcal{E}} \circ \gamma\|_{L^2}(\tau) = \mathcal{O}(\mathsf{Fr}) \ .$$

The proof results from a simple triangle inequality.

Proof. (Theorem 2.2.7). Let U and U^* be solutions of (2.21) with respective initial conditions U^0 and $P_{\mathcal{E}}U^0$, and let $\tau = \mathcal{O}(1)$ be a positive time. As in Theorem 2.2.6 the well-posedness of the linear modified PDE (2.21) implies that

$$\|U^0 - P_{\mathcal{E}}U^0\|_{L^2} = \mathcal{O}(\mathsf{Fr}) \Longrightarrow \|U - U^*\|_{L^2}(\tau) = \mathcal{O}(\mathsf{Fr}) \ .$$

Furthermore, for any time $t \geq 0$, the distance between $U(t,\cdot)$ and $U^0_{\mathcal{E}} \circ \gamma(t,\cdot)$ can be bounded from above using the following triangle inequality:

$$\|U - U^0_{\mathcal{E}} \circ \gamma\|_{L^2} = \|(U - U^*) + (U^* - U^0_{\mathcal{E}} \circ \gamma)\|_{L^2} \leq \|U - U^*\|_{L^2} + \|U^* - U^0_{\mathcal{E}} \circ \gamma\|_{L^2} \ .$$

But since U^* initially belongs to $\mathcal{E}$, we have by the assumption (2.25) that

$$\|U^* - U^0_{\mathcal{E}} \circ \gamma\|_{L^2}(\tau) = \mathcal{O}(\mathsf{Fr}) \ ,$$

and thus we get $\|U - U^0_{\mathcal{E}} \circ \gamma\|_{L^2}(\tau) = \mathcal{O}(\mathsf{Fr})$. □

As suggested before, the differential operator $\mathcal{F}$ represents the consistent part $K + G$ in addition to some diffusive or dispersive error terms introduced by the scheme. One could also neglect the operator K by setting it to zero, since it does not act over acoustic time scales and thus shouldn't be a cause for concern. Note that this would amount to linearize Equation (2.14) around the state $(\overline{h}, \overline{V})$ with $\overline{V}$ the null velocity, in which case we also have $\gamma(t,\cdot) = \mathrm{id}$ for any time t. In this situation, the condition (2.23) corresponds to (1.6) in [27], meaning that we want the modified PDE to keep the solutions close to their initial condition if the latter are nearly incompressible. This explains why we referred to (2.25) as *near stationary condition*. Again, the purpose here is to emphasize on the fact that the near stationary condition can be seen as a requirement for having accurate results as $\mathsf{Fr} \to 0$ in the linear case. Despite this simplification, we hope that this condition remains a good indicator of accuracy in the setting of nonlinear schemes.

In the next section, we will consider a wave splitting differing from the one suggested by Equation (2.14). The main reason for this is that we want a linear representation of the surface waves in order to obtain an efficient IMEX scheme. Despite this change, the incompressible set $\mathcal{E}$ will still be included in the kernel of the new acoustic operator, and we expect all the arguments mentioned until now to remain relevant for this new splitting.

2.3 Semi-discretization in time

2.3.1 Wave splitting and low Froude accuracy

In the shallow water system $(\mathcal{P}_{\mathsf{Fr}})$, time variations of the solution are given by $S - \nabla \cdot F$ where F and S are respectively the nonlinear flux and source term (2.2). We have seen that the eigenvalues of the flux Jacobian DF do not remain bounded as $\mathsf{Fr} \to 0$, and the use of an explicit time integrator would require the time steps to vanish in the low Froude limit which we cannot afford. On the other hand an implicit treatment of the

nonlinear term can be deemed too expensive at the computational level. Hence the first and foremost goal is to decompose the time variations of the solution as a contribution from a nonlinear term with slow dynamic, and a linear term embodying the fast scales. The former will then be treated explicitly, while the latter will be handled implicitly.

Under the light of the above, we introduce a new convective flux H whose eigenvalues are independent of Froude, and we recast system $(\mathcal{P}_{\mathsf{Fr}})$ under the form

$$\partial_t U + \nabla \cdot H(U,z) + [\nabla \cdot (F-H) - S](U,z) = 0 \ . \tag{2.26}$$

In the choice detailed below H will incorporate a contribution from the source term, and it is why we make it depend on z. The remaining differential operator $\nabla \cdot (F-H) - S$ accounts for the acoustic waves and we denote it by L. We have to chose H so that L is linear, and both operators $\nabla \cdot H, L$ should lead to hyperbolic systems for the well-posedness of the splitting. They should also admit a conservative writing for the water height component, and for the discharge component provided a flat bathymetry. Finally, the splitting should yield semi-implicit schemes verifying the low Froude accuracy and asymptotic preserving properties while preserving the lakes at rest. The former two points have already been motivated in Section 2.2, and preserving lakes at rest is particularly important since most geophysical flows can be seen as small perturbations around this steady state.

Taking L linear signifies the existence of two matrices $\mathbf{L}_x(z)$ and $\mathbf{L}_y(z)$ in $\mathbb{R}^{3\times 3}$ such that $L(U,z) = \mathbf{L}_x(z)\partial_x U + \mathbf{L}_y(z)\partial_y U$ up to a vector independent of U. Making use of the relation $\nabla \cdot H = \nabla \cdot F - S - L$, we have the following sufficient condition:

Proposition 2.3.1. *For H to have bounded eigenvalues, it is sufficient for matrices $\mathbf{L}_x$ and $\mathbf{L}_y$ to satisfy:*

$$\mathbf{L}_x = \begin{pmatrix} \mathcal{O}(1) & 1+\mathcal{O}(\mathsf{Fr}^2) & \mathcal{O}(\mathsf{Fr}^2) \\ O(\mathsf{Fr}^{-2}) & \mathcal{O}(1) & \mathcal{O}(1) \\ O(\mathsf{Fr}^{-2}) & \mathcal{O}(1) & \mathcal{O}(1) \end{pmatrix}, \quad \mathbf{L}_y = \begin{pmatrix} \mathcal{O}(1) & \mathcal{O}(\mathsf{Fr}^2) & 1+\mathcal{O}(\mathsf{Fr}^2) \\ O(\mathsf{Fr}^{-2}) & \mathcal{O}(1) & \mathcal{O}(1) \\ O(\mathsf{Fr}^{-2}) & \mathcal{O}(1) & \mathcal{O}(1) \end{pmatrix}$$

If we restrict ourselves to the set of well prepared data $\mathbb{W}_{\mathrm{p}}$, we can replace the condition above with:

$$\mathbf{L}_x = \begin{pmatrix} \mathcal{O}(1) & \mathcal{O}(1) & \mathcal{O}(1) \\ -z/\mathsf{Fr}^2 + \mathcal{O}(1) & \mathcal{O}(1) & \mathcal{O}(1) \\ \mathcal{O}(1) & \mathcal{O}(1) & \mathcal{O}(1) \end{pmatrix}, \quad \mathbf{L}_y = \begin{pmatrix} \mathcal{O}(1) & \mathcal{O}(1) & \mathcal{O}(1) \\ \mathcal{O}(1) & \mathcal{O}(1) & \mathcal{O}(1) \\ -z/\mathsf{Fr}^2 + \mathcal{O}(1) & \mathcal{O}(1) & \mathcal{O}(1) \end{pmatrix}$$

Proof. Up to a vector independent of U, we have by definition:

$$\nabla \cdot H(U,z) = (DF(U;n_x) - \mathbf{L}_x(z))\partial_x U + (DF(U;n_y) - \mathbf{L}_y(z))\partial_y U - S(U,z)$$

with $n_x = (1,0)^T$ and $n_y = (0,1)^T$. The expression of the flux Jacobian $DF(U;n)$ is given in (2.3). Eigenvalues along direction $n \in \mathbb{S}^2$ are given as the roots of the characteristic polynomial associated to $DH(U,z;n) = DF(U;n) - \mathbf{C}(z;n)$ where $\mathbf{C} = n_1\mathbf{L}_x + n_2\mathbf{L}_y$. It is then sufficient for the characteristic polynomial $P_{DH}(\lambda)$ to have bounded coefficients.

If we assume that all the coefficients $(c_{jk})_{j,k}$ of matrix $\mathbf{C}$ are in $\mathcal{O}(1)$ except potentially c_{21} and c_{31}, and noting that

$$P_{DH}(\lambda) = \lambda^3 - \mathrm{tr}(DH)\lambda^2 + \frac{1}{2}(\mathrm{tr}(DH^2) - \mathrm{tr}(DH)^2)\lambda - \det(DH)$$

we get the desired result by computing:

$$\begin{aligned}
\mathrm{tr}(DH) &= \mathcal{O}(1),\\
\mathrm{tr}(DH^2) &= 2(n_1 - c_{12})\Big(\frac{hn_1}{\mathsf{Fr}^2} - c_{21}\Big) + 2(n_2 - c_{13})\Big(\frac{hn_2}{\mathsf{Fr}^2} - c_{31}\Big) + \mathcal{O}(1)\\
\det(DH) = \quad & (n_2 - c_{13})(vn_1 - c_{32})\Big(\frac{hn_1}{\mathsf{Fr}^2} - c_{21}\Big) + (n_1 - c_{12})(un_2 - c_{23})\Big(\frac{hn_2}{\mathsf{Fr}^2} - c_{31}\Big)\\
& - (n_2 - c_{13})(2un_1 + vn_2 - c_{22})\Big(\frac{hn_2}{\mathsf{Fr}^2} - c_{31}\Big)\\
& - (n_1 - c_{12})(un_1 + 2vn_2 - c_{33})\Big(\frac{hn_1}{\mathsf{Fr}^2} - c_{21}\Big)\\
& + \mathcal{O}(1)
\end{aligned}$$

Indeed, under the first assumption one has $n_1 - c_{12} = \mathcal{O}(\mathsf{Fr}^2)$ and $n_2 - c_{13} = \mathcal{O}(\mathsf{Fr}^2)$, and as a result every coefficient is a $\mathcal{O}(1)$. When restricting to the set $\mathbb{W}_{\mathrm{p}}$ we have $h + z = \mathcal{O}(\mathsf{Fr}^2)$, which in turns implies that $hn_1/\mathsf{Fr}^2 - c_{21} = \mathcal{O}(1)$ and $hn_2/\mathsf{Fr}^2 - c_{31} = \mathcal{O}(1)$. □

In practice, the acoustic wave operator is obtained by linearizing system $(\mathcal{P}_{\mathsf{Fr}})$ around the lake at rest $(\overline{h}, \overline{Q}) = (-z, 0)$, similarly to what was done in (ζ, Q) coordinates in [14]. The purpose is to later get an easy way to preserve lakes at rest during the implicit step. Incorporating the source term into the linearized spatial operator, we find the non-conservative formulation for the acoustic waves:

$$\begin{cases} \dfrac{\partial h}{\partial t} + \nabla \cdot (hV) = 0 \\ \dfrac{\partial hV}{\partial t} + \dfrac{-z}{\mathsf{Fr}^2}\nabla(h + z) = 0 \end{cases}$$

As a consequence, it entails the following splitting, referred to as the lake at rest splitting (LAR), and that we will use from now on in the rest of this work:

$$L(U, z) = \begin{pmatrix} 0 & 1 & 0 \\ -z/\mathsf{Fr}^2 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \frac{\partial U}{\partial x} + \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -z/\mathsf{Fr}^2 & 0 & 0 \end{pmatrix} \frac{\partial U}{\partial y} + \begin{pmatrix} 0 \\ (-z/\mathsf{Fr}^2)\partial_x z \\ (-z/\mathsf{Fr}^2)\partial_y z \end{pmatrix} \tag{2.27}$$

$$H(U, z) = \begin{pmatrix} 0 \\ hV \otimes V + \frac{1}{2\mathsf{Fr}^2}(h + z)^2\, \mathbf{I}_2 \end{pmatrix} \tag{2.28}$$

Clearly, both sufficient conditions from proposition 2.3.1 are satisfied, and eigenvalues are bounded for the convection flux H. Indeed, we can compute them to find 0, $(V \cdot n)$ and $2(V \cdot n)$. Hence DH admits three real eigenvalues that are different as long as $V \cdot n \neq 0$. Regarding L, its eigenvalues are 0 and $\pm\sqrt{-z}/\mathsf{Fr}$, which for well prepared data does not deviate too much from the gravity waves velocity $\sqrt{h}/\mathsf{Fr}$.

The accuracy of schemes based on the LAR splitting (2.27)-(2.28) will be assessed by the mean of the near stationary condition (2.25) introduced in Theorem 2.2.7. This condition will only be checked over the discretization of the linear acoustic part, meaning that the convective part $\nabla \cdot H$ is neglected next to L. In the framework of IMEX Runge-Kutta methods detailed later, this simplification makes it easier to compute the modified PDE, which will furthermore be linear.

Definition 2.3.2 (Low Froude accuracy)**.** *Assume the bathymetry to be flat, and let Δt be the time increment independent of Froude. We call low Froude accurate any scheme for* (2.26) *such that when neglecting $\nabla \cdot H$ — or its numerical approximation in the fully discrete case, the solutions of the corresponding linear modified PDE*

$$\frac{\partial U}{\partial t} + L(U, z \equiv \mathrm{Cst}) = R_{\Delta t}(U)$$

satisfy the near stationary condition (2.25) *with $\gamma(t, \cdot) = \mathrm{id}$.*

Remark 2.3.3. *In Definition 2.3.2, $R_{\Delta t}$ denotes the linear differential operator related to the error of discretization. When also discretizing in space, this term will also depend on the mesh resolution δ. We will detail how to compute $R_{\Delta t}$ further down in the document, first in the time semi-discrete case and then in the fully discrete case. We only require the near stationary condition* (2.25) *to be satisfied for convective time steps, i.e. for time steps that don't depend on the Froude parameter. In particular this excludes small acoustic times scales $\tau_a = \mathcal{O}(\mathsf{Fr})$ that would otherwise be seen by the scheme. The reason for this restriction is that a scheme could yield accurate results for convective time steps but not for acoustic ones — such an example will be encountered in Section 2.4.5. This not a big problem, as we aim at designing schemes that are asymptotically stable, and plan to use large time steps anyway.*

2.3.2 IMEX Runge-Kutta methods

Time integration shall be performed by the mean of IMEX Runge-Kutta methods. It is usual to represent such methods thanks to double Butcher tableaux constituted from the triple $(\mathbf{A}, b, c)$ for the implicit part, and $(\tilde{\mathbf{A}}, \tilde{b}, \tilde{c})$ for the explicit part. For an s-stages update, vectors $c, \tilde{c} \in \mathbb{R}^s$ represent the fractional time steps used throughout the successive internal updates respectively for the implicit and explicit parts. Matrices $\mathbf{A}, \tilde{\mathbf{A}} \in \mathbb{R}^{s \times s}$ store the weights for the internal updates, whereas $b, \tilde{b} \in \mathbb{R}^s$ store the final update weights. We introduce the vector $e = (1, \dots, 1)^T \in \mathbb{R}^s$ which will be useful in the next lines. A standard constraint on vectors $c, \tilde{c}$ is for them to satisfy $c = \mathbf{A}e$, $\tilde{c} = \tilde{\mathbf{A}}e$.

We restrict ourselves to *diagonally implicit Runge-Kutta* methods (DIRK) for efficiency reasons. In the framework of IMEX-DIRK methods, matrix $\widetilde{\mathbf{A}}$ is lower triangular with zeros on the diagonal, and $\mathbf{A}$ is lower triangular, see Table 2.2. More concrete examples of such IMEX-DIRK Butcher tables are given in Appendix 2.D. An s-stages update associated to these generic IMEX-DIRK Butcher tableaux then reads:

$$\frac{U^{(j)} - U^0}{\Delta t} + \sum_{k=1}^{j-1} \tilde{a}_{jk} \nabla \cdot H(U^{(k)}, z) + \sum_{k=1}^{j} a_{jk} L(U^{(k)}, z) = 0, \qquad \forall 1 \leq j \leq s \qquad (2.29)$$

$$\begin{array}{c|cccc} \tilde{c}_1 & 0 & & & \\ \tilde{c}_2 & \tilde{a}_{21} & 0 & & \\ \vdots & \vdots & \vdots & \ddots & \\ \tilde{c}_s & \tilde{a}_{s1} & \tilde{a}_{s2} & \cdots & 0 \\ \hline & \tilde{b}_1 & \tilde{b}_2 & \cdots & \tilde{b}_s \end{array} \qquad \begin{array}{c|cccc} c_1 & a_{11} & & & \\ c_2 & a_{21} & a_{22} & & \\ \vdots & \vdots & \vdots & \ddots & \\ c_s & a_{s1} & a_{s2} & \cdots & a_{ss} \\ \hline & b_1 & b_2 & \cdots & b_s \end{array}$$

Figure 2.2: Butcher tableaux for the explicit convection (left) and implicit acoustic waves (right) involving s intermediate stages.

$$\frac{U^1 - U^0}{\Delta t} + \sum_{j=1}^{s} \tilde{b}_j \nabla \cdot H(U^{(j)}, z) + \sum_{j=1}^{s} b_j L(U^{(j)}, z) = 0 \tag{2.30}$$

An assumption we make is that none of the coefficients appearing in the Butcher tableaux depend on the Froude number. Also note that in our case the coefficients from $c, \tilde{c}$ don't appear since the equations are autonomous, and thus we will stop refering to these quantities from now on.

We introduce the classification by refering to Boscarino [16][15][17] that will be usefull throughout this document.

Definition 2.3.4 (RK classification)**.** *Let $(\mathbf{A}, b)$ be a diagonally implicit Runge-Kutta method. We call it of type A if matrix $\mathbf{A}$ has all its diagonal coefficients nonzero ($\forall 1 \leq i \leq s,\ a_{ii} \neq 0$), and of type CK if there is a vector $a \in \mathbb{R}^{s-1}$ and a matrix $\mathbf{A'} \in \mathbb{R}^{(s-1)\times(s-1)}$ with only nonzero diagonal entries such that*

$$\mathbf{A} = \begin{pmatrix} 0 & 0 \\ a & \mathbf{A'} \end{pmatrix} .$$

Definition 2.3.5 (Stiff accuracy)**.** *Let $(\widetilde{\mathbf{A}}, \widetilde{b}; \mathbf{A}, b)$ be a double IMEX-DIRK Butcher table of s stages. We say it is globally stiffly accurate (GSA) if for all $1 \leq i \leq s$ we have $\widetilde{b}_i = \widetilde{a}_{s,i}$ and $b_i = a_{s,i}$, that is to say if in both explicit and implicit parts the weights of the last partial update coincide with the weights of the final update. If this only holds true for the implicit part, then we say that the method is implicitly stiffly accurate (ISA).*

We will now investigate the properties of semi-discrete IMEX-DIRK schemes, which will give us valuable information regarding their compliance with the low Froude accuracy criterion, asymptotic consistency as well as L^2 stability. In particular, the notion of stiff accuracy will be helpful to prove the asymptotic consistency since it means that the final update is exactly equal to the last stage. A more general situation would be to have double Butcher tables allowing to write the final update U^1 as a convex combination of the partial updates $(U^{(j)})_{1\leq j\leq s}$.

2.3.3 Modified PDE and asymptotic consistency

Rather than studying the IMEX-DIRK method in its discrete form, it will sometimes be easier to work with the modified PDE. This is especially true for assessing the compliance with the result of Theorem 2.2.4, for which we have to check the low Froude accuracy criterion from Definition 2.3.2. We recall that the modified PDE associated to some scheme describes its behavior with more accuracy than the original system that is being approximated, at least when the refinement in time and in space is sufficiently small. Neglecting the convection part (2.28) from the LAR splitting, we can make use of the following result for a generic Butcher tableau, whose proof can be found in Appendix 2.A.

Proposition 2.3.6. *Let $p \in \mathbb{N}^*$ and let $(\mathbf{A}, b)$ be a Butcher tableau satisfying the equalities $b^T \mathbf{A}^k e = 1/(k+1)!$ for all $0 \leq k \leq p-1$. When applied to the autonomous linear wave equation*

$$\left(\frac{\partial}{\partial t} + L\right) U = 0$$

the RK method associated to $(\mathbf{A}, b)$ admits the following p-th order modified equation

$$\left(\frac{\partial}{\partial t} + L\right) U = R_{\Delta t} U \ , \tag{2.31}$$

where $R_{\Delta t}$ is the differential operator defined by:

$$R_{\Delta t} = \Delta t^p \varphi(\mathbf{A}, b; p)(-L)^{p+1}, \quad \varphi(\mathbf{A}, b; p) = \left(b^T \mathbf{A}^p e - \frac{1}{(p+1)!}\right). \tag{2.32}$$

Remark 2.3.7. *In Proposition 2.3.6, hypothesis on the Butcher tableau coincide with the order conditions in the case of an autonomous linear system. Also note that although it only involves spatial derivatives, operator $R_{\Delta t}$ arises from the leading order consistency error attributed to the time discretization.*

As a direct consequence of Proposition 2.3.6, we get that any consistent IMEX-RK time discretization of the LAR wave equation satisfies the low Froude accuracy property. Indeed, the kernels of the spatial operators in Equation (2.31) verify $\mathcal{E} \subset \ker L \subset \ker R_{\Delta t}$, even when the bathymetry is not flat. Hence if the initial data belongs to $\mathcal{E}$, it is left unchanged by (2.31).

Proposition 2.3.8. *Any consistent IMEX-RK time semi-discretization of the LAR splitting* (2.27)–(2.28) *is low Froude accurate.*

To do the proof, we will work in Fourier coordinates, where it is easier to caracterize the belonging to the set $\mathbb{W}$ of constant free surface and divergence-free discharge defined in (2.11). In fact, for any function $U \in (L^2(\mathbb{T}^2))^3$ let us denote by $\widehat{U}$ its Fourier coefficients

$$\forall k \in \mathbb{Z}^2, \ \widehat{U}(k) = \int_{\mathbb{T}^2} U(x,y) \exp(-2\mathrm{i}\pi(k_1 x + k_2 y)) \, \mathrm{d}x\mathrm{d}y \ .$$

Performing integration by parts, we get the formulas

$$\widehat{\nabla h}(k) = 2\mathbb{i}\pi k\widehat{h}(k) \ , \quad \widehat{\nabla \cdot Q}(k) = 2\mathbb{i}\pi k \cdot \widehat{Q}(k) \ , \tag{2.33}$$

leading to the following equivalence

$$U \in \mathbb{W} \Longleftrightarrow \widehat{U} \in \widehat{\mathbb{W}} \stackrel{\text{def}}{=} \left\{(\widehat{h}, \widehat{Q}) : \mathbb{Z}^2 \to \mathbb{C}^3, \ \forall k \in \mathbb{Z}^2 \setminus \{0\}, \ \widehat{h}(k) = k \cdot \widehat{Q}(k) = 0\right\} . \tag{2.34}$$

Proof. (Proposition 2.3.8) Let $U(t,\cdot) \in \mathcal{E}$ be a solution of the modified PDE (2.31) with flat bathymetry. It is sufficient to show that U is constant in time. We rewrite this equation in Fourier space:

$$\left(\frac{\partial}{\partial t} + \widehat{L}(k)\right)\widehat{U} = \widehat{R}_{\Delta t}(k)\widehat{U} \ . \tag{2.35}$$

Making use of the relations (2.33) we find that $\widehat{L}(k) = 2\mathbb{i}\pi(k_1\mathbf{L}_x + k_2\mathbf{L}_y)$, and $\widehat{R}_{\Delta t}(k)$ is proportional to $(-\widehat{L}(k))^{p+1}$. Supplementing (2.35) with some initial condition $\widehat{U}(t = 0,\cdot) = \widehat{U}^0$, for any fixed value $k \in \mathbb{Z}^2$ we get a Cauchy problem admitting a unique solution over positive times $t \in \mathbb{R}_+$. If we then take $U(t = 0,\cdot) \in \mathcal{E} = (L^2(\mathbb{T}^2))^3 \cap \mathbb{W}$, using (2.34) we also have that $\widehat{U}^0 \in \widehat{\mathbb{W}}$. Furthermore, it is straightforward to verify that $\widehat{\mathbb{W}} = \ker \widehat{L} \subset \ker \widehat{R}_{\Delta t}$, and hence the unique solution is constant in time equal to the initial data. This holds for all $k \in \mathbb{Z}^2$, therefore $U(t \geq 0,\cdot) = U(t = 0,\cdot) \in \mathcal{E}$. □

As a result, lack of low Froude accuracy cannot be imputed to IMEX-RK time integration, but will rather come from the spatial discretization, which will be the focus of Section 2.4.

Next we give a condition over the Butcher table ensuring the invariance of the space of well prepared data $\mathbb{W}_{\mathrm{p}}$ introduced in (2.13).

Proposition 2.3.9. *Let s denote the number of stages of the IMEX-DIRK method with the implicit part of type A or of type CK, and assume there exist a vector $\nu \in \mathbb{R}^s$ with positive entries and such that*

$$\sum_{i=1}^{s} \nu_i = 1, \quad b = \mathbf{A}^T\nu, \quad \tilde{b} = \widetilde{\mathbf{A}}^T\nu \ .$$

If additionally we admit that each partial update admits a decomposition in powers of Froude, we get at least formally that $U^n \in \mathbb{W}_{\mathrm{p}} \Longrightarrow U^{n+1} \in \mathbb{W}_{\mathrm{p}}$.

Remark 2.3.10. *We believe it is reasonable to assume existence of a decomposition in powers of Froude at each stage since we know that any IMEX-DIRK semi-discrete scheme is low Froude accurate, which means that the numerical approximation is kept close to $\mathcal{E}$. We will see in the next section that when discretizing in space too, not any scheme will be low Froude accurate, in which case such a condition cannot be expected to be met. Also note that in Proposition 2.3.9 the assumption on vectors $b, \tilde{b}$ encompasses globally stiffly accurate methods (GSA) for which b and $\tilde{b}$ are given by the last row from $\mathbf{A}$ and $\widetilde{\mathbf{A}}$ respectively. In fact in this situation ν is the vector from $\mathbb{R}^s$ whose all entries are zero except the last equal to one.*

Remark 2.3.11. *If $a_{11} \neq 0$, the first stage from the IMEX-DIRK method defines a partial update $U^{(1)}$ that belongs to $\mathbb{W}_{\mathrm{p}}$ even if at the start U^n only satisfies $\nabla(h_0^n + z) = 0$. In fact this can be seen in equalities* (2.69) *and* (2.70) *of the proof for $j = 1$.*

This invariance result with respect to $\mathbb{W}_{\mathrm{p}}$ enables us to get the asymptotic consistency, again formally.

Proposition 2.3.12. *The time semi-discrete scheme considered in Proposition 2.3.9 is asymptotically consistent.*

The proofs of Proposition 2.3.9 and Proposition 2.3.12 are in Appendix 2.B.

2.3.4 Asymptotic L^2 stability

Generally speaking, stability refers to the ability of the scheme, when applied to some well posed equation, to avoid blow ups or amplifications of small perturbation in the approximated solution — typically caused by limited machine precision. Among the different definitions attributed to stability, we chose here to look at the L^2 stability, that is to say the decrease of some L^2-norm of the numerical approximation, potentially under a condition on the Butcher tableau. A discrete analysis might not be straightforward, hence we rather study the modified PDE. But even in this continuous setting stability is difficult to assess, and one often substitute the reference modified PDE by a simpler version. In our case, the difficulty arises not only from the fact the overall scheme is nonlinear due to the convection part, but also because the method involves a wave splitting. In fact, this renders difficult the computing of the modified PDE itself, since coupling conditions of the double Butcher tableau have to be accounted for. In particular, these coupling conditions become increasingly many as the order of the method and the number of partial updates increase. For this reason we will rather consider the L^2 stability of the acoustic part alone with flat bathymetry and neglect the convection operator.

Let us define the acoustic energy

$$E = \frac{1}{2}\|h\|^2_{L^2(\mathbb{T}^2)} + \frac{1}{2}\|Q/c\|^2_{(L^2(\mathbb{T}^2))^2}, \qquad c = \sqrt{-z}/\mathsf{Fr} \tag{2.36}$$

and recall the quantity $\varphi(\mathbf{A}, b; p) = b^T\mathbf{A}e - 1/(p+1)!$ encountered in the definition of the consistency error in time (2.32). Whenever the method under consideration has an even order of accuracy, we can show that the associated modified PDE of same order (i.e. the one encompassing the consistent part and main leading error term) keeps the acoustic energy (2.36) constant. This is due to the presence of odd order derivatives in the leading error term, which then cancels by integration by parts and by periodicity. In fact, let $(\mathbf{A}, b)$ be a Butcher tableau with order of accuracy $p = 2k$. There holds

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}E &= (h, \partial_t h) + \frac{1}{c^2}(Q, \partial_t Q) \\ &= -(h, \nabla\cdot Q) - (Q, \nabla h) + \Delta t^p \varphi(\mathbf{A}, b; p)\Big((h, [(-L)^{p+1}U]_h) + \frac{1}{c^2}(Q, [(-L)^{p+1}U]_Q)\Big)\end{aligned}$$

The notation $[\cdot]_{h/Q}$ refers to the water height or discharge component depending on the subscript. As anticipated, the first two scalar products cancel by integration by parts and by periodicity. The same goes for the last two terms, since we have:

$$(h, [(-L)^{p+1}U]_h) = -c^{2k}(h, \Delta^k \nabla \cdot Q) = c^{2k}(\Delta^k \nabla h, Q) = -\frac{1}{c^2}(Q, [(-L)^{p+1}U]_Q)$$

Therefore in this setting the p-th order modified PDE exactly preserves the energy. Because of that we will need to go further and account for the next leading order error term which will either give dissipation or increase of energy depending upon its sign — and ultimately depending on the Butcher table. First we derive the modified PDE of order $p+1$ related to a p-th order scheme in the statement below, with the proof given in Appendix 2.A.

Proposition 2.3.13. *Consider a p-th order semi-discrete in time acoustic scheme with Butcher tableau $(\mathbf{A}, b)$ satisfying hypothesis of Proposition 2.3.6. Then its $(p+1)$-th order modified PDE is given by:*

$$\Big(\frac{\partial}{\partial t} + L\Big)U = (R_{\Delta t} + \widetilde{R}_{\Delta t})U \tag{2.37}$$

where the operator $R_{\Delta t}$ was defined in Proposition 2.3.6, and where the new operator $\widetilde{R}_{\Delta t}$ is expressed below:

$$\widetilde{R}_{\Delta t} = \Delta t^{p+1}(-\varphi(\mathbf{A}, b; p) + \varphi(\mathbf{A}, b; p+1))(-L)^{p+2}$$

Now we can give the L^2 stability result:

Proposition 2.3.14. *Let $(\mathbf{A}, b)$ be a Butcher tableau of order p applied to the acoustic system with flat bathymetry. The $(p+1)$-th order modified PDE associated with this semi-discrete scheme dissipates the acoustic energy E provided one of the following two points hold:*

1. *p is even and we have $(-1)^{p/2}(\varphi(\mathbf{A}, b; p+1) - \varphi(\mathbf{A}, b; p)) > 0$;*
2. *p is odd and we have $(-1)^{(p+1)/2}\varphi(\mathbf{A}, b; p) < 0$;*

The proof is featured in Appendix 2.B and is based on the fact that for all integer k, for all U smooth solution of the modified PDE (2.31) we have:

$$L^{2k+1}U = c^{2k}\Delta^k \begin{pmatrix} \nabla \cdot Q \\ c^2 \nabla h \end{pmatrix}, \quad L^{2(k+1)}U = c^{2(k+1)}\Delta^k \begin{pmatrix} \Delta h, \\ \nabla(\nabla \cdot Q) \end{pmatrix} . \tag{2.38}$$

Remark 2.3.15. *It is possible to have a Butcher table with an even order of accuracy p such that $\varphi(\mathbf{A}, b; p+1) = \varphi(\mathbf{A}, b; p)$. In this case the acoustic energy is exactly conserved by the $(p+1)$-th order modified PDE, and we need to incorporate additional error terms to tell whether the scheme dissipates or increases the energy. An example of such a Butcher table is given by the Crank-Nicolson time integrator.*

Remark 2.3.16. *Proposition 2.3.14 only gives part of the picture. In fact to assess the L^2 stability of a fully discretized scheme with certainty, one would also need to consider the spatial consistency error terms. However we think that the result from Proposition 2.3.14 can be usefull in itself and we illustrate this through a simplified example. We replace the modified equation by the following toy model:*

$$\partial_t u + v \cdot \nabla u = (E_{\Delta t} + E_\delta)\Delta u \ . \tag{2.39}$$

In this scalar PDE, the right hand side embodies the error one would get from first order time and spatial discretizations. Typically a situation that will be encountered later correspond to the ratio $E_\delta/E_{\Delta t}$ being equal to $K\delta\mathsf{Fr}/\Delta t$, with K independent of Froude. It is well known that when the coefficient in factor of the Laplacian is positive, the PDE (2.39) *is diffusive, and is otherwise anti-diffusive. Hence stability is conditionned to $E_{\Delta t}(1 + K\delta\mathsf{Fr}/\Delta t)$ being positive. If the time semi-discrete scheme is unconditionally unstable (i.e. $E_{\Delta t} < 0$) there are only two possible outcomes. Either $E_\delta < 0$ and the fully discretized scheme is unconditionally unstable, or $E_\delta > 0 \Rightarrow K < 0$ and the scheme is stable under the acoustic condition $\Delta t \leq -K\mathsf{Fr}\delta$. On the other hand if the time semi-discrete scheme is unconditionally stable ($E_{\Delta t} > 0$), the fully discrete method is unconditionally stable if $E_\delta > 0$. If $E_\delta < 0$, we have $K < 0$ and the scheme is stable under the reverse CFL condition $\Delta t \geq -K\mathsf{Fr}\delta$. The latter situation is not an issue as it does not contravene the notion of asymptotic stability. Put in other words, under this scenario Proposition 2.3.14 can be seen as a necessary condition for reaching asymptotical L^2 stability.*

Name	*Type*	*Order*	*Stiff accuracy*	*L^2 stability*
Forward Euler	Explicit	1	Yes	No
Heun	Explicit	2	No	No
Midpoint	Explicit	2	No	No
Backward Euler	Implicit	1	Yes	Yes
Crank-Nicolson	Implicit	2	Yes	Inconclusive
Implicit part of ARS-(2,2,2)	Implicit	2	Yes	Yes
Implicit part of JIN-(2,2,2)	Implicit	2	No	Yes

Table 2.1: Properties of the semi-discretization in time for the acoustic system for various Butcher tables. Fourth column: stiff accuracy, see Definition 2.3.5. Last column: whether the condition of Proposition 2.3.14 for L^2 stability is satisfied or not. The mentionned Butcher tables can be found in Appendix 2.D.

We check the condition of Proposition 2.3.14 for several Butcher tables, including explicit ones, to see which methods can be expected to give rise to asymptotically L^2 stable discretizations of the acoustic system. The results are gathered in Table 2.1. We see that among all of the semi-discrete explicit in time methods, none is L^2 stable, which is expected for first order methods. In fact it is known that first order explicit time integrators lead to an antidiffusive leading error term.

2.3.5 Well balanced property

The ability of a scheme to preserve the hydrostatic equilibrium $h + z = \text{Cst}$ and $V = 0$, is called well balanced property. Thanks to the choice of wave splitting (2.27)-(2.28), the well balanced property is automatically satisfied for any IMEX-DIRK method. The proof will be analougous to that of [14] and is included in Appendix 2.B for the sake of completeness.

Proposition 2.3.17. *Let $z \leq 0$ be the bathymetry profile lying in $W^{1,\infty}(\mathbb{T}^2)$. Suppose that U^0 verifies the lake at rest condition $h^0 + z \equiv K \in \mathbb{R}$, $Q^0 \equiv 0$. Then the final update U^1 produced by the semi-discrete IMEX-DIRK method $(\widetilde{\mathbf{A}}, \widetilde{b}; \mathbf{A}, b)$ is equal to U^0.*

2.4 Spatial discretization

2.4.1 Notations

We discretize the torus $\mathbb{T}^2$ with a uniform cartesian mesh $\mathcal{C}(\mathbb{T}^2)$. Every element of $\mathcal{C}(\mathbb{T}^2)$ is a rectangular cell whose sizes in the horizontal and vertical directions are respectively Δx and Δy. These cells are indexed by a unique relative integer pair $\mathcal{I} = (i, j) \in \mathbb{Z}^2$, and we note $C_{\mathcal{I}}$ or $C_{(i,j)} \in \mathcal{C}(\mathbb{T}^2)$ the cell of center (x_i, y_j) with $x_i = (i + 1/2)\Delta x$ and $y_j = (j + 1/2)\Delta y$. The scheme is initialized by projecting the initial condition U^0 and bathymetry z onto the space of cellwise-constant functions making use of some projector $\mathcal{Q} : X \mapsto \sum_{\mathcal{I}\in\mathbb{Z}^2} \mathcal{Q}_{\mathcal{I}}(X)\, \mathbb{1}_{C_{\mathcal{I}}}(\cdot)$ with X a scalar or vector field defined on $\mathbb{T}^2$ and $\mathcal{Q}_{\mathcal{I}}$ a quadrature method applied to each component of X with given accuracy over $C_{\mathcal{I}}$. As a rule of thumb, we denote the average of a given function over the cell $C_{(i,j)}$ by indexing it with the pair $(i, j) \in \mathbb{Z}^2$. Hence after the initialization step, we are provided with cellwise-constant data $\mathsf{U}^0 = \mathcal{Q}(U^0)$ and $\mathsf{z} = \mathcal{Q}(z)$ such that for all point (x, y) belonging to $C_{\mathcal{I}}$, one has $\mathsf{U}^0_{\mathcal{I}} = \mathsf{U}^0(x, y) = \mathcal{Q}_{\mathcal{I}}(U^0)$ and similarly $\mathsf{z}_{\mathcal{I}} = \mathsf{z}(x, y) = \mathcal{Q}_{\mathcal{I}}(z)$. Note that it is important to use the same quadrature for h^0 and z if we wish to correctly describe lakes at rest at the discrete level, requiring us to have $\nabla(h^0 + z) = 0 \Rightarrow (\mathsf{h}^0 + \mathsf{z})_{\mathcal{I}} = \text{Cst}$. For convenience, from now on we will denote the set of cellwise-constant scalar functions valued in $\mathbb{R}$ by $\mathbb{R}^{\mathcal{C}}$.

Knowing an approximation $\mathsf{U}^n \in (\mathbb{R}^{\mathcal{C}})^3$ to the solution of Equation $(\mathcal{P}_{\mathsf{Fr}})$ at time t^n, we would like to find an approximation $\mathsf{U}^{n+1} \in (\mathbb{R}^{\mathcal{C}})^3$ for the next iteration time $t^{n+1} = t^n + \Delta t$. Hence we need a procedure to determine the values $(\mathsf{U}^{n+1}_{\mathcal{I}})_{\mathcal{I}\in\mathbb{Z}^2}$ taken by the function U^{n+1} over each cell $C_{\mathcal{I}}$. The main discussion in the following lines aims for a strategy enabling a suitable discretization of operators $\nabla \cdot H$ and L encountered in the LAR splitting. In particular, let us stress that the low Froude accuracy, asymptotic preserving and well-balanced properties are highly desirable.

In this regard, we will highlight the usefulness of the low Froude accuracy criterion for assessing whether a scheme is accurate or not at low Froude numbers. Especially, we will compare it to the $\mathcal{E}$-invariance condition proposed by Dellacherie in Theorem 2.2.6. Interestingly, the latter allows to fix the inaccuracy of a first order upwind scheme but, contrary to our low Froude accuracy criterion, doesn't give the true orgin of the defect. A well known correction consists in centering the discretization of the pressure — or at

least its fast component in the case of a splitting, see [28][9][8] for example. Thus the pressure is discretized at second order accuracy, which renders the associated spatial error negligible with respect to the first order error in time. What is more, we have seen in the previous section that the error in time is compatible with the near stationary condition (2.25), and this gives an intuition as to why the modified schemes proposed in the litterature are accurate at low Froude numbers.

Another point of improvement brought over by the low Froude accuracy criterion is its ability to explain the good behavior of a second order scheme in time and space with centered acoustic part. In fact we will see that the associated modified PDE doesn't satisfy the $\mathcal{E}$-invariance, but that it is nearly stationary in the sense of (2.25).

Time discretization is achieved using IMEX-DIRK methods as studied in Section 2.3. Finite volumes making use of approximate Riemann solvers are used to update the convection step (2.28), whereas finite differences are involved in the propagation of surface waves (2.27).

2.4.2 Stencils for the acoustic wave operator

We start by introducing the notion of stencil, which provides a practicle way to write finite difference schemes. A stencil is a linear mapping S between the values stored in the neighborhood of a cell and the new value to be assigned to this cell. Hence every stencil S is characterized by wheights $(\omega[S]_{\mathcal{I}})_{\mathcal{I}\in\mathbb{Z}^2}$ such that for any $\mathsf{f}\in\mathbb{R}^{\mathcal{C}}$ and for all (x,y) in $\mathbb{T}^2$,

$$S\mathsf{f}(x,y)=\sum_{(i,j)\in\mathbb{Z}^2}\omega[S]_{(i,j)}\cdot\mathsf{f}(x+i\Delta x,y+j\Delta y)\ . \tag{2.40}$$

The weights do not depend on which cell the stencil is being applied to, and we will only consider stencils with a finite number of nonzero weights. It should be noted that the definition (2.40) can easily be generalized to functions that are not cellwise-constant. This will be especially usefull to study the error of a stencil by interpolating and performing a Taylor expansion. We now give a more concrete example by considering the following translation operators of $\mathcal{L}(\mathbb{R}^{\mathbb{T}^2})$ used later as building blocks.

$$\forall f\in\mathbb{R}^{\mathbb{T}^2},\forall(x,y)\in\mathbb{T}^2,\ \begin{cases} t_x^{\pm}f(x,y)=f(x\pm\Delta x/2,y)\\ t_y^{\pm}f(x,y)=f(x,y\pm\Delta y/2)\end{cases}$$

From these we can define the averaging operators

$$\mu_x=\frac{1}{2}(t_x^+ + t_x^-),\quad \mu_y=\frac{1}{2}(t_y^+ + t_y^-)$$

as well as the discrete deriatives approximating ∂_x and ∂_y

$$\overline{\partial}_x=\frac{1}{\Delta x}(t_x^+ - t_x^-),\quad \overline{\partial}_y=\frac{1}{\Delta y}(t_y^+ - t_y^-)\ .$$

The composition of any two of the previous operators with same subscript (x or y) defines a scalar stencil when restricted to the class of functions belonging to $\mathbb{R}^{\mathcal{C}}$. For instance

the stencils $t_x^{\pm}\overline{\partial}_x$ and $t_y^{\pm}\overline{\partial}_y$ are edge-centered partial derivatives, with:

$$\forall (i,j) \in \mathbb{Z}^2,\ \forall \mathsf{f} \in \mathbb{R}^{\mathcal{C}},\quad \begin{cases} t_x^-\overline{\partial}_x \mathsf{f}_{(i,j)} = \dfrac{f_{(i,j)} - \mathsf{f}_{(i-1,j)}}{\Delta x}, & t_x^+\overline{\partial}_x \mathsf{f}_{(i,j)} = \dfrac{f_{(i+1,j)} - \mathsf{f}_{(i,j)}}{\Delta x} \\ t_y^-\overline{\partial}_y \mathsf{f}_{(i,j)} = \dfrac{f_{(i,j)} - \mathsf{f}_{(i,j-1)}}{\Delta y}, & t_y^+\overline{\partial}_y \mathsf{f}_{(i,j)} = \dfrac{f_{(i,j+1)} - \mathsf{f}_{(i,j)}}{\Delta y} \end{cases}$$

The standard cell centered discrete derivatives are obtained from $\mu_x\overline{\partial}_x$ and $\mu_y\overline{\partial}_y$. It is then possible to define the centered second order discrete operator mimicking the action of the continuous acoustic wave operator L on the space of cellwise-constant functions as

$$L^\star(\mathsf{U},\mathsf{z}) = \begin{pmatrix} \mu_x\overline{\partial}_x(\mathsf{h}\mathsf{V}_x) + \mu_y\overline{\partial}_y(\mathsf{h}\mathsf{V}_y) \\ \mathsf{c}^2\mu_x\overline{\partial}_x(\mathsf{h}+\mathsf{z}) \\ \mathsf{c}^2\mu_y\overline{\partial}_y(\mathsf{h}+\mathsf{z}) \end{pmatrix} . \tag{2.41}$$

Other choices of discretization for the acoustic wave operator L will be introduced thereafter, but before that we would like to restate useful results regarding the modified PDE of the fully discretized scheme and its ability to preserve incompressible states $\mathcal{E}$. Let $\widetilde{L}$ be a discrete approximation of L accurate at order p. This means that there exists a differential operator $R_\delta[\widetilde{L}] = \mathcal{O}(\delta^p)$ representing the leading order consistency error in space, and such that for every function $U \in (C^{p+1}(\mathbb{T}^2))^3$ interpolating $\mathsf{U} \in (\mathbb{R}^{\mathcal{C}})^3$ at each cell center we have

$$\forall (i,j) \in \mathbb{Z}^2,\ \widetilde{L}\mathsf{U}_{(i,j)} = (L + R_\delta[\widetilde{L}])U(x_i,y_j) + \mathcal{O}(\delta^{p+1}) . \tag{2.42}$$

In practice $R_\delta[\widetilde{L}]$ will be obtained by performing a Taylor expansion. For instance we have $R_\delta[t_x^{\pm}\overline{\partial}_x] = \pm(\Delta x/2)\partial^2_{xx}$ and $R_\delta[t_y^{\pm}\overline{\partial}_y] = \pm(\Delta y/2)\partial^2_{yy}$, whereas the centered versions satisfy $R_\delta[\mu_x\overline{\partial}_x] = (\Delta x^2/6)\partial^3_{xxx}$ and $R_\delta[\mu_y\overline{\partial}_y] = (\Delta y^2/6)\partial^3_{yyy}$. Hence the leading error term for (2.41) is:

$$R_\delta[L^\star] = \frac{\Delta x^2}{6}\mathbf{L}_x\partial^3_{xxx} + \frac{\Delta y^2}{6}\mathbf{L}_y\partial^3_{yyy} .$$

It is important to know $R_\delta[\widetilde{L}]$ as this differential operator will appear in the modified PDE of the fully discretized scheme. Especially, when accounting for leading error terms only, the modified PDE is obtained by summing $R_{\Delta t}U$ and $-R_\delta[\widetilde{L}]$. This is the statement of the proposition below whose proof can be found in Appendix 2.A.

Proposition 2.4.1. *Let $(\mathbf{A}, b)$ be a p-th order Butcher tableau satisfying hypothesis from Proposition 2.3.6, and let $\widetilde{L}$ be a p-th order discretization of L, i.e. such that there exists a differential operator $R_\delta[\widetilde{L}]$ satisfying* (2.42)*. When applied to the acoustic wave equation with flat bathymetry, the resulting scheme admits the following p-th order modified equation:*

$$\left(\frac{\partial}{\partial t} + L\right)U = (R_{\Delta t} - R_\delta[\widetilde{L}])U \tag{2.43}$$

where $R_{\Delta t}$ has been defined in (2.32)*.*

We recall from Definition 2.3.2 that the low Froude accuracy criterion is satisfied when a scheme on the linear acoustic wave system admits a modified PDE whose solutions are nearly stationary for purely convective time steps — i.e. time increments independent of Froude. The near stationary condition (2.25) means that whenever the initial data is in the incompressible set $\mathcal{E}$, it remains unchanged up to $\mathcal{O}(\mathsf{Fr})$ perturbation terms. We believe that a scheme for the LAR splitting cannot be asymptotically consistent if it is not low Froude accurate. Regardless of the discretization of the convective part $\nabla \cdot H$, we can assess whether the discrete operator L^* leads to a low Froude accurate scheme or not. Especially, we have the sufficient condition thereafter.

Proposition 2.4.2. *Consider a scheme for the acoustic wave equation whose modified equation is given by* (2.43)*. For this modified PDE to admit stationary solutions over $\mathcal{E}$, it is sufficient to have $\mathcal{E} \subset \ker R_\delta$. In particular, such a scheme is low Froude accurate.*

Proof. Since the space $\mathcal{E}$ is encompassed in the kernel of all three operators $L, R_{\Delta t}$ and R_δ, we conclude by reusing the same arguments as for the proof of Proposition 2.3.8. □

We now give the fully discrete analog of Proposition 2.3.17 regarding the well-balancedness of the method.

Proposition 2.4.3. *Let $\mathsf{z} \leq 0$ be the discretized bathymetry profile, and let $\mathsf{U}^0 \in (\mathbb{R}^{\mathcal{C}})^3$ describe a cellwise-constant lake at rest, that is to say $\mathsf{h}^0 + \mathsf{z} \equiv K \in \mathbb{R}$, $\mathsf{Q}^0 \equiv 0$. Consider an s stages consistent IMEX-DIRK method such that for all $1 \leq j \leq s$ the map $\mathrm{id} + \Delta t a_{j,j} L^*(\cdot, \mathsf{z})$ is invertible over $(\mathbb{R}^{\mathcal{C}})^3$. It amounts to ask for each stage to admit a unique solution so that the overall scheme is well posed. Then final update U^1 of the fully discrete scheme is equal to U^0 as soon as the two conditions are fulfilled:*

1. *the convective numerical fluxes are constant over cellwise-constant lakes at rest;*
2. *cellwise-constant lakes at rest are in the kernel of discrete operator $L^*(\cdot, \mathsf{z})$;*

Proof. Once more the proof is by induction over the stages of the IMEX-DIRK method. The initialization is obvious and we focus on the recurrence. Assume $U^{(j)} = U^0$ for all $0 \leq j < k$. By assumption on the convective numerical flux and on discrete operator L^*, the k-th stage reads:

$$\mathsf{U}^{(k)} = \mathsf{U}^0 - \Delta t a_{k,k} L^*(\mathsf{U}^{(k)}, \mathsf{z})$$

Since $\mathsf{U}^0 \in \ker L^*(\cdot, \mathsf{z})$, a solution is given by $\mathsf{U}^{(k)} = \mathsf{U}^0$ and by hypothesis it is the unique one. □

All the fully discrete schemes considered in the remainder of this document will satisfy the two points from Proposition 2.4.3 and will thus be well balanced.

2.4.3 Inaccuracy of the standard upwind scheme

We begin by investigating the case of first order schemes. It is well known that a naive upwind approach fails to yield accurate results at low Froude numbers, and we want to specify the origin of failure. In Theorem 2.2.6, the suggested reason for this lack

of accuracy is a loss of incompressibility. We will see that it is indeed the case when considering an acoustic time scale, i.e. when both the time steps and the final time of the simulation scale as Fr. However we are rather interested in convective times τ_c independent of Froude, where it makes sense to use IMEX methods with large time increments. In this setting, the approximation of solution $U(\tau_c, \cdot)$ is kept nearly incompressible, but it suffers from an excessive diffusion over the discharge components. Besides, we will justify that such upwind schemes are consistent with the more restrictive and undesirable constraint

$$\lim_{\mathsf{Fr}\to 0} \left(\left\| \frac{\partial Q_x}{\partial x} \right\|_{L^2(\mathbb{T}^2)} + \left\| \frac{\partial Q_y}{\partial y} \right\|_{L^2(\mathbb{T}^2)} \right)(\tau_c) = 0 \ ,$$

disregarding of the initial condition. This underlines the fact that the $\mathcal{E}$-invariance is not enough to get asymptotic consistency, and that we also need to have the near stationary condition (2.25). However this more restrictive condition cannot hold if the discharge derivatives vanish as Fr becomes small.

In this section, the updating of the discharge during the convection step will be dealt with an HLL approximate Riemann solver with directionnal splitting, whose formula is written below for $U_L, U_R \in \mathbb{R}_+ \times \mathbb{R}^2$ and $n \in \mathbb{S}^2$.

$$H_Q^{\mathrm{HLL}}(U_L, U_R; n) = \begin{cases} \dfrac{\lambda_R H_Q(U_L; n) - \lambda_L H_Q(U_R; n) + \lambda_L \lambda_R (Q_R - Q_L)}{\lambda_R - \lambda_L} & \text{if } \lambda_L \lambda_R < 0 \\ H_Q(U_L; n) \quad \text{if } \lambda_L \text{ and } \lambda_R \text{ positive} \\ H_Q(U_R; n) \quad \text{if } \lambda_L \text{ and } \lambda_R \text{ negative} \end{cases} \tag{2.44}$$

In (2.44), H_Q is the discharge component of the convective flux (2.28) encountered in the LAR splitting

$$H_Q(U, z) = hV \otimes V + \frac{1}{2\mathsf{Fr}^2}(h + z)^2 \, \mathbf{I}_2 \ ,$$

and λ_L (resp. λ_R) is the left-most (resp. the right-most) eigenvalue:

$$\lambda_L = \min \left\{ \mathrm{Sp}_{\mathbb{R}}(DH_Q(U_L; n)) \cup \mathrm{Sp}_{\mathbb{R}}(DH_Q(U_R; n)) \right\} \ ,$$
$$\lambda_R = \max \left\{ \mathrm{Sp}_{\mathbb{R}}(DH_Q(U_L; n)) \cup \mathrm{Sp}_{\mathbb{R}}(DH_Q(U_R; n)) \right\} \ .$$

Let us recall that during the convective step we have $\partial_t h = 0$ and thus we don't need to evolve the water height. Hence the overall numerical flux in (h, Q) coordinates is $H^{\mathrm{HLL}} = (0, H_Q^{\mathrm{HLL}})$. Especially it is straightforward to see that the first point of Proposition 2.4.3 regarding the well balancedness is satisfied, thanks to H_Q being constant over lakes at rest.

Next we focus on the discretization of the LAR surface gravity wave system by the mean of stencils introduced previously. A first order spatial discretization is achieved by introducing an upwinding giving rise to some numerical diffusion. Such a diffusion is usually required to stabilize explicit methods, but might seem artificial in the framework

of semi-implicit ones. Indeed in the latter case, when factoring Remark 2.3.16 in, we expect a centered approach for the acoustic part to be stable without the need for acoustic time steps. Let us stress that here, our motivation for considering diffusive terms is not to improve the stability, but rather to illustrate how poorly low Froude numbers are handled over convective times by a naive upwind approach. We will also compare this scheme with its counterpart without additional diffusion — meaning a first order scheme in time and space except for the acoustic part, approximated at second order.

Before introducing the first order discretization for the acoustic operator, we briefly recall and comment on the acoustic wave system written in (ζ, Q)-coordinates, with $\zeta = h + z$ being the elevation of the free surface:

$$\begin{cases} \dfrac{\partial \zeta}{\partial t} + \nabla \cdot Q = 0 \\ \dfrac{\partial Q}{\partial t} + c^2 \nabla \zeta = 0 \end{cases}$$

In the above, $c = \sqrt{-z}/\mathsf{Fr}$ can be seen as an approximation to the dimensionless speed of sound $\sqrt{h}/\mathsf{Fr}$, provided that $-z$ is close to h or, equivalently, that ζ is close to zero. We motivate the choice of the (ζ, Q) coordinates, also known as pre-balanced coordinates, by remarking that adding a viscosity on the equation for the free surface doesn't change the steady state $(0, 0, 0)$, whereas in (h, Q) coordinates an additionnal viscosity term on the water height would in general modify it. In fact in the case of a lake at rest with varying bathymetry, there is no reason for the second order derivatives of h to cancel, unlike for ζ. This convenient choice of coordinates has already been used in [30] or [14] for instance. Hence we consider the discretization of L with diffusive term below:

$$L^{\mathrm{upwind}}(\mathsf{U}, \mathsf{z}) = \begin{pmatrix} \mu_x \overline{\partial}_x(\mathsf{hV}_x) + \mu_y \overline{\partial}_y(\mathsf{hV}_y) \\ \mathsf{c}^2 \mu_x \overline{\partial}_x(\mathsf{h} + \mathsf{z}) \\ \mathsf{c}^2 \mu_y \overline{\partial}_y(\mathsf{h} + \mathsf{z}) \end{pmatrix} - \frac{\mathsf{c}}{2} \begin{pmatrix} [\Delta x\, (\overline{\partial}_x)^2 + \Delta y\, (\overline{\partial}_y)^2](\mathsf{h} + \mathsf{z}) \\ \Delta x\, (\overline{\partial}_x)^2(\mathsf{hV}_x) \\ \Delta y\, (\overline{\partial}_y)^2(\mathsf{hV}_y) \end{pmatrix} \tag{2.45}$$

This discrete operator admits cellwise-constant lakes at rest in its kernel, which will allow the overall scheme to preserve this class of steady states. Provided a flat bathymetry (c constant over each cell), the second term in definition (2.45) is similar to the diffusion one would get from the upwinding of a Rusanov scheme. Furthermore, the error of consistency obtained with the choice (2.45) is directly related to this numerical diffusion, and is given by:

$$R_\delta[L^{\mathrm{upwind}}]U = -\frac{c\Delta x}{2} \frac{\partial^2}{\partial x^2} \begin{pmatrix} \zeta \\ Q_x \\ 0 \end{pmatrix} - \frac{c\Delta y}{2} \frac{\partial^2}{\partial y^2} \begin{pmatrix} \zeta \\ 0 \\ Q_y \end{pmatrix}$$

Now that the spatial discretization of the scheme has been specified, we will study the modified PDE associated with its acoustic part for a first order Butcher table $(\mathbf{A}, b)$. We first notice that the incompressible set $\mathcal{E}$ is not included in $\ker R_\delta[L^{\mathrm{upwind}}]$. In fact, a divergence-free discharge doesn't necessary have its second order derivatives $\partial^2_{xx} Q$

nor $\partial^2_{yy}Q$ equal to zero. We will justify later that a consequence of this will be an incompatibility with the near stationary condition (2.25) at convective times. Let us develop the modified PDE (2.43):

$$\left\{\begin{array}{l} \dfrac{\partial h}{\partial t} + \nabla\cdot Q = c\Big[\dfrac{\Delta x}{2}\dfrac{\partial^2}{\partial x^2} + \dfrac{\Delta y}{2}\dfrac{\partial^2}{\partial y^2} + c\Delta t\varphi(\mathbf{A},b;1)\Delta\Big]h \\ \dfrac{\partial Q}{\partial t} + c^2\nabla h = c\Big[\mathrm{diag}\Big(\dfrac{\Delta x}{2}\dfrac{\partial^2}{\partial x^2}, \dfrac{\Delta y}{2}\dfrac{\partial^2}{\partial y^2}\Big) + c\Delta t\varphi(\mathbf{A},b;1)\nabla\otimes\nabla\Big]Q \end{array}\right. . \tag{2.46}$$

The scalar function φ was introduced in (2.32). The question is to know whether any initial condition U^0 belonging to $\mathcal{E}$ results in a solution remaining equal to U^0 up to a $\mathcal{O}(\mathsf{Fr})$ term. We recall that being incompressible in physical variables translates to having the Fourier coefficients in the set $\widehat{\mathbb{W}}$ given in (2.34). For convenience, we fix $k\in\mathbb{Z}^2$ and introcuce the shorthand notation $\eta(k) = 2\mathbb{i}\pi k$. For any solution $U(t,\cdot)\in(L^2(\mathbb{T}^2))^3$ of (2.46), its Fourier coefficient $\widehat{U}$ associated to k satisfies the ODE below.

$$\frac{\partial}{\partial t}\begin{pmatrix}\widehat{h}\\ \widehat{Q}\end{pmatrix} = \begin{pmatrix}\alpha(\eta)+\beta(\eta)+\gamma(\eta)\cdot\eta & \eta^T \\ c^2\eta & \mathrm{diag}(\alpha(\eta),\beta(\eta)) + \gamma(\eta)\otimes\eta\end{pmatrix}\begin{pmatrix}\widehat{h}\\ \widehat{Q}\end{pmatrix} \tag{2.47}$$

The coefficients α,β are related to the error in space, while γ is related to the error in time and depends on the Butcher tableau $(\mathbf{A},b)$ as follows:

$$\alpha(\eta) = \frac{c\Delta x}{2}\eta_x^2 \le 0, \quad \beta(\eta) = \frac{c\Delta y}{2}\eta_y^2 \le 0, \quad \gamma(\eta) = c^2\Delta t\varphi(\mathbf{A},b;1)\eta \in \mathbb{i}\mathbb{R}^2 . \tag{2.48}$$

Due to the structure of the matrix encountered in (2.47), it is unfortunately difficult, if possible at all, to compute the related solutions for gerneric values of k. Instead we will simplify the problem by assuming that an incompressible initial condition leads to a solution that belongs to $\mathcal{E}+\mathcal{O}(\mathsf{Fr})$, meaning that it stays nearly incompressible. The simplification is based on the observation that any smooth U belonging to $\mathcal{E}+\mathcal{O}(\mathsf{Fr})$ verifies:

$$\nabla\cdot Q = \mathcal{O}(\mathsf{Fr}), \qquad c\Big[\frac{\Delta x}{2}\frac{\partial^2}{\partial x^2} + \frac{\Delta y}{2}\frac{\partial^2}{\partial y^2} + c\Delta t\varphi(\mathbf{A},b;1)\Delta\Big]h = \mathcal{O}(\Delta t/\mathsf{Fr}),$$

$$c^2\nabla h = \mathcal{O}(1/\mathsf{Fr}), \quad c\Big[\mathrm{diag}\Big(\frac{\Delta x}{2}\frac{\partial^2}{\partial x^2}, \frac{\Delta y}{2}\frac{\partial^2}{\partial y^2}\Big) + c\Delta t\varphi(\mathbf{A},b;1)\nabla\otimes\nabla\Big]Q = \mathcal{O}(\delta,\Delta t)/\mathsf{Fr} .$$

Hence we will neglect $\nabla\cdot Q$ the only term scaling as Fr and substitute (2.46) by

$$\left\{\begin{array}{l} \dfrac{\partial h}{\partial t} = c\Big[\dfrac{\Delta x}{2}\dfrac{\partial^2}{\partial x^2} + \dfrac{\Delta y}{2}\dfrac{\partial^2}{\partial y^2} + c\Delta t\varphi(\mathbf{A},b;1)\Delta\Big]h \\ \dfrac{\partial Q}{\partial t} + c^2\nabla h = c\Big[\mathrm{diag}\Big(\dfrac{\Delta x}{2}\dfrac{\partial^2}{\partial x^2}, \dfrac{\Delta y}{2}\dfrac{\partial^2}{\partial y^2}\Big) + c\Delta t\varphi(\mathbf{A},b;1)\nabla\otimes\nabla\Big]Q \end{array}\right. . \tag{2.49}$$

We believe that if the solutions of the original modified PDE (2.46) remain nearly incompressible, then they should be close to solutions of System (2.49) for $\mathsf{Fr}\ll 1$. In other words, either the modified PDE does not keep the solution close to $\mathcal{E}$, or it inherits the properties from (2.49) in the low Froude limit. We study the simplified equations through the following result.

Proposition 2.4.4. *Let $(\mathbf{A}, b)$ be a first order Butcher table, and assume the resolutions $\Delta x, \Delta y$ as well as the time τ to be independent of Fr. Consider $U \in (L^2(\mathbb{T}^2))^3$ a solution of System (2.49) with initial condition in $\mathcal{E}$. If $\Delta t = \mathcal{O}(\mathsf{Fr})$, then $U(\mathsf{Fr} \times \tau, \cdot)$ does not converge in $\mathcal{E}$ as $\mathsf{Fr} \to 0$ and we recover the result from Dellacherie regarding the loss of incompressibility at acoustic times (see Proposition 4.1 from [28]). If $\mathsf{Fr}^{-1}\Delta t$ does not vanish and as long as the CFL condition*

$$c\Delta t\varphi(\mathbf{A}, b; 1) > -\min\Big(\frac{\Delta x}{2}, \frac{\Delta y}{2}\Big) \tag{2.50}$$

holds, we have that $U(\tau, \cdot)$ admits a limit in $\mathcal{E}$, and additionally

$$\lim_{\mathsf{Fr}\to 0}\left(\left\|\frac{\partial Q_x}{\partial x}\right\|_{L^2(\mathbb{T}^2)} + \left\|\frac{\partial Q_y}{\partial y}\right\|_{L^2(\mathbb{T}^2)}\right)(\tau) = 0\ . \tag{2.51}$$

Proof. (Proposition 2.4.4). If $U \in (L^2(\mathbb{T}^2))^3$ satisfies (2.49), then its Fourier coefficients $\widehat{U}$ are solution of

$$\frac{\partial}{\partial t}\begin{pmatrix}\widehat{h}\\ \widehat{Q}\end{pmatrix} = \begin{pmatrix}\alpha(\eta)+\beta(\eta)+\gamma(\eta)\cdot\eta & 0\\ c^2\eta & \mathrm{diag}(\alpha(\eta),\beta(\eta))+\gamma(\eta)\otimes\eta\end{pmatrix}\begin{pmatrix}\widehat{h}\\ \widehat{Q}\end{pmatrix}, \tag{2.52}$$

Since the initial condition on $\widehat{h}$ is zero, we have $\widehat{h}(t,k) = 0$ for all $(t,k) \in \mathbb{R}_+ \times \mathbb{Z}^2$. Hence the discharge coefficient is determined by the ODE

$$\frac{\partial}{\partial t}\widehat{Q} = [\mathrm{diag}(\alpha(\eta),\beta(\eta)) + \gamma(\eta)\otimes\eta]\widehat{Q}\ , \tag{2.53}$$

combined to the initial condition $\widehat{Q}(0,k) = \widehat{Q}^0(k)$ with $k \cdot \widehat{Q}^0(k) = 0$.

First assume that $\mathsf{Fr}^{-1}\Delta t$ does not vanish as $\mathsf{Fr} \to 0$ and that CFL condition (2.50) holds. For $k \neq 0$ colinear to $(1,0)^T$ (resp. colinear to $(0,1)^T$), we see that $\widehat{Q}_x(\tau,k)$ (resp. $\widehat{Q}_y(\tau,k)$) decays exponentially to zero as $\mathsf{Fr} \to 0$. This is true because all the coefficients of $\mathrm{diag}(\alpha(\eta),\beta(\eta)) + \gamma(\eta)\otimes\eta$ are zero except the first diagonal term (resp. the second), which converges to $-\infty$ by assumption. Now consider the case where k has no zero coefficient. By defining the quantities

$$r_1 = c\Big(\frac{\Delta x}{2} + c\Delta t\varphi(\mathbf{A}, b; 1)\Big)\eta_1^2,\ r_2 = c\Big(\frac{\Delta y}{2} + c\Delta t\varphi(\mathbf{A}, b; 1)\Big)\eta_2^2,\ s = c^2\Delta t\varphi(\mathbf{A}, b; 1)\eta_1\eta_2$$

the eigenvalues of the matrix found in Equation (2.53) are given by

$$\lambda^{\pm} = \frac{1}{2}(r_1 + r_2) \pm \frac{1}{2}\sqrt{(r_1 - r_2)^2 + 4s^2}\ .$$

The associated exponential of matrix can be expressed as

$$\frac{1}{s(\lambda^+ - \lambda^-)}\begin{pmatrix}s & s\\ p^- & p^+\end{pmatrix}\begin{pmatrix}\exp(t\lambda^-) & 0\\ 0 & \exp(t\lambda^+)\end{pmatrix}\begin{pmatrix}p^+ & -s\\ -p^- & s\end{pmatrix}, \quad p^{\pm} = \lambda^{\pm} - r_1\ .$$

Using that $\eta_1^2, \eta_2^2 < 0$ and thanks to the CFL condition (2.50), we get $r_1, r_2 < 0$. It is then possible to show that $\lambda^{\pm} < 0$, by remarking that it is equivalent to ask

$$|r_1 + r_2| > \sqrt{(r_1 - r_2)^2 + 4s^2} \iff (r_1 + r_2)^2 > (r_1 - r_2)^2 + 4s^2 \iff r_1 r_2 > s^2\ .$$

The last inequality is always true under the assumption we made. Hence $\lambda^{\pm} \to -\infty$ in the low Froude limit, and the exponential of matrix converges to the null matrix. Gathering our findings we have shown that for any $k \in \mathbb{Z}^2$,

$$0 = \lim_{\mathsf{Fr}\to 0} k_1 \widehat{Q}_x(\Delta t, k) = \lim_{\mathsf{Fr}\to 0} k_2 \widehat{Q}_y(\Delta t, k)$$

and we recover (2.51) by using Parseval's equality. Furthermore the fact that $\widehat{h}(\tau, k) = 0$ and $k \cdot \widehat{Q}(\tau, k) \to 0$ for any $k \neq 0$ implies that $U(\tau, \cdot)$ converges in $\mathcal{E}$.

Next we take $\Delta t = \mathcal{O}(\mathsf{Fr})$ and choose for instance $k = (2, 1)^T$. Let $\widehat{Q}$ be the solution of (2.53) with initial condition $k^{\perp}$. Noting $K^{\pm} = \exp(\mathsf{Fr}\,\tau\lambda^{\pm})$ and using the exponential of matrix computed previously we get at time $t = \Delta t$:

$$k \cdot \widehat{Q} = \frac{1}{s(\lambda^+ - \lambda^-)} \begin{pmatrix} 2s + p^- \\ 2s + p^+ \end{pmatrix} \cdot \begin{pmatrix} -K^-(p^+ + 2s) \\ K^+(p^- + 2s) \end{pmatrix} = \frac{K^+ - K^-}{s(\lambda^+ - \lambda^-)} (2s + p^-)(2s + p^+)$$

By assumption, $\mathsf{Fr}\,\tau\lambda^{\pm}$ remains bounded in the low Froude limit, and we can check that $K^+ - K^-$ doesn't vanish. We also verify that the terms $2s + p^{\pm}$ and $s(\lambda^+ - \lambda^-)$ are scaling as Fr^{-2} and Fr^{-4} respectively. Hence $k \cdot \widehat{Q}(\Delta t, k)$ doesn't converge to zero as $\mathsf{Fr} \to 0$, and $U(\mathsf{Fr} \times \tau, \cdot)$ doesn't admit a limit in $\mathcal{E}$.

□

Remark 2.4.5. *In Proposition 2.4.4, the CFL condition* (2.50) *is always satisfied when* $\varphi(\mathbf{A}, b; 1)$ *is strictly positive. Hence we recover the result anticipated in Remark 2.3.16. Especially we see that when the sign of* $\varphi(\mathbf{A}, b; 1)$ *is strictly negative while having a uniform time increment, one can have* $r_1, r_2 > 0$ *for* $\mathsf{Fr} \ll 1$ *and we obtain positive eigenvalues, implying a blow up of the Fourier coefficients.*

Thanks to Proposition 2.4.4, we expect the approximation of the solutions at acoustic times to have a non vanishing compressible part, whereas at convective times an undesirable constraint on the discharge derivatives is enforced no matter what the initial condition is. Anyway, the resulting scheme will not be asymptotically consistent. We will now support this statement through a numerical illustration. We approximate System (2.47) by the mean of a Crank-Nicolson time integrator for values of k comprised in $[\![-12, 12]\!]^2 \setminus \{(0, 0)\}$ and for different values of Fr. Regarding the mesh resolution, we set $\delta = 10^{-2}$. We choose $z = -1$ for the bathymetry, and the initial condition is given by $\widehat{h}^0(k) = 0$ and $\widehat{Q}^0(k) = k^{\perp}/|k|$.

Figures 2.3 and 2.4 give the results obtained for an acoustic time step $\Delta t_a = \mathsf{Fr}\delta/4$ and a forward Euler method ($\varphi(\mathbf{A}, b; 1) = -1/2$). The plots are performed at time $t = \Delta t_a$. Figure 2.3 aims at measuring how well the incompressible constraint is satisfied as we reduce the value of Froude. Especially, for an asymptotically consistent scheme we should observe the vanishing of the quantity $|\widehat{h}| + |k \cdot \widehat{Q}|$ as $\mathsf{Fr} \to 0$. This doesn't seem to be the case, as all of the plotted Fourier coefficients remain unaffected by the value of Froude. The same can be said about the norm of $\widehat{Q}$, featured in Figure 2.4. This coincides with the result of Proposition 2.4.4.

In Figures 2.5 and 2.6, we investigate the case of a convective time step $\Delta t_c = \delta/4$ combined with a backward Euler method ($\varphi(\mathbf{A}, b; 1) = 1/2$). Results are shown at time

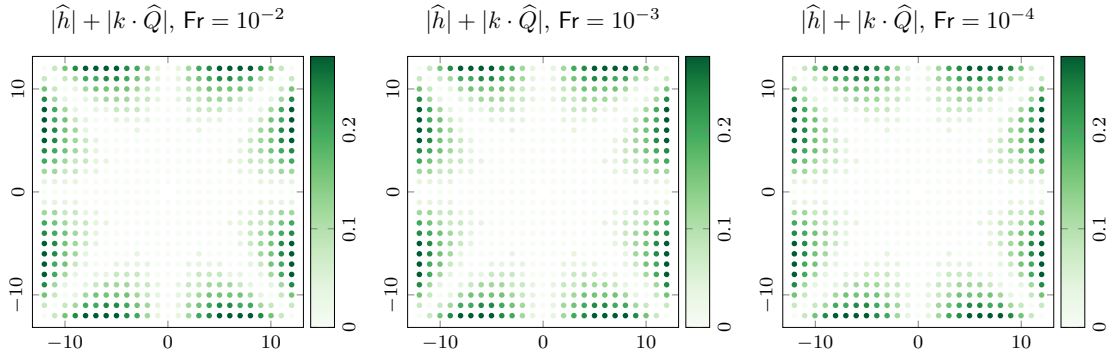

Figure 2.3: Constraint of incompressibility obtained from a solution of (2.47) at time $\Delta t_a = \mathsf{Fr}\delta/4$. Dots correspond to different values of $k \in \mathbb{Z}^2$. Left to right: decreasing values of Froude do not impact the constraint function $|\widehat{h}| + |k \cdot \widehat{Q}|$.

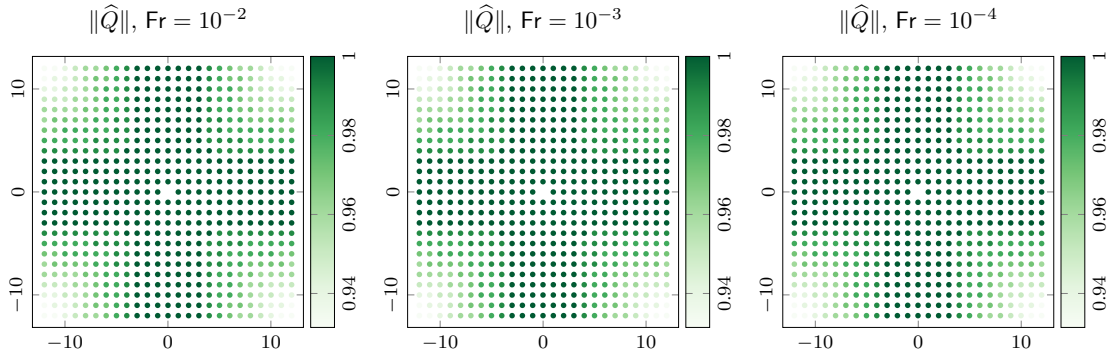

Figure 2.4: Norm of the discharge Fourier coefficients obtained from a solution of (2.47) when the time step is acoustic. Values are unchanged when reducing the Froude number.

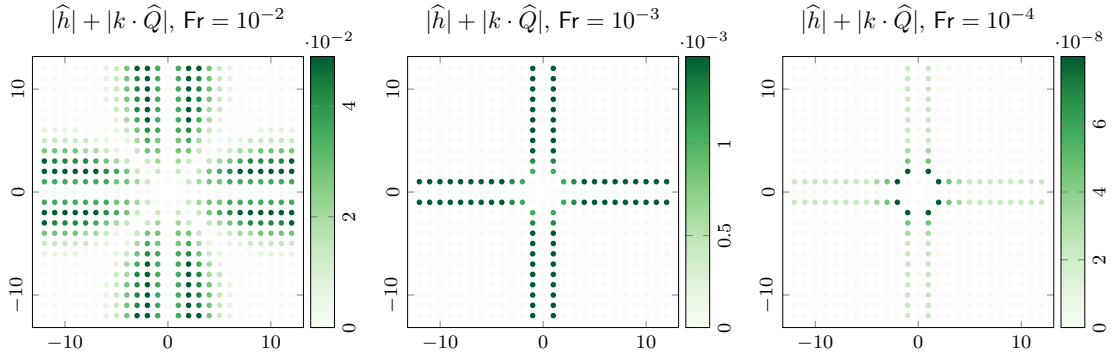

Figure 2.5: Constraint of incompressibility obtained from a solution of (2.47) at time $\Delta t_c = \delta/4$. The constraint function seems to vanish as $\mathsf{Fr} \to 0$.

$t = \Delta t_c$. In Figure 2.5, we witness a decrease of the constraint function towards zero when $\mathsf{Fr} \to 0$. In physical variables, this signifies that the solution $U(\Delta t, \cdot)$ converges

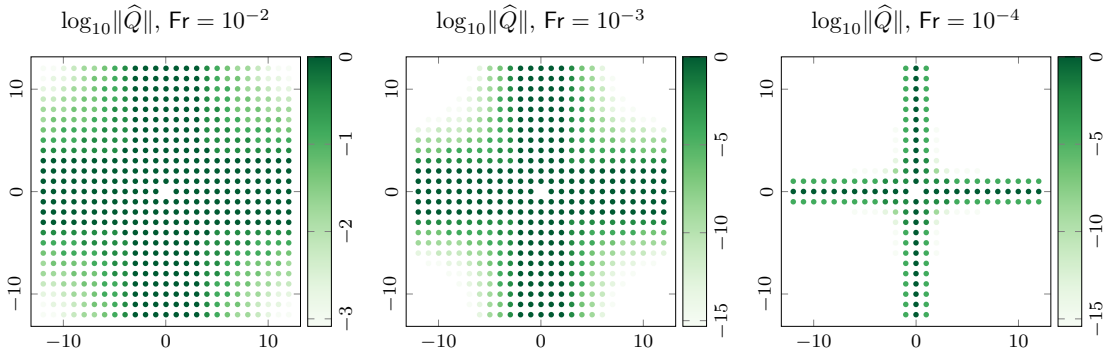

Figure 2.6: Norm of the discharge Fourier coefficients obtained from a solution of (2.47) when the time step is convective. Every mode is vanishing except along the main axes.

in $\mathcal{E}$. The Figure 2.6 showcases the vanishing of the discharge Fourier coefficients when $k_1 k_2 \neq 0$. In particular we can conjecture that

$$\lim_{\mathsf{Fr}\to 0} k_2 \widehat{Q}_y(k_1 \neq 0) = 0 \ , \quad \lim_{\mathsf{Fr}\to 0} k_1 \widehat{Q}_x(k_2 \neq 0) = 0 \ .$$

But from the previous Figure 2.5, we also believe that $k_1\widehat{Q}_x$ and $-k_2\widehat{Q}_y$ share the same limit. Therefore in the vanishing Froude limit we get that $k_1\widehat{Q}_x(k_1 \neq 0)$ and $k_2\widehat{Q}_y(k_2 \neq 0)$ both go to zero. In other words $Q_x(x,y) = Q_x(y)$ and $Q_y(x,y) = Q_y(x)$ in the limit, which coincides with the result of Proposition 2.4.4 for the simplified modified PDE (2.49).

Remark 2.4.6. *In Figures 2.3 and 2.5, the constraint function is kept to zero over the main axes $k_1 k_2 = 0$ and over the secondary axes $|k_1| = |k_2|$. In fact when either one of these equalities holds, we can take the scalar product between k and the last two equations of (2.47) in order to obtain a linear system of two equations for the unknowns $\widehat{h}$ and $k \cdot \widehat{Q}$. Since these quantities are initially zero, the unique solution is also zero.*

Now that we have a better understanding of how the solutions to the modified PDE (2.46) behave, we will confront this with numerical simulations of the Saint-Venant system using L^{upwind} with both small and large time steps. The testcase will be given by the Gresho Vortex over a varying bathymetry, which is a steady state whose derivation has been included in Appendix 2.C. The results are plotted in Figure 2.7 at time $1/2$, with the reference solution in the first row. The discharge components are more diffused by the explicit scheme, but overall we get a similar behavior compatible with our expectations regarding the vanishing of $\partial_x Q_x$ and $\partial_y Q_y$.

2.4.4 Scheme without acoustic diffusion

The defect of the first order naive scheme involving L^{upwind} is coming from the diffusive term found in the definition (2.45). In fact it is this term that shapes the leading order spatial error $R_\delta[L^{\text{upwind}}]$ scaling as $1/\mathsf{Fr}$ even when the data is well prepared. A simple solution would then be to just remove it, and this will get us to consider the second order centered operator $L^\star$ defined in (2.41). In the setting of explicit time integrators this

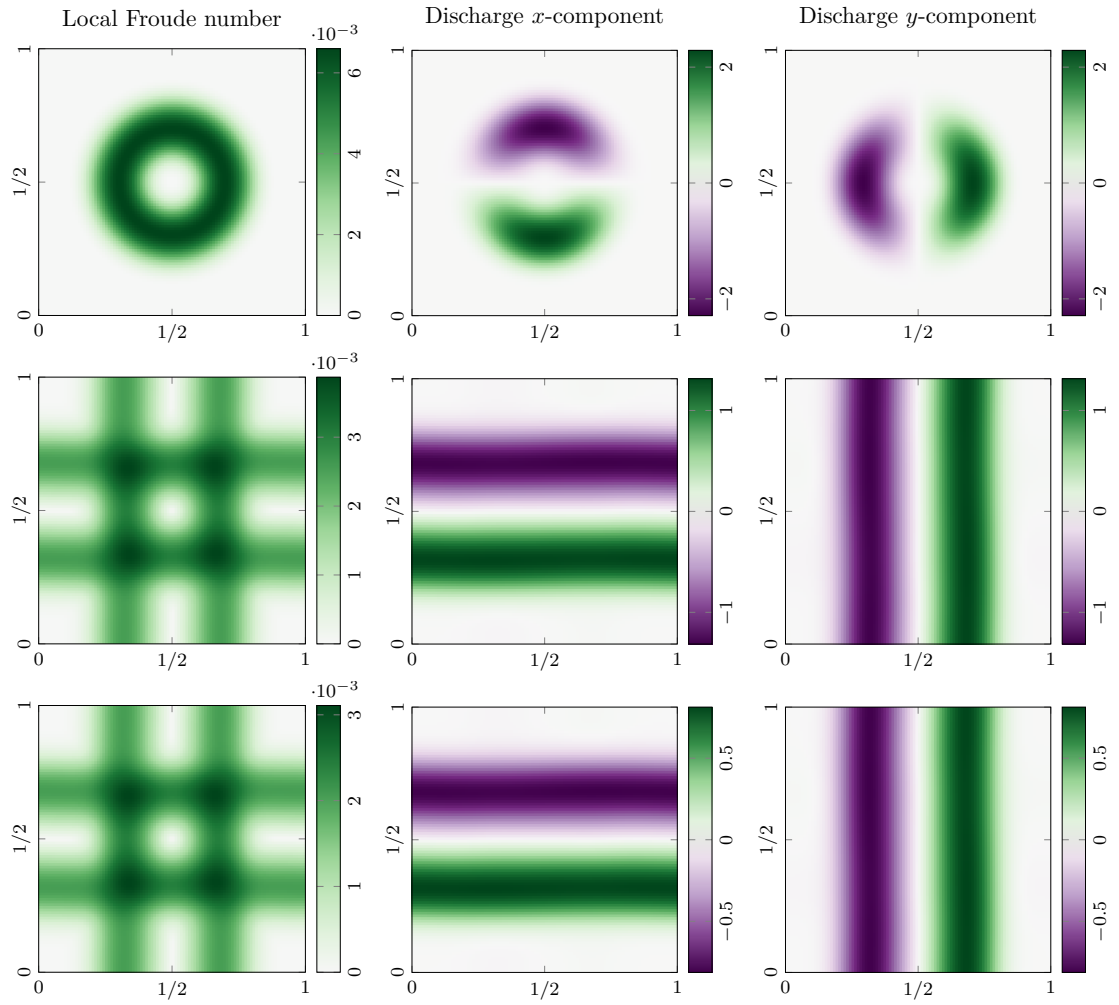

Figure 2.7: Top: initial condition. Second row: implicit scheme using large time steps. Third row: explicit scheme using small time steps.

would cause stability issues as per Remark 2.3.16, but since we focus on implicit in time methods for the acoustic part, this should not be a cause for concern. Thanks to this choice, the leading order error term will come from the time discretization only, which we keep at first order accuracy. This means that the first order modified PDE of the fully discrete scheme is stationary over $\mathcal{E}$, in particular the low Froude accuracy criterion is satisfied.

We now investigate if a scheme using $L^\star$ is asymptotically consistent when combined with discretization (2.44) for the convection. This will of course depend on the choice of Butcher tables, as it did in the time semi-discrete case. The next result states the invariance of a discrete counterpart $\mathbb{W}_{\mathrm{p}}^\star$ of the well prepared set $\mathbb{W}_{\mathrm{p}}$ introduced in (2.13),

that we define as

$$\mathbb{W}_{\mathrm{p}}^{\star} = \left\{ \sum_{k\in\mathbb{N}} \mathsf{Fr}^k \begin{pmatrix} \mathsf{h}_k \\ \mathsf{V}_k \end{pmatrix} \in (\mathbb{R}^{\mathcal{C}})^3,\ \mathsf{h}_0 + \mathsf{z} = \zeta_{\mathrm{ref}},\ \nabla^{\star}\cdot(\mathsf{hV})_0 = \mathcal{O}(\delta),\ \nabla^{\star}\mathsf{h}_1 = 0 \right\} . \quad (2.54)$$

The convenient notation for the discrete nabla operator $\nabla^{\star}$ correspond to the second order centered discretization

$$\nabla^{\star} = \begin{pmatrix} \mu_x\overline{\partial}_x \\ \mu_y\overline{\partial}_y \end{pmatrix} . \quad (2.55)$$

The choice (2.54) stems from the fact that the initial projection $\mathcal{Q}$ onto cellwise-constant functions sends every smooth element of $\mathbb{W}_{\mathrm{p}}$ into $\mathbb{W}_{\mathrm{p}}^{\star}$, meaning that we have the inclusion $\mathcal{Q}(C^1(\mathbb{T}^2)\cap\mathbb{W}_{\mathrm{p}}) \subset \mathbb{W}_{\mathrm{p}}^{\star}$. Especially, a divergence-free constraint at the continuous level cannot be discretized exactly, which is why we only ask to have $\nabla^{\star}\cdot(\mathsf{hV})_0 = \mathcal{O}(\delta)$. Since the bathymetry profile is defined relatively to some reference, we choose to define it so that $\zeta_{\mathrm{ref}} = 0$.

Proposition 2.4.7 (Invariance of $\mathbb{W}_{\mathrm{p}}^{\star}$)**.** *Let $(\widetilde{\mathbf{A}}, \widetilde{b}; \mathbf{A}, b)$ be a double Butcher tableau with s stages, such the final update is a convex combination of the partial stages, and such that the implicit part is either of type A or CK. Assume that when applied along with the discrete operators $(H^{\mathrm{HLL}}, L^{\star})$ to initial data $(\mathsf{h}^0, \mathsf{V}^0) \in \mathbb{W}_{\mathrm{p}}^{\star}$, the resulting scheme yields partial updates admitting a decomposition in powers of Froude. Then the final update $(\mathsf{h}^1, \mathsf{V}^1)$ is also in $\mathbb{W}_{\mathrm{p}}^{\star}$.*

Proof. (Proposition 2.4.7) Similarly to the time semi-discrete case, it is sufficient to show that each partial update $(\mathsf{h}^{(j)}, \mathsf{V}^{(j)})_{1\leq j\leq s}$ belongs to $\mathbb{W}_{\mathrm{p}}^{\star}$. We proceed by induction.

Initialization. By hypothesis we have that $\mathsf{U}^{(0)} := \mathsf{U}^0 \in \mathbb{W}_{\mathrm{p}}^{\star}$. In the case where $a_{1,1} = 0$ we directly get that $\mathsf{U}^{(1)} \in \mathbb{W}_{\mathrm{p}}^{\star}$ too since $\mathsf{U}^{(1)} = \mathsf{U}^{(0)}$.

Recurrence. Let $j \in \{1,\dots,s\}$ if $a_{1,1} \neq 0$, $j \in \{2,\dots,s\}$ otherwise. Assume that $\mathsf{U}^{(k)} \in \mathbb{W}_{\mathrm{p}}^{\star}$ for all k comprised between 0 and $j-1$ included. For the sake of simplicity, we replace the HLL flux H_Q^{HLL} by a standard Rusanov flux with global upwinding $\lambda_{\max} = \mathcal{O}(1)$ to ease the notations, but the arguments of the proof in the former case remain the same. In this setting the j-th partial update for the discharge reads:

$$\frac{\mathsf{Q}^{(j)} - \mathsf{Q}^0}{\Delta t} + \sum_{k=1}^{j-1} \tilde{a}_{jk} \left(\nabla^{\star}\cdot(\mathsf{hV}\otimes\mathsf{V})^{(k)} + \frac{1}{2\mathsf{Fr}^2}\nabla^{\star}(\mathsf{h}^{(k)} + \mathsf{z})^2 - \frac{\lambda_{\max}}{2} \begin{pmatrix} \Delta x \overline{\partial}_x^2 \mathsf{Q}_x^{(k)} \\ \Delta y \overline{\partial}_y^2 \mathsf{Q}_y^{(k)} \end{pmatrix} \right) + \sum_{k=1}^{j} a_{jk}\mathsf{c}^2\nabla^{\star}(\mathsf{h}^{(k)} + \mathsf{z}) = 0$$

Assume expansion $\mathsf{U}^{(j)} = \sum_{k\in\mathbb{N}} \mathsf{Fr}^k \mathsf{U}_k^{(j)}$ holds. Extracting the terms in Fr^{-2} leads to:

$$\sum_{k=1}^{j-1} \frac{\tilde{a}_{jk}}{2}\nabla^{\star}(\mathsf{h}_0^{(k)} + \mathsf{z})^2 - \sum_{k=1}^{j} a_{jk}\mathsf{z}\nabla^{\star}(\mathsf{h}_0^{(k)} + \mathsf{z}) = 0$$

Every summed term cancel for $k < j$ by assumption on the $\mathsf{U}_0^{(k)}$. Using furthermore that $a_{jj} \neq 0$, we find $\nabla^\star(\mathsf{h}_0^{(j)} + \mathsf{z}) = 0$. Similarly terms in Fr^{-1} are isolated so that we find $\nabla^\star \mathsf{h}_1^{(j)} = 0$. Next let N be such that the torus $\mathbb{T}^2$ is identified with $\cup_{0 \leq i,j \leq N} \mathcal{C}_{(i,j)}$. In particular this means that $\Delta x = \Delta y$, but what follows can be generalized to the case $\Delta x \neq \Delta y$. We extract the leading order terms from the mass update to get:

$$\frac{\mathsf{h}_0^{(j)} - \mathsf{h}_0^0}{\Delta t} = -\sum_{k=1}^{j} a_{jk} \nabla^\star \cdot \mathsf{Q}_0^{(k)} \tag{2.56}$$

A difference with the continuous case is that although the cellwise-constant function $\mathsf{h}_0^{(j)} - \mathsf{h}_0^0$ lies in the kernel of stencil $\nabla^\star$, it doesn't generally translate into $\mathsf{h}_0^{(j)} - \mathsf{h}_0^0$ being constant over the whole mesh. In fact this is wrong when $N+1$ the number of cells in one direction is even, due to the stencil $\nabla^\star$ being centered. To make up for this we can sum (2.56) over the sets $J_{\text{even}} = \{0 \leq i,j \leq N,\ i+j \text{ even}\}$ and $J_{\text{odd}} = \{0 \leq i,j \leq N,\ i+j \text{ odd}\}$, over which $\mathsf{h}_0^{(j)} - \mathsf{h}_0^0$ is constant. The case where $N+1$ is odd is treated naturally by summing over all cells and we omit it. As the computations are the same, we only detail the summation over J_{odd}:

$$(2.56) \implies |J_{\text{odd}}| \frac{\mathsf{h}_0^{(j)} - \mathsf{h}_0^0}{\Delta t} = -\sum_{k=1}^{j} a_{jk} \sum_{\mathcal{I} \in J_{\text{odd}}} (\nabla^\star \cdot \mathsf{Q}_0^{(k)})_{\mathcal{I}}$$

Since we are considering $N+1$ even and by periodicity we find:

$$\forall 1 \leq k \leq j,\ \sum_{\mathcal{I} \in J_{\text{odd}}} (\nabla^\star \cdot \mathsf{Q}_0^{(k)})_{\mathcal{I}} = 0$$

and as a consequence $\mathsf{h}_0^{(j)} = \mathsf{h}_0^0$ for every cell with index in J_{odd}. Since this holds true also for J_{even}, the equality $\mathsf{h}_0^{(j)} = \mathsf{h}_0^0$ is valid over the whole mesh. Using hypotheses $a_{jj} \neq 0$ as well as $\nabla^\star \cdot \mathsf{Q}^{(k)} = \mathcal{O}(\delta)$ for all $1 \leq k < j$, it follows directly that

$$\nabla^\star \cdot \mathsf{Q}_0^{(j)} = -\frac{1}{a_{jj}} \sum_{k=1}^{j-1} a_{jk} \nabla^\star \cdot \mathsf{Q}_0^{(k)} = \mathcal{O}(\delta) \tag{2.57}$$

We have proved that $(\mathsf{h}^{(j)}, \mathsf{V}^{(j)})$ belongs to $\mathbb{W}^\star_{\text{p}}$, and so does the final update by convex combination. □

Remark 2.4.8. *In practice, the decoupling between odd and even cells doesn't lead to a checkerboard pattern on the updated water height since in the proof we managed to get the equality* $\mathsf{h}_0^{(j)} = \mathsf{h}_0^0$*, where* h_0^0 *is taken equal to* $-\mathsf{z}$*.*

Proposition 2.4.7 is helpful to prove the asymptotic consistency of the considered scheme, which is stated below.

Proposition 2.4.9. *A fully discrete scheme satisfying hypotheses of Proposition 2.4.7 is asymptotically consistent with the limiting system* $(\mathcal{P}_0)$*.*

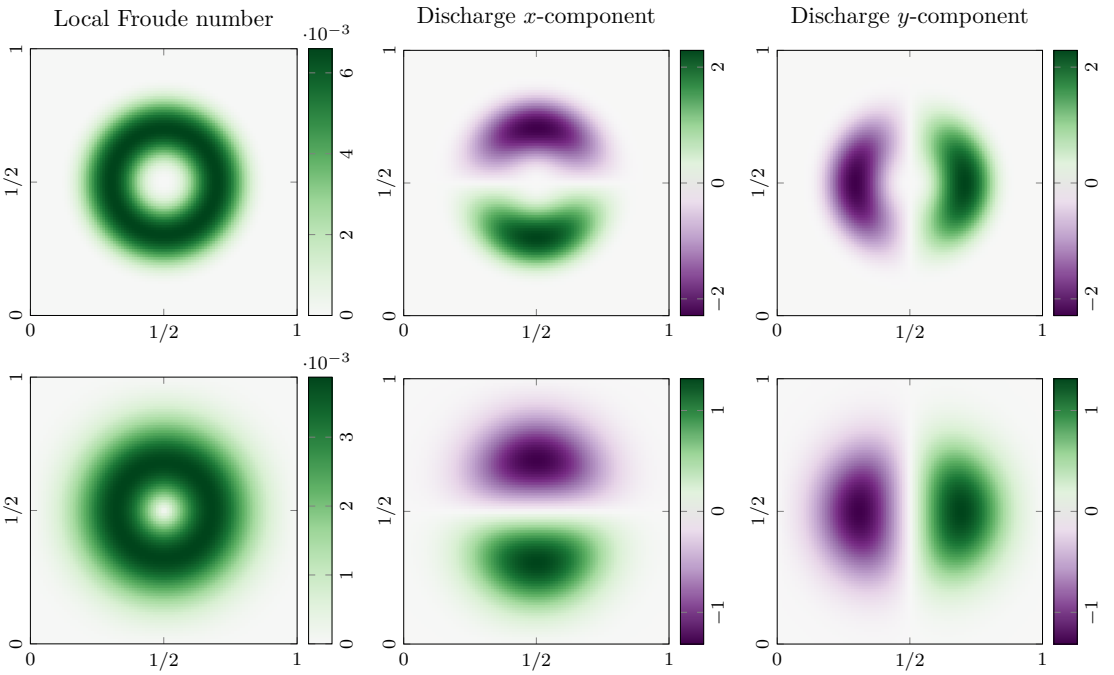

Figure 2.8: First row: reference solution. Second row: results obtained for the backward Euler method combined with operator $L^\star$.

Proof. We know from Proposition 2.4.7 that the final update U^{n+1} defines a cellwise-constant function lying in $\mathbb{W}^\star_{\mathrm{p}}$. This implies the existence of some $U \in \mathbb{W}$ such that

$$\lim_{\delta \to 0} \left\| U - \lim_{\mathsf{Fr} \to 0} \mathsf{U}^{n+1} \right\|_{L^2(\mathbb{T}^2)} = 0 \ .$$

The remainder of the proof focuses on the asymptotic consistency with respect to the velocity equation from $(\mathcal{P}_0)$. Consider the decomposition in powers of Froude $\mathsf{U}^{(j)} = \sum_{k \in \mathbb{N}} \mathsf{Fr}^k \mathsf{U}_k^{(j)}$. When extracted from the j-th partial update of the discharge, terms in Fr^0 give

$$\frac{\mathsf{Q}_0^{(j)} - \mathsf{Q}_0^0}{\Delta t} + \sum_{k=1}^{j-1} \tilde{a}_{jk} \left(\nabla^\star \cdot (\mathsf{h}\mathsf{V} \otimes \mathsf{V})_0^{(k)} - \frac{\lambda_{\max}}{2} \begin{pmatrix} \Delta x \overline{\partial}_x^2 \mathsf{Q}_{0,x}^{(k)} \\ \Delta y \overline{\partial}_y^2 \mathsf{Q}_{0,y}^{(k)} \end{pmatrix} \right) - \sum_{k=1}^{j} a_{jk} \mathsf{z} \nabla^\star \mathsf{h}_2^{(k)} = 0 \ , \tag{2.58}$$

where we used that $\nabla^\star \mathsf{h}_1^{(k)} = 0$. The update (2.58) is consistent at the continuous level with

$$\partial_t Q_0 + \nabla \cdot (hV \otimes V)_0 - z \nabla h_2 = 0 \iff \partial_t V_0 + (V_0 \cdot \nabla) V_0 + \nabla h_2 = 0 \ .$$

The equivalence holds because we have $h_0 = -z$ and $\nabla \cdot (hV)_0 = 0$. Hence we recover the desired equation on the velocity, and we conclude that the scheme is asymptotically consistent. □

In practice, the use of operator $L^\star$ combined to a first order double Butcher table offers much improved results compared to L^{upwind}. In Figure 2.8, we performed a simulation on the stationnary Gresho Vortex. The result is displayed at time 1/2, and we avoid the 1D diffusion that characterized the use of the upwind operator.

2.4.5 Second order approach

The case of second order schemes is now investigated. The goal is to compare several discretizations of the surface waves operator in order to underline the effectiveness of the near stationary condition (2.25) in predicting the behavior of second order methods. In particular we will underline that the exact $\mathcal{E}$-invariance is not mandatory to get accurate results. We recall that the $\mathcal{E}$-invariance, i.e. the ability of the modified PDE to keep solutions incompressible assuming they are so initially, was at the core of Dellacherie's original criterion. Instead, the refined near stationary condition states that it is acceptable for the solutions to deviate up to a $\mathcal{O}(\mathsf{Fr})$ term from an incompressible initial data. To this end we will compare between three stencils:

- the standard second order centered stencil which is shown to satisfy the refined criterion but not the exact $\mathcal{E}$-invariance;
- a modified version of the previous stencil that is also second order accurate, and satisfies both the refined criterion and the exact $\mathcal{E}$-invariance;
- a fourth order centered stencil that satisfies both criteria when combined with a second order time discretization;

The definition of these stencils will be specified latter, with the standard second order operator $L^\star$ already encountered in (2.41). We will see that the results between these discretizations are quite comparable in terms of accuracy and efficiency at low Froude numbers. In particular this means that the nearly stationary condition is able to explain why the second order centered scheme yields good results, even though the latter is not exactly $\mathcal{E}$-invariant.

In all considered schemes a second order semi-implicit DIRK method is selected, and the convection part is handled through the use of a MUSCL reconstruction, together with a minmod limiter to avoid spurious oscillations from appearing near areas with discontinuities. It is defined in the following way

$$\begin{aligned}
\mathsf{U}_{(i\pm 1/2\mp,j)} &= \mathsf{U}_{(i,j)} \pm \frac{\Delta x}{2}\text{minmod}(\mu_x\overline{\partial}_x\mathsf{U}_{(i,j)}, \partial_x^{\pm}\mathsf{U}_{(i,j)}) \ , \\
\mathsf{U}_{(i,j\pm 1/2\mp)} &= \mathsf{U}_{(i,j)} \pm \frac{\Delta y}{2}\text{minmod}(\mu_y\overline{\partial}_y\mathsf{U}_{(i,j)}, \partial_y^{\pm}\mathsf{U}_{(i,j)}) \ , \\
\text{minmod}(a,b) &= \frac{1}{2}(\text{sgn}(a)+\text{sgn}(b))\min(|a|,|b|) \ .
\end{aligned} \tag{2.59}$$

This reconstruction step happens just before computing the numerical fluxes. The latter will coincide with the null flux on the water height and HLL flux (2.44) on the discharge, as in the previous section.

We start by studying the second order scheme obtained by using the centered operator $L^\star$. This stencil has already been studied in conjunction with a first order DIRK method in Section 2.4.4, which allowed us to neglect the contribution of the spatial error in the modified PDE. This time we are combining $L^\star$ with a second order DIRK method, which means we can no longer neglect the error in space anymore. We will see that as a consequence, the modified PDE will not satisfy the $\mathcal{E}$-invariance criterion from Theorem 2.2.6. Despite this, the low Froude accuracy criterion will be satisfied, and it will be corroborated by accurate numerical results.

Proposition 2.4.10. *Let $(\mathbf{A}, b)$ be a second order accurate Butcher table. When associated to the discrete operator $L^\star$, the resulting scheme admits a modified PDE satisfying the low Froude accuracy property, but which is not $\mathcal{E}$-invariant.*

Proof. (Proposition 2.4.10) Consider the modified PDE of the scheme, taking the form

$$(\partial_t + L)U = (R_{\Delta t} - R_\delta[L^\star])U$$

where the detail of error operators is given thereafter

$$R_{\Delta t} = -\Delta t^2 \varphi(\mathbf{A}, b; 2) L^3, \quad R_\delta = \frac{\Delta x^2}{6}\mathbf{L}_x \partial^3_{xxx} + \frac{\Delta y^2}{6}\mathbf{L}_y \partial^3_{yyy} \ .$$

Hence the Fourier coefficients of the solution satisfy

$$\forall k \in \mathbb{Z}^2, \ \partial_t \widehat{U}(t,k) = \widehat{A}(k)\widehat{U}(t,k), \quad \widehat{A} = \mathbb{i} \begin{pmatrix} 0 & \alpha(k) & \beta(k) \\ c^2\alpha(k) & 0 & 0 \\ c^2\beta(k) & 0 & 0 \end{pmatrix} , \tag{2.60}$$

with the real coefficients α, β defined as

$$\begin{cases} \alpha(k) = -2\pi(1 + 4\pi^2 c^2 \Delta t^2 \varphi(\mathbf{A}, b; 2)|k|^2)k_1 - \dfrac{\Delta x^2}{6}(2\pi k_1)^3 \\ \\ \beta(k) = -2\pi(1 + 4\pi^2 c^2 \Delta t^2 \varphi(\mathbf{A}, b; 2)|k|^2)k_2 - \dfrac{\Delta y^2}{6}(2\pi k_2)^3 \end{cases} . \tag{2.61}$$

For convenience, we introduce the vector $\vartheta(k) = (\alpha(k), \beta(k))^T$ and diagonalize matrix tA as below.

$$\frac{1}{2|\vartheta|^2}\begin{pmatrix} 0 & -|\vartheta| & |\vartheta| \\ \vartheta^\perp & c\vartheta & c\vartheta \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 \\ 0 & -\mathbb{i}c|\vartheta|t & 0 \\ 0 & 0 & \mathbb{i}c|\vartheta|t \end{pmatrix} \begin{pmatrix} 0 & 2(\vartheta^\perp)^T \\ -|\vartheta| & \vartheta^T/c \\ |\vartheta| & \vartheta^T/c \end{pmatrix}$$

Computing the exponential of matrix, one finds the next expression

$$\frac{1}{|\vartheta|^2}\begin{pmatrix} \cos(c|\vartheta|t)|\vartheta|^2 & -\mathbb{i}\sin(c|\vartheta|t)|\vartheta|\vartheta^T/c \\ -\mathbb{i}c\sin(c|\vartheta|t)|\vartheta|\vartheta & \vartheta^\perp \otimes \vartheta^\perp + \cos(c|\vartheta|t)\vartheta \otimes \vartheta \end{pmatrix} .$$

Let $\widetilde{\vartheta}$ denote the unit vector $\vartheta/|\vartheta|$, the solutions of (2.60) can be expressed as

$$\begin{cases} \widehat{h}(t,k) = \cos(c|\vartheta|t)\widehat{h}^0(k) - \mathbb{i}\dfrac{\sin(c|\vartheta|t)}{c}\widetilde{\vartheta} \cdot \widehat{Q}^0(k) \\ \\ \widehat{Q}(t,k) = -\mathbb{i}c\sin(c|\vartheta|t)\widetilde{\vartheta}\widehat{h}^0(k) + (\widehat{Q}^0(k) \cdot \widetilde{\vartheta}^\perp)\widetilde{\vartheta}^\perp + \cos(c|\vartheta|t)(\widehat{Q}^0(k) \cdot \widetilde{\vartheta})\widetilde{\vartheta} \end{cases} . \tag{2.62}$$

We are now able to study the evolution of incompressible initial data characterized by

$$\forall k \neq 0, \quad \widehat{h}^0(k) = 0, \quad k \cdot \widehat{Q}^0(k) = 0 \ .$$

Let us remark that from (2.61) and by definition of vector ϑ, the latter can be decomposed as $\vartheta_{\Delta t} + \vartheta_\delta$ with $\vartheta_{\Delta t}$ the contribution of the consistent part and the error in time, and with ϑ_δ the contribution of the error in space. What is important to note is that $\vartheta_{\Delta t}(k)$ is colinear to k, and that $\vartheta_\delta(k)$ is neither colinear nor orthogonal to the wavenumber k. A direct consequence of this and of expression (2.62) is that, for any time $t > 0$ we do not have $k \cdot \widehat{Q}(t,k) = 0$, and the modified PDE cannot be exactly $\mathcal{E}$-invariant.

Now assume the resolution of the scheme is scale independent, meaning that $\Delta t, \Delta x, \Delta y$ do not depend on Froude. From (2.61), using $c^2 = -z/\mathsf{Fr}^2$ we see that

$$\begin{aligned} \alpha(k) &= -8\pi^3 c^2 \Delta t^2 \varphi(\mathbf{A}, b; 2)|k|^2 k_1 + \mathcal{O}(1) \\ \beta(k) &= -8\pi^3 c^2 \Delta t^2 \varphi(\mathbf{A}, b; 2)|k|^2 k_2 + \mathcal{O}(1) \end{aligned} \tag{2.63}$$

Hence there holds $\widetilde{\vartheta}(k) = k/|k| + \mathcal{O}(\mathsf{Fr}^2)$ and $\widetilde{\vartheta}(k)^\perp = k^\perp/|k^\perp| + \mathcal{O}(\mathsf{Fr}^2)$, and it ensues from (2.62) that for any time t, for any wavenumber $k \in \mathbb{Z}^2$,

$$\widehat{h}(t,k) = \mathcal{O}(\mathsf{Fr}^2), \quad \widehat{Q}(t,k) = \widehat{Q}^0(k) + O(\mathsf{Fr}^2) \ .$$

Thus the near stationary condition (2.25) is verified for a scale independent resolution which is precisely the definition of the low Froude accuracy criterion. □

Remark 2.4.11. *It is important to note that in the proof of Proposition 2.4.10, we restricted to using time steps Δt that are truly independent of Froude. On the contrary, if the updates were advanced within an acoustic time-scale ($\Delta t = \mathcal{O}(\mathsf{Fr})$), then we wouldn't have the near stationary condition* (2.25) *anymore since the contribution from the error in time* (2.63) *would scale as $c^2\Delta t^2 = \mathcal{O}(1)$. As we plan on using this scheme with large time steps afforded by the implicit time integration, this is not an issue.*

2.4.6 Second order modified stencil for exact $\mathcal{E}$-invariance

To better assess how well the second order scheme making use of $L^\star$ behaves despite not admitting an exactly $\mathcal{E}$-invariant modified PDE, we will consider and compare it with other discrete operators that are both low Froude accurate and $\mathcal{E}$-invariant. Similarly to what has been done in the first order case, a simple solution could be to discretize L with an order of accuracy strictly greater than that of the time discretization. This way, the error in space can be neglected in the modified PDE, and the solutions are stationary when the initial data is in $\mathcal{E}$. For instance we introduce the centered fourth order discretization $L^\sharp$ defined by

$$L^\sharp : (\mathsf{U}, \mathsf{z}) \longmapsto (\mathbf{L}_x \partial_x^\sharp + \mathbf{L}_y \partial_y^\sharp) \begin{pmatrix} \mathsf{h} + \mathsf{z} \\ \mathsf{Q} \end{pmatrix}, \quad \begin{cases} \partial_x^\sharp = \dfrac{1}{3}(2\mu_x^2 + \mathrm{id})\mu_x \overline{\partial}_x \\ \partial_y^\sharp = \dfrac{1}{3}(2\mu_y^2 + \mathrm{id})\mu_y \overline{\partial}_y \end{cases} . \tag{2.64}$$

Regarding the implementation of this operator, it is worthy to note tha its stencil doesn't only include neighboring cells. When generalizing to limit conditions other than periodic — such as Dirichlet, Neumann or based on Riemann invariants, one will have to use two layers of ghost cells. Anyway, this is usually required when dealing with spatial discretizations of order at least three. We illustrate the stencils of $\partial_x^\sharp, \partial_y^\sharp$ in the below.

$$\partial_x^\sharp = \frac{1}{12\Delta x} \times \begin{array}{|c|c|c|c|c|} \hline -1 & -4 & & 4 & 1 \\ \hline \end{array}\ , \quad \partial_y^\sharp = \frac{1}{12\Delta y} \times \begin{array}{|c|} \hline 1 \\ \hline 4 \\ \hline \\ \hline -4 \\ \hline -1 \\ \hline \end{array}\ .$$

Figure 2.9: Stencil discretizing the spatial derivatives intervening in definition of $L^\sharp$.

An alternative idea for getting stationary solutions while remaining second order accurate in space is to modify $L^\star$ such that the kernel of the associated leading truncation error contains all incompressible states. This way the sufficient condition from Proposition 2.4.2 will be satisfied, and with this in mind we have the following result.

Proposition 2.4.12. *Consider a scheme for the acoustic wave system obtained from combining a second order Butcher table* $(\mathbf{A}, b)$ *with a discrete operator* $L^\star + R^\star$*, where the centered discretization* $L^\star$ *was introduced in* (2.41)*, and where* $R^\star$ *is a discrete operator consistent with*

$$\frac{\Delta y^2}{6}\mathbf{L}_x\partial^3_{xyy} + \frac{\Delta x^2}{6}\mathbf{L}_y\partial^3_{xxy}\ .$$

Then such a scheme admits a modified PDE whose solutions in $\mathcal{E}$ *are stationary.*

Proof. When the bathymetry is flat, under the assumption of Proposition 2.4.12 the leading error term associated to operator $L^\star + R^\star$ is given by

$$\frac{\Delta x^2}{6}\partial_{xx}(\mathbf{L}_x\partial_x + \mathbf{L}_y\partial_y) + \frac{\Delta y^2}{6}\partial_{yy}(\mathbf{L}_x\partial_x + \mathbf{L}_y\partial_y) = \Big(\frac{\Delta x^2}{6}\partial_{xx} + \frac{\Delta y^2}{6}\partial_{yy}\Big)L\ ,$$

whose kernel clearly contains $\ker L$, and thus it contains the incompressible set $\mathcal{E}$. Hence an initial data from $\mathcal{E}$ remains constant in time. □

Discrete operators satisfying Proposition 2.4.12 can only be obtained by approximating third order crossed partial derivatives. This strategy has already been studied for instance in [41][11], where it enables to go further by designing methods that are exactly constraint-preserving at the discrete level with respect to the divergence-free condition. This means that in the incompressible case without bathymetry, it could be possible to construct

some update on the discrete velocity field V and to design some discrete divergence operator $\widetilde{\nabla}\cdot$ such that

$$\widetilde{\nabla}\cdot\mathsf{V}^n=0\Longrightarrow\widetilde{\nabla}\cdot\mathsf{V}^{n+1}=0\ . \tag{2.65}$$

The preservation of such a discrete constraint involves a truly multidimensional mechanism, and this gives an intuition as to why it is necessary to discretize crossed partial derivatives in order to satisfy (2.65). However this goes beyond the scope of the present work, as we believe that working at the level of the modified PDE is sufficient to ensure accurate results at low Froude numbers. Among the good candidates for approximating L while satisfying the condition of Proposition 2.4.12, we have operator $L^\flat$ defined here:

$$L^\flat:(\mathsf{U},\mathsf{z})\longmapsto(\mathbf{L}_x\partial_x^\flat+\mathbf{L}_y\partial_y^\flat)\begin{pmatrix}\mathsf{h}+\mathsf{z}\\ \mathsf{Q}\end{pmatrix},\quad\begin{cases}\partial_x^\flat=\mu_x\overline{\partial}_x+\dfrac{\Delta y^2}{6}(\mu_x\overline{\partial}_x)\overline{\partial}_y^2\\ \partial_y^\flat=\mu_y\overline{\partial}_y+\dfrac{\Delta x^2}{6}(\mu_y\overline{\partial}_y)\overline{\partial}_x^2\end{cases} \tag{2.66}$$

The stencils $\partial_x^\flat,\partial_y^\flat$ admit the graphical representation in Figure 2.10.

$$\partial_x^\flat=\frac{1}{2\Delta x}\times\begin{array}{|c|c|c|}\hline&&\\\hline -1&&1\\\hline&&\\\hline\end{array}+\frac{1}{12\Delta y}\times\begin{array}{|c|c|c|}\hline 1&-2&1\\\hline&&\\\hline -1&2&-1\\\hline\end{array}$$

$$\partial_y^\flat=\frac{1}{2\Delta y}\times\begin{array}{|c|c|c|}\hline&1&\\\hline&&\\\hline&-1&\\\hline\end{array}+\frac{1}{12\Delta x}\times\begin{array}{|c|c|c|}\hline -1&&1\\\hline 2&&-2\\\hline -1&&1\\\hline\end{array}$$

Figure 2.10: The stencils encountered in the definition of $L^\flat$ discretize crossed partial derivatives.

Proposition 2.4.13. *Let $(\widetilde{\mathbf{A}},\widetilde{b};\mathbf{A},b)$ be a double Butcher table satisfying assumptions of Proposition 2.4.7. We consider the numerical flux H^{HLL} on the convective part combined with MUSCL reconstruction. Regarding the choice of approximation for the acoustic part:*

1. *the use of $L^\star$ leads to a scheme that leaves the set $\mathbb{W}_{\mathrm{p}}^\star$ defined in (2.54) invariant;*
2. *the use of $L^\sharp$ allows to preserve the set $\mathbb{W}_{\mathrm{p}}^\sharp$ defined for $\nabla^\sharp=(\partial_x^\sharp,\partial_y^\sharp)^T$ as:*

$$\mathbb{W}_{\mathrm{p}}^\sharp=\left\{\sum_{k\in\mathbb{N}}\mathsf{Fr}^k\begin{pmatrix}\mathsf{h}_k\\ \mathsf{V}_k\end{pmatrix}\in(\mathbb{R}^{\mathcal{C}})^3,\ \mathsf{h}_0+\mathsf{z}=\zeta_{\mathrm{ref}},\ \nabla^\sharp\cdot(\mathsf{hV})_0=\mathcal{O}(\delta),\ \nabla^\sharp\mathsf{h}_1=0\right\}\ ;$$

3. the use of $L^\flat$ allows to preserve the set $\mathbb{W}_{\mathrm{p}}^\flat$ defined for $\nabla^\flat = (\partial_x^\flat, \partial_y^\flat)^T$ as:

$$\mathbb{W}_{\mathrm{p}}^\flat = \left\{ \sum_{k\in\mathbb{N}} \mathsf{Fr}^k \begin{pmatrix} \mathsf{h}_k \\ \mathsf{V}_k \end{pmatrix} \in (\mathbb{R}^{\mathcal{C}})^3,\ \mathsf{h}_0 + \mathsf{z} = \zeta_{\mathrm{ref}},\ \nabla^\flat \cdot (\mathsf{hV})_0 = \mathcal{O}(\delta),\ \nabla^\flat \mathsf{h}_1 = 0 \right\} ;$$

Any of the three aforementioned choices results in an asymptotically consistent scheme.

Proof. The proof of Proposition 2.4.13 follows the same lines to that of Proposition 2.4.7 and Proposition 2.4.9. This why we only focus on the one point that could potentially cause an issue, which is the slow pressure term in factor of Fr^{-2} in the discharge update

$$\frac{1}{2\mathsf{Fr}^2} \nabla^\star (\widetilde{\mathsf{h}}_0 + \widetilde{\mathsf{z}})^2 \ .$$

The tildes signify the use of values reconstructed by the MUSCL approach (2.59). In the case of a discrete lake at rest, the slopes involved in the reconstruction of the water height are equal to minus that of the bathymetry. In other words we have $\widetilde{\mathsf{h}}_0 + \widetilde{\mathsf{z}} = \mathsf{h}_0 + \mathsf{z}$, and this term is treated exactly as in the previous propositions. □

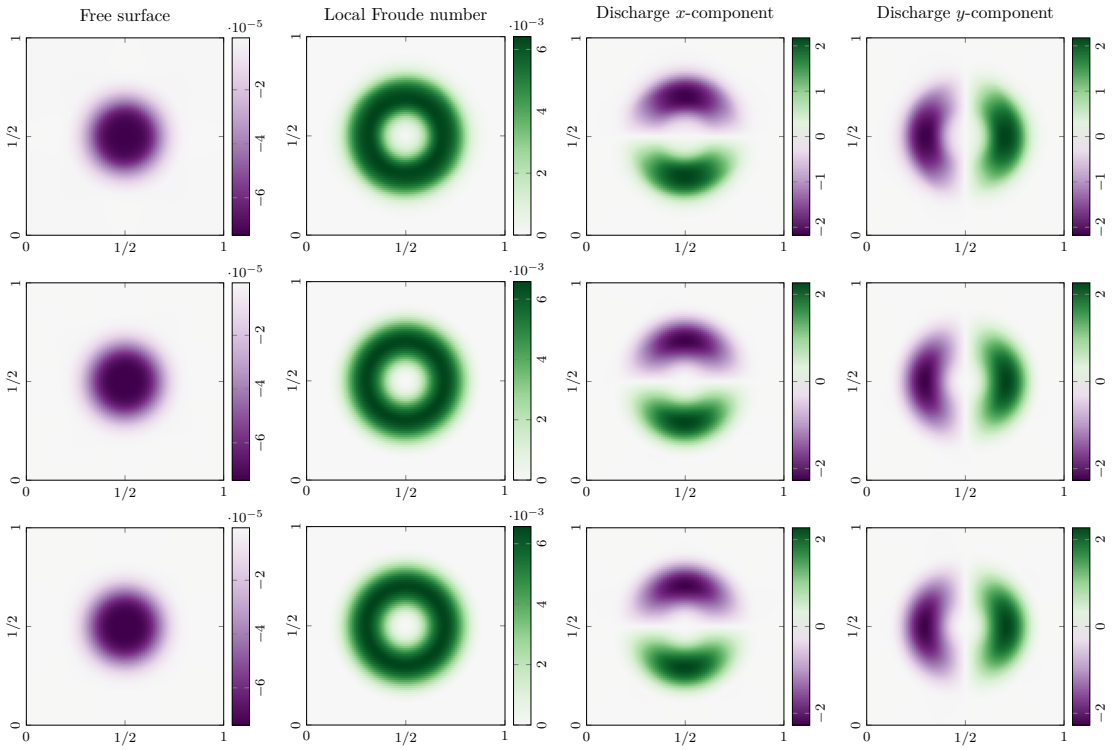

Figure 2.11: Results obtained using the ARS-(2,2,2) Butcher table combined to various acoustic operators. First row: centered second order $L^\star$, second row: modified second order $L^\flat$, third row: centered fourth order $L^\sharp$.

We compare qualitatively the schemes obtained from using the ARS-(2,2,2) Butcher table together with operators $L^\star$, $L^\flat$ and $L^\sharp$ in Figure 2.11. No real difference can be seen

between the standard centered approach of second order and the two other discretizations. It is a good point, since it means that no specific treatment is required to get accurate results at low Froude numbers.

As a more rigorous approach, we should instead look at whether the order of convergence deteriorates when the value of Fr decreases. We can see on the left cuves from Figure 2.12 that the schemes using $L^\star$, $L^\flat$ and $L^\sharp$ are unaffected by smaller values of Fr. Curves on the right give the efficiency. Idealy, we would like both the error and the CPU time to be as small as possible. We see that discrete operators $L^\flat, L^\sharp$ produce better results in this regard at Froude equal 1 and 10^{-2}, although the error for $L^\sharp$ increases when going from 64 to 128 cells per direction. When $\mathsf{Fr} = 10^{-4}$, the standard scheme using $L^\star$ seems quite competitive. We also featured an explicit scheme to highlight that it is less efficient because of the small time steps needed for stability. This scheme makes use of the same LAR wave splitting, and time integration is performed via the Heun method for both convection and acoustic steps. Note that this Butcher table is not of type A nor CK, and thus we cannot apply the result of Proposition 2.4.13 related to the asymptotic consistency.

2.5 Conclusion

In this work, we have revisited the issue of efficient and accurate schemes for the bidimensional Saint-Venant system over non-flat bottoms in the low Froude regime. The main contribution of our work is the study of the *low Froude accuracy* criterion for predicting if a scheme is able to yield accurate results when the Froude number becomes small. This criterion has been obtained by refining an existing condition proposed by Dellacherie in [28]. The difference lies in the fact that instead of asking for incompressible states to be left invariant by the modified PDE related to the discretization of the surface waves, we rather need to make sure that solutions remain close to the initial data if this latter is incompressible. This refined criterion was able to detect the origin of the loss of accuracy encountered in first order upwind schemes. In fact such schemes keep the numerical solutions close to an incompressible state, but not the good one. It seems that a general solution to cure the loss of accuracy is to consider a spatial discretization with an order of accuracy strictly greater than that of the time discretization, and was validated numerically for the implicit-explicit Euler method. When restricting to scale independent time steps, the low Froude accuracy also enables to justify the good behavior of a second order IMEX-DIRK scheme with centered discretization of the acoustic part, despite its modified PDE not being exactly $\mathcal{E}$-invariant. To validate our approach, we made several comparisons with modified schemes satisfying both the exact $\mathcal{E}$-invariance and the low Froude accuracy, and saw no real difference.

Aside from this, we have carried out an extensive study of the properties of IMEX-DIRK schemes, first in a time semi-discrete framework and then in the fully discrete case. Thanks to this, we know that any semi-discrete IMEX-DIRK scheme is low Froude accurate and that it is the spatial approximation that can lead to inaccuracies when the Froude number becomes small. Asymptotic consistency is obtained formally when the Butcher table of the acoustic part is of type A or CK, and when the final update rewrites

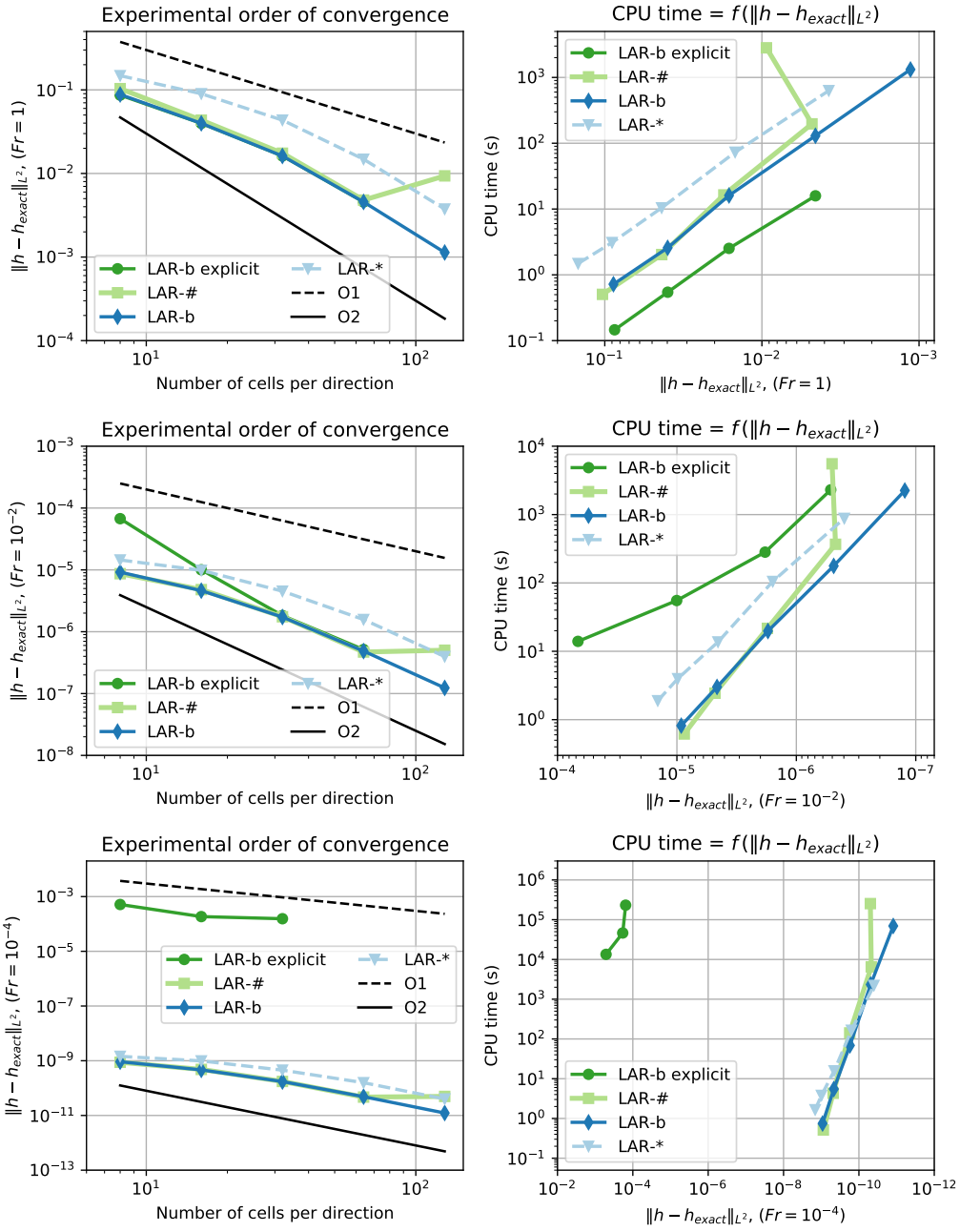

Figure 2.12: Convergence and efficiency curves for $\mathsf{Fr} = 1$ (top row), $\mathsf{Fr} = 10^{-2}$ (center row) and $\mathsf{Fr} = 10^{-4}$. Butcher table: ARS-(2,2,2).

as a convex combination of the partial updates. The L^2 stability of the surface waves time semi-discretization is obtained for arbitrary large time steps under some condition over the corresponding Butcher table, and as expected it seems that time integrators satisfying this condition are implicit ones. All the proposed schemes are preserving the hydrostatic equilibrium thanks to the choice of wave splitting already under use in [14].

There are several perspectives to this work. One of them would be to compare the dispersion relation of the surface gravity waves equation with that of the modified PDE of the scheme. This would allow to understand the behavior of the dispersion error in the low Froude regime. We have applied the low Froude accuracy criterion to first and second order methods, but we believe it can be generalized to arbitrary high orders. Thus it would be interesting to validate our findings at higher orders. We think that taking the Coriolis force into account should not cause any issue in the rotating lake regime, i.e. when the Froude number is small and the Rossby number remains of order unity. We also believe that discrete energy estimates can enable a rigorous justification of the asymptotic consistency, in opposition to the formal approach usually found in the literature and that we followed in this paper. A drawback that we did not mention so far is that the proposed second order schemes are not suited to handle discontinuities such as shocks, because they introduce spurious oscillations during the surface waves step. A MOOD approach could potentially solve this problem. This method consists to try to update the numerical solution with a high order scheme and, if oscillations are detected, the update is restarted by the mean of a first order scheme that doesn't oscillate. The hope is that the low order updates only occur every once in awhile, so that the experimental order of accuracy remains globally high. We refer to [29][53] for more details on this strategy. Another issue with the proposed schemes is that they do not warrant positivity of the water height. In practice this is not problematic when the water depth is consequent over the whole domain, such as in coastal flows, but it becomes a major hindrance when the handling of wet/dry transitions is relevant — as in littoral flows. This issue is rooted in the LAR splitting itself, as the surface waves system can lead to negative water heights at the continuous level. A solution could be to propose a hybrid approach where an IMEX-RK discretization based on the LAR splitting is used in areas where the water height is considered large, and a positivity preserving explicit scheme with hydrostatic reconstruction is used everywhere else. The CFL condition shouldn't be impacted too badly if the explicit scheme is applied in areas where the water depth is small enough, so that the speed of surface waves $\sqrt{h}/\mathsf{Fr}$ remains uniformly bounded with respect to the Froude number. A possible choice for the explicit scheme is the one studied by Duran [30], and is based on the (ζ, Q) coordinates.

Appendix

2.A Proofs related to modified PDEs

— Proof of Proposition 2.3.6 —

The proof follows the same lines as for deriving the order conditions, see for instance Boscarino [17]. Let U be the solution of (2.31) with initial condition $U(t=0,\cdot) = U^0(\cdot)$, and let U^1 be the discrete solution at time Δt obtained through the DIRK method associated to $(\mathbf{A}, b)$. We need to show that the local truncation error $U(\Delta t, \cdot) - U^1(\cdot)$ is in $\mathcal{O}(\Delta t^{p+2})$. We remark that

$$U(\Delta t) - U^1 = U(\Delta t) - (U^0 - \Delta t\, L\, \overline{U}(\Delta t) b) \tag{2.67}$$

where $\overline{U} : (\Delta t, x, y) \mapsto (U^{(1)}, \ldots, U^{(s)})(\Delta t, x, y) \in \mathbb{R}^{3\times s}$ is the matrix-valued function defined implicitly by the set of s linear PDEs corresponding to each one of the internal updates (2.29):

$$\overline{U}(\Delta t, x, y) = U^0(x, y)\, e^T - \Delta t\, L\, \overline{U}(\Delta t, x, y)\, \mathbf{A}^T \tag{2.68}$$

We perform a Taylor-Young expansion of $\overline{U}$ around $\Delta t = 0$ at order p:

$$\overline{U}(\Delta t) = U^0 e^T + \sum_{k=1}^{p} \frac{\Delta t^k}{k!} \frac{\partial^k \overline{U}}{\partial \Delta t^k}\bigg|_{\Delta t=0} + \mathcal{O}(\Delta t^{p+1})$$

Differentiating equation (2.68) $k \geq 1$ times with respect to Δt leads to the following relation:

$$\frac{\partial^k \overline{U}}{\partial \Delta t^k} = -\left(k\, L \left(\frac{\partial}{\partial \Delta t} \right)^{k-1} + \Delta t\, L \left(\frac{\partial}{\partial \Delta t} \right)^{k} \right) \overline{U}\, \mathbf{A}^T$$

Hence for $\Delta t = 0$ we get that $(\partial_{\Delta t})^k\, \overline{U}(0) = k!\, (-L)^k\, U^0 e^T (\mathbf{A}^T)^k$. The truncation error (2.67) can then be expressed in terms of Taylor expansions for $\overline{U}$ as well as U:

$$\begin{aligned} &U(\Delta t) - U^1 \\ &= \left[\left(\sum_{k=0}^{p+1} \frac{\Delta t^k}{k!} \frac{\partial^k U}{\partial t^k} \right) - U^0 + \Delta t\, L \sum_{k=0}^{p} (-\Delta t L)^k\, U b^T \mathbf{A}^k e \right]_{t=0} + \mathcal{O}(\Delta t^{p+2}) \end{aligned}$$

$$= \Delta t\left[\left(\frac{\partial}{\partial t} + L\right)U + \sum_{k=2}^{p+1} \frac{\Delta t^{k-1}}{k!}\frac{\partial^k U}{\partial t^k} + L\sum_{k=1}^{p}(-\Delta t L)^k\, U b^T \mathbf{A}^k e\right]_{t=0} + \mathcal{O}(\Delta t^{p+2})$$
$$= \Delta t\left[\left(\frac{\partial}{\partial t} + L\right)U + \frac{1}{\Delta t}\sum_{k=1}^{p}\left(\frac{1}{(k+1)!} - b^T\mathbf{A}^k e\right)(-\Delta t L)^{k+1}\, U\right]_{t=0} + \mathcal{O}(\Delta t^{p+2})$$

where we used that $e^T(\mathbf{A}^T)^k b = b^T\mathbf{A}^k e$ by symmetry of the scalar product, and in order to get to the last line we remark that equality $\Delta t^k \partial_{t^k}^k U = (-\Delta t\, L)^k U + \mathcal{O}(\Delta t^{p+2})$ holds for $2 \leq k \leq p+1$. Since terms between the brackets cancel out by hypothesis, this concludes the proof.

— Proof of Proposition 2.3.13 —

Let U be solution of (2.37), and let U^1 be the approximation of $U(\Delta t, x, y)$ obtained with the Butcher tableau $(\mathbf{A}, b)$. Taking the same steps as in the proof of proposition 2.3.6, we find that

$$U(\Delta t) - U^1$$
$$= \Delta t\left[\left(\frac{\partial}{\partial t} + L\right)U + \sum_{k=2}^{p+2} \frac{\Delta t^{k-1}}{k!}\frac{\partial^k U}{\partial t^k} + L\sum_{k=1}^{p+1}(-\Delta t L)^k\, U b^T \mathbf{A}^k e\right]_{t=0} + \mathcal{O}(\Delta t^{p+3})$$

From (2.37) we then remark that $\partial_t^k U = (R_{\Delta t} + \widetilde{R}_{\Delta t} - L)^k U$ hence we have

$$k = 2: \quad \Delta t^2 \partial_t^2 U = \Delta t^2 (L^2 U - 2R_{\Delta t} L U) + \mathcal{O}(\Delta t^{p+3})$$
$$k \geq 3: \quad \Delta t^k \partial_t^k U = \Delta t^k (-L)^k U + \mathcal{O}(\Delta t^{p+3})$$

It follows that up to $\mathcal{O}(\Delta t^{p+3})$ terms, the difference $U(\Delta t) - U^1$ is equal at time $t = 0$ to

$$\Delta t\left[\left(\frac{\partial}{\partial t} + L\right)U - \Delta t R_{\Delta t} L U + \sum_{k=1}^{p+1} \Delta t^k \Big(\frac{1}{(k+1)!} - b^T\mathbf{A}^k e\Big)(-L)^{k+1} U\right]_{t=0}$$

Using the assumption $\varphi(\mathbf{A}, b; k) = 0$ for all $0 \leq k \leq p-1$ this quantity is equal to

$$\Delta t\left[\left(\frac{\partial}{\partial t} + L\right)U + \Delta t^{p+1}\varphi(\mathbf{A}, b; p)(-L)^{p+2} U - \sum_{k=p}^{p+1} \Delta t^k \varphi(\mathbf{A}, b; k)(-L)^{k+1} U\right]_{t=0}$$

Finally the terms between the brackets cancel by hypothesis on U.

— Proof of Proposition 2.4.1 —

The proof is mostly the same than that of 2.3.6, only substituting L with L^* and $R_{\Delta t}$ with $R_{\Delta t} - R_\delta[L^*]$. Let $U \in (C^{p+1}(\mathbb{R}_+ \times \mathbb{T}^2))^3$ be a solution of (2.43) with initial condition $U(t = 0, \cdot) = U^0(\cdot)$. Let $\mathsf{U}^0 \in (\mathbb{R}^{\mathcal{C}})^3$ be the discrete initial data interpolated by U^0 at every cell center, and denote $\mathsf{U}^1 \in (\mathbb{R}^{\mathcal{C}})^3$ the update at time Δt obtained through

the DIRK method associated to $(\mathbf{A}, b)$ for the acoustic wave system. For every $(i,j) \in \mathbb{Z}^2$, we get the following relation by taking the same steps as in the proof of 2.3.6:

$$\begin{aligned} &U(\Delta t, x_i, y_j) - \mathsf{U}^1_{(i,j)} \\ &= \sum_{k=0}^{p+1} \frac{\Delta t^k}{k!} \frac{\partial^k U}{\partial t^k}(0, x_i, y_j) - U^0(x_i, y_j) - \sum_{k=0}^{p} (-\Delta t L^*)^{k+1}\, \mathsf{U}^0_{(i,j)} b^T \mathbf{A}^k e + \mathcal{O}(\Delta t^{p+2}) \end{aligned}$$

Note that because $\partial_t U = -LU + \mathcal{O}(\delta^p)$, the first sum on the right hand side becomes

$$\left[U + \Delta t \partial_t U + \sum_{k=2}^{p+1} \frac{\Delta t^k}{k!} (-L)^k U \right](0, x_i, y_j) + \mathcal{O}(\Delta t, \delta)^{p+2}$$

As for the second sum, since $U(t=0,\cdot)$ interpolates U^0 at every cell center we have that for all $(i,j) \in \mathbb{Z}^2$,

$$L^* \mathsf{U}^0_{(i,j)} = \sum_{(k,l) \in \mathrm{supp}(L^*)} \omega[L^*]_{(k,l)} \cdot U(0, x_{i+k}, y_{j+l}) = (L + R_\delta[L^*]) U(0, x_i, x_j) + \mathcal{O}(\delta^{p+1})$$

where the second equality holds thanks to L^* being consistent with $L + R_\delta[L^*]$ up to order $p+1$, and by definition of the consistency error $R_\delta[L^*]$. By recurrence, we then get $(L^*)^k \mathsf{U}^0_{(i,j)} = (L + R_\delta[L^*])^k U(0, x_i, y_j) + \mathcal{O}(\delta^{p+1})$ for all $k \in \mathbb{N}$. Hence we obtain:

$$\begin{aligned} &U(\Delta t, x_i, y_j) - \mathsf{U}^1_{(i,j)} \\ &= \quad \Delta t(\partial_t + L + R_\delta) U(0, x_i, y_j) + \sum_{k=2}^{p+1} \frac{\Delta t^k}{k!} (-L)^k U(0, x_i, y_j) \\ &\quad - \sum_{k=1}^{p} (-\Delta t)^{k+1} (L + R_\delta)^{k+1}\, U(0, x_i, y_j) b^T \mathbf{A}^k e + \mathcal{O}(\Delta t^{p+2}) \\ &= \left[\Delta t(\partial_t + L + R_\delta) + \sum_{k=1}^{p} \Big(\frac{1}{(k+1)!} - b^T \mathbf{A}^k e \Big) (-\Delta t L)^{k+1} \right] U(0, x_i, y_j) + \mathcal{O}(\Delta t, \delta)^{p+2} \\ &= \Delta t(\partial_t + L + R_\delta[L^*] - R_{\Delta t}) U(0, x_i, y_j) + \mathcal{O}(\Delta t, \delta)^{p+2} \\ &= \mathcal{O}(\Delta t, \delta)^{p+2} \end{aligned}$$

2.B Proofs of the properties in the time semi-discrete setting

This appendix gathers the proofs of the properties stated for time semi-discrete methods using the IMEX-DIRK strategy.

— Proof of Proposition 2.3.9 —

Assumption on the weights $b, \tilde{b}$ implies that the final update is a linear combination of the partial updates. Indeed in that case there holds $U^{n+1} = \sum_{j=1}^{s} \nu_j U^{(j)}$. Hence it

is sufficient to show that each partial update $U^{(0)}, \dots, U^{(s)}$ belongs to $\mathbb{W}_{\mathrm{p}}$, where we defined $U^{(0)} := U^n$ for consistency. We proceed by induction and shall assume that each $U^{(j)}$ admits an expansion in powers of Froude.

Initialization. By hypothesis we have that $U^{(0)} \in \mathbb{W}_{\mathrm{p}}$. In the case where $a_{1,1} = 0$ we directly get that $U^{(1)} \in \mathbb{W}_{\mathrm{p}}$ too since $U^{(1)} = U^{(0)}$.

Recurrence. Let $j \in \{1, \dots, s\}$ if $a_{1,1} \neq 0$, $j \in \{2, \dots, s\}$ otherwise. Assume that $U^{(k)} \in \mathbb{W}_{\mathrm{p}}$ for all k comprised between 0 and $j-1$ included. The j-th partial update for the discharge reads:

$$\frac{Q^{(j)} - Q^n}{\Delta t} + \sum_{k=1}^{j-1} \tilde{a}_{jk} \Big(\nabla \cdot (hV \otimes V)^{(k)} + \frac{1}{2\mathsf{Fr}^2} \nabla (h^{(k)} + z)^2 \Big) + \sum_{k=1}^{j} a_{jk} \frac{-z}{\mathsf{Fr}^2} \nabla (h^{(k)} + z) = 0$$

Assume expansion $U^{(j)} = \sum_{k \in \mathbb{N}} \mathsf{Fr}^k U_k^{(j)}$ holds. Extracting the terms in Fr^{-2} leads to:

$$\sum_{k=1}^{j-1} \frac{\tilde{a}_{jk}}{2} \nabla (h_0^{(k)} + z)^2 - \sum_{k=1}^{j} a_{jk} z \nabla (h_0^{(k)} + z) = 0 \tag{2.69}$$

Every summed term cancel for $k < j$ by assumption on the $U_0^{(k)}$. Using furthermore that $a_{jj} \neq 0$, we find $\nabla (h_0^{(j)} + z) = 0$. Similarly terms in Fr^{-1} are identified so that we find $\nabla h_1^{(j)} = 0$. Next we extract the leading order terms from the mass update and integrate it over the domain $\Omega = \mathbb{T}^2$:

$$\frac{h_0^{(j)} - h_0^n}{\Delta t} + \sum_{k=1}^{j} a_{jk} \nabla \cdot Q_0^{(k)} = 0 \implies |\Omega| \frac{h_0^{(j)} - h_0^n}{\Delta t} + a_{jj} \int_{\partial \Omega} Q_0^{(j)} \cdot n|_{\partial \Omega} \, \mathrm{d}\sigma = 0 \tag{2.70}$$

We have used that $h_0^{(j)} - h_0^n$ is space-independent and $\nabla \cdot Q_0^{(k)} = 0$ for $k < j$. Because we restrict to periodic boundary conditions, the integral cancel and we have $h_0^{(j)} = h_0^n$. This exactly imply the divergence-free condition $\nabla \cdot Q_0^{(j)} = 0$. To conclude $U^{(j)} \in \mathbb{W}_{\mathrm{p}}$ and thus U^{n+1} belongs to the same set.

— Proof of Proposition 2.3.12 —

Since $U^{n+1}(x, y; \mathsf{Fr}) \in \mathbb{W}_{\mathrm{p}}$ by proposition 2.3.9, we have $\lim_{\mathsf{Fr} \to 0} U^{n+1} \in \mathbb{W}$. Thus it only remains to show the asymptotic consistency with the velocity equation from $(\mathcal{P}_0)$. Again, we assume that for all $j \in \{1, \dots, s\}$, $U^{(j)} = \sum_{k \in \mathbb{N}} \mathsf{Fr}^k U_k^{(j)}$. Extracting terms in Fr^0 from the j-th partial update of the discharge we get:

$$\frac{Q_0^{(j)} - Q_0^n}{\Delta t} + \sum_{k=1}^{j-1} \tilde{a}_{jk} \Big[\nabla \cdot (hV \otimes V)_0^{(k)} + \frac{1}{2} \nabla (h_1^{(k)})^2 \Big] - \sum_{k=1}^{j} a_{jk} z \nabla h_2^{(k)} = 0$$

In the above, the leading order velocities $V_0^{(k)}$ are such that $Q_0^{(k)} = h_0^{(k)} V_0^{(k)}$ for all $1 \leq k \leq j$. Using then that $h_0^{(k)} = -z$, $\nabla h_1^{(k)} = 0$ and that $\nabla (hV \otimes V)_0^{(k)} = (h_0^{(k)} V_0^{(k)} \cdot \nabla) V_0^{(k)}$

due to the divergence-free condition on the leading order discharge, we finally get:

$$\frac{V_0^{(j)} - V_0^n}{\Delta t} + \sum_{k=1}^{j-1} \tilde{a}_{jk} (V_0^{(k)} \cdot \nabla) V_0^{(k)} + \sum_{k=1}^{j} a_{jk} \nabla h_2^{(k)} = 0$$

Collecting all the updates $1 \leq j \leq s$, this results in an IMEX-RK discretization of the velocity equation in $(\mathcal{P}_0)$, which is consistent as soon as $\sum_{j=1}^s \tilde{b}_j = \sum_{j=1}^s b_j = 1$.

— Proof of Proposition 2.3.14 —

Let $(h, Q)^T$ be a smooth solution of (2.31) on a flat bottom. The variations in time of the acoustic energy (2.36) are given by

$$\frac{\mathrm{d}}{\mathrm{d}t} E = (U, \partial_t U) = (U, (R_{\Delta t} + \widetilde{R}_{\Delta t} - L)U) \tag{2.71}$$

Error terms $R_{\Delta t} U$ and $\widetilde{R}_{\Delta t} U$ were respectively defined in propositions 2.31 and 2.37. As already mentionned, we have $(U, LU) = 0$ by integration by parts and by periodicity. Hence (2.71) simplifies to:

$$\frac{\mathrm{d}}{\mathrm{d}t} E = (U, R_{\Delta t} U + \widetilde{R}_{\Delta t} U)$$

Furthermore we know that one of the two errors $R_{\Delta t} U, \widetilde{R}_{\Delta t} U$ will cancel when taking the scalar product against U. Which one depends on the parity of p, and we will treat both cases separately. Assume for instance that we have $p = 2k$. Making use of the corresponding identity (2.38), we compute $(U, R_{\Delta t} U) = 0$, and thus there remains

$$\frac{\mathrm{d}}{\mathrm{d}t} E = \Delta t^{p+1} (\varphi(\mathbf{A}, b; p+1) - \varphi(\mathbf{A}, b; p)) \Big(\big(h, [(-L)^{2(k+1)} U]_h\big) + \big(Q, [(-L)^{2(k+1)} U]_Q\big) \Big)$$

Meanwhile we have:

$$(h, [(-L/c)^{2(k+1)} U]_h) = (h, \Delta^{k+1} h) = \begin{cases} \|\Delta^{(k+1)/2} h\|^2 & k \text{ odd} \\ -\|\Delta^{k/2} \nabla h\|^2 & k \text{ even} \end{cases}$$

Similarly we get

$$(Q, [(-L/c)^{2(k+1)} U]_Q) = -(\nabla \cdot Q, \Delta^k \nabla \cdot Q) = \begin{cases} -\|\Delta^{k/2} (\nabla \cdot Q)\|^2 & k \text{ even} \\ \|\Delta^{(k-1)/2} \nabla (\nabla \cdot Q)\|^2 & k \text{ odd} \end{cases}$$

Hence for even values $p = 2k$ the time derivative of E has same sign as $(-1)^{k+1}(\varphi(\mathbf{A}, b; p+1) - \varphi(\mathbf{A}, b; p))$, and the energy is dissipated if this quantity is negative. This is exactly the assumption made in the first point of proposition 2.3.14. Next we assume that $p = 2k + 1$. It follows from (2.38) that $(U, \widetilde{R}_{\Delta t} U) = 0$ and thus

$$\frac{\mathrm{d}}{\mathrm{d}t} E = \Delta t^p \varphi(\mathbf{A}, b; p) \Big(\big(h, [(-L)^{2(k+1)} U]_h\big) + \big(Q, [(-L)^{2(k+1)} U]_Q\big) \Big)$$

In this case the time derivative of E has same sign as $(-1)^{k+1}\varphi(\mathbf{A}, b; p)$, which concludes the proof since $k = (p-1)/2$.

— Proof of Proposition 2.3.17 —

We show by induction that each partial update $U^{(j)}$ for $0 \leq j \leq s$ is equal to U^0. Since lakes at rest belong to the kernel of both operators $\nabla \cdot H$ and L, the final update will then read $U^1 = U^0$ and the proof will be complete.

Initialization. Equality $U^{(0)} = U^0$ is true by definition.

Recurrence. Let k be such that $U^{(j)} = U^0$ for all $0 \leq j < k$. Partial update k then reads:

$$\begin{aligned} U^{(k)} &= U^0 - \Delta t \sum_{j=1}^{k-1} \widetilde{a}_{k,j} \nabla \cdot H(U^{(j)}, z) - \Delta t \sum_{j=1}^{k} a_{k,j} L(U^{(j)}, z) \\ \implies U^{(k)} &= U^0 - \Delta t a_{k,k} L(U^{(k)}, z) \end{aligned} \tag{2.72}$$

It is clear that $U^{(k)} = U^0$ is solution to the above equation. It remains to prove its unicity. Let $\widetilde{U}$ be another solution to (2.72), then by linearity of L the difference $U := \widetilde{U} - U^0$ satisfies the PDE $U = -a_{k,k}\Delta t L(U, z)$, whose unique solution is identically equal to zero. In fact there holds:

$$\begin{cases} h = -\Delta t a_{k,k} \nabla \cdot Q \\ Q = \Delta t a_{k,k} (z/\mathsf{Fr}^2) \nabla h \end{cases} \implies h = -\nabla \cdot (z \nabla h) = \lambda h$$

with $\lambda = (\Delta t a_{k,k}/\mathsf{Fr})^{-2} > 0$. Integrating this equality against h over $\mathbb{T}^2$ we get by integration by parts:

$$0 \leq \lambda \|h\|^2_{L^2(\mathbb{T}^2)} = -\int_{\mathbb{T}^2} h \nabla \cdot (z \nabla h) \, \mathrm{d}x = \int_{\mathbb{T}^2} z |\nabla h|^2 \, \mathrm{d}x \leq 0$$

The last inequality holds because z is taken negative. We conclude that necessarily $\widetilde{U} = U^0$ and the unique solution to the IMEX-DIRK k-th partial update is the lake at rest U^0.

2.C Derivation of the stationary vortex test-case

We restrict ourselves to steady states of the shallow water system with Coriolis source, thus verifying:

$$\begin{cases} \nabla \cdot (hV) = 0 \\ (V \cdot \nabla) V + \dfrac{1}{\mathsf{Ro}} V^{\perp} + \dfrac{1}{\mathsf{Fr}^2} \nabla (h + z) = 0 \end{cases} \tag{2.73}$$

The case without the Coriolis term is obtained by taking the limit of the Rossby number Ro towards $+\infty$. We rewrite this system under polar coordinates r, θ associated with the basis (e_r, e_θ) where $e_r = (\cos\theta, \sin\theta)^T$ and $e_\theta = (-\sin\theta, \cos\theta)^T$. Let $\Phi^{-1} : (r, \theta) \mapsto$

$(r\cos\theta, r\sin\theta)$ a bijective map. For conciseness, if f is a function of the cartesian variables, we denote by $\widehat{f}$ the function $f\circ\Phi$ in polar coordinates. It is then possible to establish the following relations:

$$\partial_r\widehat{f}(r,\theta) = \partial_r(f\circ\Phi^{-1})(r,\theta) = \partial_x f\partial_r\Phi_x^{-1} + \partial_y f\partial_r\Phi_y^{-1} = \cos\theta\partial_x f + \sin\theta\partial_y f \tag{2.74}$$

$$\partial_\theta\widehat{f}(r,\theta) = \partial_\theta(f\circ\Phi^{-1})(r,\theta) = \partial_x f\partial_\theta\Phi_x^{-1} + \partial_y f\partial_\theta\Phi_y^{-1} = -r\sin\theta\partial_x f + r\cos\theta\partial_y f \tag{2.75}$$

Reciprocally we find:

$$\left.\begin{aligned}\partial_x f &= \cos\theta\partial_r\widehat{f} - \frac{\sin\theta}{r}\partial_\theta\widehat{f}\\ \partial_y f &= \sin\theta\partial_r\widehat{f} + \frac{\cos\theta}{r}\partial_\theta\widehat{f}\end{aligned}\right\} \Longrightarrow \nabla f = e_r\partial_r\widehat{f} + \frac{e_\theta}{r}\partial_\theta\widehat{f} \tag{2.76}$$

Next we decompose the velocity vector alongside the vector of the polar basis (e_r, e_θ) by denoting u_r and u_θ its coordinates:

$$\widehat{V} = u_r e_r + u_\theta e_\theta$$

Hence we have:

$$\partial_x V_x = \cos\theta[\cos\theta\partial_r u_r - \sin\theta\partial_r u_\theta] - \frac{\sin\theta}{r}[\cos\theta\partial_\theta u_r - \sin\theta u_r - \sin\theta\partial_\theta u_\theta - \cos\theta u_\theta]$$

$$\partial_y V_y = \sin\theta[\sin\theta\partial_r u_r + \cos\theta\partial_r u_\theta] + \frac{\cos\theta}{r}[\sin\theta\partial_\theta u_r + \cos\theta u_r + \cos\theta\partial_\theta u_\theta - \sin_\theta u_\theta]$$

As a result, the divergence of the discharge becomes:

$$\nabla\cdot(hV) = \partial_r(\widehat{h}u_r) + \frac{\widehat{h}u_r}{r} + \frac{\partial_\theta\widehat{h}u_\theta}{r} = \frac{\partial_r(r\widehat{h}u_r)}{r} + \frac{\partial_\theta\widehat{h}u_\theta}{r}$$

and the divergence-free condition reads:

$$\partial_r(r\widehat{h}u_r) + \partial_\theta(\widehat{h}u_\theta) = 0 \tag{2.77}$$

Next, we develop the nonlinear Burgers convection term. From (2.74) and (2.75) we directly get:

$$(u_r e_r\cdot\nabla)V = u_r\partial_r\widehat{V} = u_r e_r\partial_r u_r + u_r e_\theta\partial_r u_\theta$$

$$(u_\theta e_\theta\cdot\nabla)V = \frac{u_\theta}{r}\partial_\theta\widehat{V} = \frac{u_\theta}{r}[e_r\partial_\theta u_r + e_\theta u_r + e_\theta\partial_\theta u_\theta - e_r u_\theta]$$

In the last line we made use of $\partial_\theta e_r = e_\theta$ and $\partial_\theta e_\theta = -e_r$. Regrouping the terms we get:

$$(V\cdot\nabla)V = \Big[u_r\partial_r u_r + \frac{u_\theta}{r}(-u_\theta + \partial_\theta u_r)\Big]e_r + \Big[u_r\partial_r u_\theta + \frac{u_\theta}{r}(u_r + \partial_\theta u_\theta)\Big]e_\theta \tag{2.78}$$

The Coriolis force is expressed as:

$$\frac{1}{\mathsf{Ro}}V^\perp = \frac{1}{\mathsf{Ro}}(u_r e_\theta - u_\theta e_r) \tag{2.79}$$

Collecting (2.76), (2.78) and (2.79), we decompose alongsied the basis vectors e_r and e_θ:

$$u_r\partial_r u_r - \frac{u_\theta^2}{r} + \frac{u_\theta}{r}\partial_\theta u_r - \frac{u_\theta}{\mathsf{Ro}} + \frac{1}{\mathsf{Fr}^2}\partial_r(\widehat{h}+\widehat{z}) = 0$$
$$u_r\partial_r u_\theta + \frac{u_r u_\theta}{r} + \frac{u_\theta}{r}\partial_\theta u_\theta + \frac{u_r}{\mathsf{Ro}} + \frac{1}{\mathsf{Fr}^2}\frac{\partial_\theta(\widehat{h}+\widehat{z})}{r} = 0$$

To conclude, the steady state solutions verify the following system in polar coordinates:

$$\left\{\begin{array}{l} \partial_r(r\widehat{h}u_r) + \partial_\theta(\widehat{h}u_\theta) = 0 \\ u_r\partial_r u_r - \frac{u_\theta^2}{r} + \frac{u_\theta}{r}\partial_\theta u_r - \frac{u_\theta}{\mathsf{Ro}} + \frac{1}{\mathsf{Fr}^2}\partial_r(\widehat{h}+\widehat{z}) = 0 \\ u_r\partial_r u_\theta + \frac{u_r u_\theta}{r} + \frac{u_\theta}{r}\partial_\theta u_\theta + \frac{u_r}{\mathsf{Ro}} + \frac{1}{\mathsf{Fr}^2}\frac{\partial_\theta(\widehat{h}+\widehat{z})}{r} = 0 \end{array}\right. \tag{2.80}$$

We consider the following axisymmetric solution of (2.80):

$$u_r = \partial_\theta u_\theta = \partial_\theta\widehat{h} = \partial_\theta\widehat{z} = 0 \tag{2.81}$$

The bathymetry is:

$$\hat{z}(r) = \exp(1/(r^2-1/9))\mathbf{1}_{r<1/3} - 5/2$$

We define the tangential velocity field as a C^2 piecewise polynomial:

$$u_\theta(r) = \left\{\begin{array}{ll} 5^6\times r^3(2/5-r)^3 & \text{if } 0\le r\le 2/5 \\ 0 & \text{otherwise} \end{array}\right.$$

Well prepared data obtained with $\widehat{h} = \widehat{h}_0 + \mathsf{Fr}^2\widehat{h}_2$ with $\widehat{h}_0 = -\widehat{z}$ and $\widehat{h}_2$ defined below:

$$\widehat{h}_2(r) = \int_0^r \frac{u_\theta(s)}{\mathsf{Ro}}\,\mathrm{d}s + \int_0^r \frac{u_\theta(s)^2}{s}\,\mathrm{d}s \quad \text{if } 0\le r\le 2/5, \quad \widehat{h}_2(2/5) \quad \text{if } r>2/5$$

with:

$$\int_0^r \frac{u_\theta(s)}{\mathsf{Ro}}\,\mathrm{d}s = \frac{5^6}{\mathsf{Ro}}\sum_{j=0}^{3}\binom{3}{j}\left(\frac{2}{5}\right)^{3-j}\frac{(-r)^{4+j}}{4+j}$$
$$\int_0^r \frac{u_\theta(s)^2}{s}\,\mathrm{d}s = 5^{12}\sum_{j=0}^{6}\binom{6}{j}\left(\frac{2}{5}\right)^{6-j}\frac{(-r)^{6+j}}{6+j}$$

2.D Butcher tables

We give some Butcher tables that have been investigated in the document.

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & 1 & 0 \end{array} \qquad \begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 0 & 1 \\ \hline & 0 & 1 \end{array}$$

Figure 2.D.1: Left: forward Euler. Right: backward Euler.

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & 1/2 & 1/2 \end{array} \qquad \begin{array}{c|cc} -1 & -1 & 0 \\ 2 & 1 & 1 \\ \hline & 1/2 & 1/2 \end{array}$$

Figure 2.D.2: Left: Heun. Overall IMEX scheme: JIN(2,2,2).

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \gamma & \gamma & 0 & 0 \\ 1 & \delta & 1-\delta & 0 \\ \hline & \delta & 1-\delta & 0 \end{array} \qquad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \gamma & 0 & \gamma & 0 \\ 1 & 0 & 1-\gamma & \gamma \\ \hline & 0 & 1-\gamma & \gamma \end{array}$$

Figure 2.D.3: ARS(2,2,2) obtained for the choice $\gamma = 1 - \sqrt{2}/2$ and $\delta = 1 - 1/(2\gamma)$.

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1/2 & 1/2 & 0 \\ \hline & 0 & 1 \end{array}$$

Figure 2.D.4: Explicit midpoint method.

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array}$$

Figure 2.D.5: Crank-Nicolson.

Chapter 3

Implicit kinetic schemes and iterative methods

3.1 Introduction

This chapter mainly deals with the one dimensional Saint-Venant equations in presence of a bathymetry with spatial variations, and gives a perspective towards the two dimensional case. The Saint-Venant system constitutes one of the simplest yet accurate nonlinear model for studying free surface flows with a small water height compared to the characteristic horizontal length, usually the domain width or the wavelength of the water waves. The smallness of the depth compared to the horizontal scale allows to neglect the non hydrostatic part of the pressure as a first order approximation. One can think of flows with a thin layer of water such as rivers, lakes or coastal flows which all enter this framework. As a vertically averaged model, the dimension is reduced by one which simplifies the study of the model and greatly diminishes the computational cost of numerical methods. Especially, the geometry of the fluid domain is naturally accounted for, and we don't need to keep track of the free surface, whose description is self-contained in the variables of the model.

Despite being a simplified model, the Saint-Venant system has several important properties that can be challenging to recover at the discrete level. It is a hyperbolic system of conservation laws with the possibility of shock and rarefaction waves developing in the solution, even if the initial data is smooth. The water height has to remain positive at all times, and a still free surface together with a null velocity constitutes a stationary equilibria called hydrostatic equilibrium or lake at rest. Other non stationary equilibria exist and they all follow Bernoulli's principle, but the hydrostatic equilibrium is probably the most important one since a numerical scheme preserving it usually guarantees better results even when the flow is not quite a lake at rest. Finally, we have to mention the entropy inequality satisfied by the energy of the system. A discrete counterpart to this inequality is desirable as it grants stability properties and enables to prove the convergence towards an entropy solution, see Bouchut and Lhebrard [19].

Recently kinetic solvers have been investigated [4][5][18][20][23][37][58], and they offer a favorable framework to satisfy the previous properties at the discrete level. In [10], the authors proposed an explicit kinetic scheme combined to the hydrostatic reconstruction strategy introduced in [6]. This method was shown to be positive, well balanced, and to verify a fully discrete entropy inequality with a positive error term which, unfortunately, is not always dominated by the dissipation arising from the upwinding of the numerical fluxes. As a consequence, in some cases the aforementioned scheme may increase the total energy in violation of the entropy inequality. A more general issue common to explicit finite volume schemes is that a CFL condition is required for stability. This condition can be quite restrictive when the time scale is consequent, or in presence of large wave velocities such as in the low Froude regime.

In light of these limitations, the goal of the present work is to explore an implicit kinetic approach leading to a fully discrete entropy inequality, what is more without any restriction on the time step. We also want to assess the interest and usability of such an implicit approach, in the sense that we get rid of the CFL condition in exchange of a greater algorithmic complexity. Furthermore, taking very large time steps is not always a wise choice as it can make the resolution inaccurate, so this also has to be taken into

account. In practice the update of our fully implicit scheme involves integrals without analytical expressions, and we have to simplify it by either giving up on the discrete entropy inequality, or by approximating it with an iterative strategy requiring once again a CFL condition for the sake of convergence.

The document is organized as follows. In Section 2, we recall the properties of the Saint-Venant system and the formalism of kinetic representations. In Section 3, we focus on the case of a flat bathymetry where we propose a fully implicit scheme admitting a discrete entropy inequality with no restriction on the time step. A simplified version of this scheme can be written explicitly at the macroscopic level. We detail its expression, how to implement it efficiently and perform numerical tests. In Section 4, we approximate the implicit scheme by an iterative approach and study the properties of the method. We also extend it to the case with varying bathymetry by the mean of the hydrostatic reconstruction, and validate it with numerical simulations. Finally, in Section 5 we give some perspectives towards the two dimensional case.

This project started in the context of the master internship of Antonin Leprevost. It was continued in this thesis in collaboration with Chourouk El Hassanieh and Jacques Sainte-Marie. In a way, the present work complements the LAR wave splitting approach proposed in the previous chapter, as it introduces an alternative strategy alleviating the lack of positivity and discrete entropy inequality.

3.2 Preliminaries about the Saint-Venant system

3.2.1 Properties of the model

In one spatial dimension, the Saint-Venant system reads

$$\begin{cases} \dfrac{\partial h}{\partial t} + \dfrac{\partial}{\partial x}(hu) = 0 \\ \dfrac{\partial}{\partial t}(hu) + \dfrac{\partial}{\partial x}\Big(hu^2 + \dfrac{g}{2}h^2\Big) = -gh\dfrac{\partial z}{\partial x} \end{cases} , \tag{3.1}$$

with $h(t,x) \geq 0$ the water height, $u(t,x)$ the horizontal velocity averaged on the vertical and $z(x)$ the bottom profile fixed in time. This set of equations admits the partialy conservative form

$$\partial_t U + \partial_x F(U) = S(U,z) \ , \quad U = \begin{pmatrix} h \\ hu \end{pmatrix} \ , \tag{3.2}$$

where $F(U) = (hu, hu^2 + gh^2/2)^T$ is the flux and $S(U,z) = (0, -gh\partial_x z)^T$ the source term. The eigenvalues associated to the flux jacobian are $\lambda_\pm(U) = u \pm \sqrt{gh}$ and imply the hyperbolicity of the model. We consider solutions U belonging to some convex set

$$\mathcal{U} \subset \mathbb{R}_+ \times \mathbb{R} \ ,$$

and satisfying the entropy inequality

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

given by the energy E of the system and its flux G defined as

$$E(U) = \frac{hu^2}{2} + \frac{gh^2}{2} + ghz, \quad G(U) = \Big(E(U) + \frac{gh^2}{2}\Big)u \ . \tag{3.3}$$

3.2.2 Kinetic representations

First introduced by Maxwell and Boltzmann, the kinetic theory aims at studying large systems of particles — such as those constitutive of gases and plasmas. The idea is to describe the state of a system by the mean of a distribution of particles $f(t, x, \xi)$ in the phase space related to the position x and the velocity ξ. Macroscopic quantities of interest such as the density or the momentum are recovered through integrals of the distribution function over all possible velocities. In our case the integral of f, called first moment, will yield a water height, whereas its integral against ξ, called second moment, defines a discharge. The kinetic equation ruling the evolution of the distribution function is given for all real ξ by

$$\frac{\partial f}{\partial t}(t, x, \xi) + \xi \frac{\partial f}{\partial x}(t, x, \xi) = \frac{1}{\varepsilon} Q[f](t, x, \xi) \ . \tag{3.4}$$

In the left hand side of Equation (3.4) we recognize a linear transport of particles characterized by velocity ξ. Conceptually, when two particles enter in collision they bounce and their velocities change. This mechanism is embedded on the right hand side through the frequency $1/\varepsilon$ and the collision operator Q satisfying the mass and momentum conservation constraints

$$\int_{\mathbb{R}} Q[f](t, x, \xi) \, \mathrm{d}\xi = \int_{\mathbb{R}} \xi Q[f](t, x, \xi) \, \mathrm{d}\xi = 0 \quad \text{for a.e. } (t, x) \ . \tag{3.5}$$

These constraints signify that overall, when accounting for all velocities the total mass and momentum of the system is unchanged by the collisions. There exist several choices compatible with (3.5), one of which is the Boltzmann operator resolving collisions by the mean of hard spheres mechanic (elastic collisions). However, as we don't need a very fine description of the collision process, we will make use of the simpler model introduced by Bhatnagar, Gross and Krook in [13].

The kinetic approach conveys a mesoscopic point of view, sitting between the microscopic level of individual particle dynamics and macroscopic laws. Great care was given to link these different descriptions, and it constitutes an active field of research to this day. Passing from microscopic to mesoscopic scale is achieved by taking the limit towards an infinite number of particles [32]. By doing so we get a continuum description of matter, yet the mean free path remains nonzero ($\varepsilon > 0$). On the other hand, transitioning from the kinetic level to the macroscopic level is achieved by taking the limit $\varepsilon \to 0$, meaning that particles are permanently colliding, and formally [62][36] solutions f from (3.4) reach an equilibrium $M(U_f, \xi)$ provided the kernel of operator Q satisfies the equivalence

$$Q[f] \equiv 0 \iff f(\xi) = M(U_f, \xi), \quad U_f = \int_{\mathbb{R}} \begin{pmatrix} 1 \\ \xi \end{pmatrix} f(\xi) \, \mathrm{d}\xi \ .$$

The distribution M is called maxwellian or Gibbs equilibrium, and in the hydrodynamical limit $\varepsilon \to 0$ Equation (3.4) becomes formally

$$\frac{\partial M}{\partial t} + \xi \frac{\partial M}{\partial x} = \mu(t,x,\xi) \ ,$$

with the measure μ satisfying the same conservation constraints (3.5) as $Q[f]$. For more details on this approach, we refer to the article [55] from Perthame. In order to link the hydrodynamic limit to the Saint-Venant equations (3.1), we need to choose the maxwellian so as to have the moment relations

$$\forall U \in \mathcal{U}, \ \int_{\mathbb{R}} \begin{pmatrix} 1 \\ \xi \end{pmatrix} M(U,\xi)\,\mathrm{d}\xi = U, \quad \int_{\mathbb{R}} \xi \begin{pmatrix} 1 \\ \xi \end{pmatrix} M(U,\xi)\,\mathrm{d}\xi = F(U) \ . \tag{3.6}$$

Note that although a generic distribution $f(t,x,\xi)$ allows to define a macroscopic state U_f, its integral against ξ^2 doesn't necessarily coincide with the second component of $F(U_f)$. A family of maxwellians satisfying (3.6) is given in the following lemma.

Lemma 3.2.1. *(Perthame and Simeoni [58]) Let $\chi : \mathbb{R} \to \mathbb{R}_+$ be an even shape function such that*

$$\forall \omega \in \mathbb{R}, \quad \chi(\omega) = \chi(-\omega) \geq 0, \quad \int_{\mathbb{R}} \chi(\omega)\,\mathrm{d}\omega = \int_{\mathbb{R}} \omega^2 \chi(\omega)\,\mathrm{d}\omega = 1 \ .$$

Then a maxwellian M satisfying the moment relations (3.6) *is obtained by setting for all $U \in \mathcal{U}$ and $\xi \in \mathbb{R}$*

$$M(U,\xi) = \frac{h}{c}\chi\Big(\frac{\xi - u}{c}\Big), \quad c = \sqrt{\frac{gh}{2}} \ .$$

Proof. One computes the integrals using the change of variable $\omega = (\xi - u)/c$. The odd nature of the shape function is used to compute the second and third moments.

$$\begin{aligned}
&\int_{\mathbb{R}} M(U,\xi)\,\mathrm{d}\xi = \frac{h}{c}\int_{\mathbb{R}} \chi(\omega)\, c\,\mathrm{d}\omega = h \ , \\
&\int_{\mathbb{R}} \xi M(U,\xi)\,\mathrm{d}\xi = \frac{h}{c}\int_{\mathbb{R}} (u + c\omega)\chi(\omega)\, c\,\mathrm{d}\omega = hu\int_{\mathbb{R}} \chi(\omega)\,\mathrm{d}\omega = hu \ , \\
&\int_{\mathbb{R}} \xi^2 M(U,\xi)\,\mathrm{d}\xi = \frac{h}{c}\int_{\mathbb{R}} (u + c\omega)^2\chi(\omega)\, c\,\mathrm{d}\omega = hu^2 + h\int_{\mathbb{R}} c^2\omega^2\chi(\omega)\,\mathrm{d}\omega = hu^2 + \frac{g}{2}h^2 \ .
\end{aligned}$$

□

We relate the kinetic description with the Saint-Venant system through the

Lemma 3.2.2. *U is a weak solution of* (3.1) *if and only if $M(U,\cdot)$ satisfies the moment relations* (3.6) *together with the kinetic representation*

$$\forall \xi \in \mathbb{R}, \ \partial_t M + \xi \partial_x M - g(\partial_x z)\partial_\xi M = \mu(t,x,\xi) \ , \tag{3.7}$$

where μ is subject to the conservation constraints (3.5).

The proof can be found in [10], and the last term on the left hand side allows to recover the source term by integration by parts. For numerical purposes, it will be more useful to replace (3.7) by the kinetic relaxation

$$\partial_t f + \xi \partial_x f - g(\partial_x z)\partial_\xi f = \frac{1}{\varepsilon}(M(U_f, \xi) - f) \ . \tag{3.8}$$

The right hand side of (3.8) corresponds to the BGK collision operator, which is the simplest one can think of. Yet, since the first two moments of $M(U_f, \xi)$ coincide with the ones from f, this collision operator satisfies conservation constraints (3.5) and thus leads to an hydrodynamic equilibrium compatible with the Saint-Venant system. We also remind the notion of kinetic entropy, which is useful to rewrite the entropy inequality (1.3) at the kinetic level.

Definition 3.2.3. *A kinetic entropy associated to the hydrodynamic equilibrium M is a function $H : \mathbb{R}_+ \times \mathbb{R} \to \mathbb{R}$ convex with respect to its first variable and satisfying for any admissible $U \in \mathcal{U}$*

$$\int_{\mathbb{R}} H(M(U, \xi), \xi)\, \mathrm{d}\xi = \eta(U) \ , \tag{E1}$$

and for any density $f(\xi)$

$$\int_{\mathbb{R}} H(M(U_f, \xi), \xi)\, \mathrm{d}\xi \leq \int_{\mathbb{R}} f(\xi)\, \mathrm{d}\xi \ . \tag{E2}$$

A kinetic entropy can be seen as a distribution of energy E at the kinetic level. A general framework to construct BGK models based on kinetic entropies was proposed by Bouchut in [20]. It stems from the fact that when there is no source term (in our case assume a flat bathymetry $\partial_x z = 0$), entropy inequality (1.3) is recovered at the limit $\varepsilon \to 0$ by multiplying (3.8) with $\partial_1 H(f(t, x, \xi), \xi)$. In fact doing so we get

$$\partial_t H(f, \xi) + \xi \partial_x H(f, \xi) = \frac{1}{\varepsilon} \partial_1 H(f, \xi)(M(U_f, \xi) - f) \ .$$

Using the convexity of H, the right hand side has to satisfy

$$\frac{1}{\varepsilon} \partial_1 H(f, \xi)(M(U_f, \xi) - f) \leq \frac{1}{\varepsilon}(H(M(U_f, \xi), \xi) - H(f, \xi)) \ .$$

We then integrate the resulting inequality, and using (E2) the upper bound obtained on the right hand side becomes zero

$$\partial_t \int_{\mathbb{R}} H(f, \xi)\, \mathrm{d}\xi + \partial_x \int_{\mathbb{R}} \xi H(f, \xi)\, \mathrm{d}\xi \leq \frac{1}{\varepsilon} \int_{\mathbb{R}} (H(M(U_f, \xi), \xi) - H(f, \xi))\, \mathrm{d}\xi = 0 \ .$$

Finally, in the limit $\varepsilon \to 0$ the distribution f converges to $M(U_f, \xi)$ and we recover the entropy inequality (1.3) by using (E1) and by defining the entropy flux G as the integral of $H(M(U_f, \xi), \xi)$ against ξ.

In the case of a single known entropy, a BGK model can be designed by first choosing a convex kinetic entropy, and then taking the maxwellian as the distribution minimizing the functional from (E2) under the constraints formed by the moment relations (3.6). Equality (E1) is then seen as a definition of the entropy. For the one-dimensional Saint-Venant system a commonly used kinetic entropy [58][10] is given by

$$H(f,\xi) = \frac{\xi^2}{2} f + \frac{g^2\pi^2}{6} f^3 + gzf \ , \tag{3.9}$$

whose terms from left to right correspond to the distribution kinetic energy, transverse translational energy and potential energy. As remarked by Perthame and Simeoni [58], the cubic contribution is a one-dimensional feature that is not required anymore in the two-dimensional case. They also proved that the maxwellian satifying the constrained minimization problem (E2) associated to kinetic entropy (3.9) is the half-disk maxwellian

$$\chi(\omega) = \frac{1}{\pi}\sqrt{\Big(1-\frac{\omega^2}{4}\Big)_+} \quad \Longrightarrow \quad M(U,\xi) = \frac{1}{g\pi}\sqrt{(2gh-(\xi-u)^2)_+} \ . \tag{3.10}$$

Other maxwellians can be considered, but to our knowledge they lack a kinetic entropy. Such alternatives are for instance

$$\chi(\omega) = \frac{1}{2\sqrt{3}}\mathbb{1}_{|\omega|\leq\sqrt{3}} \ , \quad \chi(\omega) = \Big(\frac{3}{20\sqrt{5}}\omega^2 + \frac{3}{4\sqrt{5}}\Big)\mathbb{1}_{|\omega|\leq\sqrt{5}} \ .$$

3.3 Kinetic schemes without source term

In this section we assume the bathymetry profile to be flat, so that no source term is present. Especially the evolution of quantities of interest only involves conservative flux variations. Since z is defined up to a constant, we can choose it so that $z \equiv 0$ and in this setting the kinetic entropy (3.9) will simplify to

$$H(f,\xi) = \frac{\xi^2}{2} f + \frac{g^2\pi^2}{6} f^3 \ . \tag{3.11}$$

3.3.1 Reminder on the explicit approach

We recall results about the explicit kinetic scheme that has been extensively studied by Audusse et. al in [10], first in its time semi-discrete form, and then in the fully discrete case. As usual Δt denotes the time step. Given some initial data $U^0 \in \mathcal{U}$, we wish to compute U^1 an approximation of the solution of the Saint-Venant system (3.1) at time Δt. To this end, the kinetic relaxation (3.8) is of great use, as it allows to replace a nonlinear system of two equations by a scalar equation with a linear transport that is easy to treat. However the presence of the BGK collision term $Q[f]$, which carries a nonlinear and integral dependence on the distribution f, is not straightforward to handle. Instead, we can perform a BGK splitting to treat the collision and transport terms in an

alternating way. Let us first focus on the collision step, which writes as

$$\begin{cases} \partial_t f = \dfrac{1}{\varepsilon}(M(U_f,\xi) - f(t,x,\xi)) \\ f(0,x,\xi) = f^0(x,\xi) \text{ with } \displaystyle\int_{\mathbb{R}} \begin{pmatrix} 1 \\ \xi \end{pmatrix} f^0(x,\xi)\,\mathrm{d}\xi = U^0(x) \end{cases} . \tag{3.12}$$

When integrating Equation (3.12) against $(1,\xi)^T$, the right hand side cancels and we find that the macroscopic quantity U_f is independent of time. So is $M(U_f,\xi) = M(U^0,\xi)$, and the unique solution of (3.12) is

$$f(t,x,\xi) = M(U^0(x),\xi) + \exp(-t/\varepsilon)(f^0(x,\xi) - M(U^0(x),\xi)) \ .$$

Regardless of the initial condition, we get a dampening towards the equilibrium at exponential rate in the hydrodynamic limit $\varepsilon \to 0$. Especially there holds

$$\forall t > 0 \ , \quad \lim_{\varepsilon \to 0} f(t,x,\xi) = M(U^0(x),\xi) \ . \tag{3.13}$$

A way of seeing the limit (3.13) is that all collisions are performed at once, and the equilibrium is reached instantaneously. Therefore the collision step is just a projection onto the space of maxwellians, and will serve as a starting point for the transport phase. The latter is simply solved via the method of characteristics

$$\begin{cases} \partial_t f + \xi \partial_x f = 0 \\ f(0,x,\xi) = M(U^0(x),\xi) \end{cases} \implies f(t,x,\xi) = M(U^0(x - t\xi),\xi) \ . \tag{3.14}$$

Hence the update defining the time semi-discrete scheme and approximating the solution U at time Δt is

$$U(\Delta t, x) \approx U^1(x) = \int_{\mathbb{R}} \begin{pmatrix} 1 \\ \xi \end{pmatrix} M(U^0(x - \Delta t \xi),\xi)\,\mathrm{d}\xi \ . \tag{3.15}$$

However a continuous representation in space is impractical, and in general we cannot compute integral (3.15) analytically. Instead we have to discretize in space too, and we approximate the transport equation over a uniform mesh constituted by cells $C_i = (x_{i-1/2}, x_{i+1/2})$ with center points $x_i = i\Delta x$ and interfaces $x_{i+1/2} = (i+1/2)\Delta x$ for $i \in \mathbb{Z}$. We assume the initial data U^0 to be cellwise constant, and denote by U_i^0 its value in cell C_i. In this simplified setting, the solution of the transport equation (3.14) at time Δt is expressed as

$$f(\Delta t, x, \xi) = \sum_{i \in \mathbb{Z}} M(U_i^0,\xi)\, \mathbb{1}_{C_i}(x - \Delta t \xi) \ . \tag{3.16}$$

To ensure that information doesn't travel more than one cell, we enforce the CFL condition $\sigma|\xi| \leq 1$ with $\sigma = \Delta t/\Delta x$, under which the sum in (3.16) only contains two terms. Note that when the maxwellian has compact support and when $(h_i^0)_{i\in\mathbb{Z}}$ and $(u_i^0)_{i\in\mathbb{Z}}$ are in $\ell^\infty(\mathbb{Z};\mathbb{R})$, we only need to consider velocities ξ in the domain $\Xi = \cup_{i\in\mathbb{Z}} \operatorname{supp} M(U_i^0,\cdot)$

which is bounded. Denoting $M_i^0 = M(U_i^0, \xi)$ and f_i^{1-} the average of $f(\Delta t, \cdot, \xi)$ over the cell C_i, we deduce the following relations from (3.16)

$$\begin{cases} f_i^{1-}(\xi > 0) = (1 - \sigma\xi)M_i^0 + \sigma\xi M_{i-1}^0 \\ f_i^{1-}(\xi < 0) = (1 + \sigma\xi)M_i^0 - \sigma\xi M_{i+1}^0 \end{cases} . \tag{3.17}$$

This can be recasted as the first order upwind scheme

$$\frac{f_i^{1-} - M_i^0}{\Delta t} + \frac{\xi}{\Delta x}\Big(\mathbb{1}_{\xi<0}(M_{i+1}^0 - M_i^0) + \mathbb{1}_{\xi>0}(M_i^0 - M_{i-1}^0)\Big) = 0 , \tag{3.18}$$

and we define the macroscopic approximation at time Δt by

$$U_i^1 = \int_{\mathbb{R}} \begin{pmatrix} 1 \\ \xi \end{pmatrix} f_i^{1-}(\xi)\, \mathrm{d}\xi . \tag{3.19}$$

Equivalently, by integrating (3.18) against $(1, \xi)^T$ we can relate $(U_i^1)_i$ to $(U_i^0)_i$ by the finite volume scheme

$$\frac{U_i^1 - U_i^0}{\Delta t} + \frac{1}{\Delta x}(F(U_i^0, U_{i+1}^0) - F(U_{i-1}^0, U_i^0)) = 0 . \tag{3.20}$$

The macroscopic numerical flux $F(U_L, U_R)$ can be written as $F_-(U_R) + F_+(U_L)$ where

$$F_-(U_R) = \int_{\mathbb{R}_-} \xi \begin{pmatrix} 1 \\ \xi \end{pmatrix} M(U_R, \xi)\, \mathrm{d}\xi , \quad F_+(U_L) = \int_{\mathbb{R}_+} \xi \begin{pmatrix} 1 \\ \xi \end{pmatrix} M(U_L, \xi)\, \mathrm{d}\xi . \tag{3.21}$$

This numerical flux can be computed explicitly with the half-disk maxwellian (3.10). It enters the framework of flux-vector splitting methods [18] since it is decomposed in a part carrying information coming from the left, and another part carrying information coming from the right. This upwinding is a way to account for the underlying structure of the solutions of Riemann problems from the kinetic level. The Roe scheme is a famous example but other exist, see for instance [49]. The interest of the kinetic approach is that it enables to prove several desirable properties, namely the positivity and discrete entropy inequality. This was shown in [10], and we recall the results and their proof in the below.

Proposition 3.3.1. *Let $\xi \in \mathbb{R}$. Under the CFL condition $\sigma|\xi| \leq 1$, the explicit kinetic scheme* (3.18) *satisfies the maximum principle given for any $i \in \mathbb{Z}$ by*

$$\min_{j\in\{-1,0,1\}} M_{i+j}^0(\xi) \leq f_i^{1-}(\xi) \leq \max_{j\in\{-1,0,1\}} M_{i+j}^0(\xi) . \tag{3.22}$$

Especially the positivity of the distribution function is preserved, that is to say

$$\forall i \in \mathbb{Z},\ M_i^0 \geq 0 \implies \forall i \in \mathbb{Z},\ f_i^{1-} \geq 0 .$$

Proof. Let $i \in \mathbb{Z}$ and fix the velocity $\xi \in \mathbb{R}$. We consider the explicit kinetic scheme (3.18) written under the form

$$f_i^{1-}(\xi) = (1 - \sigma|\xi|)M_i^0(\xi) + \sigma|\xi|\Big(\mathbb{1}_{\xi<0}M_{i+1}^0(\xi) + \mathbb{1}_{\xi>0}M_{i-1}^0(\xi)\Big) .$$

Assuming the CFL condition $\sigma|\xi| \leq 1$ holds, we see that the approximated distribution $f_i^{1-}(\xi)$ is a convex combination of $(M_{i+j}^0(\xi))_{-1\leq j\leq 1}$, and thus we get (3.22). The positivity of $f_i^{1-}(\xi)$ is a trivial consequence. □

The result of Proposition 3.3.1 is then used to recover the positivity of the updated water height h_i^1 under a CFL condition accounting for all the velocities encountered in the mesh. To be more specific, we denote by $\xi_{\max}$ the greatest kinetic velocity supported in the mesh at initial time, that is to say

$$\xi_{\max} = \sup_{i\in\mathbb{Z},\,\xi\in\mathbb{R}} \Big\{ |\xi|,\ M(U_i^0,\xi) \neq 0 \Big\} = \sup_{i\in\mathbb{Z}} \Big(|u_i^0| + \sqrt{2gh_i^0} \Big) \ . \tag{3.23}$$

For any $i \in \mathbb{Z}$ we have that $\operatorname{supp} f_i^{1-}(\cdot) \subset (-\xi_{\max}, \xi_{\max})$ as a consequence of (3.17). Then if the CFL condition $\sigma\xi_{\max} \leq 1$ holds we can apply the maximum principle (3.22) under the integral so that

$$h_i^1 = \int_{-\xi_{\max}}^{\xi_{\max}} f_i^{1-}(\xi)\,\mathrm{d}\xi \geq \int_{-\xi_{\max}}^{\xi_{\max}} \Big(\min_{j\in\{-1,0,1\}} M_{i+j}^0(\xi) \Big)\,\mathrm{d}\xi \ .$$

This lower bound is positive since we have $M_i^0 \geq 0$.

Remark 3.3.2. *The maximum principle* (3.22) *means that at the kinetic level, the explicit kinetic scheme* (3.18) *is L^∞-stable, in the sense that*

$$\forall \xi \in \mathbb{R}\ , \quad \sup_{i\in\mathbb{Z}} |f_i^{1-}(\xi)| \leq \sup_{i\in\mathbb{Z}} |M_i^0(\xi)| \ .$$

Note that despite having a maximum principle at the kinetic level, it is not true at the macroscopic level anymore. This is to be expected since the continuous Saint-Venant system doesn't satisfy such principle. Rather, we want to check whether a discrete entropy inequality is verified. This is achieved using the kinetic entropy, and we have the following statement.

Proposition 3.3.3. *(Audusse et al. [10]) Under the CFL condition $\sigma|\xi| \leq 1$, the explicit kinetic scheme* (3.18) *satisfies a fully discrete entropy inequality of the form*

$$H(f_i^{1-},\xi) \leq H(M_i^0,\xi) - \sigma\xi(H_{i+1/2}^0 - H_{i-1/2}^0)\ , \tag{3.24}$$

with the interfacial kinetic entropy $H_{i+1/2}^0$ defined by upwinding with respect to ξ

$$H_{i+1/2}^0 = \mathbb{1}_{\xi<0} H(M_{i+1}^0,\xi) + \mathbb{1}_{\xi>0} H(M_i^0,\xi)\ .$$

Proof (Proposition 3.3.3). The CFL restriction on the time step allows to write f_i^{1-} as a convex combination of the $(M_{i+j}^0)_{-1\leq j\leq 1}$. Using the convexity of H we then get

$$H(f_i^{1-},\xi) = \tag{3.25}$$
$$\begin{cases} H((1-\sigma\xi)M_i^0 + \sigma\xi M_{i-1}^0,\xi) \leq (1-\sigma\xi)H(M_i^0,\xi) + \sigma\xi H(M_{i-1}^0,\xi) & (\xi>0) \\ H((1+\sigma\xi)M_i^0 - \sigma\xi M_{i+1}^0,\xi) \leq (1+\sigma\xi)H(M_i^0,\xi) - \sigma\xi H(M_{i+1}^0,\xi) & (\xi<0) \end{cases}$$

Inequalities (3.25) depending on the sign of ξ can be combined in one

$$H(f_i^{1-},\xi) \leq$$
$$H(M_i^0,\xi) - \sigma\xi\Big(\mathbb{1}_{\xi<0}(H(M_{i+1}^0,\xi) - H(M_i^0,\xi)) + \mathbb{1}_{\xi>0}(H(M_i^0,\xi) - H(M_{i-1}^0,\xi)) \Big)\ ,$$

which is exactly (3.24). □

Corollary 3.3.4. *Under the CFL condition $\sigma\xi_{\max} \leq 1$ with $\xi_{\max}$ defined in* (3.23)*, the macroscopic explicit kinetic scheme* (3.20) *satisfies the discrete entropy inequality*

$$E(U_i^1) \leq E(U_i^0) - \sigma\left(\int_{\mathbb{R}} \xi\, H_{i+1/2}^0(\xi)\,\mathrm{d}\xi - \int_{\mathbb{R}} \xi\, H_{i-1/2}^0(\xi)\,\mathrm{d}\xi \right) .$$

where E is the energy defined in (3.3) *with $z \equiv 0$.*

Proof. As a direct consequence of relations (E1) and (E2) we have

$$E(U_i^1) = \int_{\mathbb{R}} H(M(U_i^1,\xi),\xi)\,\mathrm{d}\xi \leq \int_{\mathbb{R}} H(f_i^{1-},\xi)\,\mathrm{d}\xi .$$

Since the kinetic entropy $H(f_i^{1-},\xi)$ cancels outside of $(-\xi_{\max},\xi_{\max})$, the CFL condition $\sigma\xi_{\max} \leq 1$ allows to apply the result from Proposition 3.3.3 under the integral and conclude. □

3.3.2 Benefits of the implicit approach

The goal is now to compare the explicit kinetic scheme (3.20) with its implicit version. We will see that the latter will achieve similar stability properties in terms of water height positivity and discrete entropy dissipation, albeit unconditionally with respect to the time step which is an improvement. The implicit update reads

$$\frac{f_i^{1-} - M_i^0}{\Delta t} + \frac{\xi}{\Delta x}\Big(\mathbb{1}_{\xi<0}(f_{i+1}^{1-} - f_i^{1-}) + \mathbb{1}_{\xi>0}(f_i^{1-} - f_{i-1}^{1-}) \Big) = 0 . \tag{3.26}$$

This time, the fluxes are approximated at time Δt and will not coincide with maxwellian distributions in general. Over a mesh of N cells, we can rewrite this scheme as

$$\begin{cases} -\sigma\xi\mathbb{1}_{\xi>0} f_{i-1}^{1-} + (1+\sigma|\xi|) f_i^{1-} + \sigma\xi\mathbb{1}_{\xi<0} f_{i+1}^{1-} = M_i^0 & \forall 2 \leq i \leq N-1 \\ (1+\sigma|\xi|) f_1^{1-} + \sigma\xi\mathbb{1}_{\xi<0} f_2^{1-} = M_1^0 + \sigma\xi\mathbb{1}_{\xi>0} M_0^1 & \\ -\sigma\xi\mathbb{1}_{\xi>0} f_{N-1}^{1-} + (1+\sigma|\xi|) f_N^{1-} = M_N^0 - \sigma\xi\mathbb{1}_{\xi<0} M_{N+1}^1 & \end{cases} , \tag{3.27}$$

which will be useful to write the matrix of the system. The last two lines of system (3.27) use the quantities M_0^1 and M_{N+1}^1 defined at time Δt over ghost cells, and their purpose is to enforce the boundary conditions. For now on, we will assume that these maxwellians are known, and we will detail their treatment later in the document. We rewrite the scheme (3.27) under vector form, and to this end we introduce the vectors from $\mathbb{R}^N$ corresponding respectively to the maxwellians M^0 at initial time, the unknown distributions f^{1-} at time Δt and the ghost cells contributions B^1

$$M^0 = \begin{pmatrix} M_1^0 \\ M_2^0 \\ \vdots \\ M_N^0 \end{pmatrix} \in \mathbb{R}^N , \quad f^{1-} = \begin{pmatrix} f_1^{1-} \\ f_2^{1-} \\ \vdots \\ f_N^{1-} \end{pmatrix} \in \mathbb{R}^N , \quad B^1 = \begin{pmatrix} \xi M_0^1 \mathbb{1}_{\xi>0} \\ 0 \\ \vdots \\ 0 \\ -\xi M_{N+1}^1 \mathbb{1}_{\xi<0} \end{pmatrix} \in \mathbb{R}^N . \tag{3.28}$$

Then (3.27) is equivalent to

$$(\mathbf{I}+\sigma\mathbf{L})f^{1-} = M^0 + \sigma B^1 \ , \tag{3.29}$$

with $\mathbf{I}$ the identity matrix from $\mathbb{R}^{N\times N}$and

$$\mathbf{L} = \begin{pmatrix} |\xi| & \xi\mathbb{1}_{\xi<0} & 0 & \cdots & \\ -\xi\mathbb{1}_{\xi>0} & |\xi| & \xi\mathbb{1}_{\xi<0} & & \\ & \ddots & \ddots & \ddots & \\ & & -\xi\mathbb{1}_{\xi>0} & |\xi| & \xi\mathbb{1}_{\xi<0} \\ & \cdots & 0 & -\xi\mathbb{1}_{\xi>0} & |\xi| \end{pmatrix} \in \mathbb{R}^{N\times N} \ .$$

The mass matrix $\mathbf{I}+\sigma\mathbf{L}$ has important properties which will enable the implicit scheme (3.29) to comply with the desired stability properties.

Lemma 3.3.5. *For any value of $\xi \in \mathbb{R}$ and without any restriction on the time step, the mass matrix $\mathbf{I}+\sigma\mathbf{L}$ satisfies the following properties*

1. *it is invertible;*
2. *it is a monotone matrix (M-matrix), that is to say its inverse $(\mathbf{I}+\sigma\mathbf{L})^{-1}$ has only positive coefficients;*
3. *its inverse satisfies $\|(\mathbf{I}+\sigma\mathbf{L})^{-1}\|_{\infty,\infty} \leq 1$ with $\|\cdot\|_{\infty,\infty}$ the subordinate matrix norm associated to the vector norm $\|\cdot\|_\infty$;*

Proof. The mass matrix $\mathbf{I}+\sigma\mathbf{L}$ is strictly diagonally dominant and thus is invertible. To prove the second point we consider a vector $X = (x_1,\dots,x_N)^T \in \mathbb{R}^N$ such that $Y := (\mathbf{I}+\sigma\mathbf{L})X \geq 0$ component wise. It is sufficient to show that X has all its entries positive. Denoting k the index of the smallest component of X we have

$$0 \leq \sum_{j=1}^{N} (\mathbf{I}+\sigma\mathbf{L})_{k,j}\, x_j = (1+\sigma|\xi|)x_k - \sigma|\xi|\, x_{k\pm1} \implies x_k \geq \sigma|\xi|(x_{k\pm1} - x_k) \geq 0 \ .$$

The third point amounts to show that every coefficient of the inverse matrix is less than one in absolute value. We postpone the proof, which becomes clear when detailing the expression of the inverse, see Equation (3.33). □

As a consequence of Lemma 3.3.5 we have the

Proposition 3.3.6. *The system* (3.29) *is consistent with the continuous problem* (3.14)*, admits a unique solution, and thus defines an implicit kinetic scheme. Furthermore, the solution satisfies $f^{1-} \geq 0$.*

Proof. The consistency of the update comes from standard Taylor expansions and do not cause any difficulty. The existence and uniqueness of the solution is due to the invertibility of the mass matrix, so that the scheme is well defined. The positivity of the solution comes from the fact that $(\mathbf{I}+\sigma\mathbf{L})$ is an M-matrix and the right hand side $M^0+\sigma B^1$ has positive components. □

Proposition 3.3.7. *Assuming a flat bathymetry, the numerical scheme* (3.26) *involving the half-disk maxwellian* (3.10) *satisfies the fully discrete entropy equality*

$$H(f_i^{1-},\xi) = H(M_i^0,\xi) - \sigma\xi(H_{i+1/2}^{1-} - H_{i-1/2}^{1-}) + D_i \ , \tag{3.30}$$

with H the kinetic entropy presented in (3.11), *$H_{i\pm1/2}^{1-}$ the interfacial kinetic entropy defined by upwinding*

$$\begin{aligned} H_{i+1/2}^{1-} &= \mathbb{1}_{\xi<0}H(f_{i+1}^{1-},\xi) + \mathbb{1}_{\xi>0}H(f_i^{1-},\xi) \ , \\ H_{i-1/2}^{1-} &= \mathbb{1}_{\xi<0}H(f_i^{1-},\xi) + \mathbb{1}_{\xi>0}H(f_{i-1}^{1-},\xi) \ , \end{aligned}$$

and where D_i is a non positive term given by

$$\begin{aligned} D_i = \sigma\xi\,\frac{g^2\pi^2}{6}\Big(&\mathbb{1}_{\xi<0}(2f_i^{1-}+f_{i+1}^{1-})(f_{i+1}^{1-}-f_i^{1-})^2 - \mathbb{1}_{\xi>0}(2f_i^{1-}+f_{i-1}^{1-})(f_{i-1}^{1-}-f_i^{1-})^2\Big) \\ &- \frac{g^2\pi^2}{6}(2f_i^{1-}+M_i^0)(M_i^0-f_i^{1-})^2 \ . \end{aligned}$$

The proof makes use of the following lemma.

Lemma 3.3.8. *For all $(a,b)\in\mathbb{R}^2$, for all $\xi\in\mathbb{R}$ there holds*

$$\partial_1 H(a,\xi)(b-a) = H(b,\xi) - H(a,\xi) - \frac{g^2\pi^2}{6}(2a+b)(b-a)^2 \ .$$

Especially we recover that $\partial_1 H(a,\xi)(b-a) \le H(b,\xi) - H(a,\xi)$ which gives the convexity of $H(\cdot,\xi)$ over $\mathbb{R}_+$.

Proof (Lemma 3.3.8). We develop the left hand side

$$\begin{aligned} \partial_1 H(a,\xi)(b-a) &= \frac{\xi^2}{2}b + \frac{g^2\pi^2}{2}a^2b - \frac{\xi^2}{2}a - \frac{g^2\pi^2}{2}a^3 \\ &= H(b,\xi) + \frac{g^2\pi^2}{2}a^2b - \frac{g^2\pi^2}{6}b^3 - H(a,\xi) - \frac{g^2\pi^2}{2}a^3 + \frac{g^2\pi^2}{6}a^3 \\ &= H(b,\xi) - H(a,\xi) + \frac{g^2\pi^2}{6}\Big(3a^2b - 2a^3 - b^3\Big) \\ &= H(b,\xi) - H(a,\xi) - \frac{g^2\pi^2}{6}\Big(b^3 - a^3 - 3a^2(b-a)\Big) \ . \end{aligned}$$

We conclude by using the formula $b^3 - a^3 - 3a^2(b-a) = (2a+b)(b-a)^2$. □

Proof (Proposition 3.3.7). We multiply (3.18) by $\partial_1 H(f_i^{1-},\xi)$ in order to get

$$\partial_1 H(f_i^{1-},\xi)(f_i^{1-}-M_i^0) = -\sigma\xi\partial_1 H(f_i^{1-},\xi)\Big(\mathbb{1}_{\xi<0}(f_{i+1}^{1-}-f_i^{1-}) - \mathbb{1}_{\xi>0}(f_{i-1}^{1-}-f_i^{1-})\Big) \ . \tag{3.31}$$

We then apply Lemma 3.3.8 for $a = f_i^{1-}$ and $b\in\{f_{i-1}^{1-}, M_i^0, f_{i+1}^{1-}\}$ so that Equation (3.31) rewrites as

$$H(f_i^{1-},\xi) - H(M_i^0,\xi) + \frac{g^2\pi^2}{6}(2f_i^{1-}+M_i^0)(f_i^{1-}-M_i^0)^2 =$$

$$- \sigma\xi \mathbb{1}_{\xi<0}\Big(H(f_{i+1}^{1-},\xi) - H(f_i^{1-},\xi) - \frac{g^2\pi^2}{6}(2f_i^{1-} + f_{i+1}^{1-})(f_{i+1}^{1-} - f_i^{1-})^2\Big)$$
$$+ \sigma\xi \mathbb{1}_{\xi>0}\Big(H(f_{i-1}^{1-},\xi) - H(f_i^{1-},\xi) - \frac{g^2\pi^2}{6}(2f_i^{1-} + f_{i-1}^{1-})(f_{i-1}^{1-} - f_i^{1-})^2\Big) \ .$$

Regrouping the terms we get

$$\begin{aligned} H(f_i^{1-},\xi) - H(M_i^0,\xi) = \\ &- \sigma\xi\Big(\Big(\mathbb{1}_{\xi<0}H(f_{i+1}^{1-},\xi) + \mathbb{1}_{\xi>0}H(f_i^{1-},\xi)\Big) - \Big(\mathbb{1}_{\xi<0}H(f_i^{1-},\xi) + \mathbb{1}_{\xi>0}H(f_{i-1}^{1-},\xi)\Big)\Big) \\ &+ \sigma\xi\,\frac{g^2\pi^2}{6}\Big(\mathbb{1}_{\xi<0}(2f_i^{1-} + f_{i+1}^{1-})(f_{i+1}^{1-} - f_i^{1-})^2 - \mathbb{1}_{\xi>0}(2f_i^{1-} + f_{i-1}^{1-})(f_{i-1}^{1-} - f_i^{1-})^2\Big) \\ &- \frac{g^2\pi^2}{6}(2f_i^{1-} + M_i^0)(M_i^0 - f_i^{1-})^2 \ . \end{aligned}$$

This is exactly (3.30). □

As in the explicit case, the kinetic entropy dissipation (3.30) gives rise to a macroscopic entropy inequality upon integration, which is the Corollary 3.3.9 below. The proof is omitted since it is the same as that of Corollary 3.3.4, only with the additional term D_i that doesn't require any specific treatment.

Corollary 3.3.9. *The fully discrete entropy inequality*

$$E(U_i^1) \le E(U_i^0) - \sigma\Big(\int_{\mathbb{R}} \xi\, H_{i+1/2}^{1-}(\xi)\,\mathrm{d}\xi - \int_{\mathbb{R}} \xi\, H_{i-1/2}^{1-}(\xi)\,\mathrm{d}\xi\Big) + \int_{\mathbb{R}} D_i\,\mathrm{d}\xi$$

is satisfied by the implicit kinetic scheme (3.26).

3.3.3 Practical implementation of the fully implicit scheme

In this section we are interested in the practical aspect of how to implement the scheme (3.27). There are essentially two steps to go through and that we shall discuss. The first one is the inversion of the mass matrix $(\mathbf{I} + \sigma\mathbf{L})$, so as to get an analytic expression of the vector $f^{1-}(\xi)$. The second one is the computation of the integral of $f^{1-}(\xi)$ to define the macroscopic update U^1 as in (3.19). We will see that inverting the matrix will not cause any particular issue. On the other hand it seems to us that it is difficult, if possible at all, to compute the integral when using the half-disk maxwellian (3.10). The prospect of numerical integration using a quadrature formula to approximate this integral has one major hindrance, which is the loss of accuracy together with the higher computational cost that would be induced. We believe this can hardly be usable in practice, and instead we explore a compromise in which we substitute the half-disk maxwellian for a simpler one, namely the *index maxwellian* defined through the following shape function

$$\chi(\omega) = \frac{1}{2\sqrt{3}}\mathbb{1}_{|\omega|\le\sqrt{3}} \implies M(U,\xi) = \frac{h}{2\sqrt{3}\,c}\mathbb{1}_{|\xi-u|\le\sqrt{3}\,c} \ , \tag{3.32}$$

where we recall $c = \sqrt{gh/2}$. This means that we will be unable to prove the existence of a discrete entropy inequality. In fact Corollary 3.3.4 (and thus Corollary 3.3.9) isn't satisfied anymore for another choice of shape function χ. This is the otpimality argument of the half-disk maxwellian (3.10), which was constructed as the minimizer of the kinetic entropy 3.9, see Perthame and Simeoni [58], Lemma 2.3. Nevertheless, without any restriction on the time step the implicit scheme with index maxwellian (3.32) will still be positive, L^∞-stable at the kinetic level, and will enable us to write it explicitly at the macroscopic level, which is remarkable for a fully implicit scheme applied to a nonlinear system.

Matrix inversion. We begin by detailing the inversion of the mass matrix $(\mathbf{I} + \sigma\mathbf{L})$. To this end let us introduce the matrix $\mathbf{D} \in \mathbb{R}^{N\times N}$ as the diagonal part of $(\mathbf{I} + \sigma\mathbf{L})$, and $\mathbf{N} \in \mathbb{R}^{N\times N}$ the non diagonal part, that is to say

$$\mathbf{D}_{i,j} = \begin{cases} 1 + \sigma|\xi| & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}, \qquad \mathbf{N}_{i,j} = \begin{cases} -\sigma\xi \mathbb{1}_{\xi>0} & \text{if } i = j+1 \\ \sigma\xi \mathbb{1}_{\xi<0} & \text{if } i = j-1 \\ 0 & \text{otherwise} \end{cases}.$$

We have the following implication

$$\mathbf{I} + \sigma\mathbf{L} = \mathbf{D}(\mathbf{I} + \mathbf{D}^{-1}\mathbf{N}) \implies (\mathbf{I} + \sigma\mathbf{L})^{-1} = (\mathbf{I} + \mathbf{D}^{-1}\mathbf{N})^{-1}\mathbf{D}^{-1}.$$

One important property about matrix $\mathbf{N}$ is that it is either upper triangular or lower triangular depending on the sign of ξ, with zeros on its diagonal. Therefore matrix $\mathbf{D}^{-1}\mathbf{N}$ admits zero as its only eigenvalue of multiplicity N, and we can express the inverse via a geometric sum

$$(\mathbf{I} + \sigma\mathbf{L})^{-1} = \sum_{k\in\mathbb{N}} (-\mathbf{D}^{-1}\mathbf{N})^k \mathbf{D}^{-1}.$$

Next we use that matrix $\mathbf{D}^{-1}\mathbf{N}$ is nilpotent to get a finite sum that can be computed explicitly. In fact we have for all $1 \le i, j \le N$ and for all $k \in \mathbb{N}$

$$(\mathbf{N}^k)_{i,j} = \begin{cases} (-\sigma\xi)^k \mathbb{1}_{\xi>0} & \text{if } i - j = k \\ (\sigma\xi)^k \mathbb{1}_{\xi<0} & \text{if } j - i = k \\ 0 & \text{otherwise} \end{cases}$$

As a consequence we find

$$((\mathbf{I} + \sigma\mathbf{L})^{-1})_{i,j} = \begin{cases} \dfrac{(-\sigma\xi)^{j-i}}{(1-\sigma\xi)^{j-i+1}} \mathbb{1}_{\xi<0} & \text{if } j \ge i \\ \dfrac{(\sigma\xi)^{i-j}}{(1+\sigma\xi)^{i-j+1}} \mathbb{1}_{\xi>0} & \text{if } i \ge j \end{cases}. \tag{3.33}$$

Especially, we recover that every coefficient of the matrix inverse is positive, which gives the monotonicity (M-matrix). Furthermore, we see from (3.33) that the coefficients are less than one, which justifies the third point from Lemma 3.3.5 (maximum principle).

Integral computation. The next step is to write the solution of (3.29) at the macroscopic level, so as to define the updates h^1 and hu^1 depending on the data $(h^0, hu^0) \in \mathbb{R}^N \times \mathbb{R}^N$. To this end we need to express the integrals defining the following vectors of $\mathbb{R}^N$

$$\begin{cases} \overline{h} = \displaystyle\int_{\mathbb{R}} (\mathbf{I} + \sigma \mathbf{L})^{-1} M^0(\xi) \, \mathrm{d}\xi \\ \widetilde{h} = \displaystyle\int_{\mathbb{R}} (\mathbf{I} + \sigma \mathbf{L})^{-1} \sigma B^1(\xi) \, \mathrm{d}\xi \end{cases}, \qquad \begin{cases} \overline{hu} = \displaystyle\int_{\mathbb{R}} \xi \, (\mathbf{I} + \sigma \mathbf{L})^{-1} M^0(\xi) \, \mathrm{d}\xi \\ \widetilde{hu} = \displaystyle\int_{\mathbb{R}} \xi \, (\mathbf{I} + \sigma \mathbf{L})^{-1} \sigma B^1(\xi) \, \mathrm{d}\xi \end{cases}. \tag{3.34}$$

Then the macroscopic update consists to set $h^1 = \overline{h} + \widetilde{h}$ and $hu^1 = \overline{hu} + \widetilde{hu}$, which corresponds indeed to the solution of (3.29) integrated against 1 and ξ. We comment on the decomposition induced by (3.34). The overlined quantities refer to contributions coming from the interior of the computational domain, whereas the quantities with a tilde represent contributions coming from outside the domain. In fact, we recall that the purpose of vector B^1 is to enforce the boundary conditions. Its definition (3.28) makes use of the ghost values M_0^1 and M_{N+1}^1 that we assume to be known for now. We will see later how to determine these ghost maxwellians from the known macroscopic values U_1^0 and U_N^0 in the border cells.

We first focus on the interior contributions $\overline{h}$ and $\overline{hu}$. Developing the expression (3.33) obtained for the matrix inverse, we have for $1 \leq i \leq N$

$$\overline{h}_i = \sum_{j=i}^{N} \int_{\mathbb{R}_-} \frac{(-\sigma\xi)^{j-i}}{(1-\sigma\xi)^{j-i+1}} M_j^0(\xi) \, \mathrm{d}\xi + \sum_{j=1}^{i} \int_{\mathbb{R}_+} \frac{(\sigma\xi)^{i-j}}{(1+\sigma\xi)^{i-j+1}} M_j^0(\xi) \, \mathrm{d}\xi \, , \tag{3.35}$$

$$\overline{hu}_i = \sum_{j=i}^{N} \int_{\mathbb{R}_-} \xi \, \frac{(-\sigma\xi)^{j-i}}{(1-\sigma\xi)^{j-i+1}} M_j^0(\xi) \, \mathrm{d}\xi + \sum_{j=1}^{i} \int_{\mathbb{R}_+} \xi \, \frac{(\sigma\xi)^{i-j}}{(1+\sigma\xi)^{i-j+1}} M_j^0(\xi) \, \mathrm{d}\xi \, . \tag{3.36}$$

As already mentioned, finding analytic expressions for the above integrals (3.35) and (3.36) when dealing with the half-disk maxwellian doesn't seem an achievable goal. Instead, working with the simpler index maxwellian (3.32) makes it possible to compute these integrals, and we have the following results.

Proposition 3.3.10. *Let $1 \leq i \leq N$. We have the following analytical expressions for the water height $\overline{h}_i$ and the discharge $\overline{hu}_i$ when using the index maxwellian* (3.32)

$$\overline{h}_i = \frac{1}{2\sqrt{3}} \left(\sum_{j=i}^{N} \sqrt{\frac{2h_j^0}{g}} \, (\mathcal{A}h)_{i,j} + \sum_{j=1}^{i} \sqrt{\frac{2h_j^0}{g}} \, (\mathcal{B}h)_{i,j} \right) , \tag{3.37}$$

$$\overline{hu}_i = \frac{1}{2\sqrt{3}\,\sigma} \left(- \sum_{j=i}^{N} \sqrt{\frac{2h_j^0}{g}} \, (\mathcal{A}hu)_{i,j} + \sum_{j=1}^{i} \sqrt{\frac{2h_j^0}{g}} \, (\mathcal{B}hu)_{i,j} \right) . \tag{3.38}$$

Introducing the function $\phi : x \in \mathbb{R} \setminus \{-1\} \mapsto x/(1+x)$ and the velocities

$$\forall 1 \leq j \leq N \, , \quad a_j = u_j^0 - \sqrt{\frac{3}{2} g h_j^0} \, , \quad b_j = u_j^0 + \sqrt{\frac{3}{2} g h_j^0} \, ,$$

the triangular matrices $(\mathcal{A}h)$ *and* $(\mathcal{B}h)$ *can be written as*

$$(\mathcal{A}h)_{i,j} = \frac{\mathbb{1}_{j\geq i}}{\sigma}\Big[\ln(|1+x|) - \sum_{l=1}^{j-i}\frac{\phi(x)^l}{l}\Big]_{-\min(0,b_j)\sigma}^{-\min(0,a_j)\sigma},$$

$$(\mathcal{B}h)_{i,j} = \frac{\mathbb{1}_{i\geq j}}{\sigma}\Big[\ln(|1+x|) - \sum_{l=1}^{i-j}\frac{\phi(x)^l}{l}\Big]_{\max(0,a_j)\sigma}^{\max(0,b_j)\sigma}.$$

Similarly we obtain the formulas for $(\mathcal{A}hu)$ *and* $(\mathcal{B}hu)$ *under the form*

$$(\mathcal{A}hu)_{i,j} = \frac{\mathbb{1}_{j\geq i}}{\sigma}\Big[-(j-i+1)\ln(|1+x|) + x + \sum_{l=1}^{j-i} l\,\frac{\phi(x)^{j-i+1-l}}{j-i+1-l}\Big]_{-\min(0,b_j^n)\sigma}^{-\min(0,a_j^n)\sigma},$$

$$(\mathcal{B}hu)_{i,j} = \frac{\mathbb{1}_{i\geq j}}{\sigma}\Big[-(i-j+1)\ln(|1+x|) + x + \sum_{l=1}^{i-j} l\,\frac{\phi(x)^{i-j+1-l}}{i-j+1-l}\Big]_{\max(0,a_j^n)\sigma}^{\max(0,b_j^n)\sigma}.$$

To demonstrate this proposition, the two following lemmas are needed. Their proofs together with the proof of Proposition 3.3.10 are featured in Appendix 3.A.

Lemma 3.3.11. *Consider* $\phi : x \in \mathbb{R}\setminus\{-1\} \mapsto x/(1+x)$ *and let* $k \in \mathbb{N}$. *We have the following primitive for some constant* $C \in \mathbb{R}$

$$\int \frac{x^k}{(1+x)^{k+1}}\,\mathrm{d}x = \ln(|1+x|) - \sum_{l=1}^{k}\frac{\phi(x)^l}{l} + C\,.$$

Lemma 3.3.12. *Using the same notation as in the previous lemma, we have*

$$\int \frac{x^k}{(1+x)^k}\,\mathrm{d}x = -k\ln(|1+x|) + x + \sum_{l=1}^{k-1} l\,\frac{\phi(x)^{k-l}}{k-l} + C'\,.$$

Next we look at the exterior contributions given in (3.34) by $\widetilde{h}$ and $\widetilde{hu}$. The i-th component of vector $\widetilde{h}$ develops as

$$\Big(\int_{\mathbb{R}}(\mathbf{I}+\sigma\mathbf{L})^{-1}\sigma B^1\,\mathrm{d}\xi\Big)_i = \int_{\mathbb{R}_+}\frac{(\sigma\xi)^i}{(1+\sigma\xi)^i}M_0^1\,\mathrm{d}\xi + \int_{\mathbb{R}_-}\frac{(-\sigma\xi)^{N-i+1}}{(1-\sigma\xi)^{N-i+1}}M_{N+1}^1\,\xi\,,$$

whereas for $\widetilde{hu}$ we have

$$\Big(\int_{\mathbb{R}}\xi\,(\mathbf{I}+\sigma\mathbf{L})^{-1}\sigma B^1\,\mathrm{d}\xi\Big)_i = \frac{1}{\sigma}\int_{\mathbb{R}_+}\frac{(\sigma\xi)^{i+1}}{(1+\sigma\xi)^i}M_0^1\,\mathrm{d}\xi - \frac{1}{\sigma}\int_{\mathbb{R}_-}\frac{(-\sigma\xi)^{N-i+2}}{(1-\sigma\xi)^{N-i+1}}M_{N+1}^1\,\mathrm{d}\xi\,.$$

Hence we can reuse the previous analytic expression of Lemma 3.3.12 to compute the exterior contribution $\widetilde{h}$. However in order to compute $\widetilde{hu}$, we need to know how to explicitly write the integral of $(\mathbf{I}+\sigma\mathbf{L})^{-1}\sigma B^1$ against ξ which requires the following lemma.

Lemma 3.3.13. *Let $k \in \mathbb{N}^*$. Considering $\phi : x \mapsto x/(x+1)$ and $C \in \mathbb{R}$ we have the following expression*

$$\int \frac{x^{k+1}}{(1+x)^k}\,\mathrm{d}x = \Big(-\sum_{r=1}^{k-2}(k-r-1)\,\frac{k-r}{2}\,\frac{\phi(x)^r}{r}\Big)\mathbb{1}_{k\geq 3} + \Big(\frac{k(k-1)}{2}\ln|1+x|\Big)\mathbb{1}_{k\geq 2} \\ - (k+1)x + \frac{(1+x)^2}{2} - \Big(\sum_{q=1}^{k-1}(k-q)\frac{\phi(x)^q}{q}\Big)\mathbb{1}_{k\geq 2} + k\ln|1+x| + C\ .$$

The proof of Lemma 3.3.13 can be found in Appendix 3.A.

Boundary conditions. We now consider the issue of boundary conditions, which amounts to determine the vector B^1. The latter is entirely characterized by the ghost values M_0^1, M_{N+1}^1, which are defined in relation to the neighboring values in the border cells C_1 and C_N at time Δt. Thus we have an implicit problem whose solution is unknown. Depending on the type of boundary conditions one wishes to enforce, the relation between the ghost and border values can be nonlinear. We choose to simplify the problem by replacing the implicit ghost values at time Δt by the ghost values at the starting time which can be explicitly determined. This can be seen as a first order approximation in time since

$$U_0^1 = U_0^0 + \mathcal{O}(\Delta t)\ , \quad U_{N+1}^1 = U_{N+1}^0 + \mathcal{O}(\Delta t)\ .$$

Hence in the definition (3.34) of $\widetilde{h}$ and $\widetilde{hu}$ we will replace B^1 by

$$B^0 = \begin{pmatrix} \xi M(U_0^0,\xi)\mathbb{1}_{\xi>0} \\ \vdots \\ -\xi M(U_{N+1}^0,\xi)\mathbb{1}_{\xi<0} \end{pmatrix}\ .$$

Then need to fix the values of U_0^0 and U_{N+1}^0 for a given U_1^0 and U_N^0. We consider the case of fluvial flows as described by Bristeau and Coussin in [24]. This type of flows corresponds to the situation where the celerity of surface gravity waves $\sqrt{gh}$ is greater than the material velocity of particles u. This occurs quite frequently in real life, for instance in lakes, coastal flows as well as in rivers, and more generaly the low Froude regime clearly falls within this scope. Since the eigenvalues $u - \sqrt{gh}$ and $u + \sqrt{gh}$ have opposite sign, at each boundary we have exactly one wave entering the domain and one wave leaving it. Hence we only dispose of a single degree of freedom to set the ghost values, which generally consists in enforcing either a given water height or a discharge. The ghost state is then fully determined by asking the outward-going Riemann invariant to remain constant through the interface. We recall that for the Saint-Venant system the two Riemann invariants denoted $\mathcal{R}^{\pm}$ are given by

$$\mathcal{R}^{\pm}(U) = u \pm 2\sqrt{gh}\ .$$

Let us first treat the case where the water height is enforced at the boundary of the domain. We denote by $h_{g,l}$ the value attributed to the left ghost cell, and $h_{g,r}$ the one

attributed to the right ghost cell. Together with the condition on the outgoing Riemann invariant, we get the following systems

$$\begin{cases} h_0 = h_{g,l} \\ \mathcal{R}^-(U_0) = \mathcal{R}^-(U_1) \end{cases}, \qquad \begin{cases} h_{N+1} = h_{g,r} \\ \mathcal{R}^+(U_{N+1}) = \mathcal{R}^+(U_N) \end{cases}.$$

Using the equalities satisfied by the Riemann invariants we find

$$\begin{aligned} u_0 - 2\sqrt{gh_0} = u_1 - 2\sqrt{gh_1} &\implies q_0 = h_{g,l}(u_1 - 2(\sqrt{gh_1} - \sqrt{gh_{g,l}})) , \\ u_{N+1} + 2\sqrt{gh_{N+1}} = u_N + 2\sqrt{gh_N} &\implies q_{N+1} = h_{g,r}(u_N + 2(\sqrt{gh_N} - \sqrt{gh_{g,r}})) . \end{aligned}$$

Another possibility is to enforce the discharge at the boundary, and we denote by $q_{g,l}$ and $q_{g,r}$ the left and right ghost values. This time, the constraint on the Riemann invariant will lead to a nonlinear equation that has to be solved to determine the ghost water height. Indeed we have the systems

$$\begin{cases} q_0 = q_{g,l} \\ \mathcal{R}^-(U_0) = \mathcal{R}^-(U_1) \end{cases}, \qquad \begin{cases} q_{N+1} = q_{g,r} \\ \mathcal{R}^+(U_{N+1}) = \mathcal{R}^+(U_N) \end{cases},$$

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

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

In this case, this approach slightly differs from that of Bristeau and Coussin in [24], where the ghost value is chosen such that the resulting numerical flux at the interface coincides with the boundary discharge. Instead we do not enforce any value at the interface but directly in the ghost cell, which can be seen as a first order simplification in space. Also note that nothing prevents us from mixing the boundary conditions, for instance we can enforce a water height on the left boundary, and a discharge on the right. A common practice for channel flows is to enforce the water height at the inlet and the flux at the outlet.

Remark 3.3.14. *We briefly mention the case of* torrential flows*, which refer to a material velocity of particles $|u|$ greater than the celerity of surface gravitational waves $\sqrt{gh}$. In this situation the characteristics given by the eigenvalues $u - \sqrt{gh}$ and $u + \sqrt{gh}$ go in the same direction.*

- *Assume a torrential inflow at the boundaries, which implies that no information is leaving the domain. Then we have two degrees of freedom for setting the ghost cells and we don't need the data from the neighboring border cells in order to define the vector B^1.*
- *On the other hand if we assume that a torrential outflow holds at the boundaries, then no information enters the domain (all the waves are propagating outwards) and we expect the boundary fluxes to be entirely determined from the border cells*

C_1 and C_N. More specifically if throughout time the outflow remains in the regime given by $|u| \geq \sqrt{3}c$, then we can show that $B^1 = 0$, and therefore we also don't need to know precisely the data in the border cells. In fact, recalling the expression of B^1 which was given in (3.28) *together with the choice of maxwellian* (3.32)*, we obtain*

$$B^1 = 0 \iff \begin{cases} \forall \xi > 0 \ M_0^1(\xi) = 0 \\ \forall \xi < 0 \ M_{N+1}^1(\xi) = 0 \end{cases} \iff \begin{cases} u_0^1 + \sqrt{3}\, c_0^1 \leq 0 \\ u_{N+1}^1 - \sqrt{3}\, c_{N+1}^1 \geq 0 \end{cases} .$$

- *Finally the situation of a torrential outflow characterized by $c \leq |u| < \sqrt{3}c$ will require to know the data in the border cells. In this case we can as previously approximate B^1 by B^0 and apply the strategy described in [24] to explicitly determine U_0^0 and U_{N+1}^0.*

Numerical cost. It is important to try and keep a reasonable algorithmic complexity so that the implicit method presented in the previous lines remains usable in practice. We discuss here how to improve its computational cost by a substantial margin. In formulas (3.37) and (3.38), the sums can be seen as a matrix vector product. Thus to update each cell one needs to perform a scalar product for a total of $\mathcal{O}(N)$ operations. Since there are N cells, the complexity for applying this formula is quadratic, and we cannot hope to do better than this. Note however that we first have to assemble the matrices $(\mathcal{A}h)$, $(\mathcal{B}h)$, $(\mathcal{A}hu)$ and $(\mathcal{B}hu)$ whose coefficients involve a summation. Because of that, this step has a seemingly cubic complexity in the number of cells N. This is quite expensive and can render the method pretty much inefficient. However this complexity can be reduced to a quadratic cost by computing the coefficients in the correct order. This is better seen through the following recurrence relation allowing to define each coefficient from a previous one in $\mathcal{O}(1)$ operation.

$$(\mathcal{A}h)_{i,j} = \begin{cases} 0 & \text{if } j < i \\ \big[\ln(|1+x|)\big]_{-\min(0,b_j)\sigma}^{-\min(0,a_j)\sigma} & \text{if } i = j \\ (\mathcal{A}h)_{i+1,j} - \dfrac{1}{j-i}\big[\phi(x)^{j-i}\big]_{-\min(0,b_j)\sigma}^{-\min(0,a_j)\sigma} & \text{if } j > i \end{cases} , \tag{3.39}$$

$$(\mathcal{B}h)_{i,j} = \begin{cases} 0 & \text{if } i < j \\ \big[\ln(|1+x|)\big]_{\max(0,a_j)\sigma}^{\max(0,b_j)\sigma} & \text{if } i = j \\ (\mathcal{B}h)_{i-1,j} - \dfrac{1}{i-j}\big[\phi(x)^{i-j}\big]_{\max(0,a_j)\sigma}^{\max(0,b_j)\sigma} & \text{if } i > j \end{cases} . \tag{3.40}$$

Hence it is more efficient to assemble matrices $(\mathcal{A}h)$ and $(\mathcal{B}h)$ column wise, starting from the diagonal coefficient and moving towards the first or last row. A similar conclusion is achieved for $(\mathcal{A}hu)$ and $(\mathcal{B}hu)$, although the recurrence relation is less straightforward to obtain. We first remark that, introducing $l(i,j) = i - j + 1$ there holds

$$(\mathcal{A}hu)_{i,j} = \mathbb{1}_{j \geq i} \Big[-l(j,i)\ln|1+x| + x + \sum_{k=1}^{j-i} k\, \frac{\phi(x)^{l(j,i)-k}}{l(j,i)-k} \Big]_{-\min(0,b_j)\sigma}^{-\min(0,a_j)\sigma} ,$$

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

Performing the change of index $r = l(j,i) - k$ for matrix $(\mathcal{A}hu)$ and $s = l(i,j) - k$ for matrix $(\mathcal{B}hu)$ we find

$$(\mathcal{A}hu)_{i,j} = \Big[- l(j,i)\ln|1+x| + x + l(j,i)\sum_{r=1}^{j-i}\frac{\phi(x)^r}{r} - \sum_{r=1}^{j-i}\phi(x)^r\Big]_{-\min(0,b_j)\sigma}^{-\min(0,a_j)\sigma} ,$$

$$(\mathcal{B}hu)_{i,j} = \Big[- l(i,j)\ln|1+x| + x + l(i,j)\sum_{s=1}^{i-j}\frac{\phi(x)^s}{s} - \sum_{s=1}^{i-j}\phi(x)^s\Big]_{\max(0,a_j)\sigma}^{\max(0,b_j)\sigma} .$$

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

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

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

Then we can write that

$$(\mathcal{A}hu)_{i,j} = \begin{cases} 0 & j < i \\ \Big[x - \ln|1+x|\Big]_{-\min(0,b_j)\sigma}^{-\min(0,a_j)\sigma} & j = i \\ l(j,i)(\mathcal{VA})_{i,j} - (\mathcal{UA})_{i,j} + \Big[x - l(j,i)\ln|1+x|\Big]_{-\min(0,b_j)\sigma}^{-\min(0,a_j)\sigma} & j > i \end{cases} . \tag{3.41}$$

Similarly we introduce

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

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

so that we have

$$(\mathcal{B}hu)_{i,j} = \begin{cases} 0 & i < j \\ \Big[x - \ln|1+x|\Big]_{\max(0,a_j)\sigma}^{\max(0,b_j)\sigma} & i = j \\ l(i,j)(\mathcal{VB})_{i,j} - (\mathcal{UB})_{i,j} + \Big[x - l(i,j)\ln|1+x|\Big]_{\max(0,a_j)\sigma}^{\max(0,b_j)\sigma} & i > j \end{cases} . \tag{3.42}$$

To conclude, through relations (3.41) and (3.42) we are also able to assemble matrices $(\mathcal{A}hu)$ and $(\mathcal{B}hu)$ with a quadratic cost with respect to the number of cells, which means that the overall scheme has a $\mathcal{O}(N^2)$ complexity. A practical and fully vectorized implementation in Python is given in Appendix 3.C. We have to put this in perspective by comparing our scheme to explicit ones. The latter require a potentially more restrictive time step Δt_{exp} to be stable, but usually they only need $O(N)$ operations per iteration. Roughly speaking, the total number of iterations to reach the final time T with an explicit scheme is of order $N \times T/\Delta t_{\text{exp}}$. With our scheme, the time step Δt_{imp} might be taken larger than Δt_{exp}, but we have $O(N^2)$ operations to update the solution, hence the total number of operations until final time T is of order $N^2 \times T/\Delta t_{\text{imp}}$. Therefore, our scheme is less time consuming if we have

$$N\frac{T}{\Delta t_{\text{exp}}} \gg N^2\frac{T}{\Delta t_{\text{imp}}} \implies \Delta t_{\text{imp}} \gg N\Delta t_{\text{exp}} \ . \tag{3.43}$$

Note however that the computational time is not the only factor to account for, and one should also consider the error of the scheme. Generally, taking a very coarse resolution in time results in poorly accurate results, in which case it is not desirable to have (3.43). However there are some cases where the fast dynamics do not play an important role such as in the low Froude regime. Then it might be advantageous to consider large time steps. In a sense, when it comes to implicit methods the situation offered by our scheme is optimal. In fact, we explicitly know the inverse of the matrix and can hardly do better than just evaluating the update through a matrix-vector product. Yet we will see through the upcoming numerical results that its interest is rather limited when it comes to efficiency, at least for the considered testcases.

3.3.4 Numerical results

To assess the efficiency and interest of the implicit scheme, we perform a numerical test over a Riemann problem featuring a slowly moving shock. This configuration is achieved for nearly transcritical flow where the material velocity u is positive and satisfies $u - \sqrt{gh} \approx 0$ and $u + \sqrt{gh} \gg 1$. Hence the maximum eigenvalue severely constrains the time step, however a small time step might not be necessary to accurately resolve the slow shock. In Figure 3.3.1 we compare several schemes with an explicit time step Δt_{exp} given by the usual CFL condition, as well as the implicit scheme using a time step $\Delta t_{\text{imp}} = 10\Delta t_{\text{exp}}$. We notice that in the discharge profile, an oscillation appears downwind of the shock, which is quite pronounced for the explicit and iterative kinetic schemes, and less so for the fully implicit ones. As expected, the fast travelling rarefaction is strongly diffused by the implicit scheme with large time steps.

Despite this favorable context, the implicit scheme using larger time steps isn't very good. In fact, the 1-shock is more diffused that for the explicit kinetic scheme which is slightly faster. The conclusion might change by taking an even slower shock and by increasing the value of Δt_{imp}.

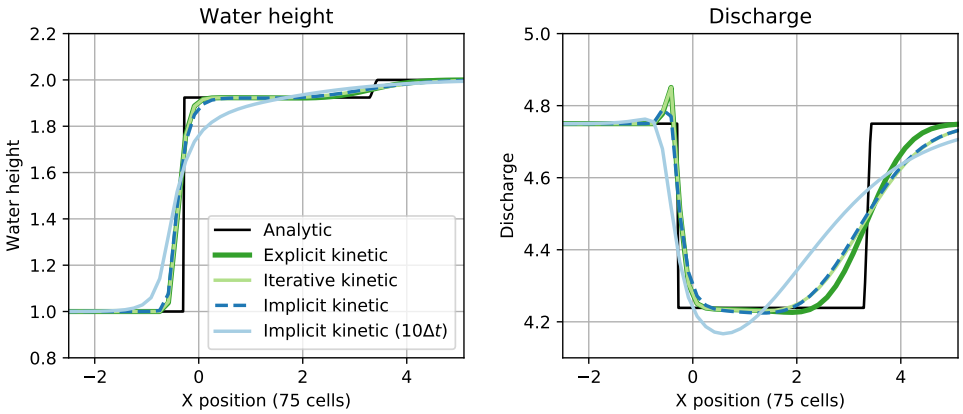

Figure 3.3.1: Slowly moving shock. The position of the discontinuity in the initial Riemann data is $x = 0$. The iterative scheme is presented in the next section.

3.4 Iterative resolution scheme

3.4.1 Case with flat bathymetry

We have proposed a modified version of the scheme (3.26) where the half-disk maxwellian has been replaced with the index maxwellian (3.32). This is a compromise between the ability to explicitly express the update at the macroscopic level and the existence of a discrete entropy inequality. We now consider a different paradigm where we don't want to be restricted by the choice of maxwellian, in exchange of what we approximate the implicit update by an iterative process. Later this will also enable us to deal with non flat bottoms. The strategy we use is a Gauss-Jacoby type decomposition which consists to introduce two matrices $\mathbf{P}$ and $\mathbf{Q}$ from $\mathbb{R}^{N\times N}$ such that $\mathbf{P}$ is invertible and such that $\mathbf{P}-\mathbf{Q}$ equals the mass matrix $\mathbf{I}+\sigma\mathbf{L}$. Then the implicit kinetic scheme (3.27) is equivalent to finding the solution $f(\xi)\in\mathbb{R}^N$ of

$$f = \mathbf{P}^{-1}\mathbf{Q}\, f + \mathbf{P}^{-1}(M^0 + \sigma B[f])\ . \tag{3.44}$$

We recall that the vector $B[f]$ stands for the boundary conditions. The left and right ghost values $U_{g,l}, U_{g,r}$ are enforced in relation with the border values $\{U_f\}_1$ and $\{U_f\}_N$, for instance using Riemann invariants as discussed previously. Hence $U_{g,l}$ and $U_{g,r}$ depend on f, and the vector $B[f]$ is defined in the usual way (3.28)

$$B[f] = \begin{pmatrix} \xi \mathbb{1}_{\xi>0} M(U_{g,l},\xi) \\ 0 \\ \vdots \\ 0 \\ -\xi \mathbb{1}_{\xi<0} M(U_{g,r},\xi) \end{pmatrix} \in \mathbb{R}^N\ .$$

In the sequel we will work with the decomposition $\mathbf{P} = (1+\alpha)\mathbf{I}$ and $\mathbf{Q} = \alpha\mathbf{I} - \sigma\mathbf{L}$, with $\alpha \geq 0$ a relaxation parameter. Instead of solving (3.44) a possibility is to study the

sequence $(f^k(\xi))_{k\in\mathbb{N}}$ from $\mathbb{R}^N$ defined by

$$f^0(\xi) = M^0 \ , \qquad \begin{cases} (1+\alpha) f^{k+1}(\xi) = (\alpha \mathbf{I} - \sigma \mathbf{L}) f^k + M^0 + \sigma B[f^k] \\ \forall 1 \leq i \leq N, \ \{U_f\}_i^k = \displaystyle\int_{\mathbb{R}} \begin{pmatrix} 1 \\ \xi \end{pmatrix} f_i^k(\xi) \, \mathrm{d}\xi \end{cases} . \tag{3.45}$$

Note that in previous section we were using the superscript "$1-$" to signify a non maxwellian distribution evaluated at time Δt. In what follows we drop this notation so as to only have the iteration index $k \in \mathbb{N}$ for simplicity. If the sequence $(f^k(\xi))_{k\in\mathbb{N}}$ converges in $\mathbb{R}^N$ to some $f^\infty(\xi)$, then this limit is necessarily the solution $f(\xi)$ of (3.44). In practice we stop the iteration process for some k large enough, hoping that f^k is close to the limit. The stopping criterion shall be discussed later, and we have the following result.

Proposition 3.4.1. *Assume that $B[f^k]$ remains constant equal to $B[M^0]$ for any k in $\mathbb{N}$. Then* (3.45) *defines an arithmetico-geometric sequence which converges if the CFL condition $\sigma|\xi| < 1 + 2\alpha$ holds for all ξ belonging to* $\operatorname{supp} M^0 \cup \operatorname{supp} B[M^0]$.

Proof. By recurrence, we can show that for any $k \in \mathbb{N}$ the support of f^k is included in $\operatorname{supp} M^0 \cup \operatorname{supp} B[M^0]$, which is why we restrict to velocities ξ belonging to this set. Denote $\mathbf{A} = \mathbf{P}^{-1}\mathbf{Q} = (1+\alpha)^{-1}(\alpha\mathbf{I} - \sigma\mathbf{L})$ and consider f the solution of

$$f = \mathbf{A} f + (1+\alpha)^{-1}(M^0 + \sigma B[M^0]) \ .$$

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

Remark 3.4.2. *As we did in the fully implicit scheme, we can replace $B[f^k]$ by $B[M^0]$ in the iterative process* (3.45)*. In fact this constitutes a first order approximation in time since we have $f^k = M^0 + \mathcal{O}(\Delta t)$. Under this simplification, the assumption $B[f^k] = B[M^0]$ from Proposition 3.4.1 becomes automatically satisfied.*

Proposition 3.4.1 points to a limitation of the iterative approach, which is that a CFL condition is now required to ensure the convergence. One might be tempted to take the relaxation parameter α large, but doing so can make the convergence slower by increasing the number of iterations needed to get close to the limit. In fact this is seen by the fact that the spectral radius of $\mathbf{A}$ gets closer to one when α is large. In particular, taking the limit $\alpha \to \infty$, the method (3.45) defines a constant sequence $f^{k+1} = f^k$ which is not usable anymore.

In practice, we do not wish to apply an iterative method at the kinetic level, since we then need to perform a numerical quadrature to approximate the macroscopic update. An issue with (3.45) is that the distribution involved in the kinetic flux (i.e. the term in factor of $\sigma\mathbf{L}$) is not a vector of maxwellians. Apparently this prevents us from retrieving a

numerical flux at the macroscopic level, in the sense that we don't know how to construct maps F_h, F_{hu} such that for any $f : \xi \in \mathbb{R} \to \mathbb{R}^N$ with compact support there holds

$$\left\{ \begin{array}{l} \displaystyle\int_{\mathbb{R}} (\mathbf{L}f(\xi))_i \, \mathrm{d}\xi = F_h(\{U_f\}_i, \{U_f\}_{i+1}) - F_h(\{U_f\}_{i-1}, \{U_f\}_i) \\ \displaystyle\int_{\mathbb{R}} \xi \, (\mathbf{L}f(\xi))_i \, \mathrm{d}\xi = F_{hu}(\{U_f\}_i, \{U_f\}_{i+1}) - F_{hu}(\{U_f\}_{i-1}, \{U_f\}_i) \end{array} \right. , \quad 2 \leq i \leq N-1 \, .$$

To bypass this issue, we propose the following modification of (3.45), where we replace all occurrences of f^k on the right hand side by a vector of maxwellians M^k, which defines a new sequence $(g^k(\xi))_{k\in\mathbb{N}}$ as

$$g^0(\xi) = M^0 \, , \qquad \left\{ \begin{array}{l} (1+\alpha) g^{k+1}(\xi) = (\alpha \mathbf{I} - \sigma \mathbf{L}) M^k + M^0 + \sigma B[M^k] \\ M^{k+1} = g^{k+1} + \Delta t Q^k \end{array} \right. . \tag{3.46}$$

This new iterative process is alternating two stages, the first one being the usual transport step, while the second one is a projection step onto the set of maxwellians. The term Q^k is a vector of collision operators each one satisfying the conservation constraints (3.5). In a sense (3.46) is an iterative BGK splitting approach. Note that the projection step doesn't modify the macroscopic quantities of interest. It is important to remark that this iterative scheme differs from (3.45) and we cannot apply the result of Proposition 3.4.1. Especially, the projection step is nonlinear in that it amounts to define $M_i^{k+1} := M(\{U_g\}_i^{k+1}, \xi)$ for $1 \leq i \leq N$. It will not be possible to prove the convergence using the argument of geometric sequence, and instead we have to show a more general contraction property. We try to get such a result later when also incorporating a source term, and skip it for now.

If the sequence $(g^k(\xi))_{k\in\mathbb{N}}$ defined by the iterative process (3.46) converges in $\mathbb{R}^N$ to a limit $\overline{g}(\xi)$, then this limit satisfies

$$(1+\alpha)\overline{g}(\xi) = (\alpha \mathbf{I} - \sigma \mathbf{L}) M(U_{\overline{g}}, \xi) + M^0(\xi) + \sigma B[\overline{g}](\xi) \, . \tag{3.47}$$

Integrating (3.47) respectively against 1 and ξ, the terms in factor of α cancel and we get at the macroscopic level

$$\forall 1 \leq i \leq N, \quad \{U_{\overline{g}}\}_i = U_i^0 - \sigma \Big(F(\{U_{\overline{g}}\}_i, \{U_{\overline{g}}\}_{i+1}) - F(\{U_{\overline{g}}\}_{i-1}, \{U_{\overline{g}}\}_i) \Big) \, ,$$

with the numerical flux $F(U_L, U_R)$ already defined in (3.21). In practice we iterate at the macroscopic level over

$$\forall 1 \leq i \leq N, \quad (1+\alpha) U_i^{k+1} = U_i^0 + \alpha U_i^k - \sigma \Big(F(U_i^k, U_{i+1}^k) - F(U_{i-1}^k, U_i^k) \Big) \, , \tag{3.48}$$

which corresponds to (3.46) integrated against $(1, \xi)^T$. We can show the following properties from the kinetic description offered by (3.46).

Proposition 3.4.3. *The distribution f^{k+1} defined by the iterative scheme* (3.46) *is positive if the CFL condition $\sigma|\xi| \leq \alpha + M_i^0/M_i^k$ holds for any ξ belonging to* $\operatorname{supp} M^k$ *and for any $1 \leq i \leq N$. In particular f^{k+1} is positive when $\sigma|\xi| \leq \alpha$.*

Remark 3.4.4. *Regarding the CFL condition from Proposition 3.4.3, we have no* a priori *knowledge on the vector of maxwellians M^k for $k > 0$. We don't think this is a concern, as nothing prevents us from modifying the value of the time step Δt over the course of the iteration process if needed. This means that we can replace Δt by Δt^k, and if the sequence $(M^k)_{k\in\mathbb{N}}$ converges to some limit then Δt^k will converge in $\mathbb{R}_+$. In practice when we stop the iterative process at some index $k_{\max}$ we advance the time of $\Delta t^{k_{\max}}$.*

Proof (Proposition 3.4.3). Let us rewrite the iterative scheme (3.46) over a given cell $1 \leq i \leq N$

$$(1+\alpha) f_i^{k+1}(\xi) = (\alpha - \sigma|\xi|) M_i^k + \sigma\xi \mathbb{1}_{\xi>0} M_{i-1}^k - \sigma\xi \mathbb{1}_{\xi<0} M_{i+1}^k + M_i^0 + \sigma B[M^k]_i \ . \tag{3.49}$$

The case where ξ is not in the set $\operatorname{supp} M^k$ is obvious since we have $f_i^{k+1}(\xi) = M_i^0 + \sigma B[M^k]_i$ which is positive. If ξ belongs to to this set and under the CFL condition $\sigma|\xi| \leq \alpha$, all the terms on the right hand side of (3.49) are positive, hence $f_i^{k+1}(\xi)$ is positive too. If α is very small this CFL becomes quite restrictive (in fact it is unusable when $\alpha = 0$), and we can improve it with the help of the term M_i^0. Notice that since

$$\sigma\xi \mathbb{1}_{\xi>0} M_{i-1}^k - \sigma\xi \mathbb{1}_{\xi<0} M_{i+1}^k + \sigma B[M^k]_i$$

is always positive, it is sufficient to have the positivity of the remaining terms

$$M_i^0 + (\alpha - \sigma|\xi|) M_i^k = \Big(\frac{M_i^0}{M_i^k} + \alpha - \sigma|\xi| \Big) M_i^k \ ,$$

which is the case as soon as $\sigma|\xi|$ is smaller than $\alpha + M_i^0/M_i^k$, thus the result. □

Proposition 3.4.5. *The kinetic entropy of the iterative process* (3.46) *satisfies the following equality*

$$\begin{aligned} &H(M_i^{k+1}, \xi) = \\ &\quad H(M_i^0, \xi) - \sigma\xi \Big(H_{i+1/2}^k - H_{i-1/2}^k \Big) + (1+\alpha)\Delta t\, \partial_1 H(M_i^k, \xi) Q_i \\ &\quad + \alpha \Big(H(M_i^k, \xi) - H(M_i^{k+1}, \xi) \Big) + (1+\alpha) \frac{g^2\pi^2}{6} (2M_i^k + M_i^{k+1})(M_i^{k+1} - M_i^k)^2 + D_i^k \ , \end{aligned} \tag{3.50}$$

with $Q_i = M_i^{k+1} - f_i^{k+1}$ a collision operator satisfying the conservation constraints (3.5)*. The interfacial kinetic entropies $H_{i\pm1/2}^k$ are defined as*

$$\begin{aligned} H_{i-1/2}^k &= \mathbb{1}_{\xi>0} H(M_{i-1}^k, \xi) + \mathbb{1}_{\xi<0} H(M_i^k, \xi) \ , \\ H_{i+1/2}^k &= \mathbb{1}_{\xi>0} H(M_i^k, \xi) + \mathbb{1}_{\xi<0} H(M_{i+1}^k, \xi) \ , \end{aligned}$$

and the dissipation $D_i^k \leq 0$ has the form

$$\begin{aligned} D_i^k = \sigma\xi \frac{g^2\pi^2}{6} \Big(&\mathbb{1}_{\xi<0} (2M_i^k + M_{i+1}^k)(M_{i+1}^k - M_i^k)^2 - \mathbb{1}_{\xi>0} (2M_i^k + M_{i-1}^k)(M_i^k - M_{i-1}^k)^2 \Big) \\ &- \frac{g^2\pi^2}{6} (2M_i^k + M_i^0)(M_i^0 - M_i^k)^2 \ . \end{aligned}$$

Before doing the proof, we want to raise the following remark.

Remark 3.4.6. *In Proposition 3.4.5, the right hand side of the kinetic entropy equality* (3.50) *contains three non conservative terms that seem to be problematic. We have the terms*

$$\alpha\Big(H(M_i^k,\xi)-H(M_i^{k+1},\xi)\Big)+(1+\alpha)\frac{g^2\pi^2}{6}(2M_i^k+M_i^{k+1})(M_i^{k+1}-M_i^k)^2$$

which can have a positive sign and do not vanish when integrating. Thus they can prevent the dissipation of entropy at the macroscopic level, but assuming the method converges, the differences $M_i^{k+1}-M_i^k$ and $H(M_i^k,\xi)-H(M_i^{k+1},\xi)$ vanish as $k\to\infty$. As a consequence, we hope that for k large enough, these two terms will become negligible before the dissipation D_i^k. On the other hand, when M_i^k is a half-disk maxwellian (3.10)*, the term*

$$(1+\alpha)\Delta t\,\partial_1 H(M_i^k,\xi)Q_i$$

does not cause any issue as it vanishes upon integration over $\xi\in\mathbb{R}$. This is intrinsically related to the form of the half-disk maxwellian M_i^k which makes $\partial_1 H(M_i^k,\xi)$ linear in ξ over the support of $M_i^k(\cdot)$. In fact we have

$$\partial_1 H(M_i^k,\xi)=\frac{\xi^2}{2}+\frac{g^2\pi^2}{2}\left(\frac{1}{g\pi}\sqrt{2gh_i^k-(\xi-u_i^k)^2}\right)^2=gh_i^k+u_i^k\xi-\frac{(u_i^k)^2}{2}\ .$$

Furthermore we remind that Q_i satisfies the conservation constraints (3.5)*, meaning that its integral against $(1,\xi)^T$ vanishes. This is why $\partial_1 H(M_i^k,\xi)Q_i$ is macroscopically zero.*

Proof. We start by rewriting the transport step from (3.46) as

$$\begin{aligned}(1+\alpha)(M_i^{k+1}-M_i^k)=&\\ (M_i^0-M_i^k)-\sigma\xi\Big(\mathbb{1}_{\xi>0}(M_i^k-M_{i-1}^k)+\mathbb{1}_{\xi<0}(M_{i+1}^k-M_i^k)\Big)+(1+\alpha)\Delta t Q_i\ .&\end{aligned}\tag{3.51}$$

To obtain this expression, we subtracted $(1+\alpha)M_i^k$ from both sides and we replaced f_i^{k+1} with $M_i^{k+1}-\Delta t Q_i$. The term Q_i is the collision operator involved in the collision step from (3.46). In the case $i=1$ (resp. $i=N$), we define M_{i-1}^k (resp. M_{i+1}^k) using the corresponding macroscopic ghost state $U_{g,l}^k$ or $U_{g,r}^k$. Similarly to the proof of Proposition 3.3.7, we multiply (3.51) by $\partial_1 H(M_i^k,\xi)$ and apply Lemma 3.3.8 for $a=M_i^k$ and $b\in\{M_{i-1}^k,M_i^k,M_{i+1}^k\}$ which leads to

$$\begin{aligned}(1+\alpha)&\Big(H(M_i^{k+1},\xi)-H(M_i^k,\xi)-\frac{g^2\pi^2}{6}(2M_i^k+M_i^{k+1})(M_i^{k+1}-M_i^k)^2\Big)=\\ &\Big(H(M_i^0,\xi)-H(M_i^k,\xi)-\frac{g^2\pi^2}{6}(2M_i^k+M_i^0)(M_i^0-M_i^k)^2\Big)\\ &+\sigma\xi\mathbb{1}_{\xi>0}\Big(H(M_{i-1}^k,\xi)-H(M_i^k,\xi)-\frac{g^2\pi^2}{6}(2M_i^k+M_{i-1}^k)(M_i^k-M_{i-1}^k)^2\Big)\\ &-\sigma\xi\mathbb{1}_{\xi<0}\Big(H(M_{i+1}^k,\xi)-H(M_i^k,\xi)-\frac{g^2\pi^2}{6}(2M_i^k+M_{i+1}^k)(M_{i+1}^k-M_i^k)^2\Big)\\ &+(1+\alpha)\Delta t\,\partial_1 H(M_i^k,\xi)Q_i\ .\end{aligned}$$

Rearranging the terms and denoting $H^k_{i+1/2} = \mathbb{1}_{\xi>0} H(M^k_i, \xi) + \mathbb{1}_{\xi<0} H(M^k_{i+1}, \xi)$ we get

$$
\begin{aligned}
&(1+\alpha) H(M_i^{k+1}, \xi) = \\
&\quad H(M_i^0, \xi) + \alpha H(M_i^k, \xi) - \sigma\xi\Big(H^k_{i+1/2} - H^k_{i-1/2}\Big) \\
&\quad + \sigma\xi \frac{g^2\pi^2}{6}\Big(\mathbb{1}_{\xi<0}(2M_i^k + M_{i+1}^k)(M_{i+1}^k - M_i^k)^2 - \mathbb{1}_{\xi>0}(2M_i^k + M_{i-1}^k)(M_i^k - M_{i-1}^k)^2\Big) \\
&\quad - \frac{g^2\pi^2}{6}(2M_i^k + M_i^0)(M_i^0 - M_i^k)^2 + (1+\alpha)\frac{g^2\pi^2}{6}(2M_i^k + M_i^{k+1})(M_i^{k+1} - M_i^k)^2 \\
&\quad + (1+\alpha)\Delta t\, \partial_1 H(M_i^k, \xi) Q_i \ .
\end{aligned}
$$

We obtain the relation we are looking for by regrouping the terms in factor of α. □

3.4.2 Stopping criterion

We want to discuss the question of the stopping criterion of (3.45) and (3.46). The usual approach is to end the iterative process whenever two successive iterates are sufficiently close to each other, as it means that the approximation might be close to the solution, and that further iterations won't make any noticeable change. Concretely we introduce a small tolerance value $\varepsilon_{\text{tol}} > 0$, and we say the algorithm converges if $\|U^{k+1} - U^k\|$ becomes smaller than this threshold for some vector norm $\|\cdot\|$. However this is not the only condition to look at, and we can combine it with another one to get a stricter criterion. In our case, we know that the Saint-Venant system dissipates the total energy

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}} E(U)(t,x)\,\mathrm{d}x \leq 0 \tag{3.52}$$

under appropriate boundary conditions. In fact this is obtained by integrating the entropy inequality (1.3) over the spatial domain and assuming the energy fluxes $G(U)$ are equal at the left and right borders. Thanks to equality (3.50) and as per Remark 3.4.6, we expect the iterative method (3.48) to verify a conservative discrete entropy inequality on the energy from a certain iteration rank $K \in \mathbb{N}$

$$k \geq K \implies E(U_i^{k+1}) \leq E(U_i^0) - \sigma\bigg(\int_{\mathbb{R}} \xi\, H^k_{i+1/2}(\xi)\,\mathrm{d}\xi - \int_{\mathbb{R}} \xi\, H^k_{i-1/2}(\xi)\,\mathrm{d}\xi\bigg)\ .$$

Hence, assuming again appropriate boundary conditions, a discrete counterpart to the total energy dissipation (3.52) is obtained by summing the discrete entropy inequality over the cells $1 \leq i \leq N$ leading to

$$\Delta x \sum_{i=1}^{N} E(U_i^{k+1}) \leq \Delta x \sum_{i=1}^{N} E(U_i^0) - \Delta t\bigg(\int_{\mathbb{R}} \xi\, H^k_{N+1/2}(\xi)\,\mathrm{d}\xi - \int_{\mathbb{R}} \xi\, H^k_{1/2}(\xi)\,\mathrm{d}\xi\bigg)\ . \tag{3.53}$$

We will call *entropy criterion* the inequality (3.53).

When presenting the numerical results in next section, we will see that a tolerance criterion can be expensive to reach compared to explicit schemes. A compromise would be to only have the entropy criterion, meaning that the iterative process stops as soon

as (3.53) is satisfied, even if the method did not converge (i.e. the approximation can be far from the true implicit update). The risk is then to have a lack of consistency with respect to the Saint-Venant system. In fact, for the sake of the example assume that the criterion (3.53) is satisfied after just one iteration. In this case, the iteration process approximates the implicit update at time Δt by

$$U_i^1 = U_i^0 - \frac{\sigma}{1+\alpha}\Big(F(U_i^0, U_{i+1}^0) - F(U_{i-1}^0, U_i^0)\Big) ,$$

which is nothing else than (3.48) with $k = 0$. We recognize the forward Euler scheme with time step $\Delta t/(1+\alpha)$, whereas we want to approximate the solution at time Δt. Hence in this setting $\alpha = 0$ is the only possibility to have the correct time stepping leading to a consistent update. Keeping this in mind, we will always set the value of α to zero when using an entropy-only stopping criterion.

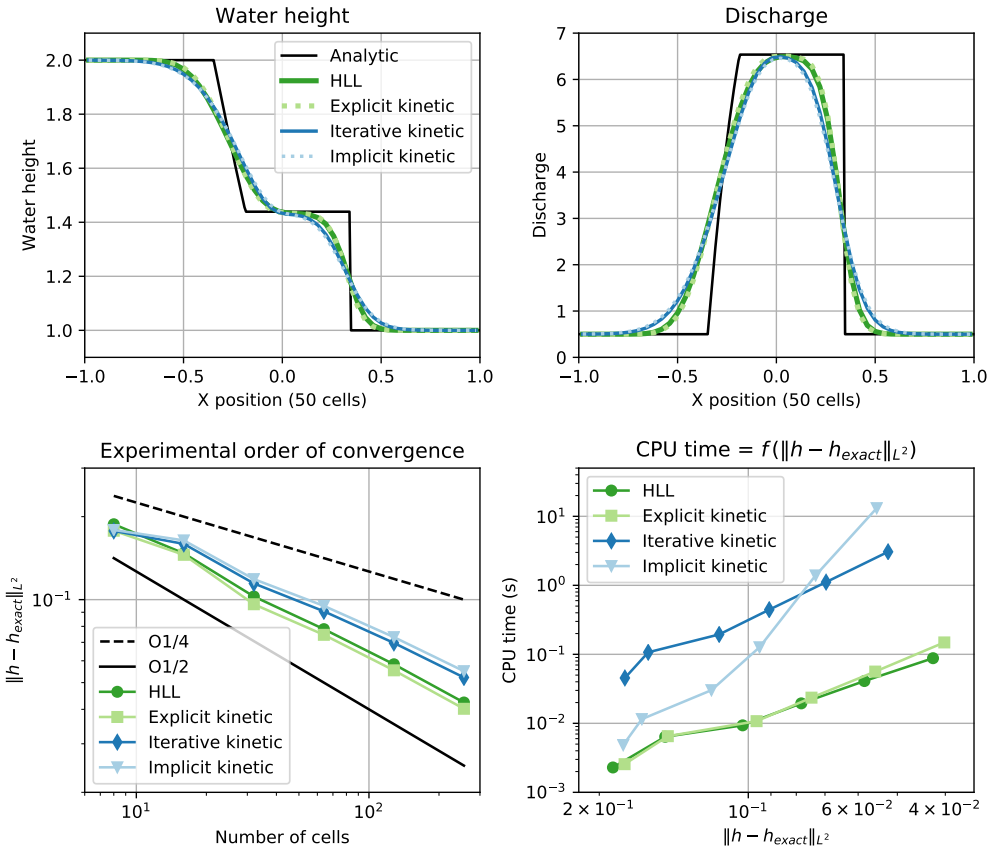

Figure 3.4.1: Comparing the iterative and fully implicit kinetic schemes with explicit strategies.

3.4.3 Numerical results over a flat bathymetry

We compare the fully implicit kinetic scheme and iterative kinetic scheme to explicit methods. The testcase is given by the Riemann problem with initial data $U^0(x) =$

$\mathbb{1}_{x<0}U_L + \mathbb{1}_{x>0}U_R$ where we define

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

The solutions consists in a 1-rarefaction and a 2-shock. The iterative kinetic scheme uses the half-disk maxwellian, and we choose the parameters $\alpha = 1$ and $\varepsilon_{\text{tol}} = 10^{-9}$. All the schemes use an explicit time step, and the results are given in Figure 3.4.1. Three aspects have to be considered, namely the accuracy, the computational cost and the stability. In the plotted curves, we see that in terms of efficiency both iterative and implicit kinetic schemes are at their disadvantage. Especially, the quadratic complexity of the fully implicit version results in a steeper slope of the efficiency curve which is a bad thing. However this is only one part of the picture, and we know for a fact that the iterative method has better stability properties, since it admits a discrete entropy inequality. This is a great advantage over the explicit methods which don't satisfy such an inequality and can increase the entropy. Concretely the greater stability comes with a higher level of diffusion which is noticeable in the discharge profile shown in Figure 3.4.1. This diffusion remains within acceptable margin, and is the price to pay to have better stability properties.

3.4.4 Hydrostatic reconstruction

The hydrostatic reconstruction is a technique introduced by Audusse et al. in [6]. It consists to take a numerical flux consistent with the Saint-Venant system, and to modify it in order to handle varying bathymetries in a well balanced way. The terminology well balanced refers to the ability of the method to preserve the lake at rest steady states characterized by $h + z \equiv \text{Cst}$ and $u = 0$, which is one of the equilibria of the model. The hydrostatic reconstruction also has the property to retain the positivity of the scheme it is applied to, meaning that if the original scheme yields positive water heights (potentially under some CFL condition), then the modified version of the scheme with hydrostatic reconstruction also gives positive water heights.

To motivate the use of a reconstruction step, let us point to the main obstacles that hinder the preservation of lakes at rest. The first one comes from the numerical diffusion introduced by the upwinding of the numerical flux. Upwinding is a common feature in finite volumes and doesn't restrict to kinetic solvers, thus we can illustrate our point with the simpler Rusanov flux, and the argument will remain the same for the kinetic flux. Let $(U_i)_{1\leq i\leq N}$ define a discrete lake at rest, that is to say $h_i + z_i = K$ and $u_i = 0$ for all $1 \leq i \leq N$ and for some $K \in \mathbb{R}$. Estimating an upper bound of the waves velocities using the global speed of upwinding $a = \max_i (|u_i| + \sqrt{gh_i})$, the first component of the flux difference used in the water height update is given by

$$\begin{aligned} F_h(U_i, U_{i+1}) - F_h(U_{i-1}, U_i) &= \frac{(hu)_{i+1} - (hu)_{i-1}}{2} - \frac{a}{2}((h_{i+1} - h_i) - (h_i - h_{i-1})) \\ &= \frac{a}{2}(z_{i+1} - 2z_i + z_{i-1}) . \end{aligned}$$

Since we don't enforce any condition on the discretized bathymetry $(z_i)_i$, this term has no reason to vanish. In fact this quantity is consistent at the continuous level with

$(a\Delta x^2/2)\partial^2_{xx} z$ which only cancels when the bathymetry profile is linear. As a consequence the value of the updated water height h_i^1 isn't the same as the starting value h_i^0, whereas a well balanced scheme would keep it constant in time. The second obstacle comes from the discharge equation, for which a lake at rest implies that the pressure variation balances exactly with the source term

$$\partial_x\Big(\frac{gh^2}{2}\Big) = -gh\partial_x z \ . \tag{3.54}$$

Put in other words, the source term admits a conservative writing in the hydrostatic equilibrium. This is not straightforward to get at the discrete level, and one of the ingredients will be to introduce an interfacial value of the bathymetry defined by

$$z_{i+1/2} = \max(z_i, z_{i+1}) \ . \tag{3.55}$$

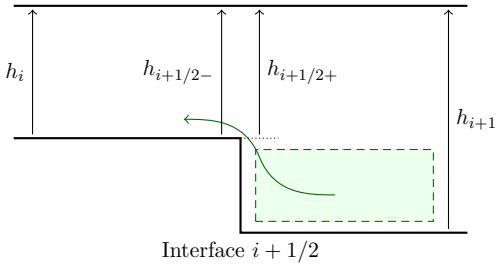

Figure 3.4.2: In standard schemes diffusion occurs even at hydrostatic equilibrium. In this example the diffusion can be interpreted as the flow from the dashed area to the left cell. The solution proposed by the hydrostatic reconstruction is to elevate the bathymetry perceived at the interface so that the reconstructed water heights $h_{i+1/2\pm}$ are the same when the free surface is flat ($h_i + z_i = K$ for all i).

The idea of the hydrostatic reconstruction is to avoid the issue of diffusion by modifying the water height in the left and right neighborhoods of each interface. For that matter we use the interfacial bathymetry (3.55) and we reconstruct the water height as below

$$h_{i-1/2+} = (h_i + z_i - z_{i-1/2})_+ \ , \quad h_{i+1/2-} = (h_i + z_i - z_{i+1/2})_+ \ , \tag{3.56}$$

with the notation $(\cdot)_+ = \max(0, \cdot)$. Figure 3.4.2 is helpful to visualize and understand how this reconstruction works. Conceptually, we say that the water comprised vertically between $\min(z_i, z_{i+1})$ and $\max(z_i, z_{i+1})$ is not able to jump and cross the interface $i+1/2$, which has the benefit of preventing the unnecessary diffusion. In the example given in Figure 3.4.2, we enforce this by raising the bathymetry in the right cell C_{i+1} and by reducing the water height h_{i+1} by the same amount, which gets us $h_{i+1/2+}$. Another way to look at this is that the numerical diffusion on the first component will now come from the the free surface elevation $h + z$ rather than the water height. We are then able to define the reconstructed vectors of conserved variables

$$\forall 1 \leq i \leq N, \quad U_{i+1/2-} = \begin{pmatrix} h_{i+1/2-} \\ h_{i+1/2-}u_i \end{pmatrix} \ , \quad U_{i-1/2+} = \begin{pmatrix} h_{i-1/2+} \\ h_{i-1/2+}u_i \end{pmatrix} \ , \tag{3.57}$$

with the value of the velocity u_i left unchanged. The numerical flux at interface $i+1/2$ is now evaluated using these modified values and incorporates the source term contribution as given below

$$\begin{cases} \mathcal{F}_{i+1/2-} = F(U_{i+1/2-}, U_{i+1/2+}) + \dfrac{g}{2}\begin{pmatrix} 0 \\ (h_i)^2 - (h_{i+1/2-})^2 \end{pmatrix} \\ \mathcal{F}_{i-1/2+} = F(U_{i-1/2-}, U_{i-1/2+}) + \dfrac{g}{2}\begin{pmatrix} 0 \\ (h_i)^2 - (h_{i-1/2+})^2 \end{pmatrix} \end{cases} , \tag{3.58}$$

where $F(U_L, U_R)$ is any consistent numerical flux. We recall the expression of the kinetic flux that we are interested in and will use throughout the whole section

$$F(U_L, U_R) = \int_{\mathbb{R}_-} \begin{pmatrix} 1 \\ \xi \end{pmatrix} \xi\, M(U_R, \xi)\, \mathrm{d}\xi + \int_{\mathbb{R}_+} \begin{pmatrix} 1 \\ \xi \end{pmatrix} \xi\, M(U_L, \xi)\, \mathrm{d}\xi \ .$$

Remark 3.4.7. *We have the following comments regarding the hydrostatic reconstruction.*

1. *When the bathymetry varies slowly, we recover a consistent discretization of the source term $-gh\partial_x z$ at interface $i+1/2$ by computing*

$$\frac{g}{2\Delta x}(h_{i+1}^2 - h_{i+1/2+}^2) - \frac{g}{2\Delta x}(h_i^2 - h_{i+1/2-}^2) \ ,$$

which comes from the difference $\big((\mathcal{F}_{hu})_{i+1/2+} - (\mathcal{F}_{hu})_{i+1/2-}\big)/\Delta x$. In fact, if we take for instance the configuration shown in Figure 3.4.2 we have

$$h_{i+1/2-} = h_i \implies \frac{g}{2\Delta x}(h_i^2 - h_{i+1/2-}^2) = 0 \ ,$$

and using that $h_{i+1/2+} = h_{i+1} + z_{i+1} - z_i$ in the case where h_{i+1} is greater than $z_{i+1} - z_i$ we can also write

$$\begin{aligned} \frac{g}{2\Delta x}(h_{i+1}^2 - h_{i+1/2+}^2) = & \quad \frac{g}{2\Delta x}(h_{i+1} + h_{i+1/2+})(z_i - z_{i+1}) \\ = & -\frac{g}{2\Delta x}(2h_{i+1} + z_{i+1} - z_i)(z_{i+1} - z_i) \ . \end{aligned}$$

This last expression is consistent with the desired term provided $|z_{i+1} - z_i| \ll h_{i+1}$. It also suggests that the hydrostatic reconstruction tends to be less consistent near wet/dry transitions. A second order extension of the hydrostatic reconstruction exists and is known to give good results, see [6].

2. *Notice that the reconstruction* (3.56) *of the water height is non conservative. This is not a problem as the reconstructed values will only be used to compute the modified numerical fluxes* (3.58)*.*

3. *Equality $(h+z)_i = (h+z)_{i+1}$ implies $h_{i+1/2-} = h_{i+1/2+}$ and we can define $h_{i+1/2} = h_{i+1/2\pm}$. Therefore over lakes at rest the difference*

$$\frac{g}{2}(h_i^2 - h_{i+1/2-}^2) - \frac{g}{2}(h_i^2 - h_{i-1/2+}^2)$$

comming from the source discretization is equal to the discrete pressure variation

$$-\frac{g}{2}(h_{i+1/2}^2 - h_{i-1/2}^2) \ ,$$

and we recover the balance (3.54) *at the continuous level.*

4. *In the hydrostatic reconstruction, the water is prevented from jumping the step characterizing the bathymetry variation. This could be relaxed by allowing the water to overcome a step $\Delta z \geq 0$ if the associated kinetic energy $hu^2/2$ is greater than the potential energy $gh\Delta z$ required for the jump to happen. Such an argument was used by Perthame and Simeoni in [58] at the kinetic level by incorporating reflections against the staircase shaped bottom. We do not take into account this phenomena.*

We end this section by recalling how to interpret the hydrostatic reconstruction at the kinetic level. Introducing the reconstructed maxwellians $M_{i+1/2\pm} = M(U_{i+1/2\pm}, \xi)$, we recover the source term discretization by the mean of the following integrals

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

Hence the explicit kinetic scheme (3.20) with hydrostatic reconstruction can be interpreted at the kinetic level as

$$\frac{f_i^{1-} - f_i^0}{\Delta t} + \frac{\xi}{\Delta x}\Big(M_{i+1/2}^0 - M_{i-1/2}^0\Big) + \frac{1}{\Delta x}(\xi - u_i^0)(M_{i-1/2+}^0 - M_{i+1/2-}^0) = 0 \ . \quad (3.59)$$

where we defined the upwinded reconstructed maxwellian

$$M_{i+1/2}^0 = \mathbb{1}_{\xi<0} M_{i+1/2+}^0 + \mathbb{1}_{\xi>0} M_{i+1/2-}^0 \ .$$

This scheme was extensively studied in [10], and in the following section we propose to look at its implicit version.

3.4.5 Iterative scheme with hydrostatic reconstruction

We are interested by the implicit version of the kinetic scheme (3.59) with hydrostatic reconstruction. It is given below at the kinetic level

$$\frac{f_i^{1-} - f_i^0}{\Delta t} + \frac{\xi}{\Delta x}\Big(M_{i+1/2}^1 - M_{i-1/2}^1\Big) + \frac{1}{\Delta x}(\xi - u_i^1)(M_{i-1/2+}^1 - M_{i+1/2-}^1) = 0 \ , \quad (3.60)$$

Having maxwellians in the implicit kinetic fluxes is important, as it allows to have a clear definition of the hydrostatic reconstruction at the kinetic scale. We have the following rewriting of (3.60) at the macroscopic level

$$U_i^1 = U_i^0 - \sigma(\mathcal{F}_{i+1/2-}^1 - \mathcal{F}_{i-1/2+}^1) \ , \quad (3.61)$$

with the implicit modified fluxes $\mathcal{F}^1_{i-1/2+}$ and $\mathcal{F}^1_{i+1/2-}$ defined following the formula (3.58) at time Δt. Neither (3.60) nor (3.61) can be solved analytically because of the underlying nonlinearity. Instead, we propose an iterative approximation based on the Gauss-Jacobi decomposition $\mathbf{P} = (1+\alpha)\mathbf{I}$ and $\mathbf{Q} = \alpha\mathbf{I} - \sigma\mathbf{L}$ already used in the case without bathymetry. Dropping the discrete time exponent to make room for the iteration index k, the iterative process reads

$$\begin{cases} (1+\alpha) f_i^{k+1} = M_i^0 + \alpha M_i^k - \sigma\xi(M^k_{i+1/2} - M^k_{i-1/2}) + \sigma(\xi - u_i^k)(M^k_{i+1/2-} - M^k_{i-1/2+}) \\ M_i^{k+1} = M\big(\{U_f\}_i^{k+1}, \xi\big) \\ M^{k+1}_{i\pm1/2} = \mathbb{1}_{\xi<0} M\big(\{U_f\}^{k+1}_{i\pm1/2+}, \xi\big) + \mathbb{1}_{\xi>0} M\big(\{U_f\}^{k+1}_{i\pm1/2-}, \xi\big) \end{cases} \tag{3.62}$$

for $1 \leq i \leq N$. We recall that for $i = 1$ (resp. $i = N$), the term $M_{1/2-}$ (resp. $M_{N+1/2+}$) makes use of a ghost value also subject to the hydrostatic reconstruction, and which was previously denoted by the mean of vector $\sigma B[M^k]$. In practice, we will implement the macroscopic iterative process obtained by integrating (3.62) against $(1, \xi)^T$ and given by

$$(1+\alpha) U_i^{k+1} = U_i^0 + \alpha U_i^k - \sigma(\mathcal{F}^k_{i+1/2-} - \mathcal{F}^k_{i-1/2+}) \ . \tag{3.63}$$

As in the case without hydrostatic reconstruction, the positivity is obtained under a CFL condition, with the possibility to adapt the time step as we iterate. This time however, we will not prove the positivity of the vector f^{k+1} but directly that of the water height h^{k+1} at the macroscopic level.

Proposition 3.4.8. *The water height h_i^{k+1} obtained from the iterative process* (3.63) *is positive if the CFL condition $\sigma|\xi| \leq \alpha + M_i^0/M_i^k$ holds for any ξ contained in* $\operatorname{supp} M^k$.

Proof. We start by remarking that $\sigma(\xi - u_i^k)(M^k_{i+1/2-} - M^k_{i-1/2+})$ is an odd function of ξ around u_i^k, hence its integral over $\xi \in \mathbb{R}$ vanishes and we have at the macroscopic level

$$(1+\alpha) h_i^{k+1} = \int_{\mathbb{R}} \Big(M_i^0 + \alpha M_i^k - \sigma\xi(M^k_{i+1/2} - M^k_{i-1/2}) \Big) \, \mathrm{d}\xi \ .$$

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

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

By definition of the water height reconstruction (3.56), we have the inequalities $h^k_{i+1/2-} \leq h_i^k$ and $h^k_{i-1/2+} \leq h_i^k$. As a consequence $M^k_{i+1/2-} \leq M_i^k$ and $M^k_{i-1/2+} \leq M_i^k$, which allows us to get the following lower bound of the integrand

$$M_i^0 + \alpha M_i^k - \sigma|\xi| M_i^k \ .$$

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

Next we propose an estimate for the kinetic entropy dissipation, which provides (3.63) with a discrete entropy inequality from some rank. We recall that since we consider a varying bottom, the kinetic entropy also depends on z. To ease the notations, we skip the dependence in ξ and write

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

Proposition 3.4.9. *The iterative kinetic scheme with hydrostatic reconstruction* (3.62) *verifies the following kinetic entropy inequality*

$$\begin{aligned} H(M_i^{k+1}, z_i) \leq & \\ & H(M_i^0, z_i) - \sigma\Big(\widetilde{G}_{i+1/2-} - \widetilde{G}_{i-1/2+}\Big) + (1+\alpha)\Delta t \partial_1 H(M_i^k, z_i) Q_i \\ & + \alpha\Big(H(M_i^k, z_i) - H(M_i^{k+1}, z_i)\Big) + (1+\alpha)\frac{g^2\pi^2}{6}(2M_i^k + M_i^{k+1})(M_i^{k+1} - M_i^k)^2 \\ & - \frac{g^2\pi^2}{6}(2M_i^k + M_i^0)(M_i^0 - M_i^k)^2 \ , \end{aligned} \tag{3.64}$$

with $Q_i = (M_i^{k+1} - f_i^{k+1})/\Delta t$ a collision term satisfying the conservation constraints (3.5)*, and where*

$$\begin{aligned} \widetilde{G}_{i+1/2-} = & \ \xi \mathbb{1}_{\xi<0} H(M_{i+1/2+}^k, z_{i+1/2}) + \xi \mathbb{1}_{\xi>0} H(M_{i+1/2-}^k, z_{i+1/2}) \\ & + \xi H(M_i^k, z_i) - \xi H(M_{i+1/2-}^k, z_{i+1/2}) \\ & + \Big(\nabla\eta(U_i^k)^T \begin{pmatrix} 1 \\ \xi \end{pmatrix} + g z_i\Big)(\xi M_{i+1/2-}^k - \xi M_i^k + (\xi - u_i^k)(M_i^k - M_{i+1/2-}^k)) \ , \end{aligned} \tag{3.65}$$

$$\begin{aligned} \widetilde{G}_{i-1/2+} = & \ \xi \mathbb{1}_{\xi<0} H(M_{i-1/2+}^k, z_{i-1/2}) + \xi \mathbb{1}_{\xi>0} H(M_{i-1/2-}^k, z_{i-1/2}) \\ & + \xi H(M_i^k, z_i) - \xi H(M_{i-1/2+}^k, z_{i-1/2}) \\ & + \Big(\nabla\eta(U_i^k)^T \begin{pmatrix} 1 \\ \xi \end{pmatrix} + g z_i\Big)(\xi M_{i-1/2+}^k - \xi M_i^k + (\xi - u_i^k)(M_i^k - M_{i-1/2+}^k)) \ , \end{aligned} \tag{3.66}$$

with the entropy $\eta(U) = \frac{hu^2}{2} + \frac{g}{2}h^2$.

Remark 3.4.10. *In Proposition 3.4.9 the difference $\widetilde{G}_{i+1/2-} - \widetilde{G}_{i-1/2+}$ is non conservative at the kinetic level, but becomes conservative when it is integrated over $\xi \in \mathbb{R}$. This is due do the fact that the last two lines of* (3.65) *and* (3.66) *are macroscopically zero, see [10] Proposition 3.1. Furthermore, we reiterate the comments made in remark 3.4.6 which are to say that in* (3.64) *the term*

$$(1+\alpha)\Delta t \partial_1 H(M_i^k, z_i) Q_i$$

is macroscopically zero for the half-disk maxwellian and when integrated, the quantity

$$\begin{aligned} & \alpha\Big(H(M_i^k, z_i) - H(M_i^{k+1}, z_i)\Big) + (1+\alpha)\frac{g^2\pi^2}{6}(2M_i^k + M_i^{k+1})(M_i^{k+1} - M_i^k)^2 \\ & - \frac{g^2\pi^2}{6}(2M_i^k + M_i^0)(M_i^0 - M_i^k)^2 \end{aligned}$$

will eventually become negative for k large enough.

Proof (Proposition 3.4.9). We start to rewrite (3.62) as

$$(1+\alpha)(M_i^{k+1}-M_i^k) = \tag{3.67}$$
$$(M_i^0-M_i^k)-\sigma\xi(M_{i+1/2}^k-M_{i-1/2}^k)+\sigma(\xi-u_i^k)(M_{i+1/2-}^k-M_{i-1/2+}^k)+(1+\alpha)\Delta t Q_i \ .$$

The strategy is to multiply (3.67) by $\partial_1 H(M_i^k, z_i)$ and to write

$$\partial_1 H(M_i^k,z_i)\Big[(1+\alpha)(M_i^{k+1}-M_i^k)-(M_i^0-M_i^k)-(1+\alpha)\Delta t Q_i\Big] = \tag{3.68}$$
$$-\sigma\partial_1 H(M_i^k,z_i)\Big[\xi(M_{i+1/2}^k-M_{i-1/2}^k)+\delta M_{i+1/2-}-\delta M_{i-1/2+}\Big] \ ,$$

where we defined

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

We apply Lemma 3.3.8 to the left hand side to get

$$\partial_1 H(M_i^k,z_i)\Big[(1+\alpha)(M_i^{k+1}-M_i^k)-(M_i^0-M_i^k)-(1+\alpha)\Delta t Q_i\Big] = \tag{3.69}$$
$$(1+\alpha)\Big(H(M_i^{k+1},z_i)-H(M_i^k,z_i)-\frac{g^2\pi^2}{6}(2M_i^k+M_i^{k+1})(M_i^{k+1}-M_i^k)^2\Big)$$
$$-\Big(H(M_i^0,z_i)-H(M_i^k,z_i)-\frac{g^2\pi^2}{6}(2M_i^k+M_i^0)(M_i^0-M_i^k)^2\Big)$$
$$-(1+\alpha)\Delta t\partial_1 H(M_i^k,z)Q_i \ .$$

Furthermore, an upper bound on the right hand side of (3.68) is obtained by applying Proposition 3.1 from [10] which directly yields

$$-\partial_1 H(M_i^k,z_i)\Big[\xi(M_{i+1/2}^k-M_{i-1/2}^k)+\delta M_{i+1/2-}-\delta M_{i-1/2+}\Big] \leq \widetilde{G}_{i-1/2+}-\widetilde{G}_{i+1/2-} \ , \tag{3.70}$$

with $\widetilde{G}_{i+1/2-}$ and $\widetilde{G}_{i-1/2+}$ defined by (3.65) and (3.66). Injecting equality (3.69) and inequality (3.70) into (3.68) we obtain the desired kinetic entropy inequality (3.64). □

The question arising naturally is under what conditions can we ensure the convergence of the sequence given by (3.63). The remainder of this section will be dedicated to answering this question. Let us fix δ, K_1, K_2 some strictly positive values and consider the set thereafter

$$\mathcal{D} = \Big\{(h,q)\in\mathbb{R}_+\times\mathbb{R},\ \delta\leq h\leq K_1,\ |q|\leq K_2 h\Big\} \ . \tag{3.71}$$

We will see that there will be convergence if the time step Δt is small enough and if the solution remains in the set $\mathcal{D}^N$. This is somewhat restrictive, as we don't have any maximum principle for the Saint-Venant system. However we believe that for a reasonable initial condition, the upper bounds $h\leq K_1$ and $|u|\leq K_2$ are realized. On the other hand, the lower bound on h means that there cannot be dry areas. This last restriction seems to be required due to the choice of the Maxwellian and its lack of regularity near

the border of the support. To establish our convergence result we will first introduce some additional notations to restate the iterative process (3.62) under vector form. We consider $R_\pm : \mathbb{R}^{2N} \to \mathbb{R}^{2N}$ the two reconstruction operators as well as $\mathcal{M}_\xi : \mathbb{R}^{2N} \to \mathbb{R}^N$ the projection onto the space of Maxwellians (hydrodynamic equilibrium at the kinetic level) defined below for any $U \in (\mathbb{R}_+ \times \mathbb{R})^N$

$$R_+(U) = \begin{pmatrix} U_{1/2+} \\ \vdots \\ U_{i-1/2+} \\ \vdots \\ U_{P-1/2+} \end{pmatrix}, \quad R_-(U) = \begin{pmatrix} U_{3/2-} \\ \vdots \\ U_{i+1/2-} \\ \vdots \\ U_{P+1/2-} \end{pmatrix}, \quad \mathcal{M}_\xi(U) = \begin{pmatrix} M(U_1, \xi) \\ \vdots \\ M(U_i, \xi) \\ \vdots \\ M(U_P, \xi) \end{pmatrix}.$$

We also introduce the matrices from $\mathbb{R}^{N\times N}$ corresponding, in the right hand side of (3.62), to the upwind linear transport and source term:

$$\begin{aligned} \mathbf{A}_\xi[U] &= \mathbf{1}_{\xi<0}|\xi|(\mathbf{J} - \mathbf{I}) + (\xi\,\mathbf{I} - \mathrm{diag}(u)), \quad \mathbf{J}_{i,j} = \delta_{i,j-1} \\ \mathbf{B}_\xi[U] &= \mathbf{1}_{\xi>0}|\xi|(\mathbf{N} - \mathbf{I}) - (\xi\,\mathbf{I} - \mathrm{diag}(u)), \quad \mathbf{N}_{i,j} = \delta_{i,j+1} \end{aligned} \tag{3.72}$$

Note that there is a dependence on the macroscopic velocity u in the source term contribution. Then the iteration (3.62) can be recasted at the macroscopic level for the water height and the momentum as

$$(1+\alpha)h^{k+1} = \int_{\mathbb{R}} \phi(U^k, \xi)\,\mathrm{d}\xi, \quad (1+\alpha)hu^{k+1} = \int_{\mathbb{R}} \xi\phi(U^k, \xi)\,\mathrm{d}\xi \tag{3.73}$$

by using the kinetic operator ϕ defined for all $U \in \mathcal{D}^N$ and $\xi \in \mathbb{R}$ as:

$$\phi(U, \xi) = \mathcal{M}_\xi(U^n) + \Big[\alpha\mathcal{M}_\xi + \sigma\mathbf{A}_\xi(\mathcal{M}_\xi \circ R_+) + \sigma\mathbf{B}_\xi(\mathcal{M}_\xi \circ R_-)\Big](U) \tag{3.74}$$

Remark 3.4.11. *The operators $R_\pm$ do not leave $\mathcal{D}^N$ invariant in general since they can lead to dry cells. To have the set $\mathcal{D}^N$ preserved, we have to assume that in each cell the water depth h_i is greater that $\delta + \max(z_{i+1/2} - z_i, z_{i-1/2} - z_i)$. Hence our approach will be limited to sufficiently deep flows. In particular, no wet/dry transition can occur.*

Remark 3.4.12. *Note that ϕ is compactly supported on $\mathcal{D}^N \times \Xi$, with the set Ξ defined as:*

$$\Xi = \bigcup_{U\in\mathcal{D}} \mathrm{supp}\, M(U, \cdot) \subset \Big[-K_2 - \sqrt{2gK_1},\, K_2 + \sqrt{2gK_1}\Big] \tag{3.75}$$

We are now able to state our convergence result.

Proposition 3.4.13. *Assume the iterative process* (3.62) *keeps the numerical approximation in $\mathcal{D}^N$. There exists a positive constant $C(K_1, K_2, 1/\delta)$ such that the scheme is granted to converge under the CFL condition*

$$\frac{\Delta t}{\Delta x} \leq C(K_1, K_2, 1/\delta)\ .$$

Remark 3.4.14. *In practice it seems that the lower bound on h from* (3.71) *is not needed. In fact we later consider Thacker's numerical test with wet/dry interfaces and despite this the method seems to be working fine. A proof without this hypothesis is currently under investigation.*

The proof of Proposition (3.4.13) makes use of the following two lemmas, where $\|\cdot\|$ denotes the infinity vector norm on $\mathbb{R}^N$.

Lemma 3.4.15. *There exists a constant $L(K_1, K_2, 1/\delta)$ such that any pair $(U, \widetilde{U})$ belonging to $\mathcal{D}^N \times \mathcal{D}^N$ satisfies*

$$\int_{\mathbb{R}} \left\|\mathcal{M}_\xi(U) - \mathcal{M}_\xi(\widetilde{U})\right\| \mathrm{d}\xi \leq L(K_1, K_2, 1/\delta)\left(\left\|h - \widetilde{h}\right\| + \left\|q - \widetilde{q}\right\|\right) . \tag{3.76}$$

In the next lemma, we have a Lipschitz property characterizing the reconstruction operators $R^h_\pm, R^{hu}_\pm$ valued respectively in $\mathbb{R}^N_+$ and $\mathbb{R}^N$, and defined for all $i \in [\![1, P]\!]$ as

$$R^h_\pm(U)_i = R_\pm(U)_{2i} = h_{i\mp 1/2\pm}, \quad R^{hu}_\pm(U)_i = R_\pm(U)_{2i+1} = h_{i\mp 1/2\pm} u_i$$

Lemma 3.4.16. *For any $U, \widetilde{U} \in \mathcal{D}^N$, we have*

$$\left\|R^h_\pm(U) - R^h_\pm(\widetilde{U})\right\| + \left\|R^{hu}_\pm(U) - R^{hu}_\pm(\widetilde{U})\right\| \leq (1 + 2K_2)\left\|h - \widetilde{h}\right\| + \left\|q - \widetilde{q}\right\| \tag{3.77}$$

All the proofs can be found in Appendix 3.B.

3.4.6 Numerical tests with varying bathymetry

We consider Thacker's testcase, also known as the parabolic bowl testcase, taken from [26]. We plot the numerical solution at two times 1/2 and 1 in Figure 3.4.3. It seems that the iterative scheme ($\alpha = 1$, $\varepsilon_{\text{tol}} = 10^{-9}$) whose stopping criterion is given by combining a tolerance and an entropy condition converges despite two moving wet/dry transitions characterizing the solution. Hence this gives hope to improve the convergence result by dropping the lower bound on h from Proposition 3.4.13.

Numerical simulations were performed for the iterative kinetic scheme with an entropy-only stopping criterion in conjunction with $\alpha = 0$. We remind that the reason for taking $\alpha = 0$ with this specific stopping criterion was discussed in Section 3.4.2, and is related to the ability of the iterative process to yield a consistent update after just a few iterations — in fact as fiew as one. Unfortunately the results we obtained weren't reliable. In fact when refining the mesh, we were witnessing oscillations together with an increase of the total energy despite reaching the maximum number of iterations allowed, which was fixed at 5000. We believe that $\alpha > 0$ could be an important condition to dissipate the energy, but in the case of the entropy-only condition the numerical approximation then doesn't converge to the solution. We also remark that poor results are obtained on the discharge profile at time $t = 1$ across all tested schemes. This corresponds to the time around which the velocity changes sign so that the fluid starts going back in the other direction. Moving to a second order accuracy method might help to circumvent this issue.

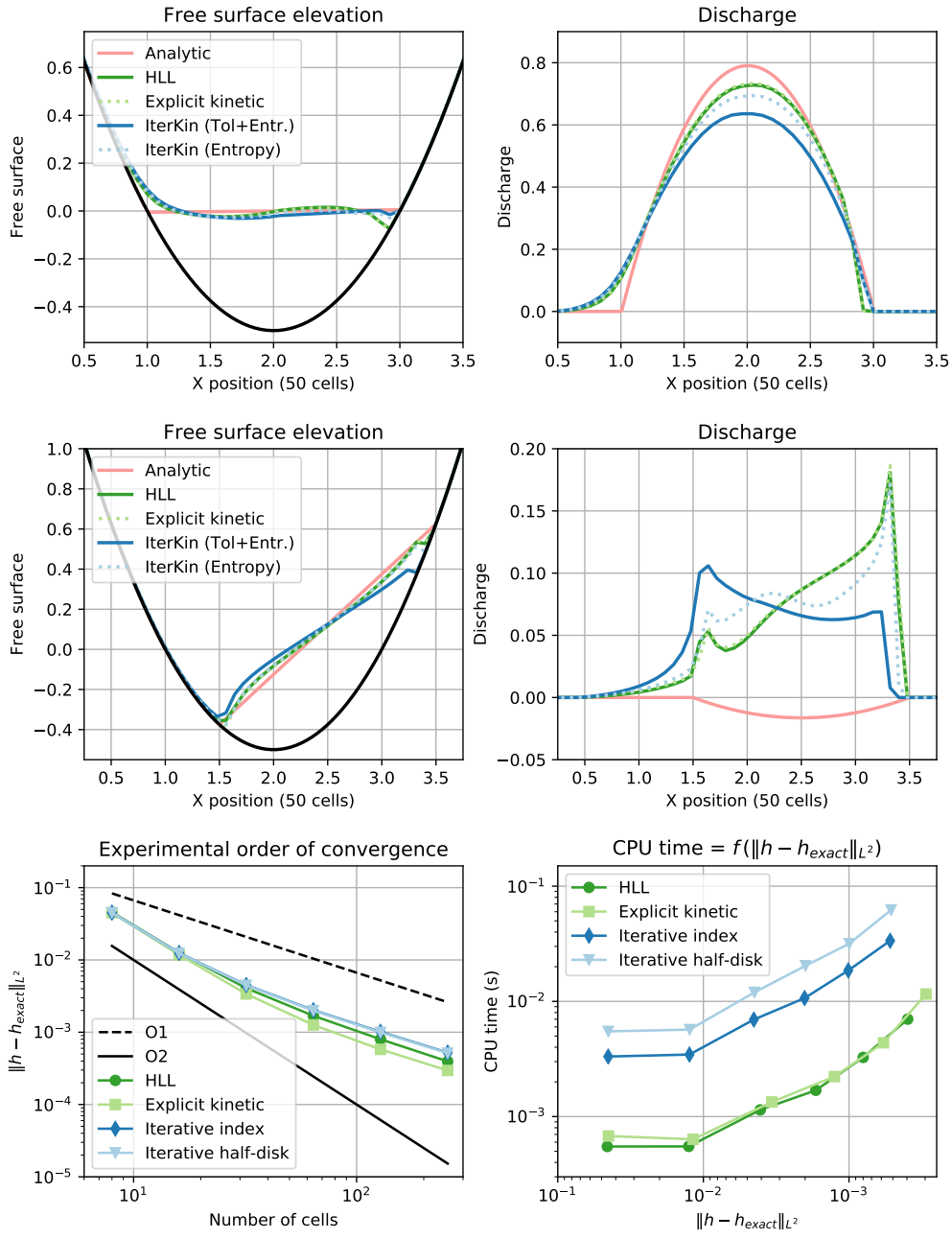

Figure 3.4.3: Thacker's testcase. From top to bottom: solution at time 1/2, solution at time 1, convergence and efficiency curves.

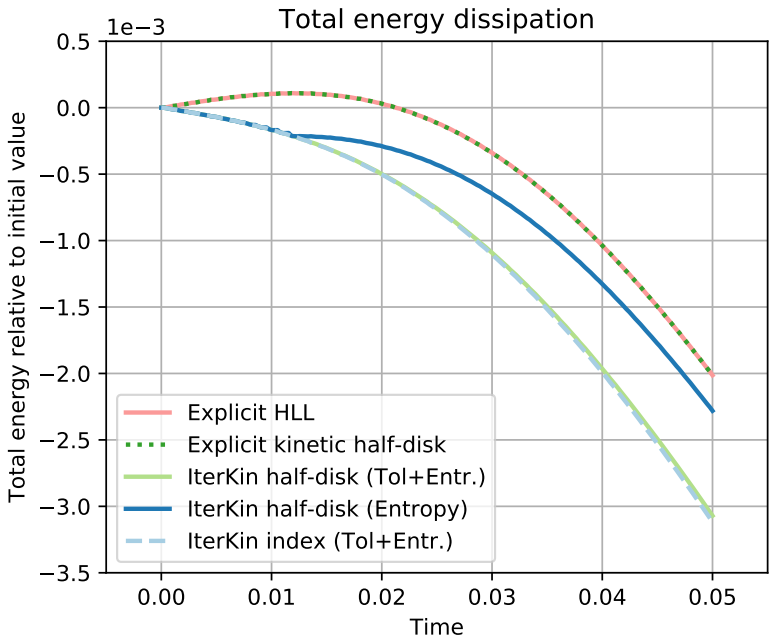

Figure 3.4.4: Dissipation of total energy is achieved by the considered iterative kinetic schemes, but not by the explicit method. The light blue and green lines represent iterative methods using a tolerance plus entropy-type stopping criterion, whereas the dark blue line is for the entropy-only iterative scheme. The three dissipate the total energy.

Given the efficiency curves shown in Figure 3.4.3, the explicit strategy seems preferable in terms of the computational cost at a prescribed accuracy. However we have to stress that among all the considered methods, the iterative kinetic scheme with half-disk maxwellian is the only one to satisfy a discrete entropy inequality, at least when the number of iterations is large enough. We remind that on the opposite, the explicit kinetic scheme with hydrostatic reconstruction does not satisfy a discrete entropy inequality without quadratic error term, however restrictive the CFL condition is, which is the Proposition 3.8 from [10]. Hence the iterative scheme can be considered an improvement over this aspect, and we illustrate this through a second numerical test where the explicit strategy increases the total energy, unlike the iterative method. More precisely we measure the variation of total energy in a configuration with a varying bottom, and where the initial condition is given by a flat free surface and a constant velocity. Periodic boundary conditions are used, and the results can be seen in Figure 3.4.4. Interestingly all the iterative methods manage to dissipate the total energy, even the scheme using the index maxwellian, for which we recall there is no proof of discrete entropy inequality. On the contrary, the explicit kinetic scheme with half-disk maxwellian increases the energy in the first few time steps, after what it decreases. The same goes for the explicit HLL scheme.

3.5 Perspectives and conclusion

3.5.1 Towards 2D: exploring the iterative method

We generalize the iterative procedure to the case of the 2D Saint-Venant system

$$\begin{cases} \dfrac{\partial h}{\partial t} + \dfrac{\partial q}{\partial x} + \dfrac{\partial r}{\partial y} = 0 \\ \dfrac{\partial q}{\partial t} + \dfrac{\partial}{\partial x}\Big(\dfrac{q^2}{h} + \dfrac{g}{2}h^2\Big) + \dfrac{\partial}{\partial y}\Big(\dfrac{qr}{h}\Big) = -gh\dfrac{\partial z}{\partial x} \\ \dfrac{\partial r}{\partial t} + \dfrac{\partial}{\partial x}\Big(\dfrac{qr}{h}\Big) + \dfrac{\partial}{\partial y}\Big(\dfrac{r^2}{h} + \dfrac{g}{2}h^2\Big) = -gh\dfrac{\partial z}{\partial y} \end{cases} . \tag{3.78}$$

A kinetic representation for System (3.78) with flat bathymetry takes the form

$$\partial_t f + \xi \cdot \nabla f = \frac{1}{\varepsilon}(M[f] - f) \tag{3.79}$$

with $\xi \in \mathbb{R}^2$ the kinetic velocity, $f(t,x,y,\xi) \in \mathbb{R}_+$ the density and $M[f]$ some hydrodynamic equilibrium associated to f and verifying

$$\forall n \in \mathbb{S}^2,\ \int_{\mathbb{R}^2} \begin{pmatrix} 1 \\ \xi \cdot n \\ (\xi \cdot n)\xi \end{pmatrix} M(U,\xi)\,\mathrm{d}\xi = \begin{pmatrix} h \\ F(U;n) \end{pmatrix} \in \mathbb{R}^4 . \tag{3.80}$$

The quantity $F(U;n)$ is the flux in direction n associated to the Saint-Venant system (3.78), and is detailed in (2.2) in dimensionless form. Defining $c = \sqrt{gh/2}$ the speed of sound, a choice compatible with the moment relations (3.80) is given by setting

$$M(U,\xi) = \frac{h}{c^2}\,\chi\Big(\frac{\xi - V}{c}\Big) ,$$

where the shape function χ defined on $\mathbb{R}^2$ has to satisfy

$$\forall (i,j) \in \{1,2\}^2, \ \int_{\mathbb{R}^2} \begin{pmatrix} 1 \\ \omega_i \omega_j \end{pmatrix} \chi(\omega) \, \mathrm{d}\omega = \begin{pmatrix} 1 \\ \delta_{ij} \end{pmatrix} \ .$$

Two possible choices can be found for instance in [1] and [5], and they correspond respectively to

$$\begin{cases} \chi_1(\omega) = \dfrac{1}{4\pi} \mathbb{1}_{|\omega| \leq 2} \\ \chi_2(\omega) = \dfrac{1}{12} \mathbb{1}_{\|\omega\|_\infty \leq \sqrt{3}} \end{cases} . \tag{3.81}$$

We will use χ_1 throughout this section. Let us remark that when averaging this shape function in one direction, say for instance along the ω_2-axis, we get a function of ω_1 that coincide with the half-disk shape function used in the 1D case (3.10)

$$\frac{1}{4\pi} \int_{\mathbb{R}} \mathbb{1}_{|\omega| \leq 2} \, \mathrm{d}\omega_2 = \frac{1}{4\pi} \int_{-\sqrt{4-\omega_1^2}}^{\sqrt{4-\omega_1^2}} \mathrm{d}\omega_2 = \frac{1}{2\pi} \sqrt{4-\omega_1^2} = \frac{1}{\pi} \sqrt{1 - \frac{\omega_1^2}{4}} \ .$$

A consequence is that applying the 2D scheme to an artificially 2D testcase (constant along one direction) should yield the same result as using the 1D scheme (3.63) in the truly one dimensional configuration.

The iterative method will be based on a BGK splitting of the kinetic representation (3.79). The first step corresponds to the collisions of particles, which amounts to relax f towards a maxwellian sharing the same moment relations when integrating against 1 and ξ. When taking the limit $\varepsilon \to 0$, the hydrodynamic equilibrium is reached immediately, and it gives us the starting point $f^0 = M(U_f^0, \xi)$. This data is then advected during the transport step below

$$\begin{cases} \partial_t f + \xi \cdot \nabla f = 0 \\ f(t=0,x,y,\xi) = M(U_f^0(x,y),\xi) \end{cases} . \tag{3.82}$$

We discretize it using an upwind approach. First we average (3.82) over a square cell $C_{i,j}$

$$\partial_t f_{i,j} + \frac{1}{|C_{i,j}|} \int_{\partial C_{i,j}} f(t,x,y,\xi) \, \xi \cdot n_{\partial C_{i,j}} \, \mathrm{d}\sigma = 0 \ . \tag{3.83}$$

We use an upwind evaluation of the flux over the interfaces, such that (3.83) is approximated by

$$\partial_t f_{i,j} + \frac{\xi_1 \Delta y}{|C_{i,j}|} (f_{i+1/2,j} - f_{i-1/2,j}) + \frac{\xi_2 \Delta x}{|C_{i,j}|} (f_{i,j+1/2} - f_{i,j-1/2}) = 0 \ ,$$

with

$$f_{i+1/2,j} = \mathbb{1}_{\xi_1 > 0} f_{i,j} + \mathbb{1}_{\xi_1 < 0} f_{i+1,j} \ , \quad f_{i,j+1/2} = \mathbb{1}_{\xi_2 > 0} f_{i,j} + \mathbb{1}_{\xi_2 < 0} f_{i,j+1} \ .$$

Let us introduce the corresponding macroscopic fluxes

$$F(U_L,U_R)=\int_{\mathbb{R}^2}\begin{pmatrix}1\\ \xi\end{pmatrix}\xi_1[\mathbb{1}_{\xi_1>0}M(U_L,\xi)+\mathbb{1}_{\xi_1<0}M(U_R,\xi)]\,\mathrm{d}\xi\ ,$$
$$G(U_L,U_R)=\int_{\mathbb{R}^2}\begin{pmatrix}1\\ \xi\end{pmatrix}\xi_2[\mathbb{1}_{\xi_2>0}M(U_L,\xi)+\mathbb{1}_{\xi_2<0}M(U_R,\xi)]\,\mathrm{d}\xi\ .$$

For a given $U^0\in\mathbb{R}_+\times\mathbb{R}^2$, the fully implicit scheme is defined by setting U^1 equal to the solution U of

$$\frac{U_{i,j}-U^0_{i,j}}{\Delta t}+\frac{F(U_{i,j},U_{i+1,j})-F(U_{i-1,j},U_{i,j})}{\Delta x}+\frac{G(U_{i,j},U_{i,j+1})-G(U_{i,j-1},U_{i,j})}{\Delta y}=0\ .$$

In the definition of the numerical fluxes F,G integrals are performed over half-planes corresponding to the natural upwinding introduced at the kinetic level. Since we are not integrating over $\mathbb{R}^2$, we cannot make use of the moment relations (3.80). Let us define

$$\alpha^{\pm}=\arcsin\Big(\min\Big(1,\max\Big(\pm1,\frac{V_1}{2c}\Big)\Big)\Big)\ ,\quad \beta^{\pm}=\arcsin\Big(\min\Big(1,\max\Big(\pm1,\frac{V_2}{2c}\Big)\Big)\Big)\ ,$$

we have the following formulas involved in the computation of the numerical flux F

$$\int_{\xi_1>0}\begin{pmatrix}1\\ \xi_2\end{pmatrix}\xi_1M(U,\xi)\,\mathrm{d}\xi=\frac{2h}{\pi}\begin{pmatrix}1\\ V_2\end{pmatrix}\Big[\frac{V_1}{2}\Big(\theta+\frac{\sin(2\theta)}{2}\Big)-\frac{2c}{3}\cos^3\theta\Big]_{\theta=\alpha^-}^{\theta=\alpha^+}\ ,$$
$$\int_{\xi_1>0}\xi_1^2M(U,\xi)\,\mathrm{d}\xi=\frac{h}{\pi}(V_1^2+4c^2)\Big[\theta+\frac{\sin(2\theta)}{2}\Big]_{\alpha^-}^{\alpha^+}-\frac{hc^2}{\pi}\Big[\frac{\sin(4\theta)}{4}+2\sin(2\theta)+3\theta\Big]_{\alpha^-}^{\alpha^+}\ .$$

Integrals over the left half plane $\xi_1<0$ are obtained from the previous ones, only substituting max by min and vice versa in the bounds $\alpha^{\pm}$. As for the numerical flux G, there holds similarly

$$\int_{\xi_2>0}\begin{pmatrix}1\\ \xi_1\end{pmatrix}\xi_2M(U,\xi)\,\mathrm{d}\xi=\frac{2h}{\pi}\begin{pmatrix}1\\ V_1\end{pmatrix}\Big[\frac{V_2}{2}\Big(\theta+\frac{\sin(2\theta)}{2}\Big)-\frac{2c}{3}\cos^3\theta\Big]_{\beta^-}^{\beta^+}\ ,$$
$$\int_{\xi_2>0}\xi_2^2M(U,\xi)\,\mathrm{d}\xi=\frac{h}{\pi}(V_2^2+4c^2)\Big[\theta+\frac{\sin(2\theta)}{2}\Big]_{\beta^-}^{\beta^+}-\frac{hc^2}{\pi}\Big[\frac{\sin(4\theta)}{4}+2\sin(2\theta)+3\theta\Big]_{\beta^-}^{\beta^+}\ .$$

A steady state for the 2D Saint-Venant system necessarily satisfies $\nabla\cdot(hV)=0$. We focus on the artificially 2D case where the quantities remain constant along the y-axis, and where we enforce

$$hV_x=K\in\mathbb{R},\quad hV_y=0\ . \tag{3.84}$$

The momentum equation results in the following definition of the bathymetry

$$\nabla\cdot(hV\otimes V)+gh\nabla\cdot((h+z)\,\mathbf{I}_2)=0\Longrightarrow g\nabla(h+z)=-(V\cdot\nabla)V=-\frac{1}{2}\partial_x\begin{pmatrix}V_x^2\\ 0\end{pmatrix}\ .$$

Integrating we get

$$g(h+z)=K'-\frac{1}{2}V_x^2\Longrightarrow z=K''-h-\frac{K^2}{2gh^2}\ . \tag{3.85}$$

We work on the square domain $[0,1] \times [0,1]$ with periodic boundary condition, and the initial condition giving rise to a steady flow is set to

$$h(x) = 1 + \mathbb{1}_{[\frac{1}{4},\frac{3}{4}]}(x)(1 + \sin(8\pi(x - 1/2) - \pi/2)), \quad K = 1/2, \quad g = 1 \ ,$$

in addition to the constraints (3.84)–(3.85) on V and z.

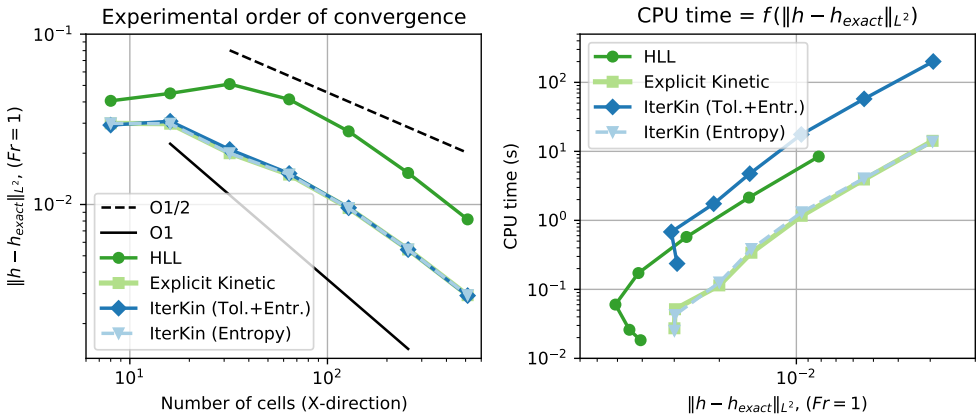

Figure 3.5.1: Left: convergence curves for different schemes. Two stopping criteria for the iterative schemes are compared: tolerance and total energy dissipation, and total energy dissipation only. Right: efficiency curve. Note that the x-axis has decreasing values for the L^2 error on the water height.

We plot the results in Figure 3.5.1. All the schemes make use of the first order hydrostatic reconstruction (3.55)–(3.58) to handle the bathymetry. We also compare to a non kinetic scheme making use of the HLL numerical flux. As in the one dimensional case, we see that the iterative scheme with tolerance criterion is less efficient. On the otherhand using an entropy-only stopping criterion in conjunction to $\alpha = 0$ is quite competitive.

3.5.2 Conclusion

In this work we have investigated the case of implicit kinetic schemes for the Saint-Venant system. Such schemes provide us with the framework to obtain a fully discrete entropy inequality without restriction on the time step. In practice, we have seen that it is possible to rewrite the implicit update explicitly when dealing with a flat bathymetry (no source term) and for a simplified choice of maxwellian. Knowing explicitly the expression of the solution on the linear system (written at the kinetic level) is the best thing one can hope for when it comes to implicit solvers applied to a nonlinear problem. Yet, the quadratic algorithmic complexity makes it less efficient to use this scheme, at least in the numerical experiments considered in the document.

Dealing with a varying bathymetry is achieved by the mean of the hydrostatic reconstruction, but this renders the scheme nonlinear at the kinetic level and we are no more able to compute the update explicitly. Instead we turn to an iterative strategy based

on a Gauss-Jacobi decomposition of the operator, and a CFL condition is required to obtain the convergence of the method. To prove this latter point we had to make strong assumptions on the boundedness of the solutions, but we believe it could be relaxed at least to cases with dry areas given the numerical results of Thacker's testcase. Most importantly, we were able to numerically validate the dissipation of total energy by the iterative kinetic scheme with hydrostatic reconstruction, which is a true improvement over the explicit case. In a future work we plan to extend this study to the two dimensional case, and already experimented an iterative kinetic scheme in this context.

Appendix

3.A Expression of the numerical updates

— Proof of Lemma 3.3.11 —

We have

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

Performing the change of variable $y = 1 - 1/(1+x)$ leads to

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

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

$$I = -\int \sum_{l=0}^{k-1} y^l\,\mathrm{d}y - \ln(|1-y|) + C = \ln(|1+x|) - \sum_{l=1}^{k}\frac{y^l}{l} + C'$$

for some $(C, C') \in \mathbb{R}^2$.

— Proof of Lemma 3.3.12 —

Yet again we make the chage of variable $y = x/(1+x) = 1 - 1/(1+x)$, and we have

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

and we use the formula $y^k - 1 = (y-1)(y^{k-1} + y^{k-2} + \ldots + y + 1)$ to get the following

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

for some $(C, C', C'') \in \mathbb{R}^3$.

— Proof of Lemma 3.3.13 —

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

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

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

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

Dividing this by $1 - y$ leads to

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

Iterating this step one more time we find

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

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

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

Finally, we simplify the double and triple sums

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

From this we also deduce that

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

As a conclusion we have the expression

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

for some $C\in\mathbb{R}$ and where we recall $y=x/(x+1)=1-1/(x+1)$.

— Proof of Proposition 3.3.10 —

We recall that the index maxwellian over cell $1\leq j\leq N$ writes

$$M_j^0=\frac{1}{2\sqrt{3}}\sqrt{\frac{2h_j^0}{g}}\,\mathbb{1}_{a_j\leq\xi\leq b_j}\ .$$

Hence to do the proof, it is sufficient to show that for $j\geq i$

$$\int_{\mathbb{R}_-}\frac{(-\sigma\xi)^{j-i}}{(1-\sigma\xi)^{j-i+1}}\mathbb{1}_{a_j\leq\xi\leq b_j}\,\mathrm{d}\xi=(\mathcal{A}h)_{i,j}\ ,\ \int_{\mathbb{R}_-}\frac{\xi\,(-\sigma\xi)^{j-i}}{(1-\sigma\xi)^{j-i+1}}\mathbb{1}_{a_j\leq\xi\leq b_j}\,\mathrm{d}\xi=-\frac{1}{\sigma}(\mathcal{A}hu)_{i,j}\ ,$$

and that for $j\leq i$

$$\int_{\mathbb{R}_+}\frac{(\sigma\xi)^{i-j}}{(1+\sigma\xi)^{i-j+1}}\mathbb{1}_{a_j\leq\xi\leq b_j}\,\mathrm{d}\xi=(\mathcal{B}h)_{i,j}\ ,\ \int_{\mathbb{R}_+}\frac{\xi\,(\sigma\xi)^{i-j}}{(1+\sigma\xi)^{i-j+1}}\mathbb{1}_{a_j\leq\xi\leq b_j}\,\mathrm{d}\xi=\frac{1}{\sigma}(\mathcal{B}hu)_{i,j}\ .$$

Since the arguments are similar, we only treat the first integral. We have

$$I = \int_{\mathbb{R}_-} \frac{(-\sigma\xi)^{j-i}}{(1-\sigma\xi)^{j-i+1}} \mathbb{1}_{a_j \le \xi \le b_j} \,\mathrm{d}\xi = \int_{\min(0,a_j)}^{\min(0,b_j)} \frac{(-\sigma\xi)^{j-i}}{(1-\sigma\xi)^{j-i+1}} \,\mathrm{d}\xi$$

The change of variable $x = -\sigma\xi$ then leads to

$$I = \frac{1}{\sigma} \int_{-\min(0,b_j)\sigma}^{-\min(0,a_j)\sigma} \frac{x^{j-i}}{(1+x)^{j-i+1}} \,\mathrm{d}x$$

We conclude by applying Lemma 3.3.11 for the value $k = j - i$.

3.B Convergence of the HR iterative kinetic scheme

Here we give the proof of convergence of the iterative kinetic scheme with hydrostatic reconstruction (3.62).

— Proof of Lemma 3.4.15 —

We start by computing the gradient of the Maxwellian with respect to h and hu:

$$\forall U \in \mathbb{R}_+ \times \mathbb{R},\ \nabla_{(h,hu)} M(U,\xi) = \frac{1}{gh\pi} \begin{pmatrix} gh - u(\xi-u) \\ (\xi-u) \end{pmatrix} \frac{\mathbb{1}_{\xi \in \operatorname{supp} M(U,\cdot)}}{\sqrt{2gh - (\xi-u)^2}}$$

More specifically, if U belongs to the set $\mathcal{D}$, the greatest derivative in absolute value between $|\partial_h M|$ and $|\partial_{hu} M|$ is bounded as

$$\left\| \nabla M(U,\xi) \right\|_\infty \le \frac{1}{\pi} L(K_1, K_2, 1/\delta) \frac{\mathbb{1}_{\xi \in \operatorname{supp} M(U,\cdot)}}{\sqrt{2gh - (\xi-u)^2}} \tag{3.86}$$

where the constant $L(K_1, K_2, 1/\delta)$ is equal to

$$L(K_1, K_2, 1/\delta) = \max\left(1 + \frac{K_2\sqrt{2gK_1}}{g\delta},\ \frac{\sqrt{2gK_1}}{g\delta}\right). \tag{3.87}$$

We remark that the upper bound (3.86) is integrable, and by performing the change of variable $\cos\theta = (\xi - u)/\sqrt{2gh}$ we find

$$\int_{u-\sqrt{2gh}}^{u+\sqrt{2gh}} \frac{\mathrm{d}\xi}{\sqrt{2gh - (\xi-u)^2}} = \pi\ .$$

It follows directly that for any pair $(U, \widetilde{U}) \in \mathcal{D} \times \mathcal{D}$ one has

$$\int_{\mathbb{R}} \left| M(U,\xi) - M(\widetilde{U},\xi) \right| \mathrm{d}\xi \le L(K_1, K_2, 1/\delta) \left(\left| h - \widetilde{h} \right| + \left| q - \widetilde{q} \right| \right)$$

and thus we have (3.76) by taking the infinity norm over the mesh cells.

— Proof of Lemma 3.4.16 —

Let U and $\widetilde{U}$ be in $\mathcal{D}^N$, and let $i \in [\![1, P]\!]$. It is enough to show the two inequalities

$$\begin{aligned}
&\left|h_{i\mp 1/2\pm} - \widetilde{h}_{i\mp 1/2\pm}\right| \leq \left|h_i - \widetilde{h}_i\right| \\
&\left|q_{i\mp 1/2\pm} - \widetilde{q}_{i\mp 1/2\pm}\right| \leq \left|q_i - \widetilde{q}_i\right| + 2K_2\left|h_i - \widetilde{h}_i\right|
\end{aligned}$$

The first inequality is obvious if both h_i and $\widetilde{h}_i$ are smaller or greater that $\max(z_i, z_{i\mp 1}) - z_i$ at the same time. In fact in the first case both reconstructed water heights are zero, whereas in the second one, their difference is equal to $h_i - \widetilde{h}_i$. There only remains the case

$$h_i < \max(z_i, z_{i\mp 1}) - z_i < \widetilde{h}_i$$

implying that

$$\left|h_{i\mp 1/2\pm} - \widetilde{h}_{i\mp 1/2\pm}\right| = \widetilde{h}_{i\mp 1/2\pm} = \widetilde{h}_i + z_i - \max(z_i, z_{i\mp 1}) \leq \widetilde{h}_i - h_i$$

Next the second inequality on the discharge is obtained by remarking that

$$\begin{aligned}
\left|q_{i\mp 1/2\pm} - \widetilde{q}_{i\mp 1/2\pm}\right| &= \left|h_{i\mp 1/2\pm}(u_i - \widetilde{u}_i) + \widetilde{u}_i(h_{i\mp 1/2\pm} - \widetilde{h}_{i\mp 1/2\pm})\right| \\
&\leq h_i\left|u_i - \widetilde{u}_i\right| + K_2\left|h_i - \widetilde{h}_i\right| = \left|q_i - (h_i - \widetilde{h}_i)\widetilde{u}_i - \widetilde{q}_i\right| + K_2\left|h_i - \widetilde{h}_i\right| \\
&\leq \left|q_i - \widetilde{q}_i\right| + 2K_2\left|h_i - \widetilde{h}_i\right|
\end{aligned}$$

where we used that $h_{i\mp 1/2\pm} \leq h_i$ as well as the first inequality to get the second line. We conclude by summing the two inequalities and by taking the maximum with respect to index i.

— Proof of Proposition 3.4.13 —

The iterative process converges as soon as we have a contraction property on the update for some $\Delta t > 0$ sufficiently small, that is to say if for any $U, \widetilde{U} \in \mathcal{D}^P$ there holds

$$\left\| \int_\Xi \frac{\phi(U,\xi) - \phi(\widetilde{U},\xi)}{1+\alpha}\,\mathrm{d}\xi \right\| + \left\| \int_\Xi \frac{\xi(\phi(U,\xi) - \phi(\widetilde{U},\xi))}{1+\alpha}\,\mathrm{d}\xi \right\| < \left\|h - \widetilde{h}\right\| + \left\|q - \widetilde{q}\right\| \tag{3.88}$$

with $\|\cdot\|$ the infinity norm on $\mathbb{R}^P$. We have the following bound:

$$\begin{aligned}
\left\| \int_\Xi \phi(U,\xi) - \phi(\widetilde{U},\xi)\,\mathrm{d}\xi \right\| \leq \alpha\left\|h - \widetilde{h}\right\| &+ \sigma \int_\Xi \left\|\mathbf{A}_\xi(\mathcal{M}_\xi \circ R_+(U) - \mathcal{M}_\xi \circ R_+(\widetilde{U}))\right\| \mathrm{d}\xi \\
&+ \sigma \int_\Xi \left\|\mathbf{B}_\xi(\mathcal{M}_\xi \circ R_-(U) - \mathcal{M}_\xi \circ R_-(\widetilde{U}))\right\| \mathrm{d}\xi
\end{aligned} \tag{3.89}$$

From the definition (3.72) of matrices $\mathbf{A}_\xi, \mathbf{B}_\xi$, we remark that the two component-wise inequalities

$$|\mathbf{A}_\xi| \leq (K_2 + \sqrt{2gK_1})(\mathbf{J} + 2\mathbf{I}) + K_2\,\mathbf{I},$$

$$|\mathbf{B}_\xi| \leq (K_2 + \sqrt{2gK_1})(\mathbf{N} + 2\mathbf{I}) + K_2\,\mathbf{I}$$

are true for any $\xi \in \Xi$. Introducing $|\!|\!| \cdot |\!|\!|$ the matrix subordinate norm associated to $\|\cdot\|$, we then get that there exist a real number $\rho(K_1, K_2)$ such that

$$|\!|\!|\mathbf{A}_\xi|\!|\!|, |\!|\!|\mathbf{B}_\xi|\!|\!| \leq \rho(K_1, K_2)\ . \tag{3.90}$$

Hence we can write that

$$\begin{aligned}\big\|\mathbf{A}_\xi(\mathcal{M}_\xi \circ R_+(U) - \mathcal{M}_\xi \circ R_+(\widetilde{U}))\big\| &\leq \rho\big\|(\mathcal{M}_\xi \circ R_+(U) - \mathcal{M}_\xi \circ R_+(\widetilde{U}))\big\| \\ \big\|\mathbf{B}_\xi(\mathcal{M}_\xi \circ R_-(U) - \mathcal{M}_\xi \circ R_-(\widetilde{U}))\big\| &\leq \rho\big\|(\mathcal{M}_\xi \circ R_-(U) - \mathcal{M}_\xi \circ R_-(\widetilde{U}))\big\|\end{aligned}$$

Combining lemmas 3.4.15 and 3.4.16 together, it then follows that

$$\begin{aligned}\int_\Xi \big\|\mathbf{A}_\xi(\mathcal{M}_\xi \circ R_+(U) - \mathcal{M}_\xi \circ R_+(\widetilde{U}))\big\|\,\mathrm{d}\xi &\leq L\,\rho\big(\big\|q - \widetilde{q}\big\| + (1 + 2K_2)\big\|h - \widetilde{h}\big\|\big) \\ \int_\Xi \big\|\mathbf{B}_\xi(\mathcal{M}_\xi \circ R_-(U) - \mathcal{M}_\xi \circ R_-(\widetilde{U}))\big\|\,\mathrm{d}\xi &\leq L\,\rho\big(\big\|q - \widetilde{q}\big\| + (1 + 2K_2)\big\|h - \widetilde{h}\big\|\big)\end{aligned}$$

where the constant L was defined in (3.87). Plugging this in (3.89) leads to:

$$\left\| \int_\Xi \phi(U, \xi) - \phi(\widetilde{U}, \xi)\,\mathrm{d}\xi \right\| \leq \alpha\big\|h - \widetilde{h}\big\| + 2\sigma L\rho(1 + 2K_2)\big(\big\|q - \widetilde{q}\big\| + \big\|h - \widetilde{h}\big\|\big) \tag{3.91}$$

In a very similar fashion, we can find a constant $L'(K_1, K_2, 1/\delta)$ such that the second term on the left handside of (3.88) satisfies

$$\begin{aligned}\left\| \int_\Xi \xi(\phi(U, \xi) - \phi(\widetilde{U}, \xi))\,\mathrm{d}\xi \right\| &\leq \alpha\big\|q - \widetilde{q}\big\| \\ &\quad + 2\sigma L'\rho(1 + 2K_2)\big(\big\|q - \widetilde{q}\big\| + \big\|h - \widetilde{h}\big\|\big)\end{aligned} \tag{3.92}$$

Indeed we can just chose $L' = (K_2 + \sqrt{2gK_1})L(K_1, K_2, 1/\delta)$ for (3.92) to be true, since we have the inclusion from remark 3.4.12. Summing the right handside of (3.91) and (3.92), we find the following upper bound for the left handside of (3.88):

$$\Big(\alpha + 2\sigma(L + L')\rho(1 + 2K_2)\Big)\big(\big\|h - \widetilde{h}\big\| + \big\|q - \widetilde{q}\big\|\big)$$

Hence the contraction property (3.88) is granted provided we have

$$\frac{\Delta t}{\Delta x} < \frac{1}{2(L + L')\rho(1 + 2K_2)}\ .$$

This is precisely the CFL condition we were looking for.

3.C Assembling matrices with Numpy

We make use of the numpy library.

```
1 import numpy as np
```

The implementation is fully vectorized, meaning no for loops are used. The matrix $(\mathcal{A}h)$ is assembled using the recursive definition (3.39) as below.

```
def AssembleMatrix_Ah(X):
    # Input: array X
    N = X.shape[0]
    Y = X/(1+X)

    Ah = np.zeros((N, N))
    #Extract upper triangular indices
    I, J = np.triu_indices(N, 0)
    Ah[I,J] = np.log(np.abs(1 + X))[J]

    S = np.zeros((N, N))
    #Extract upper triangular indices, excluding the diagonal
    I, J = np.triu_indices(N, 1)
    S[I,J] = (Y[J]**(J-I))/(J-I)
    #Cumulative sum along every column (upward direction)
    Ah[I,J] -= np.cumsum(S[::-1,:], axis=0)[N-I-1,J]

    return Ah
```

Likewise the matrix $(\mathcal{A}hu)$ is assembled using the recursive definition (3.41).

```
def AssembleMatrix_Ahu(X):
    # Input: array X
    N = X.shape[0]
    Y = X/(1+X)

    #Extract upper triangular indices
    I, J = np.triu_indices(N, 0)
    K = np.zeros((N, N))
    K[I,J] = X[J] - (J-I+1)*np.log(np.abs(1+X))[J]

    #Extract upper triangular indices, excluding the diagonal
    I, J = np.triu_indices(N, 1)
    UA = np.zeros((N, N))
    UA[I,J] = Y[J]**(J-I)
    UA = np.cumsum(UA[::-1,:], axis=0)[::-1,:]

    VA = np.zeros((N, N))
    VA[I,J] = (Y[J]**(J-I))/(J-I)
    VA = np.cumsum(VA[::-1,:], axis=0)[::-1,:]
    VA[I,J] *= J-I+1 #Multiply by l(j,i)

    Ahu = VA - UA + K
    return Ahu
```

For $(\mathcal{B}h)$ and $(\mathcal{B}hu)$ we use respectively (3.40) and (3.42).

```
def AssembleMatrix_Bh(X):
    # Input: array X
    N = X.shape[0]
    Y = X/(1+X)

    Bh = np.zeros((N, N))
```

```
    #Extract lower triangular indices
    I, J = np.tril_indices(N, 0)
    Bh[I,J] = np.log(np.abs(1 + X))[J]

    S = np.zeros((N, N))
    #Extract upper triangular indices, excluding the diagonal
    I, J = np.tril_indices(N, -1)
    S[I,J] = (Y[J]**(I-J))/(I-J)
    #Cumulative sum along every column (downward direction)
    Bh[I,J] -= np.cumsum(S, axis=0)[I,J]

    return Bh
```

```
def AssembleMatrix_Bhu(X):
    # Input: array X
    N = X.shape[0]
    Y = X/(1+X)

    #Extract lower triangular indices
    I, J = np.tril_indices(N, 0)
    K = np.zeros((N, N))
    K[I,J] = X[J] - (I-J+1)*np.log(np.abs(1+X))[J]

    #Extract upper triangular indices, excluding the diagonal
    I, J = np.tril_indices(N, -1)
    UB = np.zeros((N, N))
    UB[I,J] = Y[J]**(I-J)
    UB = np.cumsum(UB, axis=0)

    VB = np.zeros((N, N))
    VB[I,J] = (Y[J]**(I-J))/(I-J)
    VB = np.cumsum(VB, axis=0)
    VB[I,J] *= (I-J+1) #Multiply by l(i,j)

    Bhu = VB - UB + K
    return Bhu
```

The full code can be found at the following address:

`https://gitlab.com/mrigal/swimpy-1d`

Bibliography

[1] Sebastien Allgeyer et al. "Numerical approximation of the 3d hydrostatic Navier-Stokes system with free surface". In: *Mathematical Modelling and Numerical Analysis* 53 (2019), pp. 1981–2024.

[2] K.R. Arun, A.J. Das Gupta, and S. Samantaray. "Analysis of an asymptotic preserving low mach number accurate IMEX-RK scheme for the wave equation system". In: *Applied Mathematics and Computation* 411 (2021), p. 126469. ISSN: 0096-3003. DOI: https://doi.org/10.1016/j.amc.2021.126469. URL: https://www.sciencedirect.com/science/article/pii/S0096300321005580.

[3] K.R. Arun and S. Samantaray. "Asymptotic Preserving Low Mach Number Accurate IMEX Finite Volume Schemes for the Isentropic Euler Equations". In: *J Sci Comput* 82 (2020). DOI: https://doi.org/10.1007/s10915-020-01138-8.

[4] 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.

[5] Emmanuel Audusse, Marie-Odile Bristeau, and Benoît Perthame. "Kinetic Schemes for Saint-Venant Equations with Source Terms on Unstructured Grids". In: *[Research Report] RR-3989, INRIA* 53 (2000), pp. 1981–2024.

[6] Emmanuel Audusse et al. "A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows." In: *SIAM J. Sci. Comput.* 25 (2004), pp. 2050–2065.

[7] Emmanuel Audusse et al. "A multilayer Saint-Venant system with mass exchanges for Shallow Water flows. Derivation and numerical validation". In: *ESAIM: Mathematical Modelling and Numerical Analysis* 45 (Jan. 2009), pp. 169–200. DOI: 10.1051/m2an/2010036. URL: https://hal.archives-ouvertes.fr/hal-00355730.

[8] Emmanuel Audusse et al. "Analysis of modified Godunov type schemes for the two-dimensional linear wave equation with Coriolis source term on cartesian meshes". In: *Journal of Computational Physics* 373 (2018), pp. 91–129. ISSN: 0021-9991. DOI: https://doi.org/10.1016/j.jcp.2018.05.015. URL: https://www.sciencedirect.com/science/article/pii/S0021999118303073.

[9] Emmanuel Audusse et al. “Godunov type scheme for the linear wave equation with Coriolis source term”. In: LMLFN 2015 – Low Velocity Flows – Application to Low Mach and Low Froude regimes 58 (Nov. 2017). DOI: 10.1051/proc/201758001. URL: https://hal.archives-ouvertes.fr/hal-01254888.

[10] Emmanuel Audusse et al. “Kinetic entropy inequality and hydrostatic reconstruction scheme for the saint-venant system.” In: *Mathematics of Computation* 85 (2016), pp. 2815–2837.

[11] Wasilij Barsukow. “Low Mach number finite volume methods for the acoustic and Euler equations”. PhD thesis. Universität Würzburg, 2018.

[12] Christophe Berthon, Christian Klingenberg, and Markus Zenk. “An all Mach number relaxation upwind scheme”. en. In: *The SMAI journal of computational mathematics* 6 (2020), pp. 1–31. DOI: 10.5802/smai-jcm.60. URL: https://smai-jcm.centre-mersenne.org/articles/10.5802/smai-jcm.60/.

[13] P. L. Bhatnagar, E. P. Gross, and M. Krook. “A Model for Collision Processes in Gases. I. Small Amplitude Processes in Charged and Neutral One-Component Systems”. In: *Phys. Rev.* 94 (3 May 1954), pp. 511–525. DOI: 10.1103/PhysRev.94.511. URL: https://link.aps.org/doi/10.1103/PhysRev.94.511.

[14] Georgij Bispen et al. “IMEX Large Time Step Finite Volume Methods for Low Froude Number Shallow Water Flows.” In: *Communications in Computational Physics* 16 (2014), pp. 307–347. DOI: 10.4208/cicp.040413.160114a.

[15] S. Boscarino and L. Pareschi. “On the asymptotic properties of IMEX Runge-Kutta schemes for hyperbolic balance laws”. In: *Journal of Computational and Applied Mathematics* 316 (2017), pp. 60–73.

[16] S. Boscarino, L. Pareschi, and G. Russo. “Implicit-explicit Runge-Kutta schemes for hyperbolic systems and kinetic equations in the diffusion limit”. In: *SIAM J. Sci. Comput.* 35 (2013), A22–A51.

[17] Sebastiano Boscarino. *Implicit-explicit (IMEX) methods for evolutionary partial differential equations: A short Course on IMEX methods.* 2021.

[18] F. Bouchut. “Entropy satisfying flux vector splittings and kinetic BGK models”. In: *Numer. Math.* 94 (2003), pp. 623–672.

[19] F. Bouchut and X. Lhébrard. “Convergence of the the kinetic hydrostatic reconstruction scheme for the Saint Venant system with topography”. working paper or preprint. Apr. 2017. URL: https://hal-upec-upem.archives-ouvertes.fr/hal-01515256.

[20] François Bouchut. “Construction of BGK Models with a Family of Kinetic Entropies for a Given System of Conservation Laws.” In: *Journal of Statistical Physics* 95 (1999), pp. 113–170. DOI: 10.1023/A:1004525427365.

[21] François Bouchut, Christophe Chalons, and Sébastien Guisset. “An entropy satisfying two-speed relaxation system for the barotropic Euler equations. Application to the numerical approximation of low Mach number flows.” In: *Numerische Mathematik* 145 (2020), pp. 35–76. DOI: 10.1007/s00211-020-01111-5. URL: https://hal.archives-ouvertes.fr/hal-01661275.

[22] François Bouchut, Emmanuel Franck, and Laurent Navoret. "A low cost semi-implicit low-Mach relaxation scheme for the full Euler equations". In: *Journal of Scientific Computing* 83.1 (Apr. 2020), p. 24. DOI: 10.1007/s10915-020-01206-z. URL: https://hal.archives-ouvertes.fr/hal-02420859.

[23] M.-O. Bristeau, N. Goutal, and J. Sainte-Marie. "Numerical simulations of a non-hydrostatic Shallow Water model". In: *Computers & Fluids* 47.1 (2011), pp. 51–64. DOI: 10.1016/j.compfluid.2011.02.013.

[24] Marie-Odile Bristeau and Benoit Coussin. *Boundary Conditions for the Shallow Water Equations solved by Kinetic Schemes*. Research Report RR-4282. Projet M3N. INRIA, 2001. URL: https://hal.inria.fr/inria-00072305.

[25] Christophe Chalons et al. "A large time-step and well-balanced Lagrange-Projection type scheme for the shallow-water equations". working paper or preprint. July 2016. URL: https://hal.archives-ouvertes.fr/hal-01297043.

[26] Olivier Delestre et al. "SWASHES: a compilation of Shallow Water Analytic Solutions for Hydraulic and Environmental Studies". In: *International Journal for Numerical Methods in Fluids* 72.3 (May 2013). 40 pages There are some errors in the published version. This is a corrected version., pp. 269–300. DOI: 10.1002/fld.3741. URL: https://hal.archives-ouvertes.fr/hal-00628246.

[27] S. Dellacherie. "Checkerboard modes and wave equation". In: *Proceedings of the 18th Conference on Scientific Computing, Podbanske, Slovakia* 53 (2009), pp. 71–80.

[28] Stéphane Dellacherie. "Analysis of Godunov type schemes applied to the compressible Euler system at low Mach number." In: *Journal of Computational Physics* 229 (2010), pp. 978–1016. DOI: 10.1016/j.jcp.2009.09.044.

[29] Giacomo Dimarco et al. "Second order Implicit-Explicit Total Variation Diminishing schemes for the Euler system in the low Mach regime". In: *Journal of Computational Physics* 372 (Nov. 2018), pp. 178–201. DOI: 10.1016/j.jcp.2018.06.022. URL: https://hal.archives-ouvertes.fr/hal-01620627.

[30] Arnaud DURAN. "Numerical simulation of depth-averaged flow models : a class of Finite Volume and discontinuous Galerkin approaches." Theses. Université Montpellier II, Oct. 2014. URL: https://tel.archives-ouvertes.fr/tel-01109438.

[31] Enrique D. Fernández-Nieto et al. "Formal deduction of the Saint-Venant–Exner model including arbitrarily sloping sediment beds and associated energy". In: *ESAIM: Mathematical Modelling and Numerical Analysis* 51.1 (Nov. 2016), pp. 115–145. DOI: 10.1051/m2an/2016018. URL: https://doi.org/10.1051%2Fm2an%2F2016018.

[32] Isabelle Gallagher, Laure Saint-Raymond, and Benjamin Texier. *From Newton to Boltzmann: hard spheres and short-range potentials*. 2012. DOI: 10.48550/ARXIV.1208.5753. URL: https://arxiv.org/abs/1208.5753.

[33] Jean-Frédéric Gerbeau and Benoît Perthame. "Derivation of Viscous Saint-Venant System for Laminar Shallow Water; Numerical Validation". In: *Rapport de recherche Inria* (2000).

[34] F. X. Giraldo and M. Restelli. "High-order semi-implicit time-integration of a triangular discontinuous Galerkin oceanic shallow water model". In: *International journal for numerical methods in fluids* 63 (2010), pp. 1077–1102.

[35] Mathieu Girardin. "Asymptotic preserving and all-regime Lagrange-Projection like numerical schemes : application to two-phase flows in low mach regime". Theses. Université Pierre et Marie Curie - Paris VI, Dec. 2014. URL: https://tel.archives-ouvertes.fr/tel-01127428.

[36] François Golse, C. David Levermore, and Laure Saint-Raymond. "La méthode de l'entropie relative pour les limites hydrodynamiques de modèles cinétiques". fr. In: *Séminaire Équations aux dérivées partielles (Polytechnique) dit aussi "Séminaire Goulaouic-Schwartz"* (1999-2000). talk:18. URL: http://www.numdam.org/item/SEDP_1999-2000____A18_0/.

[37] N. Goutal and J. Sainte-Marie. "A kinetic interpretation of the section-averaged Saint-Venant system for natural river hydraulics". In: *Int. J. Numer. Meth. Fluids* 67.7 (2011), pp. 914–938. DOI: 10.1002/fld.2401.

[38] Hervé Guillard and Cécile Viozat. "On the behaviour of upwind schemes in the low Mach number limit". In: *Computers & Fluids* 28.1 (1999), pp. 63–86. ISSN: 0045-7930. DOI: https://doi.org/10.1016/S0045-7930(98)00017-6. URL: https://www.sciencedirect.com/science/article/pii/S0045793098000176.

[39] J. Haack, S. Jin, and J.-G. Liu. "An all-speed asymptotic-preserving method for the isentropic Euler and Navier-Stokes equations". In: *Communications in Computational Physics* 12(4) (Sept. 2012), pp. 955–980. DOI: 10.4208/cicp.250910.131011a.

[40] Amiram Harten, Peter D. Lax, and Bram van Leer. "On Upstream Differencing and Godunov-Type Schemes for Hyperbolic Conservation Laws". In: *Upwind and High-Resolution Schemes*. Ed. by M. Yousuff Hussaini, Bram van Leer, and John Van Rosendale. Berlin, Heidelberg: Springer Berlin Heidelberg, 1997, pp. 53–79. ISBN: 978-3-642-60543-7. DOI: 10.1007/978-3-642-60543-7_4. URL: https://doi.org/10.1007/978-3-642-60543-7_4.

[41] Rolf Jeltsch and Manuel Torrilhon. "On curl-preserving finite volume discretizations for shallow water equations". In: *Journal of Mathematics of Kyoto University* 46 (2006), pp. 35–53. DOI: 10.1007/s10543-006-0089-5.

[42] Shi Jin. "Asymptotic preserving (AP) schemes for multiscale kinetic and hyperbolic equations: a review". In: *Rivista di Matematica della Università di Parma. New Series* 2 (Jan. 2010).

[43] Shi Jin. "Efficient Asymptotic-Preserving (AP) schemes for some multiscale kinetic equations". In: *SIAM J. Sci. Comput.* 21 (1999), pp. 441–454. ISSN: 1095-7197.

[44] David I. Ketcheson, Randall J. LeVeque, and Mauricio J. del Razo. *Riemann Problems and Jupyter Solutions*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2020. DOI: 10.1137/1.9781611976212. eprint: https://epubs.siam.org/doi/pdf/10.1137/1.9781611976212. URL: https://epubs.siam.org/doi/abs/10.1137/1.9781611976212.

[45] Sergiu Klainerman and Andrew Majda. “Compressible and Incompressible Fluids”. In: *Communications on Pure and Applied Mathematics* 35 (1982), pp. 629–651. ISSN: 0010-3640.

[46] R. Klein. “Semi-implicit extension of a godunov-type scheme based on low mach number asymptotics I: One-dimensional flow”. In: *Journal of Computational Physics* 121.2 (1995), pp. 213–237. ISSN: 0021-9991. DOI: https://doi.org/10.1016/S0021-9991(95)90034-9. URL: https://www.sciencedirect.com/science/article/pii/S0021999195900349.

[47] Christian Klingenberg et al. “An All Speed Second Order IMEX Relaxation Scheme for the Euler Equations”. In: *Communications in Computational Physics* 28.2 (June 2020), pp. 591–620. DOI: 10.4208/cicp.oa-2019-0123. URL: https://doi.org/10.4208%2Fcicp.oa-2019-0123.

[48] D. Lannes. *From the swell to the beach: modelling shallow water waves.*

[49] Bram van Leer. “Flux-vector Splitting for the Euler Equation”. In: Jan. 1997, pp. 80–89. ISBN: 978-3-642-60543-7. DOI: 10.1007/978-3-642-60543-7_5.

[50] P.G. Lefloch. “Hyperbolic Systems of Conservation Laws, The Theory of Classical and Nonclassical Shock Waves”. In: Birkhäuser Verlag, 2002.

[51] P.L. Lions, B. Perthame, and E. Tadmor. “Kinetic Formulation of the Isentropic Gas Dynamics and p-Systems”. In: *Communications in Mathematical Physics* 163 (1994), pp. 415–431.

[52] Fabien Marche. “Derivation of a new two-dimensional viscous shallow water model with varying topography, bottom friction and capillary effects”. In: *European journal of Mechanics B/Fluids* 26 (2007), pp. 49–63.

[53] Victor Michel-Dansac and Andrea Thomann. “Large time step TVD IMEX Runge-Kutta schemes based on arbitrarily high order Butcher tableaux”. working paper or preprint. Oct. 2020. URL: https://hal.archives-ouvertes.fr/hal-02494767.

[54] Sebastian Noelle et al. *A Weakly Asymptotic Preserving Low Mach Number Scheme for the Euler Equations of Gas Dynamics*. 2014. DOI: 10.48550/ARXIV.1412.1606. URL: https://arxiv.org/abs/1412.1606.

[55] B. Perthame. *Kinetic formulation of conservation laws*. Oxford University Press, 2002.

[56] Benoit Perthame. “An Introduction to Kinetic Schemes for Gas Dynamics”. In: *An Introduction to Recent Developments in Theory and Numerics for Conservation Laws: Proceedings of the International School on Theory and Numerics for Conservation Laws, Freiburg/Littenweiler, October 20–24, 1997*. Ed. by Dietmar Kröner, Mario Ohlberger, and Christian Rohde. Berlin, Heidelberg: Springer Berlin Heidelberg, 1999, pp. 1–27. ISBN: 978-3-642-58535-7. DOI: 10.1007/978-3-642-58535-7_1. URL: https://doi.org/10.1007/978-3-642-58535-7_1.

[57] Benoît Perthame. “Boltzmann type schemes for gas dynamics and the entropy property”. In: *SIAM Journal on Numerical Analysis* 27 (1990), pp. 1405–1421.

[58] Benoît Perthame and Chiara Simeoni. "A kinetic scheme for the Saint-Venant system with a source term." In: *Calcolo* 38 (2001), pp. 201–231. DOI: 10.1007/s10092-001-8181-3.

[59] Panagiotis E. Souganidis Pierre-Louis Lions Benoît Perthame. "Existence and stability of entropy solutions for the hyperbolic systems of isentropic gas dynamics in Eulerian and Lagrangian coordinates". In: *Communications on Pure and Applied Mathematics* XLIX (1996), pp. 599–638.

[60] S. K. Godunov. "A difference method for numerical calculation of discontinuous solutions of the equations of hydrodynamics". In: *Mat. Sb. (N.S.)* 47(89) (1959), pp. 271–306.

[61] 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 lits". In: *C. R. Acad. Sci., Paris* (73 (1871)), pp. 147–154.

[62] Laure Saint-Raymond. *Hydrodynamic Limits of the Boltzmann Equation*. Springer, 2009.

[63] S. Schochet. "Fast singular limits of hyperbolic PDEs." In: *J. Differ. Eqs.* 114 (1994), pp. 476–512.

[64] D. Serre. "Systèmes hyperboliques de lois de conservation, Parties I et II". In: Paris, 1996.

[65] Andrea Thomann, Gabriella Puppo, and Christian Klingenberg. "An all speed second order well-balanced IMEX relaxation scheme for the Euler equations with gravity". In: *Journal of Computational Physics* 420 (2020), p. 109723. ISSN: 0021-9991. DOI: https://doi.org/10.1016/j.jcp.2020.109723. URL: https://www.sciencedirect.com/science/article/pii/S0021999120304976.

[66] E.F. Toro, M. Spruce, and W. Speares. "Restoration of the contact surface in the HLL-Riemann solver". In: *Shock Waves* 4 (1994), pp. 25–34. DOI: https://doi.org/10.1007/BF01414629.

[67] Seiji Ukai. "The incompressible limit and the initial layer of the compressible Euler equation". In: *Journal of Mathematics of Kyoto University* 26.2 (1986), pp. 323–331. DOI: 10.1215/kjm/1250520925. URL: https://doi.org/10.1215/kjm/1250520925.

[68] R. F. Warming and B. J. Hyett. "The modified equation approach to the stability and accuracy analysis of finite-difference methods". In: *Journal of Computational Physics* 14 (1974), pp. 159–179.

Résumé : Dans cette thèse nous étudions des discrétisations en temps implicites pour le système de Saint-Venant. Premièrement nous considérons la question du régime bas Froude dans le cas bidimensionnel. L'aptitude à transitionner vers le régime limite de manière transparente pose principalement deux problèmes, à savoir le coût de calcul associé à la gestion des échelles rapides et la bonne description de la dynamique asymptotique. Le premier point est traditionnellement traité par l'utilisation d'intégrateurs en temps implicite-explicite, tandis que le second nécessite d'avoir une erreur numérique uniformément bornée par rapport au paramètre d'échelle. En particulier, il est important pour les états quasi incompressibles de satisfaire une certaine forme de stabilité. Ceci motive le raffinement d'un critère existant permettant de prédire si un schéma est précis à bas nombre de Froude, ce que nous validons par l'intermédiaire d'exemples numériques. De plus les schémas semi-implicites proposés sont basés sur un splitting d'onde propice à la préservation de l'équilibre hydrostatique.

Nous nous concentrons ensuite sur des schémas cinétiques pour le système de Saint-Venant unidimensionnel. Dans le cas d'une bathymétrie plate, nous obtenons un schéma entièrement implicite préservant la positivité de la hauteur d'eau et admettant une inégalité d'entropie discrète sans aucune restriction sur le pas de temps. Une version simplifiée de ce schéma permet de réécrire explicitement la mise-à-jour au niveau macroscopique. Afin de prendre en compte les fonds variables, nous examinons une stratégie itérative faisant appel à la reconstruction hydrostatique. Cette approche requiert une condition CFL pour converger, en échange de quoi nous obtenons une mise-à-jour positive avec une inégalité d'entropie discrète qui dissipe toujours l'énergie du système. Ceci est une amélioration par rapport à la version entièrement explicite du schéma, qui peut parfois accroître l'énergie. Nous effectuons des tests numériques pour évaluer l'efficacité et les aspects qualitatifs des schémas proposés.

Mots-clés : équations de Saint-Venant, méthodes semi-implicites, schémas cinétiques, régime bas Froude, méthodes préservant l'asymptotique, volumes finis, inégalité d'entropie.