Implicit kinetic schemes for the Saint-Venant system

Chourouk El Hassanieh, Mathieu Rigal, Jacques Sainte-Marie

Implicit kinetic schemes for the Saint-Venant system

Chourouk El Hassanieh[1], Mathieu Rigal[2], and Jacques Sainte-Marie[1]

[1]Centre Inria de Sorbonne université, 2 rue Simone Iff, CS 42112, 75589 Paris Cedex 12 and Sorbonne Université, Univ. Paris Diderot, SPC, CNRS, LJLL, F-75005 Paris
[2]Institut de Mathématiques de Bordeaux, équipe EDP et physique-mathématique, Université de Bordeaux, 351 cours de la Libération, F-33405 Talence, France

Emails: (chourouk.el-hassanieh, jacques.sainte-marie)@inria.fr,
mathieu.rigal@math.u-bordeaux.fr

March 28, 2024

Abstract

Explicit (in time) kinetic schemes applied to the nonlinear shallow water equations have been extensively studied in the past. The novelty of this paper is to investigate an implicit version of such methods in order to improve their stability properties. In the case of a flat bathymetry we obtain a fully implicit kinetic solver satisfying a discrete entropy inequality and keeping the water height non negative without any restriction on the time step. Remarkably, a simplified version of this nonlinear implicit scheme allows to express the update explicitly which we implement in practice. The case of varying bottoms is then dealt with through an iterative solver combined with the hydrostatic reconstruction technique. We show that this scheme preserves the water height non-negativity under a CFL condition and satisfies a discrete entropy inequality without error term, which is an improvement over its explicit version. An extension of the implicit and iterative methods to the two dimensional case is also discussed. Finally we perform some numerical validations underlining the advantages and the computational cost of our strategy.

This work has been supported by a DIM MathInnov funding from Région Île-de-France.

AMS classification. 65M12, 74S10, 76M12, 35L65.
Keywords. Shallow water equations, Implicit and iterative kinetic solvers, Fully discrete entropy inequality, Well-balanced schemes, Hydrostatic reconstruction.

Contents

1 Introduction

1.1 General motivation and goals

Free surface flows are a subject of great importance whose understanding is at the core of many present challenges. These include the issue of safety related to submersion and tsunami waves affecting urbanized areas, rogue waves that can be a hazard for sailing, or the transport in rivers or lakes. Another relevant topic is that of renewable energy generated from hydraulic dams or sea buoys, and the impact of these systems on the aquatic fauna to give only a few examples. To describe such complex fluid flows, there exists numerous mathematical models, and when restricting to shallow areas where it is reasonable to neglect dispersive effects but not nonlinear ones the Saint-Venant equations [27, 18], also known as the shallow water equations, are often preferred to the incompressible Navier-Stokes system with free surface for practical applications. In fact, although the latter model is more accurate, it is also significantly more complex to study theoretically and numerically. On the other hand the Saint-Venant system has a reduced complexity since it is a vertically averaged nonlinear model belonging to the class of hyperbolic systems of balance laws. Despite being simpler, it can be successfully used to approximate various geophysical flows such as rivers, coastal areas, and oceans when completed with a Coriolis term, and granular flows when completed with friction terms.

The analytical solutions of the Saint-Venant equations satisfy important properties, among which the conservation of the total water and of the total momentum when the bathymetry is flat, the positivity of the water height, or the existence of the lake at rest steady state that becomes a nontrivial equilibrium in presence of a varying bathymetry, which is accounted for through a source term in the

momentum equation. Because the solutions of the Saint-Venant system are non-unique, it is usual to select the one obtained through a viscosity perturbation, and which satisfies additional entropy inequalities that can be interpreted as the dissipation of some energy. For the derivation of a numerical scheme approximating the Saint-Venant system, the main difficulty is to construct a structure preserving numerical method, that is to say a method able to preserve the aforementioned properties at the discrete level so that the numerical approximation is qualitatively relevant. The present paper is dedicated to this issue. More specifically we focus on a kinetic solver, which provides a good framework regarding positivity and entropy stability, and investigate the advantages of combining it with an implicit time integrator compared to the more common explicit approach.

There are several contributions in this work. In the case of the one dimensional Saint-Venant system with flat bathymetry, we propose a new, fully implicit kinetic scheme which is structure preserving without constraint on the time step. The case of varying bathymetries is dealt with by the mean of the hydrostatic reconstruction technique together with a fixed point method approximating the implicit update. Unlike its explicit version which does not always dissipate the entropy, our iterative scheme is shown to be structure preserving under a CFL condition. We then point to the possibility of extending these two numerical methods to the two dimensional Saint-Venant equations, and numerical simulations are performed to evaluate the interest of the proposed approaches.

1.2 State of the art and outline of the paper

The derivation of an efficient, robust and stable numerical scheme for the Saint-Venant system has received an extensive coverage, we refer the reader to [10, 22, 19, 29] and references therein. In particular, kinetic schemes approximating macroscopic systems of conservation laws have become popular over the last decades. Such methods consist to substitute the macroscopic system, in our case the Saint-Venant equations, by a scalar kinetic equation featuring a transport term with a collision operator, and which is easier to discretize. In [9], Bouchut developed a general theory to construct kinetic representations involving a BGK collision operator and admitting a family of kinetic entropies. As we will recall later, kinetic entropies are an important tool allowing one to recover the dissipation mechanism of the macroscopic entropy at the kinetic level. Once a kinetic representation with a kinetic entropy is known, it is then possible to design a numerical scheme satisfying a discrete entropy inequality as explained in [12].

In presence of a source term, such as the one induced by a varying bathymetry in the Saint-Venant equations, one of the challenges involves the design of a well-balanced scheme, that is to say a discretization able to preserve some characteristic stationary solutions. In [24], Perthame and Simeoni proposed a well-balanced kinetic scheme where the bathymetry is seen as a stair-shaped parametrization against which fluid particles can be reflected. A drawback of this is that the associated macroscopic update cannot be written explicitly, and instead some integrals have to be approximated by a quadrature. They also provided a Maxwellian with compact support that differs from the usual Gaussian distribution, and that has been used later in [20, 14, 4]. This allows the scheme to be positive and entropy stable under a CFL condition depending on the size of the support. In a recent work [4], some of the authors have proposed a kinetic scheme coupled with the hydrostatic reconstruction technique introduced in [3, 6] for the numerical treatment of the topography source term. With this approach an explicit writing of the macroscopic update is available, however an error term with positive sign is present in the discrete entropy inequality; as a result the entropy can increase in some test cases. Despite this, Bouchut and Lhebrard have proved the convergence of the kinetic hydrostatic reconstruction scheme towards entropy solutions of the Saint-Venant system, see [11]. The implicit version of the scheme proposed in [4] gives a stronger stability result where the discrete entropy inequality is free of the positive error term and the entropy is always dissipated.

The capability of implicit time integrators to improve the stability properties of numerical schemes has been studied in the last years. Explicit finite volume approaches for the approximation of conservation laws have to deal with a CFL constraint that can be very restrictive for some applications where large time scales and significant wave velocities have to be considered. This is for instance the case in the low Froude regime, where the surface gravity waves travel at a much larger velocity than the fluid particles. A solution consists in implicit-explicit schemes [21, 8, 28, 2] where the fast dynamics are approximated implicitly, allowing the use of larger time steps while keeping the scheme stable.

Another important question is the computational costs, in the sense that for explicit schemes, the CFL constraint implies small time steps whereas for implicit schemes, the computation of the numerical fluxes can be costly. Indeed in the explicit setting, the numerical fluxes at an interface depend on the value of the variables at the two neighboring cells whereas in the implicit context, the numerical fluxes depend on the value of the variables of all the cells. In [16], a fully implicit lattice Boltzmann scheme was proposed, with the advantage of having the computational cost of an explicit approach thanks to a sweeping algorithm. However this method requires the boundary conditions to be non periodic, and no discrete entropy inequality was provided. The novelties of the paper are:

- a fully implicit numerical scheme for the Saint-Venant system is detailed and analysed ;
- the proposed numerical scheme satisfies a fully discrete entropy inequality ;
- whereas, in general, an implicit scheme often requires to invert an operator – typically a matrix – at each time step, our fully implicit kinetic scheme offers a very favorable context since we have an explicit expression of the inverse of the operator. Hence, one can hardly imagine a truly implicit entropy stable scheme for the Saint-Venant system with a lower computational cost than a kinetic solver proposed here.

This paper is organized as follows. In the remainder of this section we recall the formulation of the Saint-Venant system, its kinetic description and the framework of its numerical approximation in the context of a kinetic solver. Especially, the notion of kinetic entropy is detailed. Then in Section 2, the implicit kinetic scheme for the Saint-Venant system with flat topography is proposed and studied in the one dimensional case. An iterative kinetic scheme is proposed in Section 3 where the topography can be taken into account through the hydrostatic reconstruction technique. In Section 4, we briefly discuss the extension of the previous approaches to the two dimensional Saint-Venant system. Finally in Section 5, numerical examples are given to evaluate the interest of the proposed schemes.

1.3 The Saint-Venant system and its kinetic interpretation

The classical Saint-Venant system for shallow water flows describes the evolution of the height of water $h(t,x) \geq 0$, and the water velocity $u(t,x) \in \mathbb{R}$ (x denotes a coordinate in the horizontal direction) over a slowly varying topography $z(x)$, and reads

$$\begin{aligned} &\partial_t h + \partial_x(hu) = 0, \\ &\partial_t(hu) + \partial_x(hu^2 + g\frac{h^2}{2}) + gh\partial_x z = 0, \end{aligned} \tag{1.1}$$

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

$$\partial_t\Big(h\frac{u^2}{2} + g\frac{h^2}{2} + ghz\Big) + \partial_x\Big(\big(h\frac{u^2}{2} + gh^2 + ghz\big)u\Big) \leq 0. \tag{1.2}$$

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

$$\eta(U) = h\frac{u^2}{2} + g\frac{h^2}{2}, \qquad G(U) = \big(h\frac{u^2}{2} + gh^2\big)u, \tag{1.3}$$

the entropy and entropy fluxes without topography. The reader can refer to [4] and references therein for a complete presentation of the description of the Saint-Venant system.

The classical kinetic Maxwellian (see e.g. [24]) is given by

$$M(U,\xi) = \frac{1}{g\pi}\Big(2gh - (\xi - u)^2\Big)_+^{1/2}, \tag{1.4}$$

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

$$\begin{aligned} \int_{\mathbb{R}} M(U,\xi)\,d\xi = h, \qquad \int_{\mathbb{R}} \xi M(U,\xi)\,d\xi = hu, \\ \int_{\mathbb{R}} \xi^2 M(U,\xi)\,d\xi = hu^2 + g\frac{h^2}{2}. \end{aligned} \tag{1.5}$$

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

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

$$\partial_t M + \xi \partial_x M - g(\partial_x z)\partial_\xi M = Q, \tag{1.6}$$

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

$$\int_{\mathbb{R}} Q d\xi = \int_{\mathbb{R}} \xi Q d\xi = 0. \tag{1.7}$$

Proof. If (1.6) and (1.7) are satisfied, we can multiply (1.6) by $(1,\xi)^T$, and integrate with respect to ξ. Using (1.5) and (1.7) and integrating by parts the term in $\partial_\xi M$, we obtain (1.1). Conversely, if (h, hu) is a weak solution to (1.1), just define Q by (1.6); it will satisfy (1.7) according to the same computations. □

The standard way to use Lemma 1.1 is to write a kinetic relaxation equation [23, 26, 15, 9, 12], like

$$\partial_t f + \xi \partial_x f - g(\partial_x z)\partial_\xi f = \frac{M - f}{\epsilon}, \tag{1.8}$$

where the distribution $f(t,x,\xi)$ is positive, where $M = M(U,\xi)$ with $U(t,x) = \int (1,\xi)^T f(t,x,\xi) d\xi$, and where $\epsilon > 0$ is a relaxation time. In the limit $\epsilon \to 0$ we recover formally the formulation (1.6), (1.7). We refer to [9] for general considerations on such kinetic relaxation models without topography, the case with topography being introduced in [24]. Note that the notion of *kinetic representation* as (1.6), (1.7) differs from the so called *kinetic formulations* where a large set of entropies is involved, see [25]. For systems of conservation laws, these kinetic formulations include non-advective terms that prevent from writing down simple approximations. In general, kinetic relaxation approximations can be compatible with just a single entropy. Nevertheless this is enough for proving the convergence as $\epsilon \to 0$, see [7].

The interest of the particular form (1.4) lies in its link with a kinetic entropy. Consider the kinetic entropy,

$$H(f,\xi,z) = \frac{\xi^2}{2} f + \frac{g^2\pi^2}{6} f^3 + gzf, \tag{1.9}$$

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

$$H_0(f,\xi) = \frac{\xi^2}{2} f + \frac{g^2\pi^2}{6} f^3. \tag{1.10}$$

Then one can check the relations

$$\int_{\mathbb{R}} H\big(M(U,\xi),\xi,z\big)\, d\xi = \eta(U) + ghz, \tag{1.11}$$

$$\int_{\mathbb{R}} \xi H\big(M(U,\xi),\xi,z\big)\, d\xi = G(U) + ghzu. \tag{1.12}$$

One has the following entropy relations.

Lemma 1.2 *Let $f(\xi) \geq 0$ satisfy $\int f(\xi) d\xi = h$ and $\int \xi f(\xi) d\xi = hu$ (assumed finite). The half-disk Maxwellian (1.4) satisfies the three properties below.*

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

$$H_0(f,\xi) \geq H_0(M(U,\xi),\xi) + \eta'(U) \cdot \begin{pmatrix} 1 \\ \xi \end{pmatrix} (f - M(U,\xi)). \tag{1.13}$$

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

$$\eta(U) = \int_{\mathbb{R}} H_0\big(M(U,\xi),\xi\big)\, d\xi \leq \int_{\mathbb{R}} H_0\big(f(\xi),\xi\big)\, d\xi. \tag{1.14}$$

(iii) There holds the kinetic entropy equality

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

and under the additional assumption $Q(t,x,\xi) \leq 0$ for all $\xi \notin \operatorname{supp} M(U(t,x),\cdot)$ we have the macroscopic inequality

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

which is exactly (1.2) owing to the relations (1.11)-(1.12).

Proof of Lemma 1.2. *(i)* A proof of inequality (1.13) is given in [4] and relies on the following relation

$$\partial_f H_0(M(U,\xi),\xi) = \begin{cases} \eta'(U)\cdot\begin{pmatrix}1\\ \xi\end{pmatrix} & \text{if } \xi \in \operatorname{supp} M(U,\cdot) \\ \geq \eta'(U)\cdot\begin{pmatrix}1\\ \xi\end{pmatrix} & \text{otherwise} \end{cases}, \tag{1.15}$$

which we will use in Section 3 to obtain discrete entropy inequalities. The property *(ii)* was proved by Perthame and Simeoni in [24]. It is simply recovered from *(i)* by integrating (1.13) over $\xi \in \mathbb{R}$ and using the fact that $f - M(U,\xi)$ is a collision term. In order to prove *(iii)*, we multiply (1.6) by $\partial_f H(M,\xi,z)$, we bound the right hand side using (1.15) yielding

$$\partial_f H(M,\xi) Q = (\partial_f H_0(M,\xi) + gz) Q \leq \Big(\eta'(U)\cdot\begin{pmatrix}1\\ \xi\end{pmatrix} + gz\Big) Q,$$

and owing to (1.7) an integration in ξ of the obtained inequality gives the result. □

1.4 Explicit kinetic schemes for the Saint-Venant system

We intend to approximate the solution $U(t,x)$ of the system (1.1) for $x \in \mathbb{R}$ and $t \geq 0$ by discrete values U_i^n, with $i \in \mathbb{Z}$ the spatial index and $n \in \mathbb{N}$ the time index. For this, we consider a grid of real points $(x_{i+1/2})_{i\in\mathbb{Z}}$ and discrete times $(t^n)_{n\in\mathbb{N}}$ satisfying

$$x_{i+1/2} - x_{i-1/2} = \Delta x_i, \qquad t^{n+1} - t^n = \Delta t^n,$$

with $\Delta x_i > 0$ the spatial step and $\Delta t^n > 0$ the time step. Defining the cell $C_i = (x_{i-1/2}, x_{i+1/2})$ of length Δx_i, we shall consider $U^n(x)$ the piecewise constant function approximating the solution at time t^n and corresponding to

$$U^n(x) = U_i^n, \quad \text{for } x \in C_i. \tag{1.16}$$

A finite volume scheme for (1.1) is a formula providing a relation between U_i^{n+1} and the set $\{U_j^n\}_{j\in\mathbb{Z}}$ of the form

$$U_i^{n+1} = U_i^n - \sigma_i(F_{i+1/2-} - F_{i-1/2+}), \tag{1.17}$$

where $\sigma_i = \Delta t^n/\Delta x_i$, and where $F_{i\pm1/2\mp}$ are two-points numerical fluxes given by

$$F_{i+1/2-} = \mathcal{F}_l(U_i^{n+p}, U_{i+1}^{n+p}, z_{i+1} - z_i), \qquad F_{i+1/2+} = \mathcal{F}_r(U_i^{n+p}, U_{i+1}^{n+p}, z_{i+1} - z_i), \tag{1.18}$$

with $p \in \{0,1\}$ and with $\mathcal{F}_l, \mathcal{F}_r$ some functions valued in $\mathbb{R}^2$ consistent with the analytic flux of the Saint-Venant system, see [10]. The value $p = 0$ classically corresponds to a first order explicit time scheme for solving (1.1) whereas $p = 1$ means an implicit time scheme. In this paper, we focus on the case $p = 1$, but we recall how to construct kinetic schemes in the classical explicit framework.

A kinetic scheme is a numerical discretization of the BGK-type kinetic equation (1.8) which then provides a numerical scheme for the Saint-Venant system using to the moment relations (1.5) after integration against $(1, \xi)$. In [24], a splitting between the transport term and the collision term is considered, requiring to solve the following problem at every time step

$$\begin{cases} \partial_t f + \xi \partial_x f - g(\partial_x z)\partial_\xi f = 0 \\ f(t^n, x, \xi) = M(U^n(x), \xi) \end{cases}, \qquad t \in (t^n, t^{n+1}). \tag{1.19}$$

Especially, we see that the initial kinetic density is projected onto the space of Maxwellians. Denoting f_i^{n+1-} an approximation of the solution of (1.19) at time t^{n+1} in cell C_i, the macroscopic update is given by

$$U_i^{n+1} = \int_{\mathbb{R}} \begin{pmatrix} 1 \\ \xi \end{pmatrix} f_i^{n+1-}(\xi)\, d\xi. \tag{1.20}$$

It remains to explain how to discretize (1.19). This is achieved by the mean of an upwind scheme accounting for the sign of the velocity ξ, and which for a flat bottom reads

$$f_i^{n+1-} = M_i - \sigma_i \xi \Big(\mathbb{1}_{\xi>0} M_i + \mathbb{1}_{\xi<0} M_{i+1} - \mathbb{1}_{\xi<0} M_i - \mathbb{1}_{\xi>0} M_{i-1} \Big). \tag{1.21}$$

In particular we see that (1.20)-(1.21) corresponds to the macroscopic update (1.17) with numerical fluxes (1.18) defined as

$$\mathcal{F}_{l/r}(U_l, U_r, \Delta z) = \int_{\mathbb{R}} \xi \begin{pmatrix} 1 \\ \xi \end{pmatrix} \left(\mathbb{1}_{\xi>0} M(U_l, \xi) + \mathbb{1}_{\xi<0} M(U_r, \xi) \right) \mathrm{d}\xi.$$

In [4], the scheme (1.21) was extended to varying bathymetries thanks to the hydrostatic reconstruction technique. The resulting discretization reads

$$\begin{aligned} f_i^{n+1-} = M_i &- \sigma_i \xi \Big(\mathbb{1}_{\xi<0}(M_{i+1/2+} - M_{i-1/2+}) + \mathbb{1}_{\xi>0}(M_{i+1/2-} - M_{i-1/2-}) \Big) \\ &+ \sigma_i (\xi - u_i)(M_{i+1/2-} - M_{i-1/2+}), \end{aligned} \tag{1.22}$$

where indices $i + 1/2\pm$ denote reconstructed values at the interface $i + 1/2$ and whose expression will be detailed later (see Section 3.2). In [4], some of the authors have proved the following properties for the schemes (1.21) and (1.22).

Proposition 1.3 (Explicit kinetic schemes [4]). *Under the CFL condition $\sigma \xi_{\max} \leq 1$, where $\xi_{\max} = \sup\{|\xi|,\ \xi \in \cup_{i\in\mathbb{Z}} \operatorname{supp} M(U_i^n, \cdot)\}$ is finite if $(U_i^n)_i \in \ell^\infty(\mathbb{Z}; \mathbb{R}^2)$, we have that*

1. *the scheme (1.21) keeps f_i^{n+1-} non-negative and satisfies a fully discrete entropy inequality of the form*

$$H_0(f_i^{n+1-}, \xi) \leq H_0(M_i^n, \xi) - \sigma_i \xi (H_{0,i+1/2}^n(\xi) - H_{0,i-1/2}^n(\xi)),$$

2. *the scheme (1.22) keeps f_i^{n+1-} non-negative and satisfies a fully discrete entropy inequality of the form*

$$H(f_i^{n+1-}, z_i, \xi) \leq H(M_i^n, z_i, \xi) - \sigma_i \xi (\widetilde{H}_{i+1/2}^n(\xi) - \widetilde{H}_{i-1/2}^n(\xi)) + D_i^n(\xi) + E_i^n(\xi),$$

where $D_i^n \leq 0$ is a dissipation and $E_i^n \geq 0$ an error term that can dominate D_i^n in many cases.

2 A fully implicit kinetic scheme over a flat bathymetry

In this section we consider the model (1.1) with a flat topography, and propose an implicit kinetic scheme based on (1.19)-(1.20) i.e. an implicit version of (1.21). Since in practice it is not possible to approximate solutions over the whole space $x \in \mathbb{R}$, we shall work in a bounded domain and will need to enforce boundary conditions. First in Section 2.1 we present and study the implicit kinetic scheme; we show that it is unconditionally positive and entropy stable by using the kinetic entropy (1.10). Next in Section 2.2 we detail how to compute explicitly the update at the kinetic level, which is possible thanks to the particular structure of the underlying matrix. Then we consider the associated macroscopic scheme in Section 2.3. Unfortunately, it is not possible to compute the integral of the kinetic update against $(1, \xi)$, instead we propose a compromise consisting to replace the standard Maxwellian by a simpler one. The question of boundary conditions is then treated in Section 2.4, and finally we discuss the practical implementation and the computational costs in Section 2.5.

2.1 Kinetic discretization and properties

In the case of a flat topography, the kinetic scheme is a *flux vector splitting* scheme [12]. We denote $P \in \mathbb{N}$ the number of cells, and the update (1.21) approaching the solution of (1.19) is replaced for all $1 \leq i \leq P$ by the implicit kinetic scheme

$$f_i^{n+1-} = M_i - \sigma\xi \left(\mathbb{1}_{\xi<0} f_{i+1}^{n+1-} + \mathbb{1}_{\xi>0} f_i^{n+1-} - \mathbb{1}_{\xi<0} f_i^{n+1-} - \mathbb{1}_{\xi>0} f_{i-1}^{n+1-} \right), \tag{2.1}$$

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

$$\begin{cases} -\sigma \mathbb{1}_{\xi>0} \xi f_{i-1}^{n+1-} + (1 + \sigma|\xi|) f_i^{n+1-} + \sigma \mathbb{1}_{\xi<0} \xi f_{i+1}^{n+1-} = M_i \quad 2 \leq i \leq P-1 \\ (1 + \sigma|\xi|) f_1^{n+1-} + \sigma \mathbb{1}_{\xi<0} \xi f_2^{n+1-} = M_1 + \sigma \mathbb{1}_{\xi>0} \xi M_0^{n+1} \\ -\sigma \mathbb{1}_{\xi>0} \xi f_{P-1}^{n+1-} + (1 + \sigma|\xi|) f_P^{n+1-} = M_P - \sigma \mathbb{1}_{\xi<0} \xi M_{P+1}^{n+1} \end{cases} \tag{2.2}$$

The quantities $M_0^{n+1} = M(U_0^{n+1}, \xi)$ and $M_{P+1}^{n+1} = M(U_{P+1}^{n+1}, \xi)$ appearing in the last two lines of (2.2) account for the imposed boundary conditions. In a first step, we assume that M_0^{n+1} and M_{P+1}^{n+1} are two known kinetic Maxwellian, their expressions will be discussed in more details in the paragraph devoted to the practical computation of the implicit variables, see paragraph 2.4.

With obvious notations, the system (2.2) consists in finding $f^{n+1} = \{f_i^{n+1-}\}_{i \in \{1,\dots,P\}}$ satisfying

$$(\mathbf{I} + \sigma \mathbf{L}) f^{n+1} = M + \sigma B^{n+1}, \tag{2.3}$$

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

$$\mathbf{L} = \begin{pmatrix} |\xi| & \xi \mathbb{1}_{\xi<0} & 0 & \dots & 0 \\ -\xi \mathbb{1}_{\xi>0} & |\xi| & \xi \mathbb{1}_{\xi<0} & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -\xi \mathbb{1}_{\xi>0} & |\xi| & \xi \mathbb{1}_{\xi<0} \\ 0 & \dots & 0 & -\xi \mathbb{1}_{\xi>0} & |\xi| \end{pmatrix}. \tag{2.4}$$

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

$$f^{n+1} = \begin{pmatrix} f_1^{n+1-} \\ \vdots \\ f_i^{n+1-} \\ \vdots \\ f_P^{n+1-} \end{pmatrix}, \quad M = \begin{pmatrix} M_1 \\ \vdots \\ M_i \\ \vdots \\ M_P \end{pmatrix} \quad \text{and} \quad B^{n+1} = \begin{pmatrix} \mathbb{1}_{\xi>0} \xi M_0^{n+1}, \\ 0 \\ \vdots \\ 0 \\ -\mathbb{1}_{\xi<0} \xi M_{P+1}^{n+1}, \end{pmatrix}. \tag{2.5}$$

The practical computation of the densities vector f^{n+1} will be discussed in paragraph 2.2. Hereafter, we focus on the properties of the numerical scheme (2.2) and the two following results hold.

Lemma 2.1 *The matrix* $\mathbf{I}+\sigma\mathbf{L}$ *defined by through (2.4)*

(i) is invertible for any σ *and* ξ*,*

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

Proposition 2.2 *The numerical scheme (2.2) satisfies the following properties*

(i) the discretization (2.2) is consistent with (1.1),

(ii) the system (2.2) – or equivalently the system (2.3) – admits an unique solution and the solution satisfies

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

To state the main result regarding the entropy stability of the fully implicit kinetic scheme we need to intrduce the function Ψ defined by

$$\Psi : \mathbb{R}^2 \ni (a,b) \longmapsto \frac{g^2\pi^2}{6}(b+2a)(b-a)^2, \tag{2.6}$$

which is positive on $\mathbb{R}^2_+$.

Proposition 2.3 *The system (2.3) admits a unique solution of positive quantities and defines an implicit kinetic scheme. Moreover, this scheme satisfies the fully discrete entropy equality*

$$\begin{aligned} H_0(f_i^{n+1-}) = H_0(M_i) &- \sigma\Big(H_{0,i+1/2}^{n+1-} - H_{0,i-1/2}^{n+1-}\Big) - \Psi(f_i^{n+1-}, M_i) \\ &+ \sigma\xi\Big(\mathbb{1}_{\xi<0}\Psi(f_i^{n+1-}, f_{i+1}^{n+1-}) - \mathbb{1}_{\xi>0}\Psi(f_i^{n+1-}, f_{i-1}^{n+1-})\Big) \end{aligned} \tag{2.7}$$

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

$$H_{0,i+1/2}^{n+1-} = \xi\,\mathbb{1}_{\xi<0}H_0(f_{i+1}^{n+1-}) + \xi\,\mathbb{1}_{\xi>0}H_0(f_i^{n+1-}), \tag{2.8}$$

$$H_{0,i-1/2}^{n+1-} = \xi\,\mathbb{1}_{\xi<0}H_0(f_i^{n+1-}) + \xi\,\mathbb{1}_{\xi>0}H_0(f_{i-1}^{n+1-}). \tag{2.9}$$

Since Ψ *given in (2.6) is positive on* $\mathbb{R}^2_+$*, the last three terms of equality (2.7) define a nonpositive dissipative term.*

Notice that the results obtained in the two Propositions 2.2 and 2.3 do not require any CFL condition. A consequence of Proposition 2.3 is that, when using the classical Maxwellian (1.4), the macroscopic scheme associated to (2.1) will satisfy a discrete entropy inequality that always dissipates the energy. In fact since the Maxwellian (1.4) minimizes the functional (1.14) we have the following upper bound on the macroscopic entropy $\eta(U_i^{n+1})$

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

We then use equality (2.7) yielding

$$\eta(U_i^{n+1}) \leq \eta(U_i^n) - \sigma\bigg(\int_{\mathbb{R}} H_{0,i+1/2}^{n+1-}(\xi)d\xi - \int_{\mathbb{R}} H_{0,i-1/2}^{n+1-}(\xi)d\xi\bigg) + \int_{\mathbb{R}} D_i(\xi)d\xi, \tag{2.10}$$

where $D_i = \mathcal{O}(\Delta t)$ is the negative dissipation term corresponding to the terms of (2.7) involving Ψ, making the inequality (2.10) consistent with the entropy inequality (1.2).

Proof of Lemma 2.1. The matrix $\mathbf{I}+\sigma\mathbf{L}$ writes

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

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

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

meaning $\mathbf{I}+\sigma\mathbf{L}$ is a monotone matrix and hence the solution of (2.3) satisfies

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

proving the result.

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

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

where all the entries of the matrix

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

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

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

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

Proof of prop. 2.2. *(i)* The four terms in parentheses in (2.1) are conservative, and are classically consistent with $\xi\partial_x f$ in (1.19).
(ii) This is a direct consequence of Lemma 2.1. □

The proof of Proposition 2.3 makes use of the following Lemma which will also be useful later.

Lemma 2.4 *The following identity holds for any real pair (a,b) and for any $\xi\in\mathbb{R}$*

$$H_0(b,\xi)=H_0(a,\xi)+\partial_f H_0(a,\xi)(b-a)+\Psi(a,b), \tag{2.11}$$

with the function Ψ defined in (2.6). Especially, we recover the convexity of $H_0(\cdot,\xi)$ on $\mathbb{R}_+$ thanks to the positivity of Ψ on $\mathbb{R}_+^2$. Equality (2.11) remains satisfied if we replace H_0 by H.

Proof of Lemma 2.4. For any (a,b) in $\mathbb{R}^2$ there holds

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

and equality (2.11) is recovered using the formula

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

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

Proof of prop. 2.3. The proof follows similar lines as what was done in the case of the fully explicit version of the kinetic scheme in [4]. Instead of multiplying Equation (2.1) by $\partial_f H_0(f_i^n)$, we multiply it by $\partial_f H_0(f_i^{n+1-})$, which leads to

$$\begin{aligned}\partial_f H_0(f_i^{n+1-})(f_i^{n+1-} - M_i) = & -\sigma\xi \mathbb{1}_{\xi<0}\partial_f H_0(f_i^{n+1-})(f_{i+1}^{n+1-} - f_i^{n+1-}) \\ & + \sigma\xi \mathbb{1}_{\xi>0}\partial_f H_0(f_i^{n+1-})(f_{i-1}^{n+1-} - f_i^{n+1-}) \ .\end{aligned} \tag{2.12}$$

In (2.12) we have three terms of the form $\partial_f H(a)(b-a)$ with $a = f_i^{n+1-}$ and $b \in \{f_{i-1}^{n+1-}, M_i, f_{i+1}^{n+1-}\}$. Taking advantage of Lemma 2.4 we can write

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

and we conclude by grouping the expressions. □

2.2 Practical computation of the implicit kinetic update

When dealing with implicit schemes, one has often to invert an operator and the key point of the numerical scheme (2.3) is the computation of the inverse of the matrix $\mathbf{I} + \sigma\mathbf{L}$, where $\mathbf{L}$ has been defined in (2.4). In our case, it will be possible to compute analytically this inverse thanks to the triangular structure of the matrix, which is due to the upwinding of the fluxes in (2.1). More precisely we decompose $(\mathbf{I} + \sigma\mathbf{L})^{-1}$ as the contributions of the left- and right-going information, which gives

$$(\mathbf{I} + \sigma\mathbf{L})^{-1} = (\mathbf{I} + \sigma\mathbf{L}^+)^{-1}\mathbb{1}_{\xi<0} + (\mathbf{I} + \sigma\mathbf{L}^-)^{-1}\mathbb{1}_{\xi>0}, \tag{2.13}$$

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

$$\mathbf{L}^+ = \begin{pmatrix} -\xi & \xi & 0 & \dots & 0 \\ 0 & -\xi & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \xi \\ 0 & \dots & & 0 & -\xi \end{pmatrix}, \qquad \mathbf{L}^- = \begin{pmatrix} \xi & 0 & & \dots & 0 \\ -\xi & \ddots & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \dots & 0 & -\xi & \xi \end{pmatrix}. \tag{2.14}$$

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

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

we can write

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

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

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

To conclude we give the analytic expression of the inverse:

$$(\mathbf{I}+\sigma\mathbf{L}^-)^{-1}_{i,j} = \begin{cases} \dfrac{(\sigma\xi)^{i-j}}{(1+\sigma\xi)^{i-j+1}} & \text{if } i \geq j \\ 0 & \text{else} \end{cases}, \tag{2.15}$$

$$(\mathbf{I}+\sigma\mathbf{L}^+)^{-1}_{i,j} = \begin{cases} \dfrac{(-\sigma\xi)^{j-i}}{(1-\sigma\xi)^{j-i+1}} & \text{if } i \leq j \\ 0 & \text{else} \end{cases}. \tag{2.16}$$

Especially we recover the properties enumerated in Lemma 2.1, since we see that all the coefficients of the inverse (2.13) are comprised between zero and one respectively when $\xi \geq 0$ and $\xi \leq 0$.

2.3 Macroscopic implicit scheme

We now focus on obtaining an explicit writing of the macroscopic update (1.19) associated to (2.1). Using the expressions (2.15)-(2.16) of the matrix inverse, we need to compute the integral of

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

against 1, ξ and ξ^2 for all $0 \leq k \leq P-1$. This seems hardly possible with the classical Maxwellian (1.4) and we use the result presented in the following remark.

Remark 2.5 *Let us recall that the half-disk Maxwellian $M(U,\xi)$ defined by (1.4) has some optimal properties presented in Lemma 1.2, which allow to obtain the discrete entropy inequality (2.10) at the macroscopic scale. Other choices of Maxwellians are possible but the discrete entropy inequality (2.7) is not granted to hold anymore. A general possibility is to choose $M(U,\xi)$ of the form*

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

with

$$c = \sqrt{\frac{gh}{2}},$$

where χ is a non-negative, compactly supported even function satisfying

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

There are a lot of possible choices for χ e.g.

$$\chi_1(z) = \frac{1}{2\sqrt{3}}\mathbb{1}_{|z|\leq\sqrt{3}}, \quad or \quad \chi_2(z) = \frac{3}{20\sqrt{5}}z^2 + \frac{3}{4\sqrt{5}}\mathbb{1}_{|z|\leq\sqrt{5}}. \tag{2.17}$$

Notice that the definition (1.4) of the half-disk Maxwellian corresponds to the choice

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

In order to compute explicitely the marcroscopic scheme i.e. without approximate quadrature formula, we now use in this section the following expression for the Maxwellian

$$M(U,\xi) = \frac{h}{c}\chi_1\Big(\frac{\xi-u}{c}\Big) = \frac{h}{2\sqrt{3}c}\mathbb{1}_{|\xi-u|\leq\sqrt{3}c} \tag{2.18}$$

and referred to as the index Maxwellian. This is the simplest choice we can make, and it will enable us to obtain analytic expressions for the aforementioned integrals. Furthermore note that it satisfies all the moment relations (1.5), which is important for consistency with the macroscopic Saint-Venant system.

In order to compute the macroscopic version of the scheme (2.2), we proceed as follows. The strategy consists to dissociate the contribution of the information coming from the interior of the computational domain, and the one coming from the boundaries

$$\left\{ \begin{array}{l} U^{\text{int}} = \displaystyle\int_{\mathbb{R}} \begin{pmatrix} 1 \\ \xi \end{pmatrix} (\mathbf{I} + \sigma \mathbf{L})^{-1} M \, d\xi \\ U^{\text{ext}} = \displaystyle\int_{\mathbb{R}} \begin{pmatrix} 1 \\ \xi \end{pmatrix} (\mathbf{I} + \sigma \mathbf{L})^{-1} \sigma B^{n+1} \, d\xi \end{array} \right. . \tag{2.19}$$

The final update is then set as $U^{n+1} = U^{\text{int}} + U^{\text{ext}}$, which coincides with definition (1.20). We postpone the details about the computation of B^{n+1} to the next section, and assume that it is known for now. First for U^{int}, we have

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

Plugging the analytic expressions (2.15) and (2.16) in the above integrals, we can express the i-th component of U^{int} as

$$\begin{aligned} U_i^{\text{int}} &= \int_{\xi<0} \sum_{j=1}^{P} \begin{pmatrix} 1 \\ \xi \end{pmatrix} (\mathbf{I} + \sigma \mathbf{L}^+)^{-1}_{i,j} M_j d\xi + \int_{\xi>0} \sum_{j=1}^{P} \begin{pmatrix} 1 \\ \xi \end{pmatrix} (\mathbf{I} + \sigma \mathbf{L}^-)^{-1}_{i,j} M_j d\xi \\ &= \int_{\xi<0} \sum_{j=i}^{P} \begin{pmatrix} 1 \\ \xi \end{pmatrix} \frac{(-\sigma\xi)^{j-i}}{(1-\sigma\xi)^{j-i+1}} M_j d\xi + \int_{\xi>0} \sum_{j=1}^{i} \begin{pmatrix} 1 \\ \xi \end{pmatrix} \frac{(\sigma\xi)^{i-j}}{(1+\sigma\xi)^{i-j+1}} M_j d\xi. \end{aligned} \tag{2.20}$$

A detailed expression of the quantities appearing in relation (2.20) is given in Appendix A. Similarly for the exterior contribution we have

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

Using definition (2.5) and equalities (2.15)–(2.16), the i-th component of U^{ext} is

$$U_i^{\text{ext}} = \int_{\xi<0} \begin{pmatrix} 1 \\ \xi \end{pmatrix} \frac{(-\sigma\xi)^{P-i+1}}{(1-\sigma\xi)^{P-i+1}} M_{P+1}^{n+1} d\xi + \int_{\xi>0} \begin{pmatrix} 1 \\ \xi \end{pmatrix} \frac{(\sigma\xi)^{i}}{(1+\sigma\xi)^{i}} M_0^{n+1} d\xi. \tag{2.21}$$

The expressions necessary to compute (2.20) and (2.21) are detailed in Appendix A and Appendix B.

2.4 Boundary conditions

In this paragraph we discuss how to enforce the boundary conditions associated with the contribution U^{ext} in (2.19). To achieve this, it will be necessary to introduce two ghosts cells numbered 0 and $P+1$ and neighbours of the cells C_1 and C_P and we have to define U_0^{n+1} and U_{P+1}^{n+1}. The problem we are facing is that these ghost quantities depend on the neighboring values in cells C_1 and C_P at time t^{n+1}, and which are themselves unknown. Hence we have an implicit problem where the relation between the ghost and border terms can be nonlinear depending on the type of boundary conditions. In practice we will avoid this issue by substituting B^{n+1} with B^n in the definition (2.19) of U^{ext}. Doing so can be interpreted as a first order approximation in time since we have

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

Note that expliciting the ghost values does not prevent the scheme from satisfying a discrete entropy as the one in Proposition 2.3. In fact, equality (2.7) still holds in the interior cells $2 \leq i \leq P-1$, and in border cells $i = 1, N$ one obtains

$$H_0(f_1^{n+1-}) = H_0(M_1) - \sigma\Big(H_{0,3/2}^{n+1-} - H_{0,1/2}^{n+1-}\Big) - \Psi(f_1^{n+1-}, M_1) \tag{2.22}$$

$$+\sigma\xi\Big(\mathbb{1}_{\xi<0}\Psi(f_1^{n+1-},f_2^{n+1-})-\mathbb{1}_{\xi>0}\Psi(f_1^{n+1-},M_0^n)\Big),$$

$$H_0(f_P^{n+1-})=H_0(M_P)-\sigma\Big(H_{0,P+1/2}^{n+1-}-H_{0,P-1/2}^{n+1-}\Big)-\Psi(f_P^{n+1-},M_P) \tag{2.23}$$

$$+\sigma\xi\Big(\mathbb{1}_{\xi<0}\Psi(f_P^{n+1-},f_{P-1}^{n+1-})-\mathbb{1}_{\xi>0}\Psi(f_P^{n+1-},M_{P+1}^n)\Big),$$

where the border entropy fluxes $H_{0,1/2}^{n+1-}$, $H_{0,P+1/2}^{n+1-}$ are given by

$$H_{0,1/2}^{n+1-}=\xi\mathbb{1}_{\xi<0}H_0(f_1^{n+1-})+\xi\mathbb{1}_{\xi>0}H_0(M_0^n),\qquad H_{0,P+1/2}^{n+1-}=\xi\mathbb{1}_{\xi<0}H_0(M_{P+1}^n)+\xi\mathbb{1}_{\xi>0}H_0(f_P^{n+1-}).$$

The benefit of expliciting the ghost Maxwellian M_0, M_{P+1} is that we can more easily determine the macroscopic ghost quantities U_0^n, U_{P+1}^n at time t^n based on U_1^n, U_P^n following the procedure described hereafter and similar to that of Bristeau and Coussin in [13]. We will focus on fluvial flows where the material velocity of particles $|u|$ is smaller than the celerity of surface gravity waves $\sqrt{gh}$; in particular low Froude flows enter this regime. Since in this case the eigenvalues $u-\sqrt{gh}$ and $u+\sqrt{gh}$ have opposite sign, at each boundary we have exactly one wave entering the domain and one wave leaving it. Hence we dispose of a single degree of freedom to set the ghost values, which generally consists in enforcing either a given water height or a discharge. The ghost state is then fully determined by asking the outgoing Riemann invariant to remain constant through the interface.

In the following we detail the handling of the left boundary condition when enforcing either the water height or the discharge; we put aside the case of the right boundary condition as it is symmetric.

Given water height. First we consider the case where the water height is enforced at the left boundary of the domain, and we denote by $h_{g,l}$ the value attributed to the ghost cell. Together with the condition on the outgoing Riemann invariant, we get the following nonlinear system which is solved analytically

$$\begin{cases} h_0^n = h_{g,l} \\ u_0^n - 2\sqrt{gh_0^n} = u_1^n - 2\sqrt{gh_1^n} \end{cases} \implies U_0^n = h_{g,l}\begin{pmatrix} 1 \\ u_1^n - 2(\sqrt{gh_1^n}-\sqrt{gh_{g,l}}) \end{pmatrix}.$$

Given flux. Another possibility is to enforce the discharge at the boundary, and we denote by $q_{g,l}$ the left ghost value. This time, the constraint on the Riemann invariant will enable to determine the ghost water height. Indeed we have the system

$$\begin{cases} q_0^n = q_{g,l} \\ u_0^n - 2\sqrt{gh_0^n} = u_1^n - 2\sqrt{gh_1^n} \end{cases}, \tag{2.24}$$

and the second equality involving the outgoing Riemann invariant requires to find the real roots of the third order polynomial in $X=\sqrt{h_0^n}$ given by

$$\mathcal{P}(X) := X^3 + \frac{u_1^n - 2\sqrt{gh_1^n}}{2\sqrt{g}}X^2 - \frac{q_{g,l}}{2\sqrt{g}} = 0. \tag{2.25}$$

Its derivative $\mathcal{P}'$ admits two real distinct roots

$$X_0 = 0 \quad \text{and} \quad X_1 = -\frac{u_1^n - 2\sqrt{gh_1^n}}{3\sqrt{g}}, \qquad \text{with} \quad \mathcal{P}(X_1) = -\frac{1}{2}X_1^3 - \frac{q_{g,l}}{2\sqrt{g}},$$

and we recal that we make the assumption that $u_1^n - \sqrt{gh_1^n} \le 0$ so that $X_1 \ge 0$. If $q_{g,l} \ge -\sqrt{g}X_1^3$, one has $\mathcal{P}(X_1) \le 0$; together with the fact that $\mathcal{P}(X) \to +\infty$ as X goes to infinity, there exists one real root larger than $X_1 \ge 0$. If additionally $q_{g,l} > 0$, this root is the only real one since $\mathcal{P}(X_0) < 0$; if $q_{g,l} \le 0$ then $\mathcal{P}(X_0) \ge 0$ and there exists a second positive root between X_0 and X_1. In this case we take the convention to retain the smaller of the two. On the other hand, when $q_{g,l} < -\sqrt{g}X_1^3$ we have that $\mathcal{P}(X_0), \mathcal{P}(X_1)$ are both positive, and there doesn't exist a positive root. In this case one can choose to enforce the water height instead.

Note that when enforcing the flux, our approach differs from that of Bristeau and Coussin in [13], 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. A common practice for channel flows is to enforce the water height at the inlet and the flux at the outlet.

Remark 2.6 *When substituting B^{n+1} with B^n in the implicit kinetic scheme (2.3), the corresponding update can be reformulated as $(\overline{\mathbf{I}+\sigma\mathbf{L}})\overline{f}^{n+1} = \overline{M}^n$ with*

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

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

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

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

2.5 Implementation and computational costs

It is important to keep a reasonable algorithmic complexity so that the implicit method presented in the previous lines and especially the formulae (2.20),(2.21) can be used in practice. We discuss here how to improve its computational cost by a substantial margin. In Appendix A, we show that the i-th component of vectors h^{int} and $(hu)^{\text{int}}$ have the form

$$\text{(2.26)} \qquad \left\{\begin{array}{l} h_i^{\text{int}} = \dfrac{1}{2\sigma\sqrt{3}}\left(\displaystyle\sum_{j=i}^{P}\sqrt{\frac{2h_j^n}{g}}(Ah)_{i,j} + \sum_{j=1}^{i}\sqrt{\frac{2h_j^n}{g}}(Bh)_{i,j}\right) \\ (hu)_i^{\text{int}} = \dfrac{1}{2\sigma^2\sqrt{3}}\left(-\displaystyle\sum_{j=i}^{P}\sqrt{\frac{2h_j^n}{g}}(Ahu)_{i,j} + \sum_{j=1}^{i}\sqrt{\frac{2h_j^n}{g}}(Bhu)_{i,j}\right) \end{array}\right.,$$

where Ah, Ahu are dense upper triangular matrices, and Bh, Bhu are dense lower triangular matrices. Therefore computing h^{int} and $(hu)^{\text{int}}$ through (2.26) is analog to performing a matrix-vector product which has a quadratic complexity $O(P^2)$, and we cannot hope to do better than that. However the coefficients (A.3)–(A.6) of the above matrices involve a summation, and at a first glance the cost to assemble them is seemingly cubic. This is quite expensive and can render the method pretty much inefficient. However this complexity can be reduced to a quadratic cost by computing the coefficients in the correct order. More specifically we show that all the matrices above can be defined through a recurrence relation allowing to compute each coefficient from a previous one in $O(1)$ operation. In fact, denoting $y = x/(1+x)$ and $z = \ln|1+x|$, the matrix Ah is given by

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

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

$$(Ah)_{i,j} = \begin{cases} 0 & \text{if } j < i \\ \big[\ln(|1+x|)\big]_{-\min(0,b_j)\sigma}^{-\min(0,a_j)\sigma} & \text{if } i = j \\ (Ah)_{i+1,j} - \frac{1}{j-i}\big[y^{j-i}\big]_{-\min(0,b_j)\sigma}^{-\min(0,a_j)\sigma} & \text{if } j > i \end{cases} . \tag{2.27}$$

Likewise, the lower triangular matrix Bh is given by

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

and can be defined by the following recurrence formula

$$(Bh)_{i,j} = \begin{cases} 0 & \text{if } i < j \\ \big[\ln(|1+x|)\big]_{\max(0,a_j)\sigma}^{\max(0,b_j)\sigma} & \text{if } i = j \\ (Bh)_{i-1,j} - \frac{1}{i-j}\big[y^{i-j}\big]_{\max(0,a_j)\sigma}^{\max(0,b_j)\sigma} & \text{if } i > j \end{cases} . \tag{2.28}$$

Hence it is more efficient to assemble matrices Ah and Bh column wise, starting from the diagonal coefficient and moving towards the first or last row. This way we only have to subtract one term to the previous coefficient so as to get the next one, and the cost of this operation is in $O(1)$. Since there are $P(P+1)/2$ coefficients to compute in total, the assembly of Ah and Bh following this strategy requires $O(P^2)$ steps.

A similar conclusion is achieved for Ahu and Bhu, although the recurrence relation is less straightforward to obtain. We first remark that, introducing $(l)_{i,j} = i - j + 1$ the relations (A.5) and (A.6) become

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

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

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

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

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

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

Then we can write that

$$(Ahu)_{i,j} = \begin{cases} 0 & j < i \\ \left[x - \ln|1+x|\right]_{-\min(0,b_j)\sigma}^{-\min(0,a_j)\sigma} & j = i \\ (l)_{j,i}(VA)_{i,j} - (UA)_{i,j} + \left[x - (l)_{j,i}\ln|1+x|\right]_{-\min(0,b_j)\sigma}^{-\min(0,a_j)\sigma} & j > i \end{cases}. \tag{2.29}$$

Similarly we introduce

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

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

so that we have

$$(Bhu)_{i,j} = \begin{cases} 0 & i < j \\ \left[x - \ln|1+x|\right]_{\max(0,a_j)\sigma}^{\max(0,b_j)\sigma} & i = j \\ (l)_{i,j}(VB)_{i,j} - (UB)_{i,j} + \left[x - (l)_{i,j}\ln|1+x|\right]_{\max(0,a_j)\sigma}^{\max(0,b_j)\sigma} & i > j \end{cases}. \tag{2.30}$$

To conclude, through relations (2.29) and (2.30) we are also able to assemble matrices Ahu and Bhu with a quadratic cost with respect to the number of cells, which means that the overall method has a $O(P^2)$ complexity.

Remark 2.7 *We have considered here the specific case of a kinetic solver and one can imagine that an implicit scheme for another finite volume solver can lead to reduced numerical costs. But it is worth noticing that since the explicit expression of the inverse of the matrix* $\mathbf{I}+\sigma\mathbf{L}$ *is accessible in the kinetic context, one can hardly find a more efficient implicit technique.*

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

Explicit scheme. Let Δt^n be the time step allowing to satisfy the CFL constraint. In order to obtain the expression of U^{n+1} from U^n, approximately $4P$ numerical fluxes have to be computed (2 numerical fluxes at each interface for each variable h and hu). The explicit kinetic scheme is fully detailed in [6, 4].

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

We conclude that the implicit strategy is less expensive when

$$\frac{\Delta t^n_{imp}}{\Delta t^n} \gg \frac{P^2}{4P} = \frac{P}{4}. \tag{2.31}$$

Note however that the computational cost is not the only factor to account for, and one should also consider the efficiency of the scheme, that is to say the relation between the error and the computational time. Generally, taking a very coarse resolution in time results in poorly accurate results, in which case it is not desirable to have (2.31). Nevertheless there are some cases where the fast dynamics

do not play an important role such as in the low Froude regime. Then it might be advantageous to consider large time steps. We will see through the upcoming numerical results from Section 5.1 that the interest of the implicit kinetic scheme is rather limited when it comes to efficiency, at least for the considered test cases. Hence the explicit strategy is preferable to the implicit one, unless we account for the greater stability offered by the latter in terms of discrete entropy inequality.

3 An iterative resolution scheme

The kinetic scheme (2.3) requires to solve a linear system and in the previous section, we have seen that it was possible to have an analytic expression for the inverse of the matrix $\mathbf{I} + \sigma\mathbf{L}$ with $\mathbf{L}$ given in (2.4). For the numerical approximation of PDEs e.g. in finite elements methods when the linear system to solve is large an iterative strategy is singled out compared to a direct inversion of the matrix. We propose to follow the same idea here, with mainly two benefits. First it will allow us to use the half disk Maxwellian (1.4), for which we recall the integrals (2.19) could not be computed analytically in the case of the fully implicit kinetic scheme. This is important as it will enable to prove some discrete entropy inequality at the macroscopic scale thanks to the existence of a kinetic entropy H_0 given by (1.10) that satisfies (1.13)-(1.14), while having an explicit writting of the update. The second advantage lies in the possibility to couple the iterative strategy with the hydrostatic reconstruction to obtain a well balanced treatment for varying bottoms, which we will discuss in Section 3.2. Before this, the simpler case of flat bathymetries is investigated as a toy problem in Section 3.1, and will give some insight.

More precisely we shall use a Gauss-Jacobi type decomposition, which consists to write the matrix $\mathbf{I} + \sigma\mathbf{L}$ as $\mathbf{D} - \mathbf{N}$, where $\mathbf{D}$ and $\mathbf{N}$ are two matrices from $\mathbb{R}^{P\times P}$ with $\mathbf{D}$ is invertible. Then the scheme (2.3) also writes

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

and one can propose an iterative resolution of the previous equation under the form

$$f^{n+1,k+1} = \mathbf{D}^{-1}\mathbf{N}f^{n+1,k} + \mathbf{D}^{-1}(M + \sigma^k B^{n+1}), \qquad \sigma^k = \frac{\Delta t^k}{\Delta x},$$

where the timestep Δt^k is allowed to vary at each subiterations. If it converges, the sequence $(f^{n+1,k})_{k\in\mathbb{N}}$ converges towards the solution of (2.3).

3.1 Case of a flat topography

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

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

when the bathymetry is flat and where $\alpha \in (0,\infty)$ is a relaxation parameter. The main goal is to put forward the differences between such iterative methods and the fully implicit kinetic scheme proposed in Section 2; the first one being that it will no longer be possible to use arbitrary large time steps since a CFL condition will be needed to ensure the convergence of the sub-iterations and the positivity of the water height; the second one being that the iterative scheme enables one to have an explicit writing of the update while using the half-disk Maxwellian. As long as we are restricted to flat bathymetries, it should also be noticed that an iterative strategy proves less advantageous than the forward Euler kinetic scheme, which is more efficient while preserving the water height positivity and being entropy stable. The iterative kinetic scheme presents however a clear advantage over the explicit scheme in terms of entropy stability in the case of varying bathymetries, which is treated in Section 3.2.

We consider two iterative schemes below; they differ only in wether or not the vector of kinetic densities $f^{n+1,k}(\xi)$ is projected onto the space of Maxwellians at every sub-iteration or just at $k = 0$. In the latter and easier setting, we shall obtain a CFL condition to ensure the convergence of sub-iterations. In the former case, this question is also addressed, although more restrictive assumptions are needed. The question of entropy stability is only investigated when the projection step occurs at every sub-iteration.

When developed, the iterative process with a single projection at $k = 0$ reads:

$$\left\{ \begin{array}{l} f^{n+1,0} = M \\ (1+\alpha) f^{n+1,k+1} = (\alpha \mathbf{I} - \sigma^k \mathbf{L}) f^{n+1,k} + M + \sigma^k B^{n+1,k} \\ \forall 1 \leq i \leq P,\ U_i^{n+1,k} = \displaystyle\int_{\mathbb{R}} \begin{pmatrix} 1 \\ \xi \end{pmatrix} f_i^{n+1,k}(\xi)\, \mathrm{d}\xi \end{array} \right. , \tag{3.1}$$

with $B^{n+1,k}$ the boundary condition associated with the macroscopic state $U^{n+1,k}$ as explained in Section 2.4, see (2.5). The following Proposition highlights the main compromise linked with such an iterative approach, which is the requirement for a CFL condition in order for the method to converge.

Proposition 3.1 *Assume that $B^{n+1,k}$ remains constant equal to B^n for any k in $\mathbb{N}$. Then (3.1) defines an arithmetico-geometric sequence which converges if the CFL condition $\sigma|\xi| < 1 + 2\alpha$ holds for all ξ belonging to* $\operatorname{supp} M \cup \operatorname{supp} B^n$.

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

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

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

Remark 3.2 *As we did in Section 2.4 for the fully implicit scheme, we can replace $B^{n+1,k}$ by B^n in the iterative process (3.1). In fact this constitutes a first order approximation in time since we have $f^{n+1,k} = M + \mathcal{O}(\Delta t)$. Under this simplification, the assumption from Proposition 3.1 that $B^{n+1,k}$ does not depend on k becomes automatically satisfied.*

In practice, we wish to apply an iterative method directly at the macroscopic level. An issue with (3.1) is that the distribution involved in the kinetic flux (i.e. the term in factor of $\sigma^k \mathbf{L}$) is not a vector of Maxwellians, which prevents us to write the recurrence relation at the macroscopic level since there is no general expression for the numerical flux. To bypass this issue, we propose the following modification of (3.1), where we replace all occurrences of $f^{n+1,k}$ on the right hand side by a vector of Maxwellians $M^{n+1,k}$, which yields

$$\left\{ \begin{array}{l} \bar{f}^{n+1,0}(\xi) = M \\ (1+\alpha)\bar{f}^{n+1,k+1}(\xi) = (\alpha \mathbf{I} - \sigma^k \mathbf{L}) M^{n+1,k} + M + \sigma^k \bar{B}^{n+1,k} \\ M^{n+1,k+1} = \bar{f}^{n+1,k+1} + \Delta t^k \bar{Q}^{n+1,k+1} \end{array} \right. . \tag{3.2}$$

This new iterative process is alternating two stages, the first one being the usual transport step, while the second one is a projection step onto the set of Maxwellians yielding $M^{n+1,k+1}$. In this sense (3.2) is an iterative BGK splitting approach where the projection step doesn't modify the macroscopic quantities of interest since the term $\bar{Q}^{n+1,k}$ is a vector of collision operators satisfying the conservation constraints (1.7). Note that the time stepping Δt^k is made dependent on k as the support of $M^{n+1,k}$ can change from iteration to iteration.

Remark 3.3 *For each value of k, each row of the density $\bar{f}^{n+1,k+1}(\xi)$ defined by (3.2) is a combination of half disks. Hence $\bar{f}_i^{n+1,k+1}(\xi)$the that is the row i of $\bar{f}^{n+1,k+1}(\xi)$ is a combination of $M_i^{n+1,k}$, M_i and possibly $B_i^{n+1,k}$ if $i = 1$ or $i = P$, see Fig. 1. This illustrates the role of the collision term $\bar{Q}^{n+1,k+1}$.*

The practical implementation of scheme (3.2) is based on its macroscopic version given by

$$(1+\alpha)U_i^{n+1,k+1} = \alpha U_i^{n+1,k} + U_i - \sigma\Big(\mathcal{F}(U_i^{n+1,k}, U_{i+1}^{n+1,k}) - \mathcal{F}(U_{i-1}^{n+1,k}, U_i^{n+1,k})\Big), \tag{3.3}$$

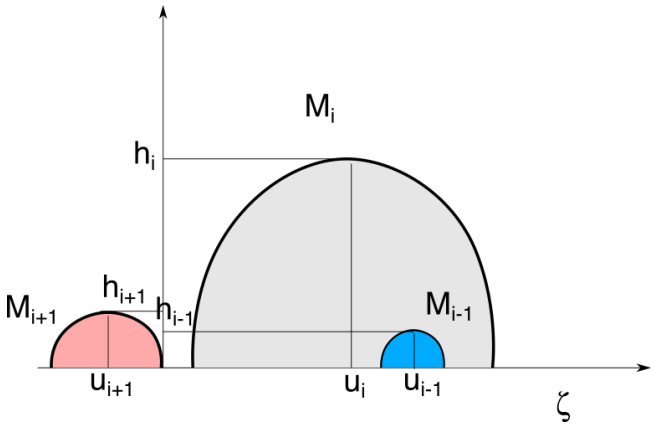

Figure 1: The maxwellians M_i, M_{i-1} and M_{i+1}. For each quantities appearing over the figure, the superscripts have been omitted.

for all $1 \leq i \leq P$, where the numerical flux $\mathcal{F}$ is defined as

$$\mathcal{F}(U_L, U_R) = \int_{\mathbb{R}} \xi \begin{pmatrix} 1 \\ \xi \end{pmatrix} \Big(\mathbb{1}_{\xi>0} M(U_L, \xi) + \mathbb{1}_{\xi<0} M(U_R, \xi) \Big) \, \mathrm{d}\xi, \tag{3.4}$$

and where the vectors $U_0^{n+1,k}, U_{P+1}^{n+1,k}$ appearing in the border cells $i \in \{1, P\}$ are respectively functions of $U_1^{n+1,k}$ and $U_P^{n+1,k}$ since the boundary conditions are imposed though a ghost cell strategy fully described in [13, 1]. As previously, one can choose to make these ghost values independent from the sub-iteration index k. Notice that if the sequence $(U^{n+1,k})_{k\in\mathbb{N}} \subset (\mathbb{R}^2)^P$ from (3.3) converges in $(\mathbb{R}^2)^P$, its limit U^{n+1} then satisfies

$$\forall 1 \leq i \leq P, \quad U_i^{n+1} = U_i^n - \sigma \Big(\mathcal{F}(U_i^{n+1}, U_{i+1}^{n+1}) - \mathcal{F}(U_{i-1}^{n+1}, U_i^{n+1}) \Big) \tag{3.5}$$

by continuity of the numerical flux (3.4).

Because of the projection step, the convergence of the iterative method (3.2) is more complex to prove than in Prop. 3.1. But the following result holds.

Proposition 3.4 *Let the sequence $\{\bar{f}^{n+1,k}(\cdot)\}_{k\in\mathbb{N}}$ be defined through (3.2), and let $\{\bar{Q}^{n+1,k}(\cdot)\}_{k\in\mathbb{N}}$ be the associated sequence of collision terms assumed to be bounded. Then under the CFL condition $2\sigma^k|\xi| < 1$ the sequence $\{\bar{f}^{n+1,k}(\cdot)\}_{k\in\mathbb{N}}$ converges.*

Proof of Prop. 3.4. First, we see that under the CFL condition $\sigma^k|\xi| \leq \alpha$, the quantity $\bar{f}^{n+1,k}$ defined by (3.2) is non-negative for any ξ.
Note that in (3.2) the convergence of $\{\bar{f}^{n+1,k}\}_{k\in\mathbb{N}}$ implies that of $\{\bar{Q}^{n+1,k}\}_{k\in\mathbb{N}}$, therefore we focus on the former sequence. The recurrence formula (3.2) also writes

$$(1+\alpha)M^{n+1,k+1} = (\alpha\mathbf{I} - \sigma^k\mathbf{L})M^{n+1,k} + M + \sigma^k\bar{B}^{n+1,k} + (1+\alpha)\Delta t^k \bar{Q}^{n+1,k+1}. \tag{3.6}$$

The previous formula is nothing else than an arithmetico-geometric-like sequence having the form

$$M^{n+1,k+1} = \mathbf{A}^{k,\alpha} M^{n+1,k} + C^{n,k},$$

with

$$\begin{aligned} \mathbf{A}^{k,\alpha} &= \frac{\alpha\mathbf{I} - \sigma^k\mathbf{L}}{1+\alpha}, \\ C^{n,k} &= \frac{M + \sigma^k\bar{B}^{n+1,k}}{1+\alpha} + \Delta t^k \bar{Q}^{n+1,k+1}, \end{aligned} \tag{3.7}$$

whose convergence is ensured as soon as the sequence

$$\sum_{j=0}^{k} (\mathbf{A}^{j,\alpha})^j C^{n,k-j},$$

converges when k grows. Since the collision terms $\bar{Q}^{n+1,k+1}$ are assumed to be bounded independently of k, we have $|C^{n,k}| \leq \bar{C}$ and the CFL condition condition ensures that

$$0 \leq \|\mathbf{A}^{k,\alpha}\|_{\infty,\infty} = \frac{\alpha + 2\sigma^k|\xi|}{1+\alpha} < 1,$$

where $\|\cdot\|_{\infty,\infty}$ is the maximum subordinate matrix norm. This allows to write

$$\sum_{j=0}^{k} |(\mathbf{A}^{j,\alpha})^j C^{n,k-j}| \leq \sum_{j=0}^{k} \|\mathbf{A}^{k,\alpha}\|_{\infty,\infty}^j |\bar{Q}^{n+1,k-j}| \leq \frac{\bar{C}}{1 - \|\mathbf{A}^{k,\alpha}\|_{\infty,\infty}}, \tag{3.8}$$

giving the absolute convergence of the series and thus proving the result. □

Next we state the following result regarding the positivity of the iterative kinetic scheme (3.2).

Proposition 3.5 *Assume that the water height vectors h^n and $h^{n+1,k}$ are positive. Then the update $f^{n+1,k+1}$ defined in the iterative scheme (3.2) is positive if for all $1 \leq i \leq P$ the CFL condition $\sigma^k|\xi| \leq \alpha + M_i/M_i^{n+1,k}$ holds for any ξ belonging to* $\operatorname{supp} M^{n+1,k}$*. As a direct consequence, the water height vector $h^{n+1,k+1}$ from scheme (3.3) is positive under these assumptions.*

We postpone the proof of Proposition 3.5 to the next section, where it is generalized to the case with varying bottom in Proposition 3.9. Finally we investigate the entropy stability of the iterative kinetic scheme (3.2). Note that when restricting to a flat bathymetry, the explicit kinetic scheme [4] already verifies a discrete entropy inequality, while being more efficient since it avoids the use of sub-iterations. The benefit of the iterative kinetic scheme compared to the fully explicit one will become clearer in the next section, where it is combined with the hydrostatic reconstruction. Nevertheless, the simpler case of a flat bottom is interesting in order to understand why the iterative kinetic scheme involving the half-disk Maxwellian (1.4) offers a favorable setting to get a discrete entropy inequality.

Proposition 3.6 *The kinetic entropy of the iterative scheme (3.2) with half-disk Maxwellian (1.4) satisfies the following inequality*

$$\begin{aligned} H_0(M_i^{n+1,k+1}) \leq \frac{H_0(M_i) + \alpha H_0(M_i^{n+1,k})}{1+\alpha} - \frac{\sigma^k \xi}{1+\alpha}\Big(H_{0,i+1/2}^{n+1,k} - H_{0,i-1/2}^{n+1,k}\Big) \\ + \eta'(U_i^{n+1,k+1}) \cdot \begin{pmatrix} 1 \\ \xi \end{pmatrix} Q_i^{n+1,k+1} + D_i^{n+1,k+1}, \end{aligned} \tag{3.9}$$

with $Q_i^{n+1,k+1} = M_i^{n+1,k+1} - f_i^{n+1,k+1}$ a collision operator verifying the conservation constraints (1.7), and with η the entropy given in (1.3). The interfacial kinetic entropies $H_{0,i\pm1/2}^{n+1,k}$ are

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

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

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

where $\Xi = \operatorname{supp} M^{n+1,k}$*, where we recall that the function* Ψ *defined in (2.6) is positive on* $\mathbb{R}_+^2$*, and with* $i \pm 1 = i - \operatorname{sgn}\xi$*. As a consequence, if for any integer* k *the CFL condition*

$$\forall \xi \in \operatorname{supp} M^{n+1,k},\ \sigma^k |\xi| \leq \alpha \tag{3.10}$$

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

$$H_0(M_i^{n+1,k+1},\xi) \leq H_0(M_i,\xi) - \sigma^k \xi \Big(\widetilde{H}_{0,i+1/2}^{n+1,k} - \widetilde{H}_{0,i-1/2}^{n+1,k} \Big) + \widetilde{D}_i^{n+1,k+1} + \widetilde{Z}_i^{n+1,k+1}. \tag{3.11}$$

Before giving the proof we have the remark below.

Remark 3.7 *Even when the CFL condition (3.10) is not satisfied, we can ensure that the scheme (3.2) satisfies a discrete entropy inequality from some rank* k *assuming the convergence of the method, which holds under the assumptions of Proposition 3.4. In fact, multiplying inequality (3.9) by* $1+\alpha$ *it is possible to write*

$$\begin{aligned} H_0(M_i^{n+1,k+1},\xi) \leq\ & \\ & H_0(M_i,\xi) - \sigma^k \xi \Big(H_{0,i+1/2}^{n+1,k} - H_{0,i-1/2}^{n+1,k} \Big) + (1+\alpha)\eta'(U_i^{n+1,k+1}) \cdot \begin{pmatrix} 1 \\ \xi \end{pmatrix} Q_i^{n+1,k+1} \\ & - \Psi(M_i^{n+1,k+1}, M_i) - \sigma^k |\xi| \mathbb{1}_\Xi \Psi(M_i^{n+1,k+1}, M_{i\pm1}^{n+1,k}) \\ & + \alpha \Big(H_0(M_i^{n+1,k},\xi) - H_0(M_i^{n+1,k+1},\xi) \Big) - (\alpha - \sigma^k |\xi| \mathbb{1}_\Xi) \Psi(M_i^{n+1,k+1}, M_i^{n+1,k}). \end{aligned} \tag{3.12}$$

In the right hand side of (3.12), the quantity

$$(1+\alpha)\eta'(U_i^{n+1,k+1}) \cdot \begin{pmatrix} 1 \\ \xi \end{pmatrix} Q_i^{n+1,k+1}$$

does not cause any issue as it vanishes upon integration over $\xi \in \mathbb{R}$*. This is due to the collision term* $Q_i^{n+1,k+1}$ *satisfying the conservation constraints (1.7), meaning that its integral against* $(1,\xi)^T$ *vanishes. Therefore in (3.12) the only problematic terms are contained in the last line, as their sign can be positive since we do not assume* $\sigma^k|\xi|\mathbb{1}_\Xi \leq \alpha$ *anymore. Nevertheless, by regularity of* $H_0(\cdot,\xi)$ *and by definition (2.6) of* Ψ*, these terms write as a* $\mathcal{O}(M_i^{n+1,k+1} - M_i^{n+1,k})$ *and vanish as* $k \to \infty$ *owing to the convergence of the method. As a consequence, from some rank* k *these two terms become negligible compared to* $-\Psi(M_i^{n+1,k+1}, M_i) < 0$ *which remains bounded away from zero, and we recover a dissipation with negative sign. (Note that if* $M^{n+1,k}$ *was converging to* M *as* $k \to \infty$*, it would imply that* M *solves the fixed point problem and thus* $M^{n+1,k} = M$ *for all* k*; putting aside this trivial case, this is why* $\Psi(M_i^{n+1,k+1}, M_i)$ *remains bounded away from zero).*

Proof of prop. 3.6. First, we remark that for any ξ there holds

$$\partial_f H_0(M_i^{n+1,k+1},\xi) Q_i^{n+1,k+1} \leq \eta'(U_i^{n+1,k+1}) \cdot \begin{pmatrix} 1 \\ \xi \end{pmatrix} Q_i^{n+1,k+1}, \tag{3.13}$$

which is implied by the relation (1.15) and by the fact that $Q_i^{n+1,k+1} = M_i^{n+1,k+1} - f_i^{n+1,k+1}$ is negative for any $\xi \notin \operatorname{supp} M_i^{n+1,k+1}$. As a consequence of (3.13), inequality (3.9) (and equivalently (3.12)) is verified if there holds

$$\begin{aligned} H_0(M_i^{n+1,k+1}) = \frac{H_0(M_i) + \alpha H_0(M_i^{n+1,k})}{1+\alpha} - \frac{\sigma^k \xi}{1+\alpha} \Big(H_{0,i+1/2}^{n+1,k} - H_{0,i-1/2}^{n+1,k} \Big) \\ + \partial_f H_0(M_i^{n+1,k+1}) Q_i^{n+1,k+1} + D_i^{n+1,k+1}, \end{aligned} \tag{3.14}$$

To prove equality (3.14) we write the sub-iteration (3.2) as

$$M_i^{n+1,k+1} = \frac{1}{1+\alpha}\Big(M_i + (\alpha - \sigma^k|\xi|\mathbb{1}_\Xi)M_i^{n+1,k} + \sigma^k|\xi|\mathbb{1}_\Xi M_{i\pm1}^{n+1,k}\Big) + Q_i^{n+1,k+1}, \tag{3.15}$$

with $i \pm 1 = i - \operatorname{sign}\xi$. Applying Lemma 2.4 for $a = M_i^{n+1,k+1}$ and $b = M_{i\pm1}^{n+1,k}, M_i, M_i^{n+1,k}$, we respectively get:

$$\begin{aligned} H_0(M_{i\pm1}^{n+1,k}) = H_0(M_i^{n+1,k+1}) + \partial_f H_0(M_i^{n+1,k+1})(M_{i\pm1}^{n+1,k} - M_i^{n+1,k+1}) \\ + \Psi(M_i^{n+1,k+1}, M_{i\pm1}^{n+1,k}), \end{aligned} \tag{3.16}$$

$$\begin{aligned} H_0(M_i^{n+1,k}) = H_0(M_i^{n+1,k+1}) + \partial_f H_0(M_i^{n+1,k+1})(M_i^{n+1,k} - M_i^{n+1,k+1}) \\ + \Psi(M_i^{n+1,k+1}, M_i^{n+1,k}), \end{aligned} \tag{3.17}$$

$$\begin{aligned} H_0(M_i) = H_0(M_i^{n+1,k+1}) + \partial_f H_0(M_i^{n+1,k+1})(M_i - M_i^{n+1,k+1}) \\ + \Psi(M_i^{n+1,k+1}, M_i). \end{aligned} \tag{3.18}$$

Performing the linear combination

$$\frac{1}{1+\alpha}\Big((3.18) + (\alpha - \sigma^k|\xi|\mathbb{1}_\Xi)(3.17)) + \sigma^k|\xi|\mathbb{1}_\Xi(3.16)\Big),$$

and using (3.15) we obtain

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

which corresponds to equality (3.14) after rearranging the terms.

Next we proceed by induction to show that the kinetic entropy is dissipated at every iteration assuming the CFL condition (3.10) holds for any integer k. The key argument is that under this CFL condition, the term $D_i^{n+1,k+1}$ defines a convex combination of negative quantities, and is thus negative. The initialization is obvious since we have $M_i^{n+1,0} = M_i$, so we focus on the recurrence. We want to show that (3.11) holds at some rank $k \geq 1$ assuming that it is satisfied at rank $k-1$. Under this assumption we can develop (3.9) as

$$\begin{aligned} &H_0(M_i^{n+1,k+1}) \leq \\ &\quad \frac{1}{1+\alpha}\Big(H_0(M_i) + \alpha\Big(H_0(M_i) - \sigma^k\xi\Big(\widetilde{H}_{0,i+1/2}^{n+1,k-1} - \widetilde{H}_{0,i-1/2}^{n+1,k-1}\Big) + \widetilde{D}_i^{n+1,k} + \widetilde{Z}_i^{n+1,k}\Big)\Big) \\ &\quad - \frac{\sigma^k\xi}{1+\alpha}\Big(H_{0,i+1/2}^{n+1,k} - H_{0,i-1/2}^{n+1,k}\Big) + \eta'(U_i^{n+1,k+1}) \cdot \begin{pmatrix}1\\ \xi\end{pmatrix} Q_i^{n+1,k+1} + D_i^{n+1,k+1}, \end{aligned}$$

with $\widetilde{Z}_i^{n+1,k}$ and $\eta'(U_i^{n+1,k+1}) \cdot \begin{pmatrix}1\\ \xi\end{pmatrix} Q_i^{n+1,k+1}$ macroscopically zero as per Remark 3.7. Therefore we have

$$\begin{aligned} &H_0(M_i^{n+1,k+1}) \leq \\ &\quad H_0(M_i) - \frac{\sigma^k\xi}{1+\alpha}\Big(H_{0,i+1/2}^{n+1,k} + \alpha\widetilde{H}_{0,i+1/2}^{n+1,k-1} - H_{0,i-1/2}^{n+1,k} - \alpha\widetilde{H}_{0,i-1/2}^{n+1,k-1}\Big) \\ &\quad + \frac{\alpha}{1+\alpha}\widetilde{Z}_i^{n+1,k} + \eta'(U_i^{n+1,k+1}) \cdot \begin{pmatrix}1\\ \xi\end{pmatrix} Q_i^{n+1,k+1} + \frac{\alpha}{1+\alpha}\widetilde{D}_i^{n+1,k} + D_i^{n+1,k+1} \end{aligned}$$

and the proof is complete by setting

$$\widetilde{H}_{0,i+1/2}^{n+1,k} = \frac{1}{1+\alpha}\Big(H_{0,i+1/2}^{n+1,k} + \alpha\widetilde{H}_{0,i+1/2}^{n+1,k-1}\Big), \quad \widetilde{D}_i^{n+1,k+1} = \frac{\alpha}{1+\alpha}\widetilde{D}_i^{n+1,k} + D_i^{n+1,k+1},$$

$$\widetilde{Z}_i^{n+1,k+1} = \frac{\alpha}{1+\alpha}\widetilde{Z}_i^{n+1,k} + \eta'(U_i^{n+1,k+1}) \cdot \begin{pmatrix} 1 \\ \xi \end{pmatrix} Q_i^{n+1,k+1}.$$

□

3.2 Case of a non flat topography

In presence of a varying bathymetry, one difficulty is to design well-balanced schemes, that is to say schemes that preserve the lake at rest equilibrium given by $(h, q) = (-z, 0)$, where we recall $z(x)$ is a parametrization of the topography at the bottom. The hydrostatic reconstruction technique introduced in [3] is a well-known strategy to circumvent this difficulty. More recently [4], it has been established that when combined with the explicit kinetic scheme, this reconstruction step can prevent the entropy from being dissipated in some testcases featuring a non-flat bathymetry, no matter how refined the time and spatial steps are. This is due to a positive error term appearing in the discrete entropy inequality and which can dominate the dissipation. In this section our goal is to show that an iterative version of this explicit kinetic scheme does not suffer from this defect provided enough sub-iterations are performed. Therefore the iterative strategy, which approximates a fully implicit kinetic scheme, can be considered an improvement over the explicit method in terms of stability. We also emphasize that compared to the fully implicit kinetic scheme proposed in Section 2, the better stability properties of the iterative scheme does not translate in the possibility to use arbitrary large time steps. In fact a CFL condition is required to ensure the convergence of the sub-iterations, as already pointed by the Propositions 3.1 and 3.4 for iterative schemes applied to the flat bottom case.

Let us briefly recall the principle of the hydrostatic reconstruction, which is based on the reconstruction of the water height at every interface. Let $U_i = (h_i, h_i u_i)^T \in \mathbb{R}^2$ denote the vector of quantities of interest over cell $1 \leq i \leq P$, with P the number of interior cells and with ghost cells corresponding to indices 0 and $P+1$. The reconstructed states are vectors from $\mathbb{R}^+ \times \mathbb{R}$ defined on the left and right neighborhoods of each cell interface as follows:

$$\forall 1 \leq i \leq P, \quad U_{i+1/2-} = \begin{pmatrix} h_{i+1/2-} \\ h_{i+1/2-}u_i \end{pmatrix}, \quad U_{i-1/2+} = \begin{pmatrix} h_{i-1/2+} \\ h_{i-1/2+}u_i \end{pmatrix}. \tag{3.19}$$

The reconstructed interfacial water heights are given by

$$h_{i-1/2+} = (h_i + z_i - z_{i-1/2})_+, \quad h_{i+1/2-} = (h_i + z_i - z_{i+1/2})_+, \tag{3.20}$$

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

$$U_i^{n+1} = U_i^n - \sigma\big(F_{i+1/2-}^{n+1} - F_{i-1/2+}^{n+1}\big), \tag{3.21}$$

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

$$\begin{cases} F_{i+1/2-}^{n+1} = \mathcal{F}(U_{i+1/2-}^{n+1}, U_{i+1/2+}^{n+1}) + \frac{g}{2}\begin{pmatrix} 0 \\ (h_i^{n+1})^2 - (h_{i+1/2-}^{n+1})^2 \end{pmatrix} \\ F_{i-1/2+}^{n+1} = \mathcal{F}(U_{i-1/2-}^{n+1}, U_{i-1/2+}^{n+1}) + \frac{g}{2}\begin{pmatrix} 0 \\ (h_i^{n+1})^2 - (h_{i-1/2+}^{n+1})^2 \end{pmatrix} \end{cases}. \tag{3.22}$$

We recall that in our case the upwinding of the numerical flux $\mathcal{F}$ is induced at the kinetic level according to definition (3.4). In (3.22), the terms in factor of $g/2$ look like pressure variations and are consistent with the topography source term when we subtract them. In fact, owing to (3.20) we formally have

$$\frac{g}{2}\Big((h_{i+1/2-}^{n+1})^2 - (h_{i-1/2+}^{n+1})^2\Big) = g\frac{h_{i+1/2-}^{n+1} + h_{i-1/2+}^{n+1}}{2}(h_{i+1/2-}^{n+1} - h_{i-1/2+}^{n+1}) \approx gh_i^{n+1}(z_{i+1/2} - z_{i-1/2}).$$

This discretization of the topography source term can be interpreted at the kinetic level through the relations

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

and we introduce the following notations for convenience

$$\begin{cases} \delta M_{i+1/2-}(\xi) = (\xi - u_i)(M(U_i,\xi) - M(U_{i+1/2-},\xi)), \\ \delta M_{i-1/2+}(\xi) = (\xi - u_i)(M(U_i,\xi) - M(U_{i-1/2+},\xi)). \end{cases} \tag{3.23}$$

An issue with the nonlinear update (3.21) is that it cannot be solved analytically. Instead we will approximate it by an iterative process with a relaxation parameter $\alpha > 0$ similar to the one from Section 3.1, and which reads

$$\forall 1 \leq i \leq P, \qquad (1+\alpha)U_i^{n+1,k+1} = U_i + \alpha U_i^{n+1,k} - \sigma^k(F_{i+1/2-}^{n+1,k} - F_{i-1/2+}^{n+1,k}), \tag{3.24}$$

where we can choose to take $U^{n+1,0} = U^n$ as the initialization. If the sequence defined through (3.24) converges, we recover the implicit scheme (3.21) by setting the macroscopic update as $U^{n+1} = \lim_{k\to\infty} U^{n+1,k}$ and $\sigma = \lim_{k\to\infty} \sigma^k$. In practice we will stop the sub-iterations for some k large enough such that $U^{n+1,k} \approx U^{n+1}$. We also comment on the fact that when the bathymetry is flat ($z \equiv$ Cst) the hydrostatic reconstruction becomes transparent, in the sense that the scheme (3.24) coincides with (3.3).

At the kinetic level, the recurrence relation (3.24) consists in introducing for any real ξ the sequence $(f^{n+1,k}(\xi))_{k\in\mathbb{N}} \subset \mathbb{R}_+^P$ initialized with $f^{n+1,0}(\xi) = M(U^n,\xi)$ and defined recursively as:

$$\begin{cases} (1+\alpha)f_i^{n+1,k+1} = M_i + \alpha M_i^{n+1,k} - \sigma^k\Big(\xi(M_{i+1/2}^{n+1,k} - M_{i-1/2}^{n+1,k}) + \delta M_{i+1/2-}^{n+1,k} - \delta M_{i-1/2+}^{n+1,k}\Big) \\ M_i^{n+1,k+1} = f^{n+1,k+1} + \Delta t^k Q^{n+1,k+1} \end{cases}, \tag{3.25}$$

where the interfacial Maxwellians $M_{i\pm1/2}$ are defined through the following upwinding

$$M_{i\pm1/2}(\xi) = \mathbb{1}_{\xi>0} M(U_{i\pm1/2-},\xi) + \mathbb{1}_{\xi<0} M(U_{i\pm1/2+},\xi) \qquad \forall 1 \leq i \leq P, \tag{3.26}$$

and where we have used the notations (3.23).

We now consider the question of whether or not the iterative kinetic scheme with hydrostatic reconstruction (3.25)-(3.26) is structure preserving. We recall that in our context, structure preserving means the ability to keep the water height positive, to preserve the lake at rest steady state and to satisfy a discrete entropy inequality. When considering a flat bathymetry, this latter point was proved by making use of Lemma 2.4, providing in particular an estimate of the discrete spatial variation of the flux, namely

$$\partial_f H(M_i)(M_i - M_{i\pm1}) = H(M_i) - H(M_{i\pm1}) + \Psi(M_i, M_{i\pm1}),$$

which involves a conservative difference and the positive function Ψ defined in (2.6) responsible for the dissipation. In presence of a varying bathymetry however, fluxes are computed after performing the hydrostatic reconstruction, and we now need an estimate for the following quantity

$$\partial_f H(M_i)(\xi(M_{i+1/2} - M_{i-1/2}) + \delta M_{i+1/2-} - \delta M_{i-1/2+}),$$

which appears when multiplying the first line of (3.25) by $\partial_f H(M_i^{n+1,k},\xi)$. Because the derivative of the kinetic entropy is multiplied by quantities all different from M_i, we can no longer use Lemma 2.4. Instead, we will need the following inequality that corresponds to the Proposition 3.1 from [4].

Proposition 3.8 (Audusse, Bouchut, Bristeau, and Sainte-Marie [4]). *One has*

$$-\partial_f H(M_i^{n+1,k},z_i)\Big[\xi(M_{i+1/2}^{n+1,k}-M_{i-1/2}^{n+1,k})+\delta M_{i+1/2-}^{n+1,k}-\delta M_{i-1/2+}^{n+1,k}\Big]\leq \widetilde{G}_{i+1/2-}^{n+1,k}-\widetilde{G}_{i-1/2+}^{n+1,k}\ , \tag{3.27}$$

with $\widetilde{G}_{i+1/2-}^{n+1,k}$ *and* $\widetilde{G}_{i-1/2+}^{n+1,k}$ *defined by*

(3.28)

$$\begin{aligned}\widetilde{G}_{i\pm1/2\mp}&=\xi\mathbb{1}_{\xi<0}H(M_{i\pm1/2+},z_{i\pm1/2})+\xi\mathbb{1}_{\xi>0}H(M_{i\pm1/2-},z_{i\pm1/2})\\&+\xi H(M_i,z_i)-\xi H(M_{i\pm1/2\mp},z_{i\pm1/2})+\Big(\eta'(U_i)\cdot\begin{pmatrix}1\\ \xi\end{pmatrix}+gz_i\Big)(\xi M_{i\pm1/2\mp}-\xi M_i+\delta M_{i\pm1/2\mp})\ .\end{aligned}$$

Moreover, the difference $\widetilde{G}_{i+1/2-}^{n+1,k}-\widetilde{G}_{i-1/2+}^{n+1,k}$ *is macroscopically conservative*

$$\int_{\mathbb{R}}(\widetilde{G}_{i+1/2-}^{n+1,k}-\widetilde{G}_{i-1/2+}^{n+1,k})\,\mathrm{d}\xi=\overline{G}_{i+1/2}^{n+1,k}-\overline{G}_{i-1/2}^{n+1,k},$$

with the macroscopic entropy flux

$$\overline{G}_{i+1/2}=\int_{\mathbb{R}}\xi\Big(\mathbb{1}_{\xi>0}H(M_{i+1/2-},z_{i+1/2})+\mathbb{1}_{\xi<0}H(M_{i+1/2+},z_{i+1/2})\Big)\,\mathrm{d}\xi. \tag{3.29}$$

We are now able to state our result.

Proposition 3.9 *The scheme (3.24) is structure preserving, in the sense that we have the three properties below:*

(i) the sub-iterations are well balanced, that is to say $U^{n+1,k}=U^n$ *for all* $k\in\mathbb{N}$ *if* U^n *is a discrete lake at rest;*

(ii) assuming that the water height vectors h^n *and* $h^{n+1,k}$ *are positive, the update* $h^{n+1,k+1}$ *defined in the iterative scheme (3.24) is also positive if for all* $1\leq i\leq P$ *the CFL condition* $\sigma^k|\xi|\leq\alpha+M_i/M_i^{n+1,k}$ *holds for any* ξ *belonging to* $\operatorname{supp}M^{n+1,k}$*;*

(iii) the kinetic entropy of the iterative process (3.25) verifies the following kinetic entropy inequality

$$\begin{aligned}H(M_i^{n+1,k+1},z_i)\leq\ &H(M_i,z_i)-\sigma^k(\widetilde{G}_{i+1/2-}^{n+1,k}-\widetilde{G}_{i-1/2+}^{n+1,k})+(1+\alpha)\Big(\eta'(U_i^{n+1,k+1})\cdot\begin{pmatrix}1\\ \xi\end{pmatrix}+gz_i\Big)Q_i^{n+1,k+1}\\&+\alpha(H(M_i^{n+1,k},z_i)-H(M_i^{n+1,k+1},z_i))\\&+(1+\alpha)(\Psi(M_i^{n+1,k},f_i^{n+1,k+1})-\Psi(M_i^{n+1,k+1},f_i^{n+1,k+1}))-\Psi(M_i^{n+1,k},M_i),\end{aligned} \tag{3.30}$$

where $Q_i^{n+1,k+1}=M_i^{n+1,k+1}-f_i^{n+1,k+1}$ *is a collision term verifying the conservation constraints (1.7), where* Ψ *defined in (2.6) is positive and where* $\widetilde{G}_{i+1/2-}^{n+1,k}$ *and* $\widetilde{G}_{i-1/2+}^{n+1,k}$ *are defined by (3.28).*

Before giving the proof we make the following remark.

Remark 3.10 *We reiterate the comments made in Remark 3.7 which are to say that in (3.30) the term*

$$\Big(\eta'(U_i^{n+1,k+1})\cdot\begin{pmatrix}1\\ \xi\end{pmatrix}+gz_i\Big)Q_i^{n+1,k+1}$$

is macroscopically zero since $Q_i^{n+1,k+1}$ *is a collision term satisfying the conservation constraints (1.7). Besides, assuming the method converges as* $k\to\infty$*, the quantity*

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

will eventually become negligible compared to $-\Psi(M_i^{n+1,k}, M_i) < 0$ *from some rank* k*. Integrating inequality (3.30) over* $\xi \in \mathbb{R}$*, this implies that there exists* $K \in \mathbb{N}$ *such that for any* $k \geq K$ *the fully discrete entropy inequality*

$$\eta(U_i^{n+1,k+1}) \leq \eta(U_i^n) - \sigma^k(\overline{G}_{i+1/2}^{n+1,k} - \overline{G}_{i-1/2}^{n+1,k}) \tag{3.31}$$

is satisfied at the macroscopic level, with $\overline{G}_{i+1/2}^{n+1,k}$ *given in (3.29). Summing inequality (3.31) over every cell* $1 \leq i \leq P$ *we obtain the dissipation of the total energy up to boundary fluxes*

$$\frac{1}{\Delta t^k}\sum_{i=1}^{P}\Big(\eta(U_i^{n+1,k+1}) - \eta(U_i^n)\Big) + \frac{1}{\Delta x}\bigg(\int_{\mathbb{R}} \xi\, H_{P+1/2}^{n+1,k}(\xi)\,\mathrm{d}\xi - \int_{\mathbb{R}} \xi\, H_{1/2}^{n+1,k}(\xi)\,\mathrm{d}\xi\bigg) \leq 0. \tag{3.32}$$

In addition to the usual tolerance criterion where the iterations are stopped whenever two successive iterates are sufficiently close to each other, we can use (3.32) as a complementary condition to ensure the dissipation of total energy.

Proof of prop. 3.9. The proof makes use of the kinetic writing (3.25) of scheme (3.24).

(i) The well-balancedness is a consequence of the hydrostatic reconstruction, see [3].

(ii) Remarking that the quantity $\delta M_{i+1/2-}^{n+1,k} - \delta M_{i-1/2+}^{n+1,k}$ appearing in the last line of (3.25) defines an odd function of $\xi - u_i^{n+1,k}$, its integral over $\xi \in \mathbb{R}$ vanishes and we have at the macroscopic level

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

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

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

By definition of the water height reconstruction (3.20), we have the inequalities $h_{i+1/2-}^{n+1,k} \leq h_i^{n+1,k}$ and $h_{i-1/2+}^{n+1,k} \leq h_i^{n+1,k}$. As a consequence $M_{i+1/2-}^{n+1,k} \leq M_i^{n+1,k}$ and $M_{i-1/2+}^{n+1,k} \leq M_i^{n+1,k}$, which allows us to bound the integrand from below by

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

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

(iii) We start to rewrite (3.25) as

$$\begin{aligned}(1+\alpha)(f_i^{n+1,k+1} - M_i^{n+1,k}) = \\ (M_i - M_i^{n+1,k}) - \sigma^k \xi (M_{i+1/2}^{n+1,k} - M_{i-1/2}^{n+1,k} + \delta M_{i+1/2-}^{n+1,k} - \delta M_{i-1/2+}^{n+1,k}).\end{aligned} \tag{3.33}$$

The strategy is to multiply (3.33) by $\partial_f H(M_i^{n+1,k}, z_i)$ and to separate the discrete time variation from the flux and source contributions as

$$\begin{aligned}&\partial_f H(M_i^{n+1,k}, z_i)\Big[(1+\alpha)(f_i^{n+1,k+1} - M_i^{n+1,k}) - (M_i - M_i^{n+1,k})\Big] = \\ &\quad - \sigma^k \partial_f H(M_i^{n+1,k}, z_i)\Big[\xi(M_{i+1/2}^{n+1,k} - M_{i-1/2}^{n+1,k}) + \delta M_{i+1/2-}^{n+1,k} - \delta M_{i-1/2+}^{n+1,k}\Big]\ .\end{aligned} \tag{3.34}$$

We apply Lemma 2.4 to the left hand side of (3.34) to get

$$\partial_f H(M_i^{n+1,k}, z_i)\Big[(1+\alpha)(f_i^{n+1,k+1} - M_i^{n+1,k}) - (M_i - M_i^{n+1,k})\Big] = \tag{3.35}$$

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

Furthermore, an upper bound on the right hand side of (3.34) is obtained by applying Proposition 3.8 which directly yields

(3.36)
$$-\partial_f H(M_i^{n+1,k},z_i)\Big[\xi(M_{i+1/2}^{n+1,k}-M_{i-1/2}^{n+1,k})+\delta M_{i+1/2-}^{n+1,k}-\delta M_{i-1/2+}^{n+1,k}\Big]\leq \widetilde{G}_{i+1/2-}^{n+1,k}-\widetilde{G}_{i-1/2+}^{n+1,k},$$

with $\widetilde{G}_{i+1/2-}^{n+1,k}$ and $\widetilde{G}_{i-1/2+}^{n+1,k}$ defined by (3.28). Injecting equality (3.35) and inequality (3.27) into (3.34) we obtain

$$\begin{aligned}(1+\alpha)H(f_i^{n+1,k+1},z_i)\leq H(M_i,z_i)-\sigma^k(\widetilde{G}_{i+1/2-}^{n+1,k}-\widetilde{G}_{i-1/2+}^{n+1,k})+\alpha H(M_i^{n+1,k},z_i)\\ +(1+\alpha)\Psi(M_i^{n+1,k},f_i^{n+1,k+1})-\Psi(M_i^{n+1,k},M_i)\end{aligned}\tag{3.37}$$

Using again Lemma 2.4 we can also write

$$\begin{aligned}H(f_i^{n+1,k+1},z_i)&=H(M_i^{n+1,k+1},z_i)+\partial_f H(M_i^{n+1,k+1},z_i)(f_i^{n+1,k+1}-M_i^{n+1,k+1})\\&\quad+\Psi(M_i^{n+1,k+1},f_i^{n+1,k+1})\\&\geq H(M_i^{n+1,k+1},z_i)-\Big(\eta'(U_i^{n+1,k+1})\cdot\begin{pmatrix}1\\ \xi\end{pmatrix}+gz_i\Big)(M_i^{n+1,k+1}-f_i^{n+1,k+1})\\&\quad+\Psi(M_i^{n+1,k+1},f_i^{n+1,k+1}),\end{aligned}\tag{3.38}$$

where we used $\partial_f H=\partial_f H_0+gz$ and (3.13) to get (3.38). Combining this inequality with (3.37) we finally get

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

After rearranging the terms and using $Q_i^{n+1,k+1}=-(f_i^{n+1,k+1}-M_i^{n+1,k+1})$ we obtain the desired kinetic entropy inequality (3.30).

□

The iterative kinetic scheme with hydrostatic reconstruction is validated through the numerical experiments performed in Section 5.2.

4 Towards the two dimensional Saint-Venant system

Being able to simulate the two dimensional Saint-Venant system is important if one wishes to deal with more relevant applications. In this Section, we investigate the possibility to extend the previous methods to the higher dimension, while retaining the good stability properties obtained in the one dimensional case. The following lines should be regarded as more exploratory, and in particular we leave for a future work the in-depth study of the two dimensional schemes that we briefly discuss below. In Section 4.1 we give the two dimensional Saint-Venant system and its kinetic interpretation. Then in Section 4.2 we derive a fully implicit scheme over flat bathymetry, and finally in Section 4.3 we propose an iterative kinetic scheme with hydrostatic reconstruction. In all cases a Cartesian mesh is considered.

4.1 The two dimensional system and its kinetic representation

We consider the two dimensional Saint-Venant system written, with obvious notations, under the form

$$\left\{\begin{array}{l} \dfrac{\partial h}{\partial t}+\dfrac{\partial hu}{\partial x}+\dfrac{\partial hv}{\partial y}=0 \\ \dfrac{\partial hu}{\partial t}+\dfrac{\partial}{\partial x}\Big(hu^2+\dfrac{g}{2}h^2\Big)+\dfrac{\partial}{\partial y}(huv)=-gh\dfrac{\partial z}{\partial x} \\ \dfrac{\partial hv}{\partial t}+\dfrac{\partial}{\partial x}(huv)+\dfrac{\partial}{\partial y}\Big(hv^2+\dfrac{g}{2}h^2\Big)=-gh\dfrac{\partial z}{\partial y} \end{array}\right. . \tag{4.1}$$

with (u,v) the horizontal velocity vector. A straightforward extension of Lemma 1.1 allows to get a kinetic interpretation of system (4.1) as studied in [6, 1]. To build the two dimensional Gibbs equilibrium, we define the function

$$\chi(z_1,z_2)=\frac{1}{4\pi}\mathbb{1}_{z_1^2+z_2^2\leq 4}, \tag{4.2}$$

and we have

$$\forall U=(h,hu,hv)^T,\ \forall \xi\in\mathbb{R}^2,\quad M(U,\xi)=\frac{h}{c^2}\chi\Big(\frac{\xi_1-u}{c},\frac{\xi_2-v}{c}\Big)=\frac{1}{2g\pi}\mathbb{1}_{|\xi-(u,v)^T|_2\leq\sqrt{2gh}}, \tag{4.3}$$

with the velocity $c=\sqrt{\frac{g}{2}h}$. Let us remark that when averaging this Maxwellian function in the direction ξ_1 (resp. ξ_2), we get a function of ξ_2 (resp. ξ_1) that coincides with the half-disk Maxwellian (1.4) used in the one dimensional case. We then state the following lemma.

Lemma 4.1 *If the topography $z(x,y)$ is Lipschitz continuous, then $U=(h,hu,hv)^T$ is a weak solution to the Saint-Venant system (4.1) if and only if $M(U,\xi)$ satisfies the kinetic equation*

$$\partial_t M+\xi\cdot\nabla_{(x,y)}M-g\nabla_{(x,y)}z\cdot\nabla_{(\xi_1,\xi_2)}M=Q, \tag{4.4}$$

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

$$\int_{\mathbb{R}^2}Q\,\mathrm{d}\xi_1\mathrm{d}\xi_2=\int_{\mathbb{R}^2}\xi_1 Q\,\mathrm{d}\xi_1\mathrm{d}\xi_2=\int_{\mathbb{R}^2}\xi_2 Q\,\mathrm{d}\xi_1\mathrm{d}\xi_2=0. \tag{4.5}$$

Proof of Lemma 4.1. The proof relies on simple computations. Classically, the integral of Equation (4.4) over $\xi\in\mathbb{R}^2$ gives the first line of the system (4.1) whereas the last two lines of this system are obtained by taking the scalar product of (4.4) against $\xi\in\mathbb{R}^2$ and by integrating over $\mathbb{R}^2$. □

As in the one dimensional case, the Maxwellian (4.3) is associated with a kinetic entropy $H(f,\xi)$ satisfying the two dimensional analogous of Lemma 1.2, see [5]. This kinetic entropy is given by

$$H(f,\xi)=\frac{|\xi|^2}{2}f+gzf. \tag{4.6}$$

4.2 Fully implicit kinetic scheme over a flat bathymetry

Let us consider a bounded rectangular domain $\Omega=(0,L_x)\times(0,L_y)$ and its Cartesian discretization using P cells in both x- and y-directions, such that the total number of cells is P^2. We denote $(P_{i,j})_{0\leq i,j\leq P}$ the vertices with coordinates $(x_i,y_j)^T$ given by

$$x_i=(i+1/2)\Delta x,\qquad y_j=(j+1/2)\Delta y,$$

where $\Delta x=L_x/P$, $\Delta y=L_y/P$. We use the following notations:

- for all $(i,j)\in[\![1,P]\!]^2$, $C_{i,j}$ is the rectangular cell centered on $P_{i,j}$ with area $|C_{i,j}|=\Delta x\Delta y$,

- $\partial C_{i,j}$ is the boundary of $C_{i,j}$,
- indices (i,j) such that $i \in \{0, P+1\}$ or $j \in \{0, P+1\}$ indicate a ghost cell outside of Ω.

We define the piecewise constant functions $U^n(x,y)$ and $z(x,y)$ on cells $C_{i,j}$ as

$$U^n(x,y) = U^n_{i,j}, \quad z(x,y) = z_{i,j}, \quad \text{for } (x,y) \in C_{i,j}, \tag{4.7}$$

with $U^n_{i,j} = (h^n_{i,j}, (hu)^n_{i,j}, (hv)^n_{i,j})^T$ such that they approximate the cell averages as below

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

In the case of a flat topography, the integral over $C_{i,j}$ of the convective part of the kinetic equation (4.4) gives

$$\begin{aligned}\int_{C_{i,j}} \Big(\frac{\partial M}{\partial t} + \xi\cdot\nabla_{(x,y)}M\Big)\,\mathrm{d}x\mathrm{d}y &\approx |C_{i,j}|\frac{\partial M_{i,j}}{\partial t} + \int_{\partial C_{i,j}} (\xi\cdot n_{\partial C_{i,j}})M_{i,j}\,\mathrm{d}\ell \\ &\approx |C_{i,j}|\frac{\partial M_{i,j}}{\partial t} + \xi_1\Delta y\Big(\mathbb{1}_{\xi_1>0}(M_{i,j}-M_{i-1,j}) + \mathbb{1}_{\xi_1<0}(M_{i+1,j}-M_{i,j})\Big) \\ &\quad + \xi_2\Delta x\Big(\mathbb{1}_{\xi_2>0}(M_{i,j}-M_{i,j-1}) + \mathbb{1}_{\xi_2<0}(M_{i,j+1}-M_{i,j})\Big),\end{aligned} \tag{4.8}$$

with $M_{i,j}(t) = M(U_{i,j}(t),\xi)$ and $n_{\partial C_{i,j}}(\ell)$ the normal to $\partial C_{i,j}$ at the contour abscissa $\ell \in \partial C_{i,j}$ taken outward to $C_{i,j}$. From the spatial semi-discretization (4.8) of the convection operator, we deduce an implicit Euler scheme for the kinetic interpretation (4.4)

$$f^{n+1}_{i,j} = M^n_{i,j} - \sigma_x\xi_1(f^{n+1}_{i+1/2,j} - f^{n+1}_{i-1/2,j}) - \sigma_y\xi_2(f^{n+1}_{i,j+1/2} - f^{n+1}_{i,j-1/2}), \tag{4.9}$$

where we used $\sigma_x = \Delta t^n/\Delta x$, $\sigma_y = \Delta t^n/\Delta y$ and the interfacial kinetic densities

$$f^{n+1}_{i+1/2,j} = \mathbb{1}_{\xi_1>0}f^{n+1}_{i,j} + \mathbb{1}_{\xi_1<0}f^{n+1}_{i+1,j}, \qquad f^{n+1}_{i,j+1/2} = \mathbb{1}_{\xi_2>0}f^{n+1}_{i,j} + \mathbb{1}_{\xi_2<0}f^{n+1}_{i,j+1}.$$

Denoting f the vector of interior values from $\mathbb{R}^{P^2}$

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

the kinetic scheme (4.9) also writes

$$(\mathbf{I}_{P^2} + \mathbf{L}_{P^2})f^{n+1} = M + B^{n+1}, \tag{4.10}$$

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

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

where $\mathbf{D}, \mathbf{N}^\pm$ are $P\times P$ matrices defined by $\mathbf{N}^+ = -\sigma_y\xi_2\mathbb{1}_{\xi_2<0}\mathbf{I}_P$, $\mathbf{N}^- = \sigma_y\xi_2\mathbb{1}_{\xi_2>0}\mathbf{I}_P$ and

$$\mathbf{D} = \sigma_x|\xi_1|\begin{pmatrix} 1 & -\mathbb{1}_{\xi_1<0} & 0 & \ldots & 0 \\ -\mathbb{1}_{\xi_1>0} & 1 & -\mathbb{1}_{\xi_1<0} & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -\mathbb{1}_{\xi_1>0} & 1 & -\mathbb{1}_{\xi_1<0} \\ 0 & \ldots & 0 & -\mathbb{1}_{\xi_1>0} & 1 \end{pmatrix} + \sigma_y|\xi_2|\mathbf{I}_P.$$

Likewise, the vector $B^{n+1}(\xi) \in \mathbb{R}^P$ accounting for the boundary conditions is given by

$$B^{n+1} = \begin{pmatrix} \mathbf{N}^- & \mathbf{D}_1^- & \mathbf{D}_1^+ & 0 \\ 0 & \mathbf{D}_2^- & \mathbf{D}_2^+ & \vdots \\ \vdots & \vdots & \vdots & 0 \\ 0 & \mathbf{D}_P^- & \mathbf{D}_P^+ & \mathbf{N}^+ \end{pmatrix} \begin{pmatrix} M_{\text{bottom}} \\ M_{\text{left}} \\ M_{\text{right}} \\ M_{\text{top}} \end{pmatrix}, \quad \begin{cases} M_{\text{bottom}} & = (M_{1,0}^{n+1}, \dots, M_{P,0}^{n+1})^T \\ M_{\text{left}} & = (M_{0,1}^{n+1}, \dots, M_{0,P}^{n+1})^T \\ M_{\text{right}} & = (M_{P+1,1}^{n+1}, \dots, M_{P+1,P}^{n+1})^T \\ M_{\text{top}} & = (M_{1,P+1}^{n+1}, \dots, M_{P,P+1}^{n+1})^T \end{cases},$$

with $(\mathbf{D}_k^-)_{i,j} = \delta_{1,i}\delta_{k,j}\sigma_x|\xi_1|\mathbb{1}_{\xi_1>0}$ and $(\mathbf{D}_k^+)_{i,j} = \delta_{P,i}\delta_{k,j}\sigma_x|\xi_1|\mathbb{1}_{\xi_1<0}$ for all $(i,j,k) \in [\![1,P]\!]^3$. Since the matrix $\mathbf{I}_{P^2} + \mathbf{L}_{P^2}$ has the same structure as the matrix $\mathbf{I} + \sigma\mathbf{L}$ studied in Lemma 2.1, and since the kinetic representation (4.4) admits the kinetic entropy (4.6), the results of Propositions 2.2 and 2.3 are still valid. The practical computation of the inverse of the matrix $\mathbf{I}_{P^2} + \mathbf{L}_{P^2}$, of the numerical fluxes at the macroscopic level, and the implementation of the scheme is left for a later work.

4.3 Iterative kinetic scheme with hydrostatic reconstruction

The iterative kinetic scheme (3.2) and its version with hydrostatic reconstruction (3.25) can be extended to the two dimensional case. First we need to detail the reconstruction step over a Cartesian mesh. Note that it is also possible to apply it on unstructured meshes as detailed in [6]. The idea is to apply formulas (3.19)-(3.20) along the normal of a given edge separating two neighboring cells. More precisely we have

$$\forall 1 \le i,j \le P, \quad U_{i\pm1/2\mp,j} = \begin{pmatrix} 1 \\ u_{i,j} \\ v_{i,j} \end{pmatrix} h_{i\pm1/2\mp,j}, \quad U_{i,j\pm1/2\mp} = \begin{pmatrix} 1 \\ u_{i,j} \\ v_{i,j} \end{pmatrix} h_{i,j\pm1/2\mp}, \tag{4.11}$$

with the reconstructed interfacial water heights given by

$$h_{i\pm1/2\mp,j} = (h_{i,j} + z_{i,j} - z_{i\pm1/2,j})_+, \quad h_{i,j\pm1/2\mp} = (h_{i,j} + z_{i,j} - z_{i,j\pm1/2})_+, \tag{4.12}$$

and with the interfacial bathymetry $z_{i+1/2,j} = \max(z_{i,j}, z_{i+1,j})$ and $z_{i,j+1/2} = \max(z_{i,j}, z_{i,j+1})$. It is then possible to consider the Maxwellian densities associated with the reconstructed interfacial states (4.11) through an upwinding as follows

$$M_{i+1/2,j}(\xi) = \mathbb{1}_{\xi_1>0} M\big(U_{i+1/2-,j}, \xi\big) + \mathbb{1}_{\xi_1<0} M\big(U_{i+1/2+,j}, \xi\big),$$
$$M_{i,j+1/2}(\xi) = \mathbb{1}_{\xi_2>0} M\big(U_{i,j+1/2-}, \xi\big) + \mathbb{1}_{\xi_2<0} M\big(U_{i,j+1/2+}, \xi\big).$$

An implicit hydrostatic reconstruction kinetic scheme stemming from (4.9) consists to find macroscopic states $(U_{i,j}^{n+1})_{1\le i,j\le P}$ such that

$$\begin{aligned} M_{i,j}^{n+1} = M_{i,j}^n &- \sigma_x\xi_1(M_{i+1/2,j}^{n+1} - M_{i-1/2,j}^{n+1}) - \sigma_y\xi_2(M_{i,j+1/2}^{n+1} - M_{i,j-1/2}^{n+1}) + Q_{i,j}^{n+1} \\ &+ \sigma_x(\xi_1 - u_{i,j}^{n+1,k})(M_{i+1/2-,j}^{n+1,k} - M_{i-1/2+,j}^{n+1,k}) + \sigma_y(\xi_2 - v_{i,j}^{n+1,k})(M_{i,j+1/2-}^{n+1,k} - M_{i,j-1/2+}^{n+1,k}), \end{aligned}$$

where $Q_{i,j}^{n+1}$ is a collision term verifying (4.5), and where the last line of the above expression accounts for the topography source term. As in the one dimensional case, it is not possible to compute explicitly this update, instead we approximate it by a fixed point method

$$\begin{aligned} (1+\alpha) f_{i,j}^{n+1,k+1} = M_{i,j}^n &+ \alpha M_{i,j}^{n+1,k} - \sigma_x\xi_1(M_{i+1/2,j}^{n+1,k} - M_{i-1/2,j}^{n+1,k}) - \sigma_y\xi_2(M_{i,j+1/2}^{n+1,k} - M_{i,j-1/2}^{n+1,k}) \\ &+ \sigma_x(\xi_1 - u_{i,j}^{n+1,k})(M_{i+1/2-,j}^{n+1,k} - M_{i-1/2+,j}^{n+1,k}) + \sigma_y(\xi_2 - v_{i,j}^{n+1,k})(M_{i,j+1/2-}^{n+1,k} - M_{i,j-1/2+}^{n+1,k}), \end{aligned} \tag{4.13}$$

where $\alpha > 0$ is a relaxation parameter, combined with the projection step

$$M_{i,j}^{n+1,k+1} = f_{i,j}^{n+1,k+1} + Q_{i,j}^{n+1,k}. \tag{4.14}$$

At the macroscopic level, an iterative scheme is obtained by integrating (4.13) against $(1, \xi)$ which is implemented in practice. It reads

(4.15)
$$(1+\alpha)U_{i,j}^{n+1,k+1} = U_{i,j}^{n} + \alpha U_{i,j}^{n+1,k} - \sigma_x\big(F_{x,i+1/2-,j}^{n+1,k} - F_{x,i-1/2+,j}^{n+1,k}\big) - \sigma_y\big(F_{y,i,j+1/2-}^{n+1,k} - F_{y,i,j-1/2+}^{n+1,k}\big),$$

where we have the following definition of the numerical fluxes

$$F_{x,i\pm1/2\mp,j}^{n+1,k} = \int_{\mathbb{R}^2} \xi_1 \begin{pmatrix} 1 \\ \xi_1 \\ \xi_2 \end{pmatrix} M_{i\pm1/2,j}^{n+1,k} \, \mathrm{d}\xi_1 \mathrm{d}\xi_2 + \frac{g}{2} \begin{pmatrix} 0 \\ (h_{i,j}^{n+1,k})^2 - (h_{i\pm1/2\mp,j}^{n+1,k})^2 \\ 0 \end{pmatrix},$$

$$F_{y,i,j\pm1/2\mp}^{n+1,k} = \int_{\mathbb{R}^2} \xi_2 \begin{pmatrix} 1 \\ \xi_1 \\ \xi_2 \end{pmatrix} M_{i,j\pm1/2}^{n+1,k} \, \mathrm{d}\xi_1 \mathrm{d}\xi_2 + \frac{g}{2} \begin{pmatrix} 0 \\ 0 \\ (h_{i,j}^{n+1,k})^2 - (h_{i,j\pm1/2\mp}^{n+1,k})^2 \end{pmatrix}.$$

The results obtained in Section 3 for the 1d Saint-Venant remain valid in the two dimensional setting since, the kinetic representation (4.4) of the Saint-Venant equations with Maxwellian (4.3) possesses a convex kinetic entropy H given by (4.6).

The complete analysis of the 2d scheme is left for a forthcoming paper but a numerical experiment consisting of the two dimensional parabolic bowl [17] is presented in Section 5.3.

5 Numerical simulations

In this section we evaluate the behavior and the efficiency of the proposed kinetic schemes through numerical experiments. In Section 5.1, we focus on the fully implicit kinetic scheme over a flat bathymetry that corresponds to (2.1). Then, in Section 5.2, the iterative kinetic scheme with hydrostatic reconstruction is investigated numerically. In particular, we exhibit a testcase where our iterative method succeeds to dissipate the total energy while its explicit version increases it. Finally in Section 5.3 a two dimensional test case is proposed.

5.1 Fully implicit kinetic scheme in the flat bottom case

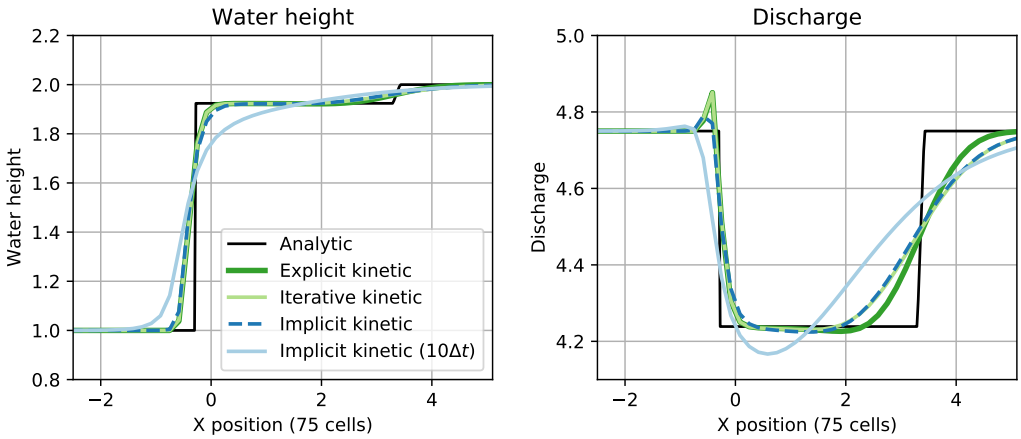

Figure 2: Slow moving shock approximated by various kinetic schemes, including explicit, implicit and iterative strategies. The initial condition is given by a Riemann data with discontinuity at position $x = 0$.

Slow moving shock. To assess the efficiency and interest of the implicit scheme (2.2), we perform a numerical test involving a Riemann problem with a slowly moving shock over a flat bottom. This configuration is achieved for a nearly transcritical flow where the material velocity u is positive and satisfies $u - \sqrt{gh} \approx 0$ and $u + \sqrt{gh} \gg 1$. Hence the maximum eigenvalue severely constrains the time step, however a small time step might not be necessary to accurately resolve the slow shock. We set the gravitational acceleration g to 10 and the Riemann problem corresponds to the initial state $U_L = (1, 4.75)^T$ for $x < 0$ and $U_R = (2, 4.75)^T$ for $x > 0$. In Figure 2 we compare several schemes with an explicit time step Δt_{exp} given by the CFL condition $\Delta t_{\text{exp}} \leq 0.45 \frac{\Delta x}{\lambda_{\max}}$, as well as the implicit kinetic scheme using a time step $\Delta t_{\text{imp}} = 10 \Delta t_{\text{exp}}$. We also use the iterative scheme (3.3) with parameter $\alpha = 1$, and plot the results at time $t = 0.5$. We notice that in the discharge profile, an oscillation appears downwind of the left-going shock, which is quite pronounced for the explicit and iterative kinetic schemes, and less so for the fully implicit ones. As expected the implicit scheme using Δt_{imp} strongly diffuses the fast traveling rarefaction. On the other hand the slow shock seems to be slightly less impacted by the large time steps, however it is still less diffused when using Δt_{exp}. Despite requiring ten times less iterations to reach the final time, the use of large time increments for the implicit kinetic scheme only results in around two percents faster computations compared to the explicit strategy which is due to the high quadratic cost of the implicit method. We believe that it is not possible to lower this cost when it comes to unconditionally stable methods, because the associated stencil has to cover the entire computational domain.

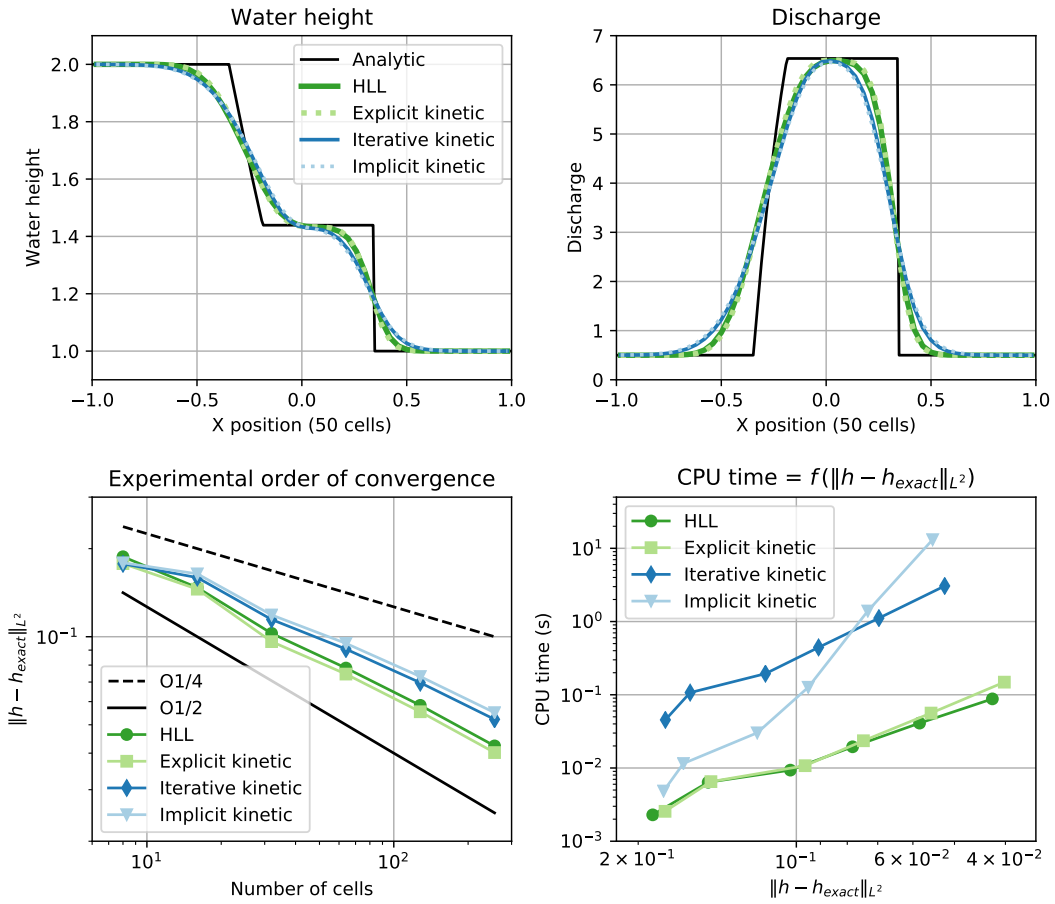

Figure 3: Comparing implicit, iterative and explicit kinetic solvers on a Riemann problem.

Riemann problem. We compare the fully implicit kinetic scheme and iterative kinetic scheme to explicit methods. The 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 gravity constant g is set to 100 and the solution consists in a 1-rarefaction and a 2-shock. The iterative kinetic scheme uses the half-disk Maxwellian, and we choose the parameters $\alpha = 1$ and $\varepsilon_{\text{tol}} = 10^{-9}$ for the stopping criterion. All the schemes use an explicit time step, and the results are given in Figure 3 at time $t = 0.025$. Three aspects have to be considered, namely the accuracy, the computational cost and the stability. Over Fig. 3, we see that in terms of efficiency both iterative and implicit kinetic schemes are at their disadvantage. Especially, the quadratic complexity of the fully implicit version results in a steeper slope of the efficiency curve. However this is only one part of the picture, and we know from Proposition 3.6 and Remark 3.7 that the iterative kinetic scheme (3.3) satisfies a discrete entropy inequality without restriction on the time step, assuming enough iterations are performed. Concretely the greater stability comes with a higher level of diffusion which is noticeable in the first two plots of Figure 3. This increased diffusion remains within acceptable margin, and is the price to pay to have better stability properties.

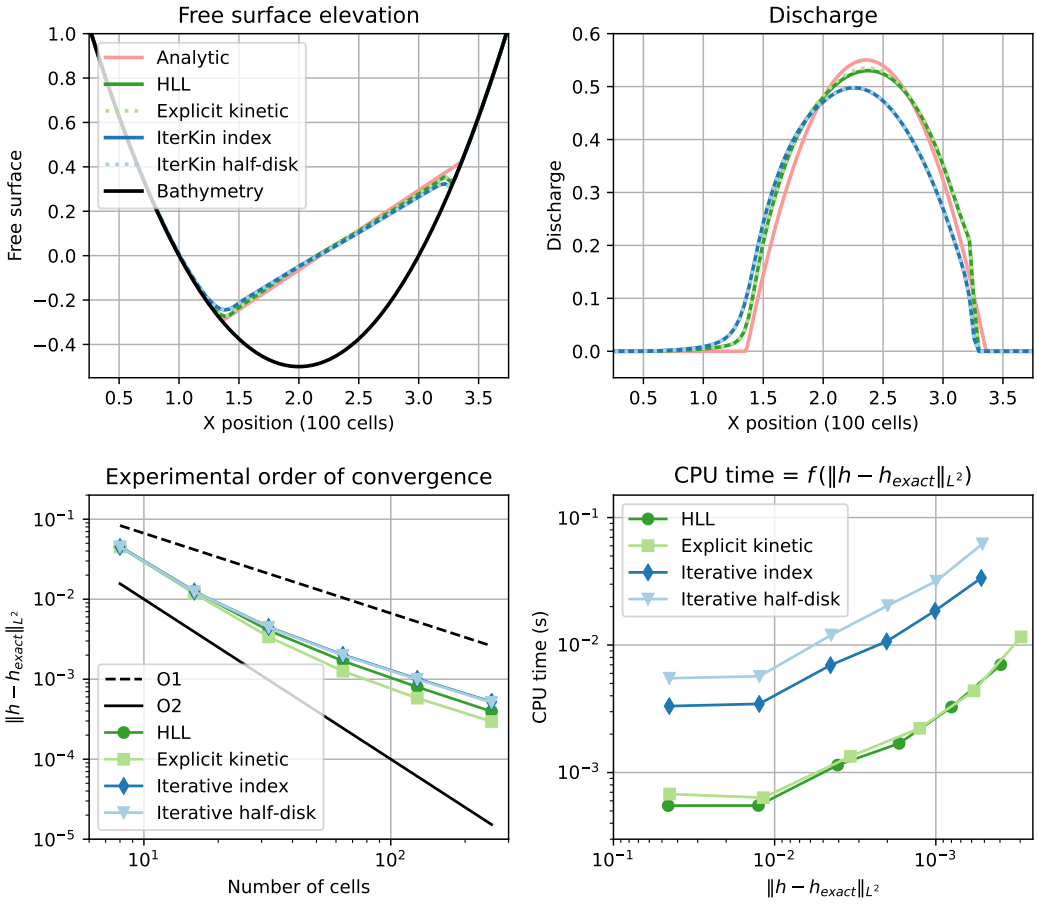

Figure 4: Parabolic bowl approximated by explicit and iterative kinetic schemes. First row: elevation and discharge at time 0.75, second row: convergence and efficiency curves. The stopping criteria used in the two kinetic iterative schemes combines the standard tolerance condition with tolerance $\varepsilon = 10^{-9}$ and the entropy condition (3.32).

5.2 Iterative kinetic scheme with hydrostatic reconstruction

Parabolic bowl. Now we consider the Thacker's testcase, also known as the parabolic bowl testcase, taken from [17]. The analytic solution in the domain $\Omega = (0, L)$ corresponds to

$$\begin{cases} h(t,x) = -\frac{h_0}{a}\left(\left[\left(x - \frac{L}{2}\right) + \frac{1}{2}\cos(2Bt)\right]^2 - 1\right)\mathbb{1}_{x\in W(t)} \\ u(t,x) = B\sin(2Bt)\mathbb{1}_{x\in W(t)} \\ z(x) = h_0\left(\frac{1}{a^2}\left(x - \frac{L}{2}\right)^2 - 1\right) \end{cases}, \quad W(t) = \frac{L}{2} - \frac{1}{2}\cos(2Bt) + (-a, a),$$

where we set $h_0 = 0.5$, $a = 1$, $L = 4$, $B = \frac{1}{2a}\sqrt{2gh_0}$ and $g = 10$. We plot the numerical solution at time $t = 0.75$s in Figure 4. This testcase is relevant as it provides us with a non trivial analytical solution enabling to plot convergence curves, and it is known to be challenging numerically, as it presents a varying bottom together with an evolving wet/dry front and a discontinuous velocity profile. It is interesting to note that the different choice of Maxwellian used in the two iterative kinetic schemes has very little impact on the approximation. In both cases we obtain a convergence with first order accuracy, and unsurprisingly the numerical cost is higher than for fully explicit methods due to the number of sub-iterations required to update the solution. One should also note that the use of the half-disk Maxwellian is slightly more expensive than the simpler index Maxwellian. Besides, in this testcase the iterative kinetic scheme with index Maxwellian was always able to fulfill the entropy condition (3.32) after some iterations, which we only proved rigorously for the half-disk Maxwellian. Hence despite using the wrong Maxwellian, it seems that the iterative kinetic scheme in question still has better stability properties than fully explicit methods. This will be further corroborated with the next testcase.

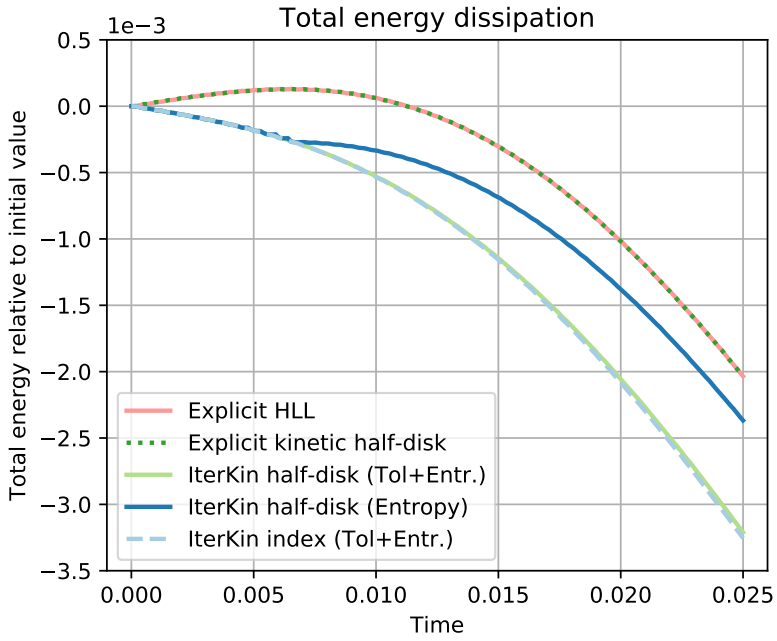

Figure 5: Evolution of the relative total energy obtained for various explicit and iterative kinetic schemes.

Total energy dissipation. To assess the good entropy stability property of our iterative kinetic scheme with hydrostatic reconstruction (3.22)-(3.24), we wish to compare it to its explicit counterpart in a testcase where this latter method fails to dissipate the entropy. To construct such a testcase,

we can use the discrete entropy inequality verified by the explicit kinetic scheme established in [4], Theorem 3.6. It reads as

$$\begin{aligned} H(f_i^{n+1}, z_i) \leq H(M_i^n, z_i) &- \sigma(\widetilde{G}^n_{i+1/2-} - \widetilde{G}^n_{i-1/2+}) \\ &- C_1\sigma|\xi|\Big(\mathbb{1}_{\xi<0}(M_{i+1/2+} + M_{i+1/2-})(M_{i+1/2+} - M_{i+1/2-})^2 \\ &+ \mathbb{1}_{\xi>0}(M_{i-1/2+} + M_{i-1/2-})(M_{i-1/2+} - M_{i-1/2-})^2\Big) \\ &+ C_2(\sigma\xi_{\max})^2 M_i\Big((M_i - M_{i+1/2-})^2 + (M_i - M_{i-1/2+})^2\Big), \end{aligned}$$

with the kinetic entropy fluxes $\widetilde{G}^n_{i+1/2-}, \widetilde{G}^n_{i-1/2+}$ defined in (3.28) and with $C_1, C_2 > 0$ two constants. Since terms in factor of C_1 have negative sign, they play the role of a dissipation; on the other hand the ones in factor of C_2 are a positive error contribution. A possibility is thus to choose an initial condition such that the dissipative term cancels, but not the error one. In the dissipation, quantities of the form $M_{i\pm1/2+} - M_{i\pm1/2-}$ vanish when the free surface elevation $h + z$ and the velocity u are constant across the mesh. At the same time, terms of the form $(M_i - M_{i\pm1/2\mp})$ appearing in the positive error term vanish if the bathymetry is taken flat. Hence we consider an initial condition given by a flat free surface, a non-flat bathymetry and constant velocity which is nonzero (otherwise one would get a lake at rest steady state). More precisely we set the spatial domain to $\Omega = (0, 1)$ with periodic boundary conditions, with $g = 10$ and we have initially

$$\forall x \in (0,1), \quad h^{\text{in}}(x) = -z(x), \quad u^{\text{in}}(x) = 1, \quad z(x) = -5 + \frac{1}{2}\Big(1 + \cos\Big(5\pi\Big(x - \frac{1}{2}\Big)\Big)\Big)\mathbb{1}_{|x-\frac{1}{2}|\leq\frac{1}{5}}.$$

The results can be seen in Figure 5, where we plot the time evolution of the total energy for various schemes over a mesh of $P = 100$ cells.

Interestingly all the iterative methods manage to dissipate the total energy at each time step, even the scheme using the index Maxwellian (2.18), for which there is no proof of discrete entropy inequality. On the contrary, as expected the explicit kinetic scheme with half-disk Maxwellian increases the energy in the first few time steps, after what it decreases. The same goes for the explicit HLL scheme, and as a result these two explicit methods are not entropy stable for this testcase. For comparison we also added in dark blue the iterative kinetic scheme with $\alpha = 0$ and whose sub-iterations stop as soon as the entropy condition (3.32) is verified. We can see that after some time this scheme becomes less dissipative than iterative kinetic methods using the standard tolerance condition; moreover this time roughly corresponds to the time at which the fully explicit schemes stop increasing the entropy.

5.3 Two dimensional simulation

We investigate the two dimensional iterative kinetic scheme with hydrostatic reconstruction given in Section 4.3. The exact solution in the domain $\Omega = (0, L)^2$ can be found in [17], and reads

$$\begin{cases} h(t,x,y) = \frac{2\eta h_0}{a^2}\big(\big[\big(x - \frac{L}{2}\big)\cos(\omega t) + \big(y - \frac{L}{2}\big)\sin(\omega t)\big] - z(x,y)\big)_+ \\ u(t,x,y) = -\eta\omega\sin(\omega t)\mathbb{1}_{h(t,x,y)>0} \\ v(t,x,y) = \eta\omega\cos(\omega t)\mathbb{1}_{h(t,x,y)>0} \end{cases}, \qquad z(x,y) = -h_0\Big(1 - \frac{r^2}{a^2}\Big),$$

with $r = \sqrt{(x - L/2)^2 + (y - L/2)^2}$, $L = 1$, $h_0 = 0.1$, $a = 0.25$, $\eta = 0.2$, $\omega = \sqrt{2gh_0}/a$ and $g = 4$. The initial condition is obtained by setting $t = 0$ in the above expressions. The results are obtained with the 2D version of the iterative kinetic scheme (4.15) with the Maxwellian defined by (4.2)-(4.3) and are displayed in Figures 6 and 7. We see that when increasing the tolerance value to $\varepsilon_{\text{tol}} = 10^{-5}$, the experimental order of convergence of the iterative scheme decreases, which illustrates that a large tolerance error prevents the sub-iterations to converge to the implicit update. On the other hand, the smaller ε_{tol} is, the more iterations are needed to reach the stopping criteria which translates to an increase in computational time.

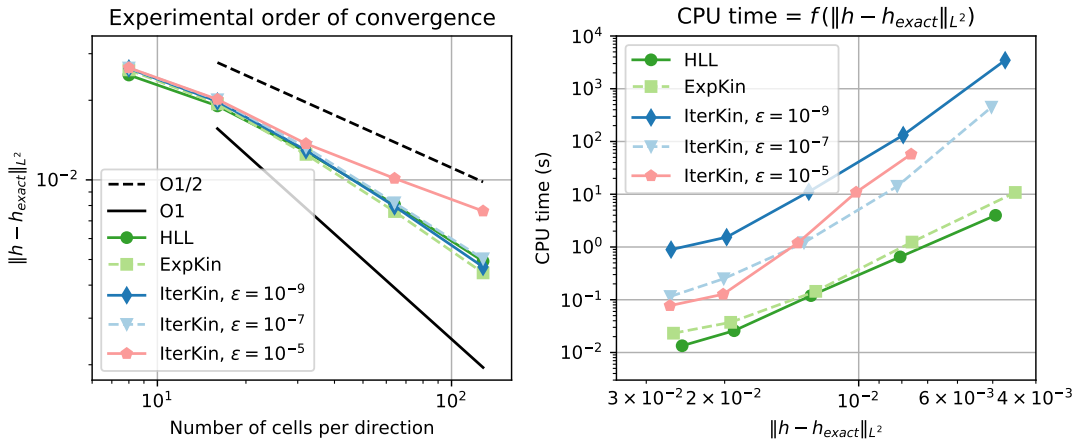

Figure 6: Convergence and efficiency curves obtained with the 2D parabolic bowl testcase with final time $t = 4\pi/\omega$. Different tolerances ε are compared for the iterative kinetic scheme. A CFL constant of $1/2$ was used, except for the case $\varepsilon = 10^{-9}$ where we set it to $1/10$.

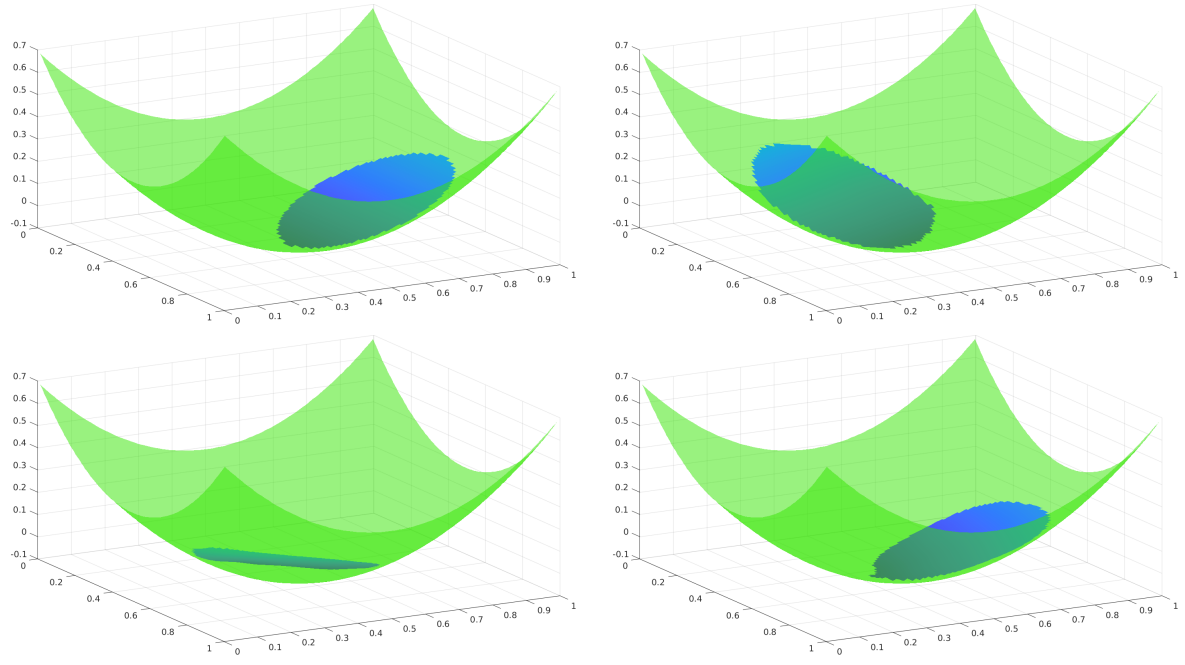

Figure 7: Numerical approximation of the 2D parabolic bowl using the iterative kinetic scheme with $\varepsilon = 10^{-7}$ over a 100×100 mesh. From left to right and top to bottom: initial condition, approximation at time $t = 1.1708$, $t = 2.3416$ and $t = 3.5124$.

Acknowledgments

The authors wish to express their warmest thanks to François Bouchut for many fruitful discussions. Antonin Leprevost and Bilal Al Taki have contributed preliminary versions of the work.

References

[1] Sebastien Allgeyer, Marie-Odile Bristeau, David Froger, Raouf Hamouda, V. Jauzein, Anne Mangeney, Jacques Sainte-Marie, Fabien Souillé, and Martin Valée, *Numerical approximation of the 3d hydrostatic Navier-Stokes system with free surface*, ESAIM, Math. Model. Numer. Anal. **53** (2019), no. 6, 1981–2024 (English).

[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*, Applied Mathematics and Computation **411** (2021), 126469.

[3] Emmanuel Audusse, François Bouchut, Marie-Odile Bristeau, Rupert Klein, and Benoît Perthame, *A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows*, SIAM J. Sci. Comput. **25** (2004), no. 6, 2050–2065 (English).

[4] Emmanuel Audusse, François Bouchut, Marie-Odile Bristeau, and Jacques Sainte-Marie, *Kinetic entropy inequality and hydrostatic reconstruction scheme for the Saint-Venant system*, Math. Comput. **85** (2016), no. 302, 2815–2837 (English).

[5] Emmanuel Audusse and Marie-Odile Bristeau, *A 2d Well-balanced Positivity Preserving Second Order Scheme for Shallow Water Flows on Unstructured Meshes*, Research Report RR-5260, INRIA, 2004.

[6] Emmanuel Audusse and Marie-Odile Bristeau, *A well-balanced positivity preserving "second-order" scheme for shallow water flows on unstructured meshes*, J. Comput. Phys. **206** (2005), no. 1, 311–333 (English).

[7] F. Berthelin and F. Bouchut, *Relaxation to isentropic gas dynamics for a BGK system with single kinetic entropy*, Methods Appl. Anal. **9** (2002), no. 2, 313–327 (English).

[8] Georgij Bispen, Koottungal Revi Arun, Maria Lukàcŏvà-Medvid'ovà, and Sebastian Noelle, *IMEX Large Time Step Finite Volume Methods for Low Froude Number Shallow Water Flows.*, Communications in Computational Physics **16** (2014), 307–347.

[9] F. Bouchut, *Construction of BGK models with a family of kinetic entropies for a given system of conservation laws*, J. Stat. Phys. **95** (1999), no. 1-2, 113–170 (English).

[10] François Bouchut, *Nonlinear stability of finite volume methods for hyperbolic conservation laws and well-balanced schemes for sources.*, Front. Math., Basel: Birkhäuser, 2004 (English).

[11] François Bouchut and Xavier Lhébrard, *Convergence of the kinetic hydrostatic reconstruction scheme for the Saint Venant system with topography*, Math. Comput. **90** (2021), no. 329, 1119–1153 (English).

[12] Bouchut, F., *Entropy satisfying flux vector splittings and kinetic BGK models*, Numer. Math. **94** (2003), no. 4, 623–672 (English).

[13] Marie-Odile Bristeau and Benoit Coussin, *Boundary Conditions for the Shallow Water Equations solved by Kinetic Schemes*, Research Report RR-4282, INRIA, 2001, Projet M3N.

[14] Marie-Odile Bristeau, Nicole Goutal, and Jacques Sainte-Marie, *Numerical simulations of a non-hydrostatic shallow water model*, Comput. Fluids **47** (2011), no. 1, 51–64 (English).

[15] F. Coron and B. Perthame, *Numerical passage from kinetic to fluid equations*, SIAM J. Numer. Anal. **28** (1991), no. 1, 26–42 (English).

[16] David Coulette, Emmanuel Franck, Philippe Helluy, Michel Mehrenberger, and Laurent Navoret, *High-order implicit palindromic discontinuous galerkin method for kinetic-relaxation approximation*, Computers & Fluids **190** (2019), 485–502.

[17] Olivier Delestre, Carine Lucas, Pierre-Antoine Ksinant, Frédéric Darboux, Christian Laguerre, T.-N.-Tuoi Vo, François James, and Stéphane Cordier, *SWASHES: a compilation of shallow water analytic solutions for hydraulic and environmental studies*, Int. J. Numer. Methods Fluids **72** (2013), no. 3, 269–300 (English).

[18] J.-F. Gerbeau and B. Perthame, *Derivation of viscous Saint-Venant system for laminar shallow water; numerical validation*, Discrete Contin. Dyn. Syst., Ser. B **1** (2001), no. 1, 89–102 (English).

[19] Laurent Gosse, *Computing qualitatively correct approximations of balance laws. Exponential-fit, well-balanced and asymptotic-preserving*, SIMAI Springer Ser., vol. 2, Milano: Springer, 2013 (English).

[20] N. Goutal and J. Sainte-Marie, *A kinetic interpretation of the section-averaged Saint-Venant system for natural river hydraulics*, Int. J. Numer. Methods Fluids **67** (2011), no. 7, 914–938 (English).

[21] J. Haack, S. Jin, and J.-G. Liu, *An all-speed asymptotic-preserving method for the isentropic euler and navier-stokes equations*, Communications in Computational Physics **12(4)** (2012), 955–980.

[22] Shi Jin and Lorenzo Pareschi, *Asymptotic-preserving (ap) schemes for multiscale kinetic equations: a unified approach*, Hyperbolic Problems: Theory, Numerics, Applications (Basel) (Heinrich Freistühler and Gerald Warnecke, eds.), Birkhäuser Basel, 2001, pp. 573–582.

[23] B. Perthame, *Boltzmann type schemes for gas dynamics and the entropy property*, SIAM J. Numer. Anal. **27** (1990), no. 6, 1405–1421 (English).

[24] B. Perthame and C. Simeoni, *A kinetic scheme for the Saint-Venant system with a source term*, Calcolo **38** (2001), no. 4, 201–231 (English).

[25] Benoît Perthame, *Kinetic formulation of conservation laws*, Oxf. Lect. Ser. Math. Appl., vol. 21, Oxford: Oxford University Press, 2002 (English).

[26] Perthame, B., *Second-order Boltzmann schemes for compressible Euler equations in one and two space dimensions*, SIAM J. Numer. Anal. **29** (1992), no. 1, 1–19 (English).

[27] Saint-Venant, *Théorie du mouvement non permanent des eaux, avec application aux crues des rivières et à l'introduction des marées dans leur lit.*, C. R. Acad. Sci., Paris **73** (1871), 147–154 (French).

[28] Andrea Thomann, Gabriella Puppo, and Christian Klingenberg, *An all speed second order well-balanced imex relaxation scheme for the euler equations with gravity*, Journal of Computational Physics **420** (2020), 109723.

[29] Yulong Xing and Chi-Wang Shu, *A survey of high order schemes for the shallow water equations*, J. Math. Study **47** (2014), no. 3, 221–249 (English).

A Expression of the numerical fluxes

The optimal choice for the Maxwellian is given by (1.4). Unfortunately the explicit expression for the numerical fluxes appearing in (2.20) is hardly possible with the choice (1.4) and the use of approximate quadrature formula for the integrals in (1.4) will degrade the accuracy of the scheme and increase the

computational costs. Hence, we we choose M defined by the first expression in (2.17) and relation (2.20) becomes

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

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

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

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

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

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

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

Lemma A.2 *Using the same notation as in the previous lemma, we have*

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

Proof of Lemma A.1. We have

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

.

We pose $y = 1 - \frac{1}{1+x}$

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

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

$$I = -\int \sum_{l=0}^{k-1} y^l dy - \ln(|1-y|) + C \qquad\qquad C \in \mathbb{R}$$

$$= \ln(|1+x|) - \sum_{l=1}^{k} \frac{y^l}{l} + C' \qquad\qquad C' \in \mathbb{R}$$

□

Proof of Lemma A.2. We already have denoted $y = \frac{x}{1+x} = 1 - \frac{1}{1+x}$

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

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

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

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

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

$$\begin{aligned} Ah_{i,j} =& \int_{\min(0,a_j^n)}^{\min(0,b_j^n)} \frac{(-\sigma\xi)^{j-i}}{(1-\sigma\xi)^{j-i+1}} d\xi = -\frac{1}{\sigma} \int_{-\min(0,a_j^n)\sigma}^{-\min(0,b_j^n)\sigma} \frac{(x)^{j-i}}{(1+x)^{j-i+1}} dx \\ =& \frac{1}{\sigma} \Big[\ln(|1+x|) - \sum_{l=1}^{j-i} \frac{y^l}{l} \Big]_{-\min(0,a_j^n)\sigma}^{-\min(0,b_j^n)\sigma}. \end{aligned} \tag{A.3}$$

$$\begin{aligned} Bh_{i,j} =& \int_{\max(0,a_j^n)}^{\max(0,b_j^n)} \frac{(\sigma\xi)^{i-j}}{(1+\sigma\xi)^{i-j+1}} d\xi = \frac{1}{\sigma} \int_{\max(0,a_j^n)\sigma}^{\max(0,b_j^n)\sigma} \frac{(x)^{i-j}}{(1+x)^{i-j+1}} dx \\ =& \frac{1}{\sigma} \Big[\ln(|1+x|) - \sum_{l=1}^{i-j} \frac{y^l}{l} \Big]_{\max(0,a_j^n)\sigma}^{\max(0,b_j^n)\sigma} \end{aligned} \tag{A.4}$$

And similarly we obtain the formulas for Ahu and Bhu under the form

$$Ahu_{i,j} = \frac{1}{\sigma} \Big[-(j-i+1)\ln(|1+x|) + x + \sum_{l=1}^{j-i} l \frac{y^{j-i+1-l}}{j-i+1-l} \Big]_{-\min(0,b_j^n)\sigma}^{-\min(0,a_j^n)\sigma} \tag{A.5}$$

$$Bhu_{i,j} = \frac{1}{\sigma} \Big[-(i-j+1)\ln(|1+x|) + x + \sum_{l=1}^{i-j} l \frac{y^{i-j+1-l}}{i-j+1-l} \Big]_{\max(0,a_j^n)\sigma}^{\max(0,b_j^n)\sigma} \tag{A.6}$$

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

$$h_i^{\text{int}} = \frac{1}{2\sigma\sqrt{3}} \Big(\sum_{j=i}^{P} \sqrt{\frac{2h_j^n}{g}} \Big[\ln(|1+x|) - \sum_{l=1}^{j-i} \frac{y^l}{l} \Big]_{-\min(0,b_j^n)\sigma}^{-\min(0,a_j^n)\sigma} \tag{A.7}$$

$$+ \sum_{j=1}^{i} \sqrt{\frac{2h_j^n}{g}} \Big[\ln(|1+x|) - \sum_{l=1}^{i-j} \frac{y^l}{l} \Big]_{\max(0,a_j^n)\sigma}^{\max(0,b_j^n)\sigma} \Big) \tag{A.8}$$

(A.9)

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

$$+\sum_{j=1}^{i}\sqrt{\frac{2h_j^n}{g}}\Big[-(i-j+1)\ln(|1+x|)+x+\sum_{k=1}^{i-j}(i-j+1-k)\frac{y^k}{k}\Big]_{\max(0,a_j^n)\sigma}^{\max(0,b_j^n)\sigma}\Big) \tag{A.10}$$

B Computations of the fluxes involving the boundary conditions

We assume the ghost quantities U_0^{n+1} and U_{P+1}^{n+1} at time t^{n+1} to be known. The exterior contribution given in (2.19) also writes

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

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

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

or equivalently

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

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

As explained in Section 2.4, in practice we will replace the unknown values of U_0^{n+1}, U_{P+1}^{n+1} with that of U_0^n, U_{P+1}^n. The expression of h_i^{ext} can then be established by the mean of Lemma A.2. We have now to find an analytic expression for the quantity $\int\frac{x^{k+1}}{(1+x)^k}dx$ in order to obtain the final expression $(hu)_i^{\text{ext}}$. The following lemma holds.

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

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

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

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

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

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

Dividing this by $1-y$ leads to

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

Iterating this one more time we find

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

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

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

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 deduce that

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

As a conclusion we have the expression

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

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