An entropy stable vectorial BGK scheme for the Saint-Venant system with hydrostatic reconstruction and adaptive time stepping

Denise Aregba-Driollet* Stéphane Brull† Mathieu Rigal‡§

April 7, 2026

Abstract

In this paper, we discretize the Saint-Venant system through kinetic schemes with hydrostatic reconstruction and adaptive time stepping, aiming to satisfy a fully discrete entropy inequality while preserving the water height non-negativity and the hydrostatic equilibrium. This method is based on vectorial BGK models with finitely many velocities, where the contribution of the bathymetry is accounted as a force term involving a shift matrix. Over a flat bathymetry, we obtain a fully discrete entropy inequality for all entropies. When the bathymetry is variable, we quantify precisely the variation of energy of the system. To ensure that the latter is always dissipated, an implicit integrator can be used. More precisely, we consider some local-in-space and dynamical-in-time convex combination between the forward and backward Euler strategies, improving accuracy by limiting unnecessary diffusion and reducing the computational cost. Finally, we explain how this scheme can be implemented by means of a fixed point method, and we evaluate its interest through numerical experiments.

Keywords: Saint-Venant system, Discrete BGK approximation, Fully discrete entropy inequality, Adaptive time stepping, Fixed point method.

1 Introduction

The study of coastal oceanography and fluvial hydrology relies on the ability to accurately and efficiently model free surface flows over a variable bathymetry. In this regard, the Saint-Venant system [8], also known as the nonlinear shallow water equations, presents an interesting compromise between precision and

*Université de Bordeaux, CNRS, Bordeaux INP, IMB, UMR 5251, 33400 Talence, France.
†Université de Bordeaux, CNRS, Bordeaux INP, IMB, UMR 5251, 33400 Talence, France.
‡INSA Toulouse, UMR 5219, 135 Avenue de Rangueil, 31077 Toulouse Cedex 4
§Corresponding author. *Email address: mathieu.rigal@insa-toulouse.fr*

reduced complexity. It is a hyperbolic system of nonlinear balance laws, which in the one-dimensional setting of horizontal space is given by

$$\begin{cases} \partial_t h + \partial_x(hu) = 0, \\ \partial_t(hu) + \partial_x\Big(hu^2 + \dfrac{1}{2}gh^2\Big) + gh\partial_x Z = 0, \end{cases} \tag{1}$$

where h is the water height, hu the vertical average of the horizontal discharge, $Z(x)$ the bathymetry and g the gravitational acceleration. This model constitutes a good approximation of the incompressible free surface Euler equations in the shallow regime, where it is able to reliably describe nonlinear effects governing wave dynamics. At the same time, the Saint-Venant system is easier to study and to discretize than the Euler equations; yet its numerical resolution presents some challenges that have to be carefully considered in order to produce relevant approximations of solutions. Our goal is to derive and study stable discrete BGK schemes overcoming these challenges.

One of the issues is to preserve the non-negativity of the water height, sometimes referred to as the robustness property. In particular, this is a critical point to deal with wetting and drying that occurs at the shoreline. Another aspect is the existence of steady states, such as the hydrostatic equilibrium (or lake at rest) given by

$$hu = 0, \qquad h\partial_x(h + Z) = 0. \tag{2}$$

A scheme able to preserve the above relation at the discrete level is said to be well-balanced, or to satisfy the C-property. Well-balancedness is very important because it prevents a loss of accuracy that typically occurs when approximating flows close to an equilibrium with a naive method. A third problem arises from the fact that the Saint-Venant system is ill-posed due to the non uniqueness of weak solutions, and is usually completed by a set of entropy inequalities as a criterion to select the solution arising from the vanishing viscosity limit. Designing a scheme that satisfies just one such inequality at the discrete level already proves to be difficult — not to mention satisfying all of them.

The aforementioned issues have received extensive coverage in the literature. To obtain the well-balanced property, it is required that the convective flux balances exactly with the source term in areas corresponding to an equilibrium. For the lake at rest, this was achieved in 1994 by Bermudez and Vazquez [9] through an upwinding of the source term. Rather than satisfying this balance constraint in the cells, Jin [24] suggested enforcing it at the interfaces by evaluating the source term with the interface value of the unknowns provided by a Godunov or Roe type flux. Another pioneering work is due to Audusse et al. [5], where both robustness and well-balanced properties are obtained through the *hydrostatic reconstruction* of the water height. This approach is very attractive as it can be implemented in existing codes effortlessly, and we shall incorporate it to our scheme. Alternatively, Liang and Marche [26] use a *pre-balanced* writing of the Saint-Venant system to ensure well-balancedness. We can also mention the work of Bollermann et al. [14] involving an FVEG predictor-corrector

scheme, with the robustness property achieved by means of a draining time after which the outgoing fluxes are canceled. For high order methods, we refer e.g. to [25, 7, 28, 29]. Although we will not consider them, it is also worth mentioning improved versions of the hydrostatic reconstruction [20, 11] and the recent article from Berthon and Michel-Dansac [12] introducing a fully well-balanced reconstruction able to preserve moving flows equilibria in addition to the lake at rest. Prior to this, fully well-balanced schemes were obtained e.g. in [29, 10] without resorting to a reconstruction technique. Another class of methods able to achieve well-balancedness and non-negativity is given by Lagrange-projection schemes, see for instance [19, 18].

The question of entropy stability also received a lot of attention, and it is known that obtaining a fully discrete entropy inequality is a difficult task, in particular when combined to the other properties discussed above. Such inequalities are of great importance since they enable convergence of the numerical approximation towards entropy solutions. In this regard, a strategy that seems adapted revolves around kinetic formulations of the underlying hyperbolic system, replacing the latter by a (somewhat simpler) set of transport equations with collision source term. In [30], Perthame and Simeoni considered a continuous kinetic model whose Maxwellian is compactly supported. Using a kinetic entropy, they were able to recover a fully discrete entropy inequality. The scheme is also non-negative; however, the update does not admit an analytic expression and has to be approximated by a costly quadrature, whose error also degrades the well-balancedness property. Reusing the same kinetic representation, Audusse et al. [6] discretized the bathymetry by means of the hydrostatic reconstruction. They obtained a fully discrete entropy inequality, but in some cases a positive error term increases the energy of the system. Yet, this scheme was shown to converge to entropy solutions by Bouchut and Lhébrard [16]. Nevertheless, a fully discrete entropy inequality without positive error is preferable as it ensures stability of the scheme regardless of the resolution of the mesh. To overcome this defect of stability, Berthon et al. [11] improved the hydrostatic reconstruction by adding numerical viscosity to enforce entropy stability in the framework of an approximate Godunov scheme. Another solution proposed in [23] consists of switching to an implicit-in-time approximation. The drawbacks of this approach are that it is more costly than explicit methods, and that it increases the diffusion even during time steps and in cells where it is not necessary from a stability standpoint. In this paper, we propose a workaround through an adaptive time stepping method, which automatically chooses the appropriate "amount of impliciting" needed at each interface.

Kinetic models referred to in the previous paragraph are based on a continuous distribution of velocities. Another family of kinetic models exists, relying on a finite set of velocities. Compared to the continuous setting, macroscopic quantities are recovered through moment relations involving a discrete sum instead of an integral which can be expensive to evaluate numerically. This advantage motivates us to explore the discrete framework. For general homogeneous systems of conservation laws, discrete kinetic formulations where used in [3, 4] to derive schemes converging towards entropy solutions under a monotonicity

condition on the Maxwellian functions. The presence of a non-conservative force term was dealt with in [1, 2] for the bitemperature Euler system. In [21], an implicit discontinuous Galerkin scheme was proposed for a discrete kinetic-relaxation model. The latter was shown to be compatible with entropy inequalities through a Chapman-Enskog expansion, and the resulting scheme enjoys an explicit computational cost thanks to a sweeping algorithm. To our knowledge, the entropy stability of discrete kinetic schemes combined with the hydrostatic reconstruction has not been addressed in the past.

In this paper we obtain the following results:

- we show how the previous studies [6, 23] can be adapted to the framework of vectorial kinetic representations with finitely many velocities [27, 3, 4, 1, 2];
- using this discrete kinetic setting, we obtain an exact quantification of the dissipation of entropy which improves the estimate from [6, 23]; furthermore we show that the scheme dissipates all the entropies (and not just the energy) in the flat bathymetry case;
- we discuss how the choice of time discretization can overcome the weaker dissipation induced by the hydrostatic reconstruction; in particular we obtain a compromise between stability and accuracy through an adaptive time stepping algorithm.

The paper is organized as follows. Section 2 first details the macroscopic model studied in this work, and then provides an interpretation at the kinetic level via a discrete set of vectorial equations. In Section 3, we investigate the case of flat bathymetries and derive a scheme dissipating all entropies. In Section 4, a semi-discrete scheme based on the hydrostatic reconstruction is proposed for the variable bathymetry case. An entropy dissipation estimate is proved with an exact quantification of the dissipation. Section 5 is then devoted to the time discretization. In order to ensure entropy stability while limiting unnecessary diffusion and reducing the computational cost, we consider some adaptive combination between the implicit and explicit Euler methods. Section 6 illustrates the properties of the scheme through numerical experiments. Finally, Section 7 concludes this work and gives several perspectives.

2 The physical models

2.1 The Saint-Venant equations

The motion of water over a variable bathymetry is described by the Saint-Venant system (1). Denoting the discharge $q = hu$, the vector of unknowns $U = (h, q)$, the flux function $F(U) = (q, \frac{1}{h}q^2 + \frac{1}{2}gh^2)$, and the vectorial source term $S(h, Z) = (0, -gh\partial_x Z)$, this model can be classically written as

$$\partial_t U + \partial_x F(U) = S(h, Z). \tag{3}$$

The source term on the right-hand side accounts for the impact of the bathymetry on the flow; since its first component is zero, the water height h is always a conserved quantity. When the bathymetry is flat ($\partial_x Z = 0$), the second component of S vanishes and the discharge q is also conserved. In this homogeneous setting, Equation (3) admits the dissipative entropy-entropy flux pair (E_0, Q_0) given by

$$E_0(U) = \frac{1}{2}\left(hu^2 + gh^2\right), \quad Q_0(U) = (E_0(U) + \frac{1}{2}gh^2)u, \tag{4}$$

where E_0 can be interpreted as the energy of the system. In the more general setting of non-flat bathymetries ($\partial_x Z \neq 0$), this entropy-entropy flux pair can be generalized by (E, Q), defined as

$$E(U, Z) = E_0(U) + ghZ, \quad Q(U, Z) = Q_0(U) + ghuZ. \tag{5}$$

Introducing the convex set of physical admissible states

$$(h, hu) \in \mathcal{D} :=]0, +\infty[\times\mathbb{R}, \tag{6}$$

one can show that for all values of $Z \in \mathbb{R}$, the entropy $E(\cdot, Z)$ is strictly convex on $\mathcal{D}$. Besides, a weak solution of system (1) is said admissible if it satisfies the entropy inequality

$$\partial_t E(U, Z) + \partial_x Q(U, Z) \leq 0. \tag{7}$$

Next, we shall recall how the Saint-Venant model and the above entropy inequality can be interpreted at the kinetic level.

2.2 Vectorial BGK system with discrete velocities

The idea behind kinetic models is to describe the flow at a finer scale by tracking the distribution of particles according to their velocity and accounting for collisions. In the discrete vectorial setting adopted here, this amounts to consider a set of $L \geq 2$ velocities $\lambda_1, \ldots, \lambda_L \in \mathbb{R}$ and the associated distributions $f_1(t, x), \ldots, f_L(t, x) \in \mathbb{R}^2$, with each of them satisfying a kinetic equation

$$\partial_t f_l + \lambda_l \partial_x f_l + N(f_l) = \frac{1}{\varepsilon}\left(\mathcal{M}_l(U_f) - f_l\right), \quad 1 \leq l \leq L. \tag{8}$$

In the above, $\varepsilon > 0$ is a relaxation parameter, the macroscopic vector U_f is recovered by summing over the kinetic distributions

$$U_f(t, x) := \sum_{l=1}^{L} f_l(t, x) = \begin{pmatrix} h_f(t, x) \\ q_f(t, x) \end{pmatrix}, \tag{9}$$

and the quantity

$$N(f_l) = (g\partial_x Z) P_{\text{shift}} \begin{pmatrix} f_{l1} \\ f_{l2} \end{pmatrix}, \qquad P_{\text{shift}} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \tag{10}$$

describes the influence of the bathymetry on the distribution $f_l = (f_{l1}, f_{l2})$. As they collide, the particles can change speed: this behavior is modelled by the source term on the right-hand side of (8), which couples the distributions f_l between them. This source term corresponds to the BGK operator introduced in [13]; it involves an equilibrium distribution $\mathcal{M}_l : \mathcal{D} \to \mathbb{R}^2$ which will be called the Maxwellian function by analogy with the kinetic theory of gases.

The definition of $\mathcal{M}_l$ is chosen such that the macroscopic state U_f defined by (8)–(9) approximates the solutions of the Saint-Venant system (3) as ϵ becomes small. More precisely, we require the following *compatibility conditions* to be satisfied

$$\forall U \in \mathcal{D}, \quad \sum_{l=1}^{L} \mathcal{M}_l(U) = U, \quad \sum_{l=1}^{L} \lambda_l \mathcal{M}_l(U) = F(U). \tag{11}$$

As proposed in [15], a suitable choice is to take the Maxwellian functions as linear combinations between the identity and the macroscopic flux F:

$$\mathcal{M}_l(U) = \alpha_l U + \beta_l F(U), \quad 1 \leq l \leq L, \tag{12}$$

where α_l and β_l are real numbers. In particular, the compatibility conditions (11) are satisfied as soon as there holds

$$\sum_{l=1}^{L} \alpha_l = 1, \quad \sum_{l=1}^{L} \beta_l = 0, \quad \sum_{l=1}^{L} \lambda_l \alpha_l = 0, \quad \sum_{l=1}^{L} \lambda_l \beta_l = 1.$$

Furthermore, in the model (8)–(9) we can check that the quantity $N(f_l)$ leads to the non-conservative bathymetry term in the macroscopic system (3), that is:

$$\sum_{l=1}^{L} N(f_l) = -S(h_f, Z).$$

Hence, by taking the sum of all the equations of (8), using the notation (9) and denoting $w = \sum_{l=1}^{L} \lambda_l f_l$, we obtain the system

$$\begin{cases} \partial_t U_f + \partial_x w - S(h_f, Z) = 0, \\ \partial_t w + \sum_{l=1}^{L} \lambda_l^2 \partial_x f_l + \sum_{l=1}^{L} \lambda_l N(f_l) = \frac{1}{\varepsilon}(F(U_f) - w). \end{cases}$$

If f has a limit when $\varepsilon \to 0$ in a sufficiently strong sense, then $f = \mathcal{M}(U_f)$, $w = F(U_f)$, and as a consequence U_f is a solution of (3).

Remark 2.1. *It can be illuminating to compare* (8) *to the continuous BGK model developed in [30]:*

$$\frac{\partial f}{\partial t} + \xi \frac{\partial f}{\partial x} - g \partial_x Z(x) \frac{\partial f}{\partial \xi} = \frac{1}{\varepsilon}(M(f, \xi) - f). \tag{13}$$

The scalar kinetic unknown $f(t,x,\xi)$ is linked to the macroscopic quantities (h,q) via moment relations w.r.t. the velocity variable $\xi \in \mathbb{R}$, and which are the continuous analogue of (9)*:*

$$h = \int_{\mathbb{R}} f d\xi, \quad hu = \int_{\mathbb{R}} \xi f d\xi.$$

The bathymetry source term $-S(h,Z)$ is then recovered by integrating the force term $-g\partial_x Z(x)\partial_\xi f$ against $(1,\xi)$ and by performing an integration by parts.

In contrast, the framework considered here is such that the velocity variable belongs to a finite discrete set, and the distribution f_l is valued in $\mathbb{R}^2$. The force term is now modelled by $N(f_l)$, and the structure of the shift matrix P_{shift} given in (10) *"mimics" the integration by parts used in [30] to recover the non-conservative contribution of the bathymetry. Such a shift matrix was first introduced in the context of the one-dimensional bitemperature Euler equations [1], where non-conservative terms are represented at the kinetic level by an electric force. An extension to the two-dimensional case was also proposed in [2].*

An important notion used throughout this paper is that of the *subcharacteristic condition*. We say that this latter is satisfied by the Maxwellian functions $(\mathcal{M}_l)_{1\leq l\leq L}$ in some subset $\mathcal{D}_0 \subset \mathcal{D}$ if there holds

$$\forall U \in \mathcal{D}_0, \quad \forall 1 \leq l \leq L, \quad \sigma(\mathcal{M}_l'(U)) \subset]0,+\infty[. \tag{14}$$

By requiring all the eigenvalues of $\mathcal{M}_l'(U)$ to be positive, (14) ensures that each $\mathcal{M}_l : \mathcal{D}_0 \to \mathcal{M}_l(\mathcal{D}_0)$ is one-to-one. In the next section, we will see how this condition enables one to recover the macroscopic entropy inequality (7) from the kinetic level.

Before this, we would like to provide an example of vectorial BGK model. As we have two compatibility relations in (11), the minimal model features two velocities ($L = 2$). Without loss of generality, we assume that $\lambda_1 < \lambda_2$, and the associated Maxwellian functions can be defined as

$$\mathcal{M}_1(U) = \frac{\lambda_2 U - F(U)}{\lambda_2 - \lambda_1}, \quad \mathcal{M}_2(U) = \frac{-\lambda_1 U + F(U)}{\lambda_2 - \lambda_1}. \tag{15}$$

In that case, we can easily check that (11) is verified, and the subcharacteristic condition (14) is satisfied if and only if $\lambda_1 < \mu(U) < \lambda_2$ for each eigenvalue $\mu(U)$ of $F'(U)$, that is

$$\lambda_1 + \sqrt{gh} < u < \lambda_2 - \sqrt{gh}. \tag{16}$$

The interested reader can refer e.g. to [4] where additional discrete BGK models have been studied.

2.3 Kinetic entropies

It is possible to interpret the entropy inequality (7) by means of kinetic entropies, which can be seen as distributions of macroscopic entropies in the velocity space

$\{\lambda_1, \ldots, \lambda_L\}$. Based on [15], we first recall how these kinetic entropies are constructed in the framework of discrete vectorial BGK models.

Assuming that the subcharacteristic condition (14) holds in some subset $\mathcal{D}_0 \subset \mathcal{D}$, we know that the Maxwellian functions $\mathcal{M}_l : \mathcal{D}_0 \to \mathcal{M}_l(\mathcal{D}_0)$ are one-to-one; in particular, for every $f_l \in \mathcal{M}_l(\mathcal{D}_0) \subset \mathbb{R}^2$, the quantity $\mathcal{M}_l^{-1}(f_l)$ is the unique macroscopic state $U_l \in \mathcal{D}_0$ such that

$$\mathcal{M}_l(U_l) = f_l.$$

Next we can define the homogeneous kinetic entropy $H_{l,0}$ for all $1 \leq l \leq L$ by

$$H_{l,0}(f_l) = G_{l,0}(\mathcal{M}_l^{-1}(f_l)), \quad G_{l,0}(U) = \alpha_l E_0(U) + \beta_l Q_0(U), \tag{17}$$

where $G_{l,0}$ is referred to as the microscopic entropy. Owing to the compatibility conditions (11), the following sums enable to recover the homogeneous macroscopic entropy and its flux (4):

$$\sum_{l=1}^{L} H_{l,0}(\mathcal{M}_l(U)) = E_0(U), \qquad \sum_{l=1}^{L} \lambda_l H_{l,0}(\mathcal{M}_l(U)) = Q_0(U). \tag{18}$$

Relation (18) can then be extended to the nonhomogeneous case (i.e. when the bathymetry Z is non-flat) by defining for all $1 \leq l \leq L$ and all $Z \in \mathbb{R}$:

$$H_l(f_l, Z) = H_{l,0}(f_l) + gZ f_{l1}. \tag{19}$$

Using once again the compatibility conditions (11), the first component $\mathcal{M}_{l1}$ of the Maxwellian function verifies $\sum_l (1, \lambda_l)\mathcal{M}_{l1}(U) = (h, hu)$, and the moment relations (18) become

$$\sum_{l=1}^{L} H_l(\mathcal{M}_l(U), Z) = E(U, Z), \qquad \sum_{l=1}^{L} \lambda_l H_l(\mathcal{M}_l(U), Z) = Q(U, Z), \tag{20}$$

where $E(U, Z)$ and $Q(U, Z)$ are given in (5).

In order to recover the macroscopic entropy inequality (7), we need to recall some important properties. The interest of the kinetic entropies (19) is that they remain strictly convex on $\mathcal{M}_l(\mathcal{D}_0)$ and give rise to a constrained minimization principle as stated below.

Proposition 2.2 (Adapted from Bouchut [15])**.** *For any $Z \in \mathbb{R}$, the functions $H_l(\cdot, Z) : \mathcal{M}_l(\mathcal{D}_0) \to \mathbb{R}$ defined through* (19) *are strictly convex. For all $U \in \mathcal{D}_0$ and for all $(f_1, \ldots, f_L) \in \prod_{1 \leq l \leq L} \mathcal{M}_l(\mathcal{D}_0)$ such that $\sum_{l=1}^{L} f_l = U$, one also has*

$$\sum_{l=1}^{L} H_l(\mathcal{M}_l(U), Z) \leq \sum_{l=1}^{L} H_l(f_l, Z). \tag{21}$$

Proof. As shown in [15], the result holds when $Z = 0$, that is, $H_{l,0}(\cdot)$ given in (17) is strictly convex over $\mathcal{M}_l(\mathcal{D}_0)$ and we have

$$\sum_{l=1}^{L} H_{l,0}(\mathcal{M}_l(U)) \leq \sum_{l=1}^{L} H_{l,0}(f_l). \tag{22}$$

The strict convexity in the general case $Z \neq 0$ follows directly by remarking that $H_l(f_l, Z)$ only differs from $H_{l,0}(f_l)$ by a linear term in f_l. Furthermore, owing to definition (19) and to (22), we have

$$\sum_{l=1}^{L} H_l(\mathcal{M}_l(U), Z) = \sum_{l=1}^{L} H_{l,0}(\mathcal{M}_l(U)) + ghZ \leq \sum_{l=1}^{L} H_{l,0}(f_l) + ghZ,$$

which yields (21) after using $h = \sum_l f_{l1}$. □

The following proposition gives another useful result. In particular, it is remarkable that neither $H'_{l,0}(\mathcal{M}_l(U))$ nor $\nabla_f H_l(\mathcal{M}_l(U), Z)$ depend on the velocity index l.

Proposition 2.3. *We have the following identities for all* $(U, Z) \in \mathcal{D}_0 \times \mathbb{R}$ *and for all* $1 \leq l \leq L$*:*

$$G'_{l,0}(U) = E'_0(U)\mathcal{M}'_l(U), \qquad H'_{l,0}(\mathcal{M}_l(U)) = E'_0(U) = (-\frac{1}{2}u^2 + gh, u),$$
$$\nabla_f H_l(\mathcal{M}_l(U), Z) = \nabla_U E(U, Z) = (-\frac{1}{2}u^2 + g(h + Z), u)^T.$$

Proof. Since $Q'_0(U) = E'_0(U)F'(U)$, it holds that

$$\begin{aligned} G'_{l,0}(U) &= \alpha_l E'_0(U) + \beta_l Q'_0(U) = E'_0(U)(\alpha_l I + \beta_l F'(U)) \\ &= E'_0(U)\mathcal{M}'_l(U). \end{aligned}$$

Moreover, we have

$$H'_{l,0}(f_l) = G'_{l,0}(\mathcal{M}_l^{-1}(f_l)) \left(\mathcal{M}'_l(\mathcal{M}_l^{-1}(f_l))\right)^{-1}.$$

Therefore, this yields

$$H'_{l,0}(\mathcal{M}_l(U)) = G'_{l,0}(U) \left(\mathcal{M}'_l(U)\right)^{-1} = E'_0(U).$$

In the nonhomogeneous case, definition (19) of the kinetic entropy H_l implies

$$\nabla_f H_l(\mathcal{M}_l(U), Z) = \nabla H_{l,0}(\mathcal{M}_l(U)) + gZ \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \nabla_U E(U, Z).$$

□

Finally, one can recover the macroscopic entropy inequality (7) thanks to the moment relations (20) and the minimization principle stated in Proposition 2.2. For this, we first assume that for each velocity index $1 \leq l \leq L$ the function $(t,x) \mapsto H_l(f_l(t,x), Z(x))$ is at least C^1. We can then take the scalar product between $\nabla_f H_l(f_l, Z)$ and the discrete BGK model (8) to get

$$\begin{aligned}\partial_t H_l(f_l,Z) + \lambda_l \nabla_f H_l(f_l,Z)\cdot\partial_x f_l + \nabla_f H_l(f_l,Z)\cdot N(f_l)\\ = \nabla_f H_l(f_l,Z)\cdot\frac{\mathcal{M}_l - f_l}{\epsilon}.\end{aligned}$$

By using

$$\begin{aligned}\lambda_l \nabla_f H_l(f_l,Z)\cdot\partial_x f_l &= \lambda_l\left(\partial_x H_l(f_l,Z) - \partial_Z H_l(f_l,Z)\partial_x Z\right)\\ &= \lambda_l \partial_x H_l(f_l,Z) - g f_{l1}\lambda_l \partial_x Z,\end{aligned}$$

it comes that

$$\begin{aligned}\partial_t H_l(f_l,Z) + \lambda_l\partial_x H_l(f_l,Z) - g f_{l1}\lambda_l\partial_x Z + \nabla_f H_l(f_l,Z)\cdot N(f_l)\\ = \nabla_f H_l(f_l,Z)\cdot\frac{\mathcal{M}_l - f_l}{\epsilon}.\end{aligned}$$

By convexity of $H_l(\cdot, Z)$ we obtain

$$\begin{aligned}\partial_t H_l(f_l,Z) + \lambda_l\partial_x H_l(f_l,Z) - g f_{l1}\lambda_l\partial_x Z + \nabla_f H_l(f_l,Z)\cdot N(f_l)\\ \leq \frac{H_l(\mathcal{M}_l,Z) - H_l(f_l,Z)}{\epsilon}.\end{aligned} \tag{23}$$

However, in the setting of weak solutions the assumption $H_l(f_l, Z) \in C^1$ is not satisfied in general, and the scalar product between $\nabla_f H_l(f_l, Z)$ and $(\partial_t + \lambda_l \partial_x) f_l$ performed in the first step is ill-defined. This is in particular the case in presence of a shock discontinuity. To circumvent this issue, we can instead see (23) as a kinetic entropy condition which, in addition to (20) and (21), is sufficient to recover the macroscopic entropy inequality (7). In fact, when the condition (23) is satisfied, we can sum it over $1 \leq l \leq L$ to get

$$\begin{aligned}\partial_t \sum_l H_l(f_l,Z) + \partial_x \sum_l (\lambda_l H_l(f_l,Z)) - \sum_l \Big(g f_{l1}\lambda_l\partial_x Z - \nabla_f H_l(f_l,Z)\cdot N(f_l)\Big)\\ \leq 0,\end{aligned}$$

where we used the minimization principle (21) to get the null upper bound. Letting $\epsilon \to 0$ we formally have $f_l \to \mathcal{M}_l(U)$ with $U(t,x) := \lim_{\epsilon\to 0}\sum_l f_l(t,x)$ a weak solution of the Saint-Venant system (3). Owing to the moment relations (20) this yields

$$\begin{aligned}&\partial_t E(U,Z) + \partial_x Q(U,Z)\\ &- g\sum_{l=1}^{L}\left(\mathcal{M}_{l1}(U)\lambda_l\right)\partial_x Z + \sum_{l=1}^{L}\nabla_f H_l(\mathcal{M}_l(U),Z)\cdot N(\mathcal{M}_l(U)) \leq 0.\end{aligned}$$

Hence, we recover (7) by remarking that the remaining sums on the left-hand side cancel each other, given that $\sum_l (\mathcal{M}_{l1}(U)\lambda_l) = hu$ and that Proposition 2.3 implies

$$\sum_{l=1}^{L} \nabla_f H_l(\mathcal{M}_l(U), Z) \cdot N(\mathcal{M}_l(U)) = \sum_{l=1}^{L} \nabla_U E(U, Z) \cdot N(\mathcal{M}_l(U)) = ghu\partial_x Z.$$

We summarize this result below.

Proposition 2.4. *Let Z be Lipschitz continuous and let U be a weak solution of* (3) *obtained as a limit of the discrete BGK model* (8)*, with convex kinetic entropy $H_l(\cdot, Z)$ defined by* (19)*. Then, as soon as the kinetic entropy condition* (23) *is verified, U is an admissible solution of* (3)*, meaning*

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

where the entropy-entropy flux pair (E, Q) is defined in (5)*. Furthermore, as long as the kinetic entropy $H_l(f_l(t, x), Z(x))$ remains C^1 in time and space, the condition* (23) *is automatically satisfied.*

It is worth noting that the proof of Proposition 2.4 provided above can also be done by taking the scalar product between the discrete BGK model (8) and the gradient $\nabla H_{l,0}(f_l)$ instead of $\nabla_f H_l(f_l, Z)$. However, to adapt the previous computations to numerical discretizations of (8) in the general setting of non-flat bottoms, it is more interesting to consider the nonhomogeneous kinetic entropy. In fact, we shall see later how this analysis can be adapted to kinetic schemes in order to obtain fully discrete entropy inequalities.

To this end, we stress that a key ingredient is the kinetic entropy condition (23), relying on the convexity of the kinetic entropy $H_l(\cdot, Z) : \mathcal{M}_l(\mathcal{D}_0) \to \mathbb{R}$. Since this latter function is smooth, we know that any second order Taylor expansion admits a non-negative remainder, which we state below.

Lemma 2.5. *A consequence of the convexity of the kinetic entropies $H_l(\cdot, Z) : \mathcal{M}_l(\mathcal{D}_0) \to \mathbb{R}$ defined through* (19) *is that the associated map $\psi_l : \mathcal{M}_l(\mathcal{D}_0)^2 \times \mathbb{R} \to \mathbb{R}$ given by*

$$\psi_l(a, b, Z) = H_l(b, Z) - H_l(a, Z) - \nabla_f H_l(a, Z) \cdot (b - a) \tag{24}$$

is non-negative.

The characterization provided by the above Lemma will prove useful in Sections 4 and 5 to quantify precisely the entropy dissipation rate of numerical discretizations. Before that, in Section 3 we will consider the simpler case of flat bottoms to introduce the discretization strategy. We also emphasize that, although we only consider the one-dimensional setting, we can easily extend this work to the multidimensional case, see [4, 2].

3 Fully discrete kinetic scheme over flat bottom

In this section we set $\partial_x Z = 0$, and we explain how the discrete vectorial BGK models presented above allow to design fully discrete schemes which preserve the positivity of h and which dissipate any entropy associated to the Saint-Venant system with flat bathymetry — including the energy E_0 defined in (4). Since the bathymetry is defined up to a constant, we can furthermore suppose that $Z = 0$ without loss of generality.

3.1 Numerical scheme

First, we introduce mesh-related notations. We assume that the Saint-Venant system (1) is solved on the spatial domain $(0, \ell)$, discretized by means of N cells of size Δx. The cells centers are given by $x_i = (i - \frac{1}{2})\Delta x$ for $1 \leq i \leq N$, and the interfaces are denoted by $x_{i+1/2}$. The time step is written Δt, and the discrete times are defined through the increment $t^{n+1} = t^n + \Delta t$ with $t^0 = 0$. Noting $(hu)_i^n = h_i^n u_i^n$, the quantity $U_i^n = (h_i^n, (hu)_i^n)^T$ is an approximation of the vector of unknowns averaged on cell i at time t^n, while Z_i denotes the average of the bathymetry on that same cell. Similar notations are adopted for the kinetic distribution f_l and the Maxwellian function $\mathcal{M}_l$.

The strategy is to design numerical fluxes thanks to the kinetic formalism introduced in the previous section. Owing to the compatibility conditions (11), we recall that there holds $\sum_{1\leq l\leq L} \lambda_l \mathcal{M}_l(U) = F(U)$ for all $U \in \mathcal{D}_0$. Hence, a numerical flux for the macroscopic equations can be obtained by approximating the transport term $\lambda_l \mathcal{M}_l(U)$ at the interfaces and summing over $1 \leq l \leq L$. More precisely, we define the numerical scheme as follows: the initial data $U_0 \in \mathcal{D}_0$ being given, we consider U_i^0 as the average of U_0 on the cell $]x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}[$. Then, for all $n \geq 0$ we set

$$\begin{cases} f_{l,i}^n = \mathcal{M}_l(U_i^n), \\ f_{l,i}^{n+\frac{1}{2}} = f_{l,i}^n - \dfrac{\Delta t}{\Delta x}\left(\lambda_l^+ (f_{l,i}^n - f_{l,i-1}^n) + \lambda_l^- (f_{l,i+1}^n - f_{l,i}^n)\right), \\ U_i^{n+1} = \sum_{1\leq l\leq L} f_{l,i}^{n+\frac{1}{2}}, \end{cases} \tag{25}$$

with

$$\lambda_l^+ = \max(\lambda_l, 0), \quad \lambda_l^- = \min(\lambda_l, 0).$$

We see that in the above definition of $f_{l,i}^{n+\frac{1}{2}}$, the numerical flux is upwinded with respect to the sign of the velocity λ_l. This scheme can also be written with the macroscopic variable U only

$$U_i^{n+1} = U_i^n - \frac{\Delta t}{\Delta x}\left(\mathcal{F}(U_i^n, U_{i+1}^n) - \mathcal{F}(U_{i-1}^n, U_i^n)\right), \tag{26}$$

where we denoted the macroscopic numerical flux by

$$\mathcal{F}(U_i, U_{i+1}) := \sum_{1\leq l\leq L} \lambda_l \Big(\mathbb{1}_{\lambda_l<0} \mathcal{M}_l(U_{i+1}) + \mathbb{1}_{\lambda_l>0} \mathcal{M}_l(U_i) \Big). \tag{27}$$

In particular, the update (26) is conservative and we have $\mathcal{F}(U,U) = F(U)$, so that consistency holds.

We can further detail the scheme (26). With equilibrium functions $\mathcal{M}_l$ under the form (12), denoting $F_j^n = F(U_j^n)$ and

$$\nu = \sum_{l=1}^{L} |\lambda_l| \alpha_l, \quad \mu = \sum_{l=1}^{L} |\lambda_l| \beta_l, \tag{28}$$

the corresponding update can be written as follows

$$\begin{aligned} \frac{U_i^{n+1} - U_i^n}{\Delta t} + \frac{F_{i+1}^n - F_{i-1}^n}{2\Delta x} - \frac{\nu}{2\Delta x}\left(U_{i+1}^n - 2U_i^n + U_{i-1}^n\right) \\ - \frac{\mu}{2\Delta x}\left(F_{i+1}^n - 2F_i^n + F_{i-1}^n\right) = 0, \end{aligned} \tag{29}$$

that is,

$$\begin{aligned} \frac{h_i^{n+1} - h_i^n}{\Delta t} + \frac{1}{2\Delta x}\left(h_{i+1}^n u_{i+1}^n - h_{i-1}^n u_{i-1}^n\right) - \frac{\nu}{2\Delta x}\left(h_{i+1}^n - 2h_i^n + h_{i-1}^n\right) \\ - \frac{\mu}{2\Delta x}\left(h_{i+1}^n u_{i+1}^n - 2h_i^n u_i^n + h_{i-1}^n u_{i-1}^n\right) = 0, \end{aligned}$$

and

$$\begin{aligned} &\frac{q_i^{n+1} - q_i^n}{\Delta t} + \frac{\frac{1}{2}g((h_{i+1}^n)^2 - (h_{i-1}^n)^2) + h_{i+1}^n (u_{i+1}^n)^2 - h_{i-1}^n (u_{i-1}^n)^2}{2\Delta x} \\ &- \frac{\nu}{2\Delta x}\left(h_{i+1}^n u_{i+1}^n - 2h_i^n u_i^n + h_{i-1}^n u_{i-1}^n\right) - \frac{\mu g}{4\Delta x}\left((h_{i+1}^n)^2 - 2(h_i^n)^2 + (h_{i-1}^n)^2\right) \\ &- \frac{\mu}{2\Delta x}\left(h_{i+1}^n (u_{i+1}^n)^2 - 2h_i^n (u_i^n)^2 + h_{i-1}^n (u_{i-1}^n)^2\right) = 0. \end{aligned}$$

In the particular case of the two velocities model (15) with symmetric velocities $\lambda_2 = \lambda = -\lambda_1$, one has $\nu = \lambda$, $\mu = 0$ in the above expressions.

Remark 3.1. *Note that in the definition* (27) *of the flux* $\mathcal{F}$*, different velocities* $\lambda_l = \lambda_{l,i+1/2}^n$ *can be used locally at each interface* $x_{i+1/2}$ *and at each time step* t^n*. Doing so is advantageous since it improves the accuracy of the scheme by optimizing its diffusion. The consequence is that definition* (12) *of the Maxwellian changes from one interface/time step to the other through the coefficients* $\alpha_l = \alpha_{l,i+1/2}^n$ *and* $\beta_l = \beta_{l,i+1/2}^n$*. Hence, the framework of local velocities entails that we do not approximate the BGK equation* (8) *on the whole domain, rather the kinetic formalism provides a strategy to construct and study the macroscopic numerical flux* (27) *locally in time and space. In particular, when local velocities are used the scheme* (29) *becomes of the form*

$$\begin{aligned} \frac{U_i^{n+1} - U_i^n}{\Delta t} + \frac{F_{i+1}^n - F_{i-1}^n}{2\Delta x} - \frac{\nu_{i+1/2}^n (U_{i+1}^n - U_i^n) - \nu_{i-1/2}^n (U_i^n - U_{i-1}^n)}{2\Delta x} \\ - \frac{\mu_{i+1/2}^n (F_{i+1}^n - F_i^n) - \mu_{i-1/2}^n (F_i^n - F_{i-1}^n)}{2\Delta x} = 0. \end{aligned}$$

Furthermore, we can say that the subcharacteristic condition is satisfied locally at $(t^n, x_{i+1/2})$ if (14) holds on the set $\mathcal{D}_0 = \{U_i^n, U_{i+1}^n\}$. For simplicity, we keep working with global velocities in the remainder of Section 3.

Finally, we need to specify the handling of the boundary condition; for this we have to define the ghost quantities U_0^n and U_{N+1}^n which appear in (26) respectively when $i = 1$ and $i = N$. For the sake of simplicity, we restrict ourselves to periodic boundary conditions $U_{N+1}^n = U_1^n$ and $U_0^n = U_N^n$, or to cases where the solution remains constant at the boundaries, with homogeneous Neumann conditions enforced for both h and hu by setting $U_0^n = U_1^n$ and $U_{N+1}^n = U_N^n$. This simplification is valid in the setting of short time simulations where the waves present in the domain do not have the time to reach the boundaries. Note however that it is possible to design more general boundary conditions using Riemann invariants as described in [17], with degrees of freedom determined from the type of flow: fluvial regime, torrential inflow or torrential outflow. Since the results obtained thereafter do not depend on the choice of boundary conditions, they also hold in this more general setting.

3.2 Entropy dissipation

We show that under some assumptions, the update (26) keeps h_i^n non-negative and dissipates any entropy of the Saint-Venant system (1), that is, any convex function $\eta : \mathcal{D} \to \mathbb{R}$ such that there exists an entropy flux $\Phi : \mathcal{D} \to \mathbb{R}$ verifying $\Phi' = \eta' F'$. As a particular case, we can take $\eta = E = E_0$ and $\Phi = Q = Q_0$ given in (4), and we remark that definition (17) of the kinetic entropy $H_{l,0}$ associated to E_0 can be extended to general entropies η as follows:

$$H_{l,\eta}(f_l) = G_{l,\eta}(\mathcal{M}_l^{-1}(f_l)), \quad G_{l,\eta}(U) = \alpha_l \eta(U) + \beta_l \Phi(U). \tag{30}$$

Similarly to $H_{l,0}$, the general kinetic entropy $H_{l,\eta} : \mathcal{M}_l(\mathcal{D}_0) \to \mathbb{R}$ is convex and gives rise to a minimization principle (22), see [15]. Thanks to this, we are able to state the proposition below.

Proposition 3.2. *Assume that the subcharacteristic condition (14) is satisfied and that for all $l \in \{1, \dots, L\}$ the CFL condition $\lambda_l \Delta t / \Delta x \leq 1$ holds. Then for all $n \geq 0$ and all i, $h_i^n > 0$ and a discrete entropy inequality holds for any entropy-entropy flux pair (η, Φ):*

$$\frac{\eta(U_i^{n+1}) - \eta(U_i^n)}{\Delta t} + \frac{\varphi_{i+\frac{1}{2}}^n - \varphi_{i-\frac{1}{2}}^n}{\Delta x} \leq 0$$

with $\varphi_{i\pm\frac{1}{2}}^n$ given below in terms of the kinetic entropy (30):

$$\varphi_{i+\frac{1}{2}}^n = \varphi(U_i^n, U_{i+1}^n) = \sum_{l=1}^{L} \left(\lambda_l^+ H_{l,\eta}(\mathcal{M}_l(U_i^n)) + \lambda_l^- H_{l,\eta}(\mathcal{M}_l(U_{i+1}^n)) \right).$$

Moreover, φ is consistent with the entropy flux Φ: $\varphi(U, U) = \Phi(U)$.

In order to prove the positivity property, we need the following Lemma.

Lemma 3.3. *If the subcharacteristic condition* (14) *is verified for some* $U = (h, q) \in \mathcal{D}_0$*, then there holds* $(\mathcal{M}_l(U))_1 > 0$*.*

Proof of Lemma 3.3. We recall that the eigenvalues of $F'(U)$ are $u \pm \sqrt{gh}$. Since $\mathcal{M}_l$ is of the form (12), the eigenvalues of $\mathcal{M}_l'(U)$ are $\alpha_l + \beta_l u \pm |\beta_l|\sqrt{gh}$. If they are positive then $\alpha_l + \beta_l u > 0$. But we can also write that

$$(\mathcal{M}_l(U))_1 = \alpha_l h + h\beta_l u = h(\alpha_l + \beta_l u),$$

and the result holds since $U \in \mathcal{D}_0$ implies $h > 0$. □

Proof of Proposition 3.2. The conservation of the positivity of h is a direct consequence of Lemma 3.3 and the monotonicity of the upwind scheme which ensures that $f_{l,i,1}^{n+\frac{1}{2}} > 0$. By convexity of $H_{l,\eta}$, as $f_{l,i}^{n+\frac{1}{2}}$ is a convex combination of $f_{l,i}^n$ and $f_{l,i\pm1}^n$, one has

$$H_{l,\eta}(f_{l,i}^{n+\frac{1}{2}}) \leq H_{l,\eta}(f_{l,i}^n) - \frac{\Delta t}{\Delta x}\left(\varphi_{l,i+\frac{1}{2}}^n - \varphi_{l,i-\frac{1}{2}}^n\right)$$

with $\varphi_{l,i+\frac{1}{2}}^n = \lambda_l^+ H_{l,\eta}(\mathcal{M}_l(U_i^n)) + \lambda_l^- H_{l,\eta}(\mathcal{M}_l(U_{i+1}^n))$. Summing over l we obtain

$$\sum_{l=1}^L H_{l,\eta}(f_{l,i}^{n+\frac{1}{2}}) \leq \sum_{l=1}^L H_{l,\eta}(f_{l,i}^n) - \frac{\Delta t}{\Delta x}\left(\varphi_{i+\frac{1}{2}}^n - \varphi_{i-\frac{1}{2}}^n\right).$$

Moreover, $\sum_{l=1}^L H_{l,\eta}(f_{l,i}^n) = \eta(U_i^n)$ and for all $U \in \mathcal{D}_0$ and all $f = (f_1, \ldots, f_L)$ such that $\sum_l f_l = U$, one has

$$\eta(U) \leq \sum_{l=1}^L H_{l,\eta}(f_l).$$

As $\sum_l f_{l,i}^{n+\frac{1}{2}} = U_i^{n+1}$ and $f_{l,i}^n = \mathcal{M}_l(U_i^n)$, we obtain

$$\eta(U_i^{n+1}) \leq \eta(U_i^n) - \frac{\Delta t}{\Delta x}\left(\varphi_{i+\frac{1}{2}}^n - \varphi_{i-\frac{1}{2}}^n\right).$$

Moreover, thanks to definition (30) of the general kinetic entropy we have

$$\varphi(U, U) = \sum_{l=1}^L \lambda_l H_{l,\eta}(\mathcal{M}_l(U)) = \Phi(U).$$

□

In definition (27) of the numerical flux, the upwinding with respect to the sign of the velocity λ_l is a crucial ingredient to ensure entropy dissipation. In what follows, we will see how to deal with the presence of a variable bathymetry, and in particular how to generalize the result of Proposition 3.2 to this framework.

4 Semi-discrete scheme in space over variable bottom

We now consider the Saint-Venant system (3) with a general bathymetry, meaning that one can possibly have $\partial_x Z \neq 0$. The goal is to obtain a numerical approximation remaining consistent with admissible solutions (i.e. satisfying a discrete analog of (7)), while at the same time preserving the hydrostatic equilibria (also known as lake-at-rest steady states) defined in (2). However, unlike what was achieved in the previous section we do not try to dissipate all the entropies, as this task is much harder in presence of a variable bottom. An intermediate step which we address here is to first consider the semi-discrete case, allowing to focus on the impact of the spatial discretization on these two issues. The fully discrete case is treated in Section 5.

4.1 The hydrostatic reconstruction and its kinetic interpretation

Let us denote $U_i(t) \in \mathbb{R}^2$ the approximation of U on the i^{th} cell at time t, defined as the solution of a coupled ODE system of the form

$$\frac{\mathrm{d}}{\mathrm{d}t}U_i(t) + \frac{1}{\Delta x}(\mathcal{F}_{i+1/2} - \mathcal{F}_{i-1/2}) = \frac{1}{\Delta x}S_i, \qquad 1 \leq i \leq N,$$

where the definition of the fluxes $\mathcal{F}_{i\pm1/2}(t)$ and of the source term $S_i(t)/\Delta x$ is specified bellow. The difficulty lies in obtaining a well-balanced discretization of the Saint-Venant system in presence of a non-flat bathymetry. By well-balanced, we mean that the approximation should preserve the discrete hydrostatic equilibrium defined as:

$$\exists C \in \mathbb{R}, \quad \forall t \in \mathbb{R}, \quad \forall\, 1 \leq i \leq N, \quad h_i(t)(h_i(t) + Z_i - C) = u_i(t) = 0. \quad (31)$$

In the above, the condition on h_i means that either the free surface is flat ($h_i + Z_i = C$), either the water height vanishes ($h_i = 0$). We briefly explain why, despite the apparent simplicity of this family of solutions, numerical schemes usually fail to preserve them. For the sake of the example, consider $\mathcal{F}$ the numerical flux (27) based on the two velocities model (15) with symmetric velocities $\lambda_2 = \lambda = -\lambda_1$, and consider the bathymetry source term to be of the form $S_i = (0, (S_2)_i)^T$. If the discrete stationary state (31) is satisfied at all times without dry areas (i.e. $h_i > 0\ \forall i$), one has

$$\frac{\mathrm{d}}{\mathrm{d}t}h_i(t) = \frac{\lambda}{2\Delta x}(h_{i+1} - 2h_i + h_{i-1}).$$

Injecting the identity $h_i = C - Z_i$ in the numerical diffusion we get

$$\frac{\mathrm{d}}{\mathrm{d}t}h_i = -\frac{\lambda}{2\Delta x}(Z_{i+1} - 2Z_i + Z_{i-1}). \quad (32)$$

For a general bathymetry Z, the right-hand side of (32) has no reason to vanish, and the water height can vary despite the discrete lake at rest. Proceeding similarly with the discharge equation we have

$$\frac{\mathrm{d}}{\mathrm{d}t}q_i + \frac{1}{2\Delta x}\Big(\frac{g}{2}h_{i+1}^2 - \frac{g}{2}h_{i-1}^2\Big) = \frac{1}{\Delta x}(S_2)_i.$$

To keep the discharge $q_i = h_i u_i$ constant in time equal to zero, we need to define the second component of the source term discretization $(S_2)_i/\Delta x$ such that it balances exactly with the discrete spatial variation of the pressure $\frac{g}{2}h^2$ in the setting of a lake at rest.

The hydrostatic reconstruction [5] addresses these two problems by modifying the water height in the neighborhood of each interface in the following way:

$$U_{i+1/2,-} = \begin{pmatrix} h_{i+1/2,-} \\ q_{i+1/2,-} \end{pmatrix}, \qquad \begin{cases} h_{i+1/2,-} = \max(0, h_i + Z_i - Z_{i+1/2}), \\ q_{i+1/2,-} = h_{i+1/2,-}u_i, \end{cases} \tag{33}$$

$$U_{i-1/2,+} = \begin{pmatrix} h_{i-1/2,+} \\ q_{i-1/2,+} \end{pmatrix}, \qquad \begin{cases} h_{i-1/2,+} = \max(0, h_i + Z_i - Z_{i-1/2}), \\ q_{i-1/2,+} = h_{i-1/2,+}u_i, \end{cases} \tag{34}$$

where the interfacial bathymetry $Z_{i+1/2}$ corresponds to

$$Z_{i+1/2} = \max(Z_i, Z_{i+1}). \tag{35}$$

With these definitions, the ODE satisfied by the approximation $U_i(t)$ reads

$$\frac{\mathrm{d}}{\mathrm{d}t}U_i + \frac{1}{\Delta x}\Big(\mathcal{F}(U_{i+1/2,-}, U_{i+1/2,+}) - \mathcal{F}(U_{i-1/2,-}, U_{i-1/2,+})\Big) = \frac{1}{\Delta x}S_i, \tag{36}$$

with the numerical flux $\mathcal{F}$ given in (27) now based on local velocities $\lambda_l = \lambda_{l,i+1/2}(t)$ (see Remark 3.1), and where the right-hand side is defined through

$$S_i = S_{i+1/2,-} - S_{i-1/2,+} = \frac{g}{2}\begin{pmatrix} 0 \\ h_{i+1/2,-}^2 - h_i^2 \end{pmatrix} - \frac{g}{2}\begin{pmatrix} 0 \\ h_{i-1/2,+}^2 - h_i^2 \end{pmatrix}. \tag{37}$$

The quantity $S_i/\Delta x$ is shown to be consistent with the bathymetry source term $S(h, Z) = (0, -gh\partial_x Z)$. In fact, we have

$$\begin{aligned} \frac{1}{\Delta x}S_i &= \begin{pmatrix} 0 \\ 1 \end{pmatrix} g\frac{h_{i-1/2,+} + h_{i+1/2,-}}{2} \cdot \frac{h_{i+1/2,-} - h_{i-1/2,+}}{\Delta x} \\ &= -\begin{pmatrix} 0 \\ 1 \end{pmatrix} gh_i\frac{Z_{i+1/2} - Z_{i-1/2}}{\Delta x} + O(\Delta x), \end{aligned}$$

where the second equality holds if Z is continuous (since it implies $Z_i - Z_{i\pm1/2} = O(\Delta x)$).

It will be useful to interpret the hydrostatic reconstruction at the kinetic level in order to quantify the entropy dissipation in the next section. To this

end, we introduce the interfacial Maxwellian

$$\mathcal{M}_{l,i+1/2} = \begin{cases} \mathcal{M}_l(U_{i+1/2,+}) = \mathcal{M}_{l,i+1/2,+} & \text{if } \lambda_{l,i+1/2} < 0, \\ \mathcal{M}_l(U_{i+1/2,-}) = \mathcal{M}_{l,i+1/2,-} & \text{if } \lambda_{l,i+1/2} > 0. \end{cases} \tag{38}$$

Then the numerical fluxes from (36) are simply expressed under the form

$$\mathcal{F}(U_{i+1/2,-}, U_{i+1/2,+}) = \sum_{1\le l\le L} \lambda_{l,i+1/2}\mathcal{M}_{l,i+/2}.$$

On the other hand, if we introduce

$$\begin{cases} \delta\mathcal{M}_{l,i+1/2,-} = \frac{g}{2}P_{\text{shift}}(h_i\mathcal{M}_l(U_i) - h_{i+1/2,-}\mathcal{M}_l(U_{i+1/2,-})), \\ \delta\mathcal{M}_{l,i-1/2,+} = \frac{g}{2}P_{\text{shift}}(h_i\mathcal{M}_l(U_i) - h_{i-1/2,+}\mathcal{M}_l(U_{i-1/2,+})), \end{cases} \tag{39}$$

with the shift matrix P_{shift} given in (10), we are able to recover exactly the source term contribution (37) since there holds:

$$S_{i+1/2,-} = -\sum_{1\le l\le L} \delta\mathcal{M}_{l,i+1/2,-}, \qquad S_{i-1/2,+} = -\sum_{1\le l\le L} \delta\mathcal{M}_{l,i-1/2,+}. \tag{40}$$

Finally, (36) becomes

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t}U_i + \frac{1}{\Delta x}\sum_{l=1}^{L}\Big(\lambda_{l,i+1/2}\mathcal{M}_{l,i+1/2} - \lambda_{l,i-1/2}\mathcal{M}_{l,i-1/2}\Big) \\ &\qquad = -\frac{1}{\Delta x}\sum_{l=1}^{L}\Big(\delta\mathcal{M}_{l,i+\frac{1}{2},-} - \delta\mathcal{M}_{l,i-\frac{1}{2},+}\Big). \end{aligned} \tag{41}$$

4.2 Exact quantification of the entropy dissipation

In Proposition 3.2, we have shown that when the bathymetry is flat, the fully discrete explicit scheme is entropy dissipative under a CFL condition. In the present section, we adapt this result for the semi-discrete scheme (36) in presence of a variable bathymetry, focusing exclusively on the entropy given by the energy E defined in (5). More precisely, we aim at quantifying exactly the dissipation induced by the numerical fluxes and source term from (36). This is achieved in Theorem 4.1, where we adapt and improve the result obtained in [6] to the vectorial BGK framework with finitely many velocities. The improvement means that instead of a discrete entropy *inequality*, we establish a discrete entropy *equality* with a dissipative term whose expression, involving the functions ψ_l defined in (24), is known explicitly. Hence, while the result from [6] might underestimate the rate of entropy dissipation, our result quantifies it precisely.

Theorem 4.1. *The solution* $(U_1, \dots, U_N)(t)$ *of* (36) *verifies the following equality:*

$$\begin{aligned} &\nabla_U E(U_i, Z_i)\cdot\Big(\mathcal{F}(U_{i+1/2,-}, U_{i+1/2,+}) - \mathcal{F}(U_{i-1/2,-}, U_{i-1/2,+}) - S_i\Big) \\ &\qquad = Q_{i+1/2} - Q_{i-1/2} - D_i, \end{aligned} \tag{42}$$

where the discrete entropy fluxes and the dissipation terms are given by

$$Q_{i\pm1/2} = \sum_{1\le l\le L} \lambda_{l,i\pm1/2} H_l(\mathcal{M}_{l,i\pm1/2}, Z_{i\pm1/2}), \tag{43}$$

$$\begin{aligned} D_i = & \sum_{\lambda_l<0} \lambda_l\Big(-\delta h_{i+1/2,-}\cdot\Big(\mathcal{M}_{l,i+1/2,+} - \mathcal{M}_{l,i+1/2,-}\Big) + \psi_{l,i+1/2,-}\Big) \\ & + \sum_{\lambda_l>0} \lambda_l\Big(-\delta h_{i-1/2,+}\cdot\Big(\mathcal{M}_{l,i-1/2,+} - \mathcal{M}_{l,i-1/2,-}\Big) - \psi_{l,i-1/2,+}\Big), \end{aligned} \tag{44}$$

with $H_l, Z_{i\pm1/2}, S_i, \mathcal{M}_{l,i\pm1/2}$ defined respectively in (19), (35), (37) *and* (38), *where the non-negative function ψ_l defined in* (24) *is evaluated as*

$$\begin{cases} \psi_{l,i+1/2,-} := \psi_l(\mathcal{M}_{l,i+1/2,-}, \mathcal{M}_{l,i+1/2,+}, Z_{i+1/2}), \\ \psi_{l,i-1/2,+} := \psi_l(\mathcal{M}_{l,i-1/2,+}, \mathcal{M}_{l,i-1/2,-}, Z_{i-1/2}), \end{cases} \tag{45}$$

and where the terms $\delta h_{i\pm1/2,\mp}$ correspond to

$$\begin{cases} \delta h_{i+1/2,-} := g(h_i + Z_i - h_{i+1/2,-} - Z_{i+1/2}, 0)^T, \\ \delta h_{i-1/2,+} := g(h_i + Z_i - h_{i-1/2,+} - Z_{i-1/2}, 0)^T. \end{cases} \tag{46}$$

Moreover, if the subcharacteristic condition (14) *holds locally in*

$$\mathcal{D}_0 = \{U_{i-1/2,-}, U_{i+1/2,+}\},$$

then the dissipation D_i is indeed non-positive.

Corollary 4.2. *Under the subcharacteristic condition, the Theorem 4.1 ensures that despite the use of the hydrostatic reconstruction, the semi-discrete scheme always contributes to dissipate the entropy. In fact, taking the scalar product of* (36) *with $\nabla_U E(U_i, Z_i)$ and using equality* (42) *we get:*

$$\partial_t E(U_i, Z_i) + \frac{1}{\Delta x}\Big(Q_{i+1/2} - Q_{i-1/2}\Big) = \frac{1}{\Delta x} D_i \le 0. \tag{47}$$

The remainder of this section is dedicated to the proof of Theorem 4.1.

Proof of Theorem 4.1. To make the computations more concise, we adopt the following notations:

$$\begin{aligned} &E_i' = \nabla_U E(U_i, Z_i), && \mathcal{F}_{i\pm1/2} = \mathcal{F}(U_{i\pm1/2,-}, U_{i\pm1/2,+}), \\ &H_{l,i}' = \nabla_f H_l(\mathcal{M}_{l,i}, Z_i), && H_{l,i+1/2,-}' = \nabla_f H_l(\mathcal{M}_{l,i+1/2,-}, Z_{i+1/2}), \\ &H_{l,i} = H_l(\mathcal{M}_{l,i}, Z_i), && H_{l,i-1/2,+}' = \nabla_f H_l(\mathcal{M}_{l,i-1/2,+}, Z_{i-1/2}). \end{aligned}$$

Note that with these notations, we do not distinguish between Maxwellians (and kinetic entropies) defined using the kinetic velocities at the right interface ($\lambda_l = \lambda_{l,i+1/2}$) or the left one ($\lambda_l = \lambda_{l,i-1/2}$). This is not an issue as the two

interfaces are treated separately so that these terms never interact. In fact, to prove equality (42) it is sufficient to show that

$$\begin{cases} E_i' \cdot \Big(\mathcal{F}_{i+1/2} - S_{i+1/2,-} - F(U_i) \Big) + Q(U_i, Z_i) = Q_{i+1/2} - D_{i+1/2,-}, \\ E_i' \cdot \Big(\mathcal{F}_{i-1/2} - S_{i-1/2,+} - F(U_i) \Big) + Q(U_i, Z_i) = Q_{i-1/2} + D_{i-1/2,+}, \end{cases} \tag{48}$$

for some $D_{i+1/2,-} + D_{i-1/2,+} = D_i$, and where Q is the entropy flux (5). The result will follow by subtracting the two equalities (48).

We begin by recalling an important consequence of Proposition 2.3, which is that E_i' is equal to $H_{l,i}'$ for any velocity index $1 \leq l \leq L$. Using the definitions (27) and (38) of the numerical flux $\mathcal{F}$ and the interfacial Maxwellian $\mathcal{M}_{l,i+1/2}$, as well as relation (40) for the kinetic interpretation of the source term, we have that the left-hand sides from (48) are respectively equal to

$$\begin{aligned} &\sum_{1 \leq l \leq L} H_{l,i}' \cdot \Big(\lambda_l \mathcal{M}_{l,i+1/2} + \delta \mathcal{M}_{l,i+1/2,-} - \lambda_l \mathcal{M}_{l,i} \Big) + \lambda_l H_{l,i}, \\ &\sum_{1 \leq l \leq L} H_{l,i}' \cdot \Big(\lambda_l \mathcal{M}_{l,i-1/2} + \delta \mathcal{M}_{l,i-1/2,+} - \lambda_l \mathcal{M}_{l,i} \Big) + \lambda_l H_{l,i}. \end{aligned}$$

We decompose the rest of the proof in three steps. First, we show that the summands can be developed under the form

$$\begin{aligned} H_{l,i}' \cdot \Big(\lambda_l \mathcal{M}_{l,i+1/2} + \delta \mathcal{M}_{l,i+1/2,-} - \lambda_l \mathcal{M}_{l,i} \Big) + \lambda_l H_{l,i} &= \mathcal{K}_{l,i+1/2,-} - D_{l,i+1/2,-}, \\ H_{l,i}' \cdot \Big(\lambda_l \mathcal{M}_{l,i-1/2} + \delta \mathcal{M}_{l,i-1/2,+} - \lambda_l \mathcal{M}_{l,i} \Big) + \lambda_l H_{l,i} &= \mathcal{K}_{l,i-1/2,+} + D_{l,i-1/2,+}. \end{aligned} \tag{49}$$

Second, we check that we recover (43)-(44), i.e.

$$\sum_{1 \leq l \leq L} \mathcal{K}_{l,i \pm 1/2,\mp} = Q_{i \pm 1/2} \quad \text{and} \quad \sum_{1 \leq l \leq L} (D_{l,i+1/2,-} + D_{l,i-1/2,+}) = D_i.$$

Finally we prove that $D_i \leq 0$ under the local subcharacteristic condition.

Step 1: kinetic entropy equality. Looking at the contribution from interface $i + 1/2$ we can write from (38):

$$\begin{aligned} &H_{l,i}' \cdot \Big(\lambda_l \mathcal{M}_{l,i+1/2} + \delta \mathcal{M}_{l,i+1/2,-} - \lambda_l \mathcal{M}_{l,i} \Big) \\ &\quad = H_{l,i}' \cdot \Big(\lambda_l \mathbb{1}_{\lambda_l < 0} \mathcal{M}_{l,i+1/2,+} + \lambda_l \mathbb{1}_{\lambda_l > 0} \mathcal{M}_{l,i+1/2,-} + \delta \mathcal{M}_{l,i+1/2,-} - \lambda_l \mathcal{M}_{l,i} \Big) \\ &\quad = \lambda_l \mathbb{1}_{\lambda_l < 0} H_{l,i}' \cdot \Big(\mathcal{M}_{l,i+1/2,+} - \mathcal{M}_{l,i+1/2,-} \Big) \\ &\qquad + H_{l,i}' \cdot \Big(\lambda_l \mathcal{M}_{l,i+1/2,-} - \lambda_l \mathcal{M}_{l,i} + \delta \mathcal{M}_{l,i+1/2,-} \Big). \end{aligned} \tag{50}$$

Similarly at interface $i-1/2$ there holds

$$
\begin{aligned}
&H'_{l,i}\cdot\Big(\lambda_l\mathcal{M}_{l,i-1/2}+\delta\mathcal{M}_{l,i-1/2,+}-\lambda_l\mathcal{M}_{l,i}\Big) \qquad (51)\\
&\quad=H'_{l,i}\cdot\Big(\lambda_l\mathbb{1}_{\lambda_l<0}\mathcal{M}_{l,i-1/2,+}+\lambda_l\mathbb{1}_{\lambda_l>0}\mathcal{M}_{l,i-1/2,-}+\delta\mathcal{M}_{l,i-1/2,+}-\lambda_l\mathcal{M}_{l,i}\Big)\\
&\quad=-\lambda_l\mathbb{1}_{\lambda_l>0}H'_{l,i}\cdot\Big(\mathcal{M}_{l,i-1/2,+}-\mathcal{M}_{l,i-1/2,-}\Big)\\
&\qquad+H'_{l,i}\cdot\Big(\lambda_l\mathcal{M}_{l,i-1/2,+}-\lambda_l\mathcal{M}_{l,i}+\delta\mathcal{M}_{l,i-1/2,+}\Big).
\end{aligned}
$$

Using again the third equality of Proposition 2.3, there holds $H'_{l,i}=(-\frac{1}{2}u_i^2+gh_i,u_i)$; thanks to the fact that the hydrostatic reconstruction (33)-(34) leaves the velocity u unchanged, this leads to the identities

$$
H'_{l,i}=H'_{l,i+1/2,-}+\delta h_{i+1/2,-},\qquad H'_{l,i}=H'_{l,i-1/2,+}+\delta h_{i-1/2,+},
$$

with the vectors $\delta h_{i\pm1/2,\mp}$ defined in (46). Hence, equalities (50)-(51) become

$$
\begin{aligned}
&H'_{l,i}\cdot\Big(\lambda_l\mathcal{M}_{l,i+1/2}+\delta\mathcal{M}_{l,i+1/2,-}-\lambda_l\mathcal{M}_{l,i}\Big) \qquad (52)\\
&\quad=\lambda_l\mathbb{1}_{\lambda_l<0}H'_{l,i+1/2,-}\cdot\Big(\mathcal{M}_{l,i+1/2,+}-\mathcal{M}_{l,i+1/2,-}\Big)\\
&\qquad+\lambda_l\mathbb{1}_{\lambda_l<0}\delta h_{i+1/2,-}\cdot\Big(\mathcal{M}_{l,i+1/2,+}-\mathcal{M}_{l,i+1/2,-}\Big)\\
&\qquad+H'_{l,i}\cdot\Big(\lambda_l\mathcal{M}_{l,i+1/2,-}-\lambda_l\mathcal{M}_{l,i}+\delta\mathcal{M}_{l,i+1/2,-}\Big),
\end{aligned}
$$

$$
\begin{aligned}
&H'_{l,i}\cdot\Big(\lambda_l\mathcal{M}_{l,i-1/2}+\delta\mathcal{M}_{l,i-1/2,+}-\lambda_l\mathcal{M}_{l,i}\Big) \qquad (53)\\
&\quad=-\lambda_l\mathbb{1}_{\lambda_l>0}H'_{l,i-1/2,+}\cdot\Big(\mathcal{M}_{l,i-1/2,+}-\mathcal{M}_{l,i-1/2,-}\Big)\\
&\qquad-\lambda_l\mathbb{1}_{\lambda_l>0}\delta h_{i-1/2,+}\cdot\Big(\mathcal{M}_{l,i-1/2,+}-\mathcal{M}_{l,i-1/2,-}\Big)\\
&\qquad+H'_{l,i}\cdot\Big(\lambda_l\mathcal{M}_{l,i-1/2,+}-\lambda_l\mathcal{M}_{l,i}+\delta\mathcal{M}_{l,i-1/2,+}\Big).
\end{aligned}
$$

The first scalar product appearing on the right-hand side of (52) is then developed in terms of $\psi_{l,i+1/2,-}$ given in (45), which we recall results from the definition (24):

$$
\begin{aligned}
&\lambda_l\mathbb{1}_{\lambda_l<0}H'_{l,i+1/2,-}\cdot\Big(\mathcal{M}_{l,i+1/2,+}-\mathcal{M}_{l,i+1/2,-}\Big) \qquad (54)\\
&\quad=\lambda_l\mathbb{1}_{\lambda_l<0}\Big(H_{l,i+1/2,+}-H_{l,i+1/2,-}-\psi_{l,i+1/2,-}\Big)\\
&\quad=\lambda_l\Big(\mathbb{1}_{\lambda_l<0}H_{l,i+1/2,+}+\mathbb{1}_{\lambda_l>0}H_{l,i+1/2,-}\Big)\\
&\qquad\qquad-\lambda_lH_{l,i+1/2,-}-\lambda_l\mathbb{1}_{\lambda_l<0}\psi_{l,i+1/2,-},
\end{aligned}
$$

where we used the notation $H_{l,i+1/2,+}=H_l(\mathcal{M}_{l,i+1/2,+},Z_{i+1/2})$. Likewise, using $\psi_{l,i-1/2,+}=\psi_l(\mathcal{M}_{l,i-1/2,+},\mathcal{M}_{l,i-1/2,-},Z_{i-1/2})$, the first term in the right-

hand side of equality (53) can be developed as

$$
\begin{aligned}
&- \lambda_l \mathbb{1}_{\lambda_l>0} H'_{l,i-1/2,+} \cdot \Big(\mathcal{M}_{l,i-1/2,+} - \mathcal{M}_{l,i-1/2,-} \Big) \\
&\quad = -\lambda_l \mathbb{1}_{\lambda_l>0} \Big(H_{l,i-1/2,+} - H_{l,i-1/2,-} + \psi_{l,i-1/2,+} \Big) \\
&\quad = \lambda_l \Big(\mathbb{1}_{\lambda_l<0} H_{l,i-1/2,+} + \mathbb{1}_{\lambda_l>0} H_{l,i-1/2,-} \Big) \\
&\qquad\qquad - \lambda_l H_{l,i-1/2,+} - \lambda_l \mathbb{1}_{\lambda_l>0} \psi_{l,i-1/2,+}.
\end{aligned}
\tag{55}
$$

Adding $\lambda_l H_{l,i}$ to equations (52)-(53) and using (54)-(55), we obtain equalities (49) where the kinetic entropy-fluxes are given by

$$
\begin{cases}
\mathcal{K}_{l,i+1/2,-} = \lambda_l \mathbb{1}_{\lambda_l<0} H_l(\mathcal{M}_{l,i+1/2,+}, Z_{i+1/2}) \\
\qquad\qquad + \lambda_l \mathbb{1}_{\lambda_l>0} H_l(\mathcal{M}_{l,i+1/2,-}, Z_{i+1/2}) \\
\qquad\qquad + \lambda_l H_l(\mathcal{M}_{l,i}, Z_i) - \lambda_l H_l(\mathcal{M}_{l,i+1/2,-}, Z_{i+1/2}) \\
\qquad\qquad + H'_{l,i} \cdot (\lambda_l \mathcal{M}_{l,i+1/2,-} - \lambda_l \mathcal{M}_{l,i} + \delta \mathcal{M}_{l,i+1/2,-}), \\
\mathcal{K}_{l,i-1/2,+} = \lambda_l \mathbb{1}_{\lambda_l<0} H_l(\mathcal{M}_{l,i-1/2,+}, Z_{i-1/2}) \\
\qquad\qquad + \lambda_l \mathbb{1}_{\lambda_l>0} H_l(\mathcal{M}_{l,i-1/2,-}, Z_{i-1/2}) \\
\qquad\qquad + \lambda_l H_l(\mathcal{M}_{l,i}, Z_i) - \lambda_l H_l(\mathcal{M}_{l,i-1/2,+}, Z_{i-1/2}) \\
\qquad\qquad + H'_{l,i} \cdot (\lambda_l \mathcal{M}_{l,i-1/2,+} - \lambda_l \mathcal{M}_{l,i} + \delta \mathcal{M}_{l,i-1/2,+}),
\end{cases}
\tag{56}
$$

and where the dissipative terms correspond to

$$
\begin{cases}
D_{l,i+1/2,-} = \lambda_l \mathbb{1}_{\lambda_l<0} \big(\delta h_{i+1/2,-} \cdot \big(\mathcal{M}_{l,i+1/2,-} - \mathcal{M}_{l,i+1/2,+} \big) + \psi_{l,i+1/2,-} \big), \\
D_{l,i-1/2,+} = \lambda_l \mathbb{1}_{\lambda_l>0} \big(\delta h_{i-1/2,+} \cdot \big(\mathcal{M}_{l,i-1/2,-} - \mathcal{M}_{l,i-1/2,+} \big) - \psi_{l,i-1/2,+} \big).
\end{cases}
\tag{57}
$$

Step 2: macroscopic entropy flux and dissipation. The equality $\sum_l (D_{l,i+1/2,-} + D_{l,i-1/2,+}) = D_i$ obviously coincides with (44). Hence, we focus on showing that $\sum_l \mathcal{K}_{l,i\pm1/2,\mp}$ is equal to $Q_{i\pm1/2}$ defined in (43). Considering interface $i+1/2$ and recalling that $H'_{l,i} = E'_i$, we obtain after summing (56) over $1 \le l \le L$ and subtracting by $Q_{i+1/2}$:

$$
\begin{aligned}
\sum_{1\le l\le L} \mathcal{K}_{l,i+1/2,-} - Q_{i+1/2} &= Q(U_i, Z_i) - Q(U_{i+1/2,-}, Z_{i+1/2}) \\
&\quad + E'_i \cdot \Big(F(U_{i+1/2,-}) - F(U_i) - S_{i+1/2,-} \Big).
\end{aligned}
$$

The right-hand side is shown to be zero by remarking that

$$
E'_i \cdot S_{i+1/2,-} = \frac{g}{2} h^2_{i+1/2,-} u_i - \frac{g}{2} h_i^2 u_i, \qquad E'_i \cdot F(U_i) = Q(U_i, Z_i) + \frac{g}{2} h_i^2 u_i,
$$

together with

$$\begin{aligned}
E_i' \cdot F(U_{i+1/2,-}) &= E_{i+1/2,-}' \cdot F(U_{i+1/2,-}) + (E_i' - E_{i+1/2,-}') \cdot F(U_{i+1/2,-}) \\
&= Q(U_{i+1/2,-}, Z_{i+1/2}) + \frac{g}{2} h_{i+1/2,-}^2 u_i \\
&\qquad + g(h_i + Z_i - h_{i+1/2,-} - Z_{i+1/2}) h_{i+1/2,-} u_i \\
&= Q(U_{i+1/2,-}, Z_{i+1/2}) + \frac{g}{2} h_{i+1/2,-}^2 u_i.
\end{aligned}$$

To obtain the last equality we used that either $h_{i+1/2,-} = 0$, or $h_{i+1/2,-} > 0$, which by definition (33) of the reconstructed water height implies that $h_i + Z_i - h_{i+1/2,-} - Z_{i+1/2} = 0$. Since the steps for interface $i - 1/2$ are exactly the same we omit them.

Step 3: Dissipative terms are non-positive. Finally, we show that the terms $D_{l,i\pm1/2,\mp}$ defined in (57) are non-positive under a local subcharacteristic condition. Given that the function ψ_l is non-negative as per Lemma 2.5, it is sufficient to prove the following inequalities:

$$\begin{aligned}
\delta h_{i+1/2,-} \cdot \Big(\mathcal{M}_{l,i+1/2,+} - \mathcal{M}_{l,i+1/2,-} \Big) \leq 0, \\
\delta h_{i-1/2,+} \cdot \Big(\mathcal{M}_{l,i-1/2,+} - \mathcal{M}_{l,i-1/2,-} \Big) \geq 0.
\end{aligned}$$

Using the same argument as in the second step, we have that either $h_{i+1/2-} = h_i + Z_i - Z_{i+1/2}$, or that $h_{i+1/2-}$ cancels. In the first case there is $\delta h_{i+1/2,-} = 0$ owing to (46), and the desired inequality follows trivially. In the second case the first component of $\mathcal{M}_{l,i+1/2,-}$ also cancels since by (12) it is proportional to $h_{i+1/2,-}$, and we then have

$$\begin{aligned}
&g(h_i + Z_i - h_{i+1/2,-} - Z_{i+1/2}) \Big(\mathcal{M}_{l,i+1/2,+} - \mathcal{M}_{l,i+1/2,-} \Big)_1 \\
&\qquad = g(h_i + Z_i - Z_{i+1/2}) \Big(\mathcal{M}_{l,i+1/2,+} \Big)_1 \\
&\qquad \leq 0.
\end{aligned}$$

The above inequality holds since Proposition 3.3 implies $(\mathcal{M}_{l,i+1/2,+})_1 \geq 0$ under the local subcharacteristic condition $\sigma(\mathcal{M}_l'(U_{i+1/2,+})) \subset]0, +\infty[$ and since by definition (33) of $h_{i+1/2,-}$ there holds

$$h_{i+1/2,-} = 0 \implies h_i + Z_i - Z_{i+1/2} \leq 0. \tag{58}$$

We can treat the case of interface $i - 1/2$ similarly, and the proof is complete. □

5 Fully discrete scheme

In the previous section, we obtained a semi-discrete entropy equality featuring an error term with dissipative properties provided that the subcharacteristic

condition (14) is satisfied. However, accounting for the dissipation implied by the spatial discretization alone is not enough to ensure that the fully discrete scheme is entropy stable, since the impact of the error in time can enter in competition with this stabilizing effect. This is particularly true when using an explicit kinetic scheme such as the one studied in [6]. In fact, in some cases explicit approximations are not entropy stable, independently of the choice of time step and mesh refinement.

In order to overcome this lack of stability, an implicit approach has been studied in [23]. However, such implicit methods usually come with a high computational cost together with an over-diffusive behavior. Rather, to avoid these issues we propose here an adaptive time-stepping technique. More precisely, we consider a local-in-space convex combination between a fully implicit and explicit time discretizations of the Saint-Venant equations (3). This convex combination will be tuned in each cell to optimize the numerical viscosity by cancelling the entropy created when using the explicit discretization.

5.1 Adaptive time discretization

The starting point is the semi-discrete scheme defined through (36), which relies on the numerical flux $\mathcal{F}$ induced by an upwinding at the kinetic level (27), and combined with the hydrostatic reconstruction studied in the previous section. We then need to replace the continuous approximation $t \mapsto (U_1, \dots, U_N)(t)$ by a discrete time-stepping procedure. This is achieved through an adaptive strategy which automatically combines the forward and backward Euler methods to ensure entropy dissipation while limiting the diffusion of the scheme. To this end, we introduce a vector of parameters $\theta = (\theta_1, \dots, \theta_N) \in [0,1]^N$ where we recall N is the number of interior cells. We can then define the following convex combination between the current state U^n and its update U^{n+1} in each cell:

$$\forall 1 \leq i \leq N, \quad U_i^\theta = \theta_i U_i^{n+1} + (1-\theta_i) U_i^n, \tag{59}$$

where the exponent n refers to the discrete time $t^n = n\Delta t$ with Δt the time step. This convex combination is then used to compute the numerical fluxes and the source term, such that an interface $i+1/2$ is treated with the forward Euler method when $\theta_i = \theta_{i+1} = 0$, and with the backward Euler method when these parameters are both equal to one. With this in mind, the adaptive update reads

$$\frac{U_i^{n+1} - U_i^n}{\Delta t} + \frac{1}{\Delta x}\Big(\mathcal{F}(U^\theta_{i+1/2,-}, U^\theta_{i+1/2,+}) - \mathcal{F}(U^\theta_{i-1/2,-}, U^\theta_{i-1/2,+})\Big) = \frac{1}{\Delta x} S_i^\theta, \tag{60}$$

where the source term S_i^θ and the reconstructed states $U^\theta_{i\pm1/2,\mp}$ correspond to the definitions given in Section 4.1, only substituting $U(t)$ by the combined state U^θ defined in (59). We take the convention that ghost cells are always evaluated explicitly. The update (60) is obviously conservative for the water height and, when the bathymetry is flat, for the discharge as well.

As we will see below, the choice of the vector θ will be made such that the error in time, contributing in some cases to increase the entropy, is compensated by the spatial error which, as a consequence of Corollary 4.2, is always dissipative. The components of θ are allowed to vary between different cells and from one iteration to the other, hence they are also part of the unknowns. As soon as one of the θ_i is non-zero, we are faced with a system of nonlinear coupled equations on U^{n+1} which we do not try to solve analytically. Instead, we will approximate the update through a fixed point method that we detail in Section 5.2. Before, we justify that the scheme (60) is well-posed for Δt small enough regardless of the values taken by the components of θ.

Proposition 5.1. *Assume that there holds $U^n \in \mathcal{D}^N$ with $\mathcal{D} =]0, +\infty[\times\mathbb{R}$ the set of wet states. Furthermore, assume that there holds $h_i^n + Z_i \geq Z_{i\pm1}$ for all i, that is, no interface is partially wet. Then there exists $\Delta t^* > 0$ such that for all $\theta \in [0,1]^N$ and for all $0 \leq \Delta t < \Delta t^*$ the update (60) admits a unique solution $(U_i^{n+1})_{1\leq i\leq N}$. Moreover, the map $(\theta, \Delta t) \mapsto U^{n+1}$ is in $C^1([0,1]^N \times [0, \Delta t^*[; \mathcal{D}^N)$.*

Proof of Proposition 5.1. Remark that multiplying (60) by Δt, it is possible to recast the equations obtained for all $1 \leq i \leq N$ under the vector form

$$0_{(\mathbb{R}^2)^N} = \phi(\theta, \Delta t, U^{n+1}) = \begin{pmatrix} U_1^{n+1} - U_1^n \\ \vdots \\ U_N^{n+1} - U_N^n \end{pmatrix} + \frac{\Delta t}{\Delta x}\Big(\text{function of } (\theta_i U_i^{n+1})_{1\leq i\leq N}\Big). \tag{61}$$

Since there is no partially wet interface, the hydrostatic reconstruction is linear, in particular the maximum function used in the reconstructed states (33)-(34) never activates; as a consequence ϕ is in $C^1(\mathbb{R}^N \times \mathbb{R} \times \mathcal{D}^N)$. Moreover, Equation (61) admits a trivial solution $U^{n+1} = U^n \in \mathcal{D}^N$ for all $\theta \in [0,1]^N$ when $\Delta t = 0$, with the Jacobian $D_{(U_1,\dots,U_N)}\phi(\theta, 0, U^n) = \mathrm{Id}$ invertible. By the implicit function theorem, there exists $\Delta t^* > 0$ such that the map $(\theta, \Delta t) \mapsto U^{n+1}$ is well-defined of class $C^1([0,1]^N \times [0, \Delta t^*[; \mathcal{D}^N)$. □

We remark that the discrete kinetic scheme with adaptive time stepping (60) can be made compatible with dry states $h = 0$ by using the desingularization method proposed in [25]. This latter consists of defining the velocity u as q/h in cells where $h > \Delta x$, sets it to zero when $h = 0$, and smoothly connects these states for $0 < h < \Delta x$. In practice, thanks to this modification we were able to successfully simulate test cases featuring vanishing water heights and partially wet interfaces, see Sections 6.2 and 6.4. However, we restrict the entropy stability analysis performed below to wet states, i.e. to $U \in \mathcal{D}$. This analysis will provide us with a criterion to select the components of the vector θ. This is done by extending the exact quantification of entropy dissipation from Corollary 4.2 to the presence of the additional error in time.

Proposition 5.2. *The update U^{n+1} of the fully discrete scheme* (60) *satisfies*

$$\frac{1}{\Delta t}\Big(E(U_i^{n+1}, Z_i) - E(U_i^n, Z_i)\Big) + \frac{1}{\Delta x}\Big(Q^\theta_{i+1/2} - Q^\theta_{i-1/2}\Big) = \frac{1}{\Delta t}\mathcal{E}_i^\theta + \frac{1}{\Delta x} D_i^\theta, \tag{62}$$

where $Q^\theta_{i\pm1/2}, D_i^\theta$ correspond to expressions (43)-(44) *evaluated with the combined state U^θ given in* (59)*, and where the time error $\mathcal{E}_i^\theta$ is given by*

$$\mathcal{E}_i^\theta = E(U_i^{n+1}, Z_i) - E(U_i^n, Z_i) - \nabla_U E(U_i^\theta, Z_i) \cdot (U_i^{n+1} - U_i^n). \tag{63}$$

Proof of Proposition 5.2. Taking the scalar product between $\nabla_U E(U_i^\theta, Z_i)$ and the adaptive update (60), and applying the result from Theorem 4.1 with $U = U^\theta$, we have

$$\frac{1}{\Delta t}\nabla_U E(U_i^\theta, Z_i) \cdot (U_i^{n+1} - U_i^n) + \frac{1}{\Delta x}\Big(Q^\theta_{i+1/2} - Q^\theta_{i-1/2}\Big) = \frac{1}{\Delta x} D_i^\theta.$$

We obtain equality (62) by remarking that due to definition (63) of the time error there holds

$$\nabla_U E(U_i^\theta, Z_i) \cdot (U_i^{n+1} - U_i^n) = E(U_i^{n+1}, Z_i) - E(U_i^n, Z_i) - \mathcal{E}_i^\theta.$$

□

By convexity of the macroscopic entropy (5), definition (63) implies that $\mathcal{E}_i^\theta$ is non-negative when $\theta_i = 0$, and non-positive when $\theta_i = 1$. Moreover, if the subcharacteristic condition (14) is satisfied locally at each interface, there holds $D_i^\theta \leq 0$ for all θ. Hence, the right-hand side $\frac{1}{\Delta t}\mathcal{E}^\theta + \frac{1}{\Delta x} D^\theta$ from (62) might be positive for $\theta \equiv 0$, but is always non-positive for $\theta \equiv 1$. A similar property was leveraged in [23] to stabilize the explicit scheme from [6] by impliciting it.

The other key takeaway here is that it is the hydrostatic reconstruction which is responsible for the lack of entropy stability of the explicit version of the scheme (60). Indeed, the reconstruction *weakens* the rate of entropy dissipation associated to the spatial discretization, making it unable to compensate for the error in time which has the wrong sign. This is seen by the fact that over flat bathymetries (i.e. when no reconstruction is performed in (33)-(34)), Proposition 3.2 gives the entropy stability in the fully explicit case ($\theta \equiv 0$) under a CFL condition; but for any non-flat bathymetry (i.e. when the hydrostatic reconstruction comes into play), one can construct counterexamples where the entropy is increased by the explicit scheme. We provide such a counterexample as a test case in Section 6.1.

Finally, we remark that to make the scheme (60) entropy stable, it is not always necessary to enforce $\theta \equiv 1$; in many situations it is enough to have $\theta_i \in [0, 1[$. This allows to improve the accuracy of the scheme, since lower values of θ mean less numerical diffusion. Hence, an optimal choice would be that of the smallest $\theta_i \geq 0$ making $\frac{1}{\Delta t}\mathcal{E}_i^\theta + \frac{1}{\Delta x} D_i^\theta$ non-positive. However, this definition suffers from the fact that the components of θ are coupled through

the numerical fluxes $\mathcal{F}$: reducing the value of θ_i might increase the entropy in neighboring cells $i \pm 1$ and make the scheme unstable. In the next section, we provide an algorithm which approximates a suitable vector $0 \leq \theta \leq 1$ with low values avoiding the problematic behavior mentioned above. In particular, this algorithm tries to use the forward Euler method whenever and wherever possible, which also reduces the computational cost associated with implicit time-stepping.

5.2 Fixed point approximation of the adaptive update

In general, the update (60) cannot be computed analytically. Instead, we will construct a sequence $(U^{n+1}(k), \theta(k))_{k\in\mathbb{N}}$ approximating (U^{n+1}, θ) from a fixed point method, where the vector of parameters θ makes the update entropy-stable. We also discuss how to choose the time step Δt to ensure the non-negativity of the water height h, as well as a stopping criteria based on tolerance and entropy dissipation arguments.

For conciseness, we will note $\sigma = \Delta t/\Delta x$. We assume given an approximation $U^n \in \mathcal{D}^N$ of the Saint-Venant system (1) at time t^n, and we would like to approximate the update U^{n+1} defined in (60). Supposing for a moment that $\theta_i(k)$ — to be defined later — is converging to some $0 \leq \theta_i \leq 1$, the core of the fixed point method is based on the following sub-iteration with respect to index $k \in \mathbb{N}$:

$$(1+r)U_i^{n+1}(k+1) = U_i^n + rU_i^{n+1}(k) - \sigma\Big(\mathcal{F}^\theta_{i+1/2} - \mathcal{F}^\theta_{i-1/2} - S_i^\theta\Big)(k), \quad (64)$$

where $r > 0$ is a relaxation parameter, and where the combined state, the numerical flux and source term are given by

$$\begin{cases} U_i^\theta(k) = \theta_i(k)U_i^{n+1}(k) + (1-\theta_i(k))U_i^n, \\ \mathcal{F}^\theta_{i+1/2}(k) = \mathcal{F}(U^\theta_{i+1/2,-}(k), U^\theta_{i+1/2,+}(k)), \\ S_i^\theta(k) = (0,1)^T\Big(\frac{g}{2}(h^\theta_{i+1/2,-}(k))^2 - \frac{g}{2}(h^\theta_{i-1/2,+}(k))^2\Big). \end{cases} \quad (65)$$

The initial state is $U^{n+1}(0) = U^n$ and $\theta(0) = 0$. We see that if the sequence $(U^{n+1}(k))_{k\in\mathbb{N}}$ defined in (64) converges, then its limit is the solution U^{n+1} of the adaptive update (60). Besides, we give a CFL condition to ensure non-negativity of the water height computed in (64).

Proposition 5.3. *Assume that $h_i^n, h_i^{n+1}(k) \geq 0$, and that the subcharacteristic condition holds locally at each interface, i.e.*

$$\forall U \in \{U^\theta_{i+1/2,-}(k), U^\theta_{i+1/2,+}(k)\}, \quad \sigma(\mathcal{M}_l'(U)) \subset]0, +\infty[. \quad (66)$$

Then the $(k+1)$-th sub-iteration $h_i^{n+1}(k+1)$ defined as the first component of (64) is kept non-negative under the CFL conditions

$$\frac{\Delta t}{\Delta x}(1-\theta_i(k))A(U_i^n) \leq 1, \qquad \frac{\Delta t}{\Delta x}\theta_i(k)A(U_i^{n+1}(k)) \leq r, \quad (67)$$

where the map A is defined as

$$A(U_i) = \left[\sum_{\lambda_l>0} \lambda_{l,i-1/2} \begin{pmatrix} \alpha_{l,i-1/2} \\ \beta_{l,i-1/2} \end{pmatrix} - \sum_{\lambda_l<0} \lambda_{l,i+1/2} \begin{pmatrix} \alpha_{l,i+1/2} \\ \beta_{l,i+1/2} \end{pmatrix} \right] \cdot \begin{pmatrix} 1 \\ u_i \end{pmatrix} \geq 0. \tag{68}$$

It is interesting to note that for the second inequality (67) to be satisfied when at least one cell is treated implicitly (i.e. $\theta_i(k) > 0$), we need to set the relaxation parameter r to be strictly positive. In practice, we will pick $r = 1$ for $k \geq 1$, and $r = 0$ for the initial step $k = 0$. The reason for this distinction is that we want the first sub-iteration of (64) to coincide with the forward Euler scheme — recalling that $\theta(0) \equiv 0$. Moreover, a difficulty is that for a fixed Δt the CFL condition (67) might be satisfied at sub-iteration k but not $k+1$. To circumvent this issue, it is possible to make Δt dependent on k: at each sub-iteration the value of $\Delta t(k)$ is set to the largest positive quantity verifying the CFL condition.

Proof of Proposition 5.3. Noting $\sigma = \Delta t/\Delta x$, we write the first component of sub-iteration k from (64) as

$$\begin{aligned} (1+r)h_i^{n+1}(k+1) &= h_i^n + rh_i^{n+1}(k) - \sigma\Big(\mathcal{F}^\theta_{h,i+1/2}(k) - \mathcal{F}^\theta_{h,i-1/2}(k)\Big), \qquad (69)\\ &= h_i^n + rh_i^{n+1}(k) - \sigma\Big(\sum_{\lambda_l<0} \lambda_{l,i+1/2}\mathcal{M}^\theta_{l1,i+1/2,+} + \sum_{\lambda_l>0} \lambda_{l,i+1/2}\mathcal{M}^\theta_{l1,i+1/2,-}\Big) \\ &\qquad + \sigma\Big(\sum_{\lambda_l<0} \lambda_{l,i-1/2}\mathcal{M}^\theta_{l1,i-1/2,+} + \sum_{\lambda_l>0} \lambda_{l,i-1/2}\mathcal{M}^\theta_{l1,i-1/2,-}\Big), \end{aligned}$$

with $\mathcal{M}^\theta_{l1,i+1/2,\pm}$ referring to the first component of $\mathcal{M}_l(U^\theta_{i+1/2,\pm})$. Under the local subcharacteristic condition (66), Lemma 3.3 implies that this quantity is positive. Applying the same argument at interface $i-1/2$, we can bound (69) from below by omitting the positive sums, i.e.

$$\begin{aligned} &(1+r)h_i^{n+1}(k+1) \\ &\quad \geq h_i^n + rh_i^{n+1}(k) - \sigma \sum_{\lambda_l>0} \lambda_{l,i+1/2}\mathcal{M}^\theta_{l1,i+1/2,-} + \sigma \sum_{\lambda_l<0} \lambda_{l,i-1/2}\mathcal{M}^\theta_{l1,i-1/2,+}. \end{aligned}$$

Using definition (12) of the Maxwellian, the fact that $\mathcal{M}_{l1} \geq 0$ is proportional to the water height h, and that by definition (33)-(34) of the reconstructed states there holds $h_{i\pm1/2,\mp} \leq h_i$, this yields

$$\begin{aligned} (1+r)h_i^{n+1}(k+1) \geq h_i^n + rh_i^{n+1}(k) &- \sigma \sum_{\lambda_l>0} \lambda_{l,i+1/2} \begin{pmatrix} \alpha_{l,i+1/2} \\ \beta_{l,i+1/2} \end{pmatrix} \cdot U_i^\theta(k) \\ &+ \sigma \sum_{\lambda_l<0} \lambda_{l,i-1/2} \begin{pmatrix} \alpha_{l,i-1/2} \\ \beta_{l,i-1/2} \end{pmatrix} \cdot U_i^\theta(k). \end{aligned}$$

Finally, using the definition (65) of the combined state $U_i^\theta(k)$, this inequality is rewritten equivalently in terms of the map A given in (68) as follows:

$$\begin{aligned}(1+r)h_i^{n+1}(k+1) \\ \geq \Big(1-\sigma(1-\theta_i(k))A(U_i^n)\Big)h_i^n + \Big(r-\sigma\theta_i(k)A(U_i^{n+1}(k))\Big)h_i^{n+1}(k).\end{aligned} \tag{70}$$

Hence, it is sufficient to verify the CFL conditions (67) to ensure the non-negativity of $h_i^{n+1}(k+1)$. □

Next we show that the sub-iteration process (64) satisfies an entropy equality analogous to (62).

Proposition 5.4. *The recurrence relation* (64) *implies the following equality*

$$\begin{aligned}&\frac{1}{\Delta t}\Big(E(U_i^{n+1}(k+1),Z_i)-E(U_i^n,Z_i)\Big)+\frac{1}{\Delta x}\Big(Q_{i+1/2}^\theta(k)-Q_{i-1/2}^\theta(k)\Big)\\ &\qquad=\frac{1}{\Delta t}\mathcal{E}_i^\theta(k)+\frac{1}{\Delta x}D_i^\theta(k),\end{aligned} \tag{71}$$

where $Q_{i+1/2}^\theta(k)$ and $D_i^\theta(k)$ correspond respectively to (43) *and* (44) *evaluated with $U=U^\theta(k)$, and where the time error $\mathcal{E}_i^\theta(k)$ now depends on the states U_i^n and $U_i^{n+1}(k)$ notably through $U_i^\theta(k)$, and on $U_i^{n+1}(k+1)$:*

$$\begin{aligned}\mathcal{E}_i^\theta(k) = E(U_i^{n+1}(k+1),Z_i)-E(U_i^n,Z_i) \\ -\nabla_U E(U_i^\theta(k),Z_i)\cdot\Big((1+r)U_i^{n+1}(k+1)-U_i^n-rU_i^{n+1}(k)\Big).\end{aligned} \tag{72}$$

Similarly to Proposition 5.2, the above result is obtained by taking the scalar product between $\nabla_U E(U_i^\theta(k),Z_i)$ and the sub-iteration step (64). The main difference lies in the presence of additional terms in factor of the relaxation coefficient r; since these terms are straightforward to handle we do not detail the proof. An important remark is that if the sequence $(U^{n+1}(k))_{k\in\mathbb{N}}$ converges, then from some rank k large enough the time error (72) behaves similarly as (63): it is non-negative when $\theta_i(k)=0$ and non-positive when $\theta_i(k)=1$. This is true because the entropy $E(U,Z)$ is continuous and strictly convex for $h>0$, and because the terms in factor of r cancel each other in the limit $k\to+\infty$.

It remains to explain how the sequence $(\theta(k))_{k\in\mathbb{N}}$ is defined to ensure the dissipation of entropy while limiting the amount of diffusion. We recall that we initialize it by zero: $\theta(0)\equiv 0$. After computing $U^{n+1}(k+1)$ through (64), we can quantify the entropy dissipation (or lack of) in every cell by evaluating the right-hand side of (71). In cells i where this quantity is non-positive, we simply set $\theta_i(k+1)=\theta_i(k)$. In every other cells, we set $\theta_i(k+1)=\max(\theta_i(k),\min(1,\bar\theta))$ where $\bar\theta\in\mathbb{R}$ is the solution of the following linear equation:

$$0=\frac{1}{\Delta t}\Big(\delta_t E_i-\nabla_U E_i^{\rm lin}(\bar\theta)\cdot\delta_t U_i\Big)+\frac{1}{\Delta x}D_i^\theta(k), \tag{73}$$

where we defined the discrete time variations

$$\delta_t E_i = E(U_i^{n+1}(k+1),Z_i)-E(U_i^n,Z_i),\qquad \delta_t U_i = U_i^{n+1}(k+1)-U_i^n,$$

and where $\nabla_U E_i^{\mathrm{lin}}$ is the linearized of $\bar{\theta} \mapsto \nabla_U E(U_i^n + \bar{\theta}\delta_t U_i, Z_i)$ around $\theta_i(k)$:

$$\nabla_U E_i^{\mathrm{lin}}(\bar{\theta}) = \nabla_U E(U_i^n + \theta_i(k)\delta_t U_i, Z_i) + (\bar{\theta} - \theta_i(k))\mathrm{Hess}_U E(U_i^n + \theta_i(k)\delta_t U_i, Z_i)\delta_t U_i.$$

Note that in (73), the expression on the right is an approximation of the right-hand side of (62). An important remark is that the Hessian of the entropy appearing above is symmetric positive definite. Hence, the definition (73) of $\bar{\theta}$ is always well-posed except when $\delta_t U_i = 0$. We can exclude this case since it implies that $\delta_t E_i = 0$ and thus the entropy is dissipated in cell i — which we recall is treated by setting $\theta_i(k+1) = \theta_i(k)$. Besides, for all i the sequence $(\theta_i(k))_k$ that we constructed is increasing and bounded from above by one, hence it converges in $[0, 1]$.

Algorithm 1: Fixed point method for approximating U^{n+1} and θ

Input: $\Delta x > 0, \quad \epsilon_{\mathrm{tol}} > 0, \quad U^n \in (\mathbb{R}_+ \times \mathbb{R})^N$

First try: forward Euler update

1 Initialize $\delta \leftarrow 0, \quad k \leftarrow 0, \quad \theta(k) \leftarrow 0_{\mathbb{R}^N}, \quad U^\theta(k) \leftarrow U^n$

2 Set $\Delta t(k)$ to satisfy the CFL condition (67)

3 Compute $U^{n+1}(k+1)$ from (64) with $r = 0$, $\theta = \theta(k)$, $\Delta t = \Delta t(k)$

4 $\forall i$, set $\mathcal{J}_i(k) = \frac{1}{\Delta t}\mathcal{E}_i^\theta(k) + \frac{1}{\Delta x} D_i^\theta(k)$ the dissipation rate from Prop. 5.4

Adaptive time stepping to ensure entropy stability

5 **while** ($\exists i$ *such that* $\Delta t \mathcal{J}_i(k) > \epsilon_{\mathrm{tol}}$) *or* ($\delta > \epsilon_{\mathrm{tol}}$) **do**

6 $\quad \forall i$, $\theta_i(k+1) \leftarrow \max(\theta_i(k), \min(1, \bar{\theta}))$, with $\bar{\theta}$ solving (73) if $\Delta t \mathcal{J}_i(k) > \epsilon_{\mathrm{tol}}$, and $\bar{\theta} = \theta_i(k)$ otherwise

7 $\quad k \leftarrow k + 1$

8 $\quad \forall i$, set $U_i^\theta(k) = \theta_i(k) U_i^{n+1}(k) + (1 - \theta_i(k)) U_i^n$

9 $\quad$ Compute $\Delta t(k)$, $U^{n+1}(k+1)$ and $\mathcal{J}_i(k)$ as in lines 2–4 with $r = 1$

10 $\quad \delta \leftarrow \|U^{n+1}(k+1) - U^{n+1}(k)\|_\infty / \|U^{n+1}(k)\|_\infty$

11 **return** $U^{n+1}(k+1)$ and $\theta(k)$

To conclude, we mention the stopping criteria. In addition to the standard tolerance condition

$$|U^{n+1}(k+1) - U^{n+1}(k)|_\infty \leq \epsilon_{\mathrm{tol}} |U^{n+1}(k)|_\infty,$$

where $|\cdot|_\infty$ is the ∞-norm on $(\mathbb{R}^2)^N$, we can also check whether the entropy is being dissipated in each cell, i.e.

$$\forall 1 \leq i \leq N, \quad \frac{1}{\Delta t}\mathcal{E}_i^\theta(k) + \frac{1}{\Delta x} D_i^\theta(k) \leq 0,$$

with $\mathcal{E}_i^\theta(k), D_i^\theta(k)$ defined as in Proposition 5.4. The overall fixed point method is summarized in Algorithm 1.

6 Numerical experiments

We now assess the interest of the proposed adaptive scheme and the usefulness of the quantification of entropy dissipation (62) through numerical experiments. For all the simulations considered below, we use the gravitational acceleration $g = 9.81$ [m/s^2]; the tolerance for the fixed point method approximating the adaptive time stepping is set to $\epsilon_{\text{tol}} = 10^{-13}$. In practice, for the fixed point method to converge it seems enough to take the time step equal to $\Delta t = 0.45\delta t$, where δt is the maximum positive value verifying the constraints (67). Note that δt can evolve at each sub-iteration k of the fixed point method. We focus on the case of two velocities corresponding to (15)-(16), with local speeds updated at each time step n and sub-iteration k through the formula:

$$\lambda^-_{i+1/2} = \min\Big(u_i - 1.125\sqrt{gh_i},\ u_{i+1} - 1.125\sqrt{gh_{i+1}}\Big),$$
$$\lambda^+_{i+1/2} = \max\Big(u_i + 1.125\sqrt{gh_i},\ u_{i+1} + 1.125\sqrt{gh_{i+1}}\Big).$$

We recall that this choice allows to verify the subcharacteristic condition at every interface, i.e. $\sigma(\mathcal{M}'_{l,i\pm1/2}(U_i^{n+1}(k))) \subset]0, +\infty[$, which is a crucial property to ensure both entropy dissipation and positivity of the water height.

6.1 Entropy dissipation

To justify the advantages of using the adaptive discrete kinetic scheme with hydrostatic reconstruction, we consider a test case designed in such a way that the forward Euler version of the scheme is unstable during the first time steps, while the backward Euler version is over-diffusive. This is achieved by considering the following initial condition over the spatial domain $[0, 1]$ with periodic boundary conditions:

$$Z(x) = -5 + \frac{1}{2}\Big(1 + \cos\Big(5\pi\Big(x - \frac{1}{2}\Big)\Big)\Big)\mathbb{1}_{|x-\frac{1}{2}|\le\frac{1}{5}}, \quad h_{\text{in}}(x) = q_{\text{in}}(x) = -Z(x).$$

The motivation behind this initial condition is that there holds $h_{\text{in}} + Z \equiv \text{Cst}$, implying that the spatial dissipation D_i^θ from (44) vanishes initially for $\theta \equiv 0$. Furthermore, this initial condition is not a steady state due to $q_{\text{in}} = -Z$ being non-constant: this causes the error in time $\mathcal{E}_i^\theta$ from (63) to be strictly positive for the forward Euler scheme, and strictly negative for the backward Euler scheme. On the other hand, thanks to the definition (73) of the parameters $\theta(k)$, the adaptive scheme is expected make the overall error $\frac{1}{\Delta t}\mathcal{E}_i^\theta + \frac{1}{\Delta x}D_i^\theta$ vanish in this context, leading to a more accurate approximation.

We perform simulations until final time $t = 2.5 \cdot 10^{-2}$ [s] with a mesh refinement of $\Delta x = 10^{-2}$ [m]. The results can be found in Figure 1 and corroborate the expected behavior for each scheme. In particular, the plot 1a shows the

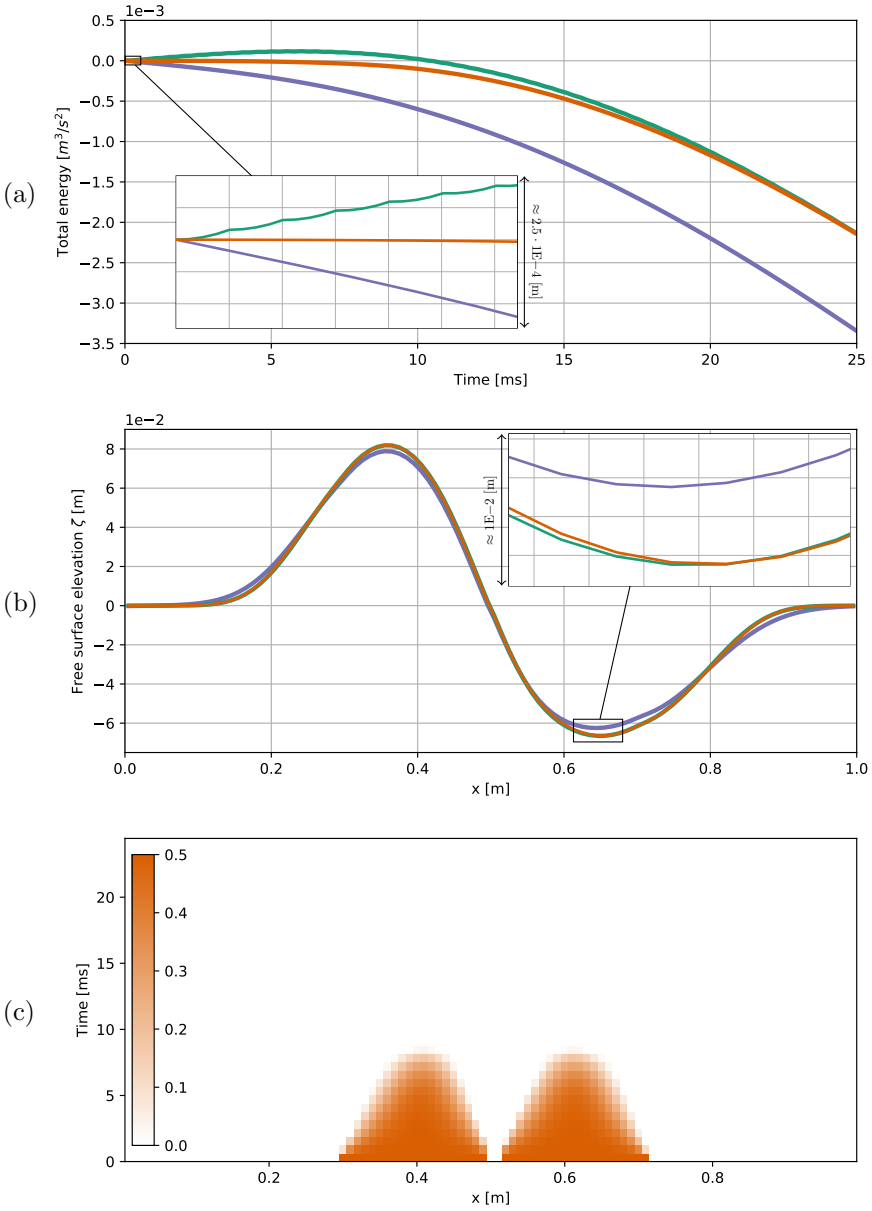

Figure 1: The first two plots represent (a) the time evolution of the total energy $\int_0^1 E(U,Z)\,\mathrm{d}x$ and (b) the free surface elevation at time $t = 25$ [ms]. Three schemes are compared: forward Euler (green), backward Euler (purple) and adaptive scheme (orange). The plot (c) represents the vector of parameters θ in the (x,t)-plane used for the adaptive method. It also indicates areas where the forward Euler scheme is unstable.

evolution of the approximated total energy $\int_0^1 E(U,Z)\,\mathrm{d}x$. We see that the adaptive scheme never increases this quantity and manages to keep it nearly constant at the beginning of the simulation. While stable, the backward Euler scheme is much more dissipative from the first iteration. On the other hand, at the start the forward Euler scheme increases the total entropy, and thus violates the discrete entropy inequality. We stress that the proper behavior of smooth periodic solutions is for the total energy to be conserved — in this regard the adaptive scheme offers much improved results.

In plot 1b we see the final free surface elevation: the wave amplitude is the smallest and greatest respectively for the backward and forward Euler schemes, and the adaptive scheme remains very close to the fully explicit method.

Finally, in plot 1c we see that after some time, the adaptive scheme sets the vector of parameters θ to zero — meaning that it reduces to the forward Euler method — until the end of the simulation. It is also worth noticing that the maximum value of θ is about $1/2$ in this test case. Hence, the backward Euler scheme is never used.

6.2 Perturbed lake at rest

We now underline the importance of the well-balanced property to accurately approximate solutions of the Saint-Venant system. Many flows in natural environments can be seen as a perturbation of the lake at rest equilibrium. Thus, we consider the following initial condition over the spatial domain $[0, 25]$:

$$h_{\text{init}}(x) = \max(0.18 + \delta(x) - Z(x), 0), \quad q_{\text{init}}(x) = 0,$$

where the bathymetry corresponds to

$$Z(x) = \begin{cases} 0 & \text{if } |x-10| > 2, \\ 0.2 - 0.05(x-10)^2 & \text{otherwise,} \end{cases}$$

and where $\delta(x)$ is a Gaussian perturbation defined as

$$\delta(x) = \frac{2a}{\sqrt{2\pi}\sigma} \exp\Big(-\frac{(x-12.5)^2}{2\sigma^2}\Big), \quad \sigma = \sqrt{2.5}, \quad a \in \{10^{-j} \mid j = 1, \dots, 17\}.$$

Note that this initial condition features a dry area, see Figure 2. Periodic conditions are enforced at the boundaries; the final time is set to $t = 10$ [s] and a spatial step of $\Delta x = 5 \cdot 10^{-2}$ [m] is used.

The parameter a allows modulating the perturbation in such a way that for all times $t > 0$, the difference $U(t,\cdot) - U_{\text{init}}$ converges to zero as $a \to 0$. In Figure 3 we check whether this property is verified at the discrete level for two different schemes. The first one is the adaptive kinetic scheme with hydrostatic reconstruction — which is well-balanced; the second one is the explicit kinetic scheme without hydrostatic reconstruction — which is not well-balanced. In this latter case, we simply discretize the bathymetry source term by centered differences, i.e. $gh_i(Z_{i+1} - Z_{i-1})/(2\Delta x)$.

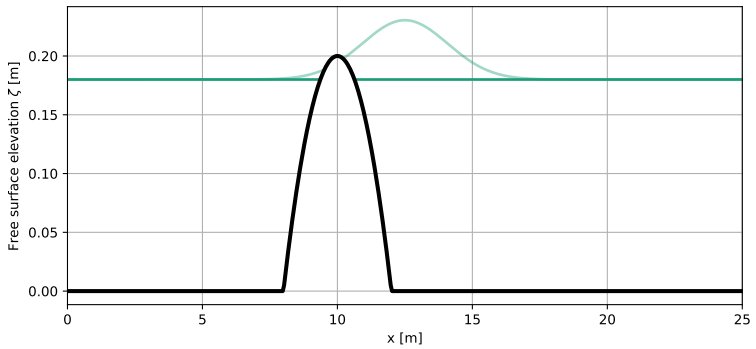

Figure 2: Bathymetry (black) and initial condition for $a = 10^{-1}$ (light green) and $a = 10^{-17}$ (dark green).

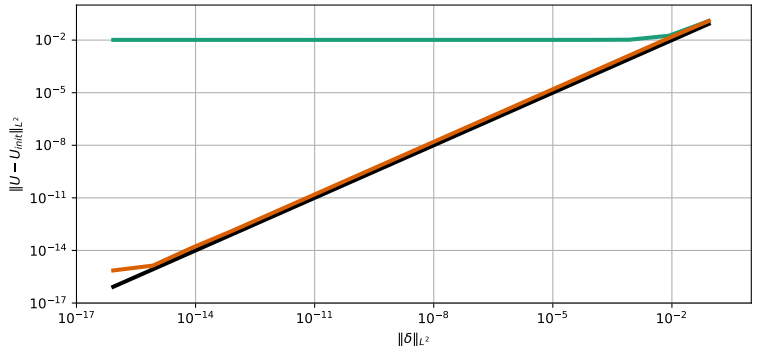

Figure 3: Deviation $\|U - U_{\text{init}}\|_{L^2}$ as a function of $\|\delta\|_{L^2}$ plotted at time $t = 10$ [s]. The adaptive kinetic scheme with hydrostatic reconstruction corresponds to the orange line, while the green line stands for the fully explicit kinetic scheme without hydrostatic reconstruction (source term discretized by centered differences). For reference, the black line is the identity.

We clearly see the interest of the well-balanced property in Figure 3: while for Δx fixed the first scheme converges to the lake at rest up to machine precision as a becomes small, the second scheme fails to preserve the approximation near the initial datum without taking $\Delta x \to 0$. Note that for the non-well-balanced scheme there is no point in using the adaptive time stepping since we aren't able to ensure that the associated spatial error dissipates the entropy.

6.3 Transcritical shock

The transcritical shock test case is interesting as it provides a (quasi) analytic solution featuring a stationary shock over a variable bathymetry. The analytical solution is determined explicitly up to the location of the jump which verifies an implicit equation — obtained through the Rankine-Hugoniot condition — that has to be approximated. The procedure to compute such solutions is described in Section 3.1.5 from [22].

The spatial domain and the bathymetry are the same as in Section 6.2. The initial condition is set to

$$h_{\text{init}}(x) = \begin{cases} h_{\text{init}}(0) - Z(x), & x < 10, \\ 0.33 - Z(x), & x > 10, \end{cases} \qquad q_{\text{init}}(x) = 0.18,$$

where $h_{\text{init}}(0) \approx 0.414$ is the largest real root of the Bernoulli relation:

$$h_{\text{init}}(0)^3 + \Big(Z(0) - \frac{q_{\max}^2}{2gh_{\max}^2} - h_{\max} - Z_{\max}\Big) h_{\text{init}}(0)^2 + \frac{1}{2g} q_{\text{init}}(0)^2 = 0,$$

with $Z_{\max} = 0.2$, $h_{\max} = (q_{\max}^2/g)^{1/3}$ and $q_{\max} = 0.18$. Note that this choice of $h_{\max}$ allows having $\frac{\mathrm{d}}{\mathrm{d}x} h_{\text{init}} \neq 0$ at the top of the bump ($x = 10$).

We compare the asymptotic steady state to the approximation obtained with the adaptive scheme at time $t = 25$ with $\Delta x = 5 \cdot 10^{-2}$. The free surface elevation is plotted in Figure 4a and shows good agreement with the analytic solution. In Figure 4b, we notice a very localized discrepancy with the computed discharge. As the mesh gets refined, the spike gets narrower, but its amplitude seems to remain unaffected. A consequence is that the approximation doesn't converge to the asymptotic steady state in infinity norm, but it does in the L^2 norm. Evidence of this is provided in Figure 4c, with an experimental order of accuracy of $1/2$ which is explained by the non-smooth aspect of the solution. We also emphasize that in this test case, the adaptive scheme reduces to the forward Euler scheme except in a very few cells at the beginning of the simulation. As a consequence the convergence curves of the adaptive and explicit schemes are hardly distinguishable, and we only plotted the former one.

6.4 Thacker's test case

To finish, we consider Thacker's experiment, also known as the parabolic bowl test case. It consists in a periodically oscillating flow over a parabolic bathymetry. The solution can be computed explicitly; in particular the free surface

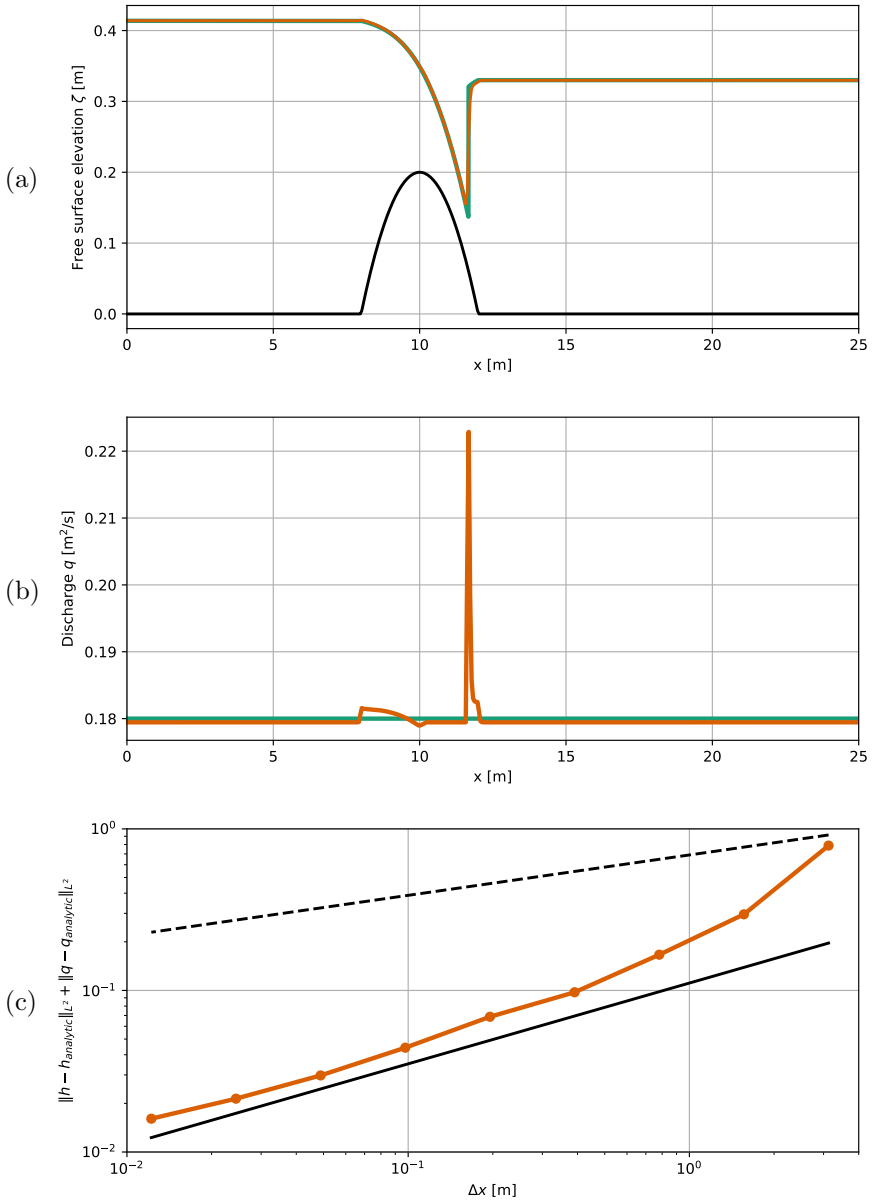

Figure 4: Transcritical shock test case over a bump. The first two plots feature the water height (a) and discharge (b) at time 25 [s], comparing the adaptive approximation (orange) with the analytic asymptotic steady state (green). The third plot (c) gives the L^2 error $\|h - h_{\text{analytic}}\|_{L^2} + \|q - q_{\text{analytic}}\|_{L^2}$ as a function of Δx (orange); for reference, orders $1/4$ and $1/2$ are given through the dashed and full lines in black.

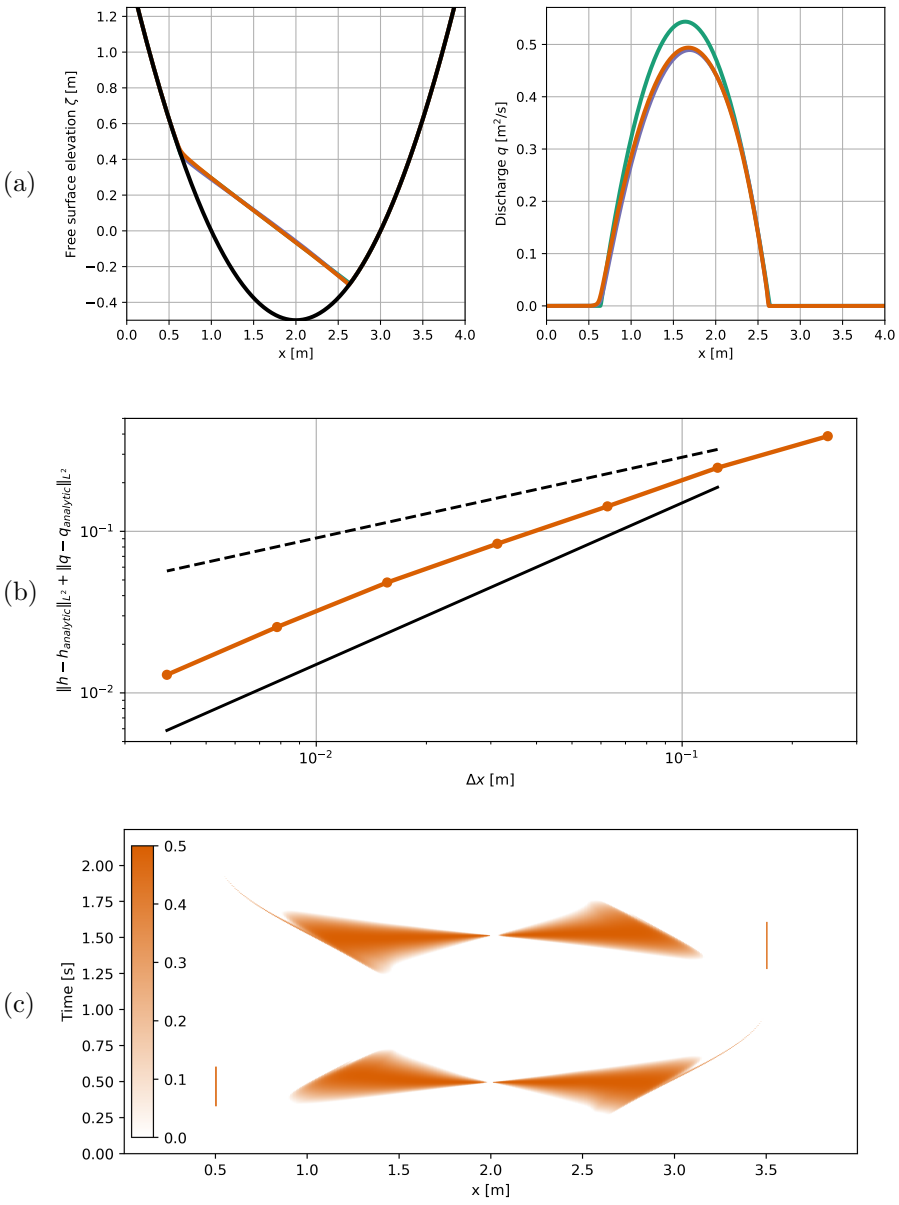

Figure 5: Thacker's test case. The plots on the first row (a) give the free surface elevation and discharge at time $t = 2.25$ [s], with the adaptive kinetic scheme in orange, backward Euler scheme in purple and analytic solution in green. Underneath, (b) corresponds to the L^2 error of the adaptive scheme, with orders $1/2$ and 1 given through the dashed and full lines in black. Finally, (c) gives the vector of parameters θ in the (x, t)-plane. In particular, we see that the forward Euler scheme is not entropy stable for this test case.

remains a plane at all times, see Section 4.2.1 from [22]. More precisely, we define

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

with $h_0 = 0.5$, $a = 1$, $L = 4$, and $B = \frac{1}{2a}\sqrt{2gh_0}$. The initial condition is deduced from the above expressions by setting $t = 0$.

A difficulty of this test case is that it presents a moving wet/dry transition which can be challenging to capture numerically. Using the adaptive kinetic scheme and its version with the backward Euler time integrator, we perform simulations with $\Delta x = 8 \cdot 10^{-3}$ until time $t = 2.25$ [s], which is slightly over one period of oscillation. In Figure 5a, we see that the free surface elevation closely matches the analytic solution, while the discharge is underestimated for both time stepping methods. Figure 5b gives the L^2-error of the adaptive scheme as a function of Δx: the approximation converges to the analytical one with a first order accuracy, which is the expected theoretical order. Finally, Figure 5c provides the values of θ in the (x,t)-plane. At maximum, its components can reach the value $1/2$, with what seems to be a repeating pattern indicating that the forward Euler scheme needs to be stabilized at regular intervals.

7 Conclusions and perspectives

In the present paper, we have developed a vectorial BGK scheme with finitely many velocities for the Saint-Venant system, relying on the hydrostatic reconstruction to account for variable bathymetries in a well-balanced way. At the kinetic level, the contribution of the bathymetry is treated through a force term involving a shift matrix. The spatial discretization of the scheme is shown to always dissipate the entropy — given by the energy of the system — under a subcharacteristic condition. As in [23], an implicit time discretization can then be used to obtain a fully discrete entropy inequality; however, in our case we get an exact quantification of the rate of dissipation through an entropy equality. Moreover, by using a local-per-cell and dynamical-in-time convex combination between the forward and backward Euler methods, the computational cost and the numerical viscosity of the scheme are optimized while ensuring the entropy dissipation. It is worth noting that over flat bathymetries, our method is also able to dissipate any entropy of the Saint-Venant system, and not only its energy.

Several perspectives can be proposed to this work. We can think of adapting the *hydrodynamic* reconstruction proposed in [12] to the framework of kinetic schemes. This reconstruction is interesting as it allows preserving the more general Bernoulli equilibria that encompass moving steady states. In particular, the underlying question would be to know whether this improved reconstruction

is entropy dissipative or not. Another promising perspective would be to replace the implicit time integration by the more efficient forward Euler method with added numerical viscosity to stabilize the resulting scheme. Such a technique has been studied in [11] in the framework of Godunov schemes, and was shown to be compatible with a fully discrete entropy inequality.

Funding

While working on this project, Mathieu Rigal has received funding supports from Institut des Mathématiques pour la Planète Terre, ANR BOURGEON (grant number ANR-23-CE40-0014-01), and ERC BLOC.

Data availability statement

The research code associated with this article is available in Gitlab, under the reference `swimpy-1d` [31].

References

[1] D. Aregba-Driollet, J. Breil, S. Brull, B. Dubroca, and E. Estibals. Modelling and numerical approximation for the nonconservative bitemperature Euler model. *ESAIM Math. Model. Numer. Anal.*, 52(4):1353–1383, 2018.

[2] D. Aregba-Driollet, S. Brull, and C. Prigent. A discrete velocity numerical scheme for the two-dimensional bitemperature Euler system. *SIAM J. Numer. Anal.*, 60(1):28–51, 2022.

[3] D. Aregba-Driollet and R. Natalini. Discrete kinetic schemes for systems of conservation laws. In *Hyperbolic problems: theory, numerics, applications, Vol. I (Zürich, 1998)*, volume 129 of *Internat. Ser. Numer. Math.*, pages 1–10. Birkhäuser, Basel, 1999.

[4] D. Aregba-Driollet and R. Natalini. Discrete kinetic schemes for multidimensional systems of conservation laws. *SIAM J. Numer. Anal.*, 37(6):1973–2004, 2000.

[5] E. Audusse, F. Bouchut, M.-O. Bristeau, R. Klein, and B. Perthame. A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows. *SIAM J. Sci. Comput.*, 25(6):2050–2065, 2004.

[6] E. Audusse, F. Bouchut, M.-O. Bristeau, and J. Sainte-Marie. Kinetic entropy inequality and hydrostatic reconstruction scheme for the Saint-Venant system. *Math. Comput.*, 85(302):2815–2837, 2016.

[7] E. Audusse and M.-O. Bristeau. A well-balanced positivity preserving "second-order" scheme for shallow water flows on unstructured meshes. *J. Comput. Phys.*, 206(1):311–333, 2005.

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

[9] A. Bermudez and M. E. Vazquez. Upwind methods for hyperbolic conservation laws with source terms. *Computers & Fluids*, 23(8):1049–1071, 1994.

[10] C. Berthon and C. Chalons. A fully well-balanced, positive and entropy-satisfying Godunov-type method for the shallow-water equations. *Math. of Comp.*, 85(299):1281–1307, 2016.

[11] C. Berthon, A. Duran, F. Foucher, K. Saleh, and J.D.D. Zabsonré. Improvement of the hydrostatic reconstruction scheme to get fully discrete entropy inequalities. *J. Sci. Comput.*, 80:924–956, 2019.

[12] C. Berthon and V. Michel-Dansac. A fully well-balanced hydrodynamic reconstruction. *Journ. Num. Math.*, 32(3):275–299, 2024.

[13] P.L. Bhatnagar, E.P. Gross, and M. Krook. A model for collision processes in gases. *Phys. Rev.*, 94:511–525, May 1954.

[14] A. Bollermann, S. Noelle, and M. Lukáčová-Medvid'ová. Finite volume evolution Galerkin methods for the shallow water equations with dry beds. *Communications in Computational Physics*, 10(2):371–404, 2011.

[15] F. Bouchut. Construction of BGK models with a family of kinetic entropies for a given system of conservation laws. *J. Statist. Phys.*, 95(1-2):113–170, 1999.

[16] F. Bouchut and X. Lhébrard. Convergence of the kinetic hydrostatic reconstruction scheme for the Saint Venant system with topography. *Math. Comp.*, 90(329):1119–1153, 2021.

[17] M.-O. Bristeau and B. Coussin. Boundary Conditions for the Shallow Water Equations solved by Kinetic Schemes. Research Report RR-4282, INRIA, 2001. Projet M3N.

[18] C. Caballero-Cárdenas, M.J. Castro, T. Morales de Luna, and M.L. Muñoz-Ruiz. Implicit and implicit-explicit Lagrange-projection finite volume schemes exactly well-balanced for 1d shallow water system. *Applied Mathematics and Computation*, 443:127784, 2023.

[19] M.J. Castro Diaz, C. Chalons, and T. Morales de Luna. A fully well-balanced Lagrange-projection type scheme for the shallow-water equations. *SIAM Journal on Numerical Analysis*, 56:3071–3098, 2018.

[20] G. Chen and S. Noelle. A new hydrostatic reconstruction scheme based on subcell reconstructions. *SIAM Journal on Numerical Analysis*, 55(2):758–784, 2017.

[21] D. Coulette, E. Franck, P. Helluy, M. Mehrenberger, and L. Navoret. High-order implicit palindromic discontinuous Galerkin method for kinetic-relaxation approximation. *Computers & Fluids*, 190:485–502, 2019.

[22] O. Delestre, C. Lucas, P.-A. Ksinant, F. Darboux, C. Laguerre, T.N.T. Vo, F. James, and S. Cordier. SWASHES: a compilation of Shallow Water Analytic Solutions for Hydraulic and Environmental Studies. *International Journal for Numerical Methods in Fluids*, 72(3):269–300, May 2013.

[23] C. El Hassanieh, M. Rigal, and J. Sainte-Marie. Implicit kinetic schemes for the Saint-Venant system. *ESAIM: M2AN*, 59(4):1863–1908, 2025.

[24] S. Jin. A steady-state capturing method for hyperbolic systems with geometrical source terms. *ESAIM: M2AN*, 35(4):631 – 645, 2001.

[25] A. Kurganov and G. Petrova. A second-order well-balanced positivity preserving central-upwind scheme for the Saint-Venant system. *Commun. Math. Sci..*, 5(1):133–160, 2007.

[26] Q. Liang and F. Marche. Numerical resolution of well-balanced shallow water equations with complex source terms. *Advances in Water Resources*, 32(6):873–884, 2009.

[27] R. Natalini. A discrete kinetic approximation of entropy solutions to multidimensional scalar conservation laws. *J. Differential Equations*, 148(2):292–317, 1998.

[28] S. Noelle, N. Pankratz, G. Puppo, and J.R. Natvig. Well-balanced finite volume schemes of arbitrary order of accuracy for shallow water flows. *Journal of Computational Physics*, 213(2):474–499, 2006.

[29] S. Noelle, Y. Xing, and C.-W. Shu. High-order well-balanced finite volume weno schemes for shallow water equation with moving water. *Journal of Computational Physics*, 226(1):29–58, 2007.

[30] B. Perthame and C. Simeoni. A kinetic scheme for the Saint-Venant system with a source term. *Calcolo*, 38(4):201–231, 2001.

[31] M. Rigal. `swimpy-1d`. `https://gitlab.com/mrigal/swimpy-1d/`, 2021. Python code.