WEAK CONSISTENCY OF NON LINEAR CONVECTION OPERATORS ON STAGGERED MESHES. APPLICATION TO A QUASI -SECOND ORDER STAGGERED SCHEME FOR THE TWO-DIMENSIONAL SHALLOW WATER EQUATIONS

T Gallouët, R Herbin, J.-C Latché, Y Nasseri, N Therme

WEAK CONSISTENCY OF NON LINEAR CONVECTION OPERATORS ON STAGGERED MESHES. APPLICATION TO A QUASI - SECOND ORDER STAGGERED SCHEME FOR THE TWO-DIMENSIONAL SHALLOW WATER EQUATIONS

T. Gallouët[1], R. Herbin[2], J.-C. Latché[3], Y. Nasseri[4] and N. Therme[5]

Abstract. In this paper a Lax-Wendroff type result of consistency is given for convection operators on staggered meshes. It is applied to a class of second order finite volume schemes developed to obtain approximate solutions of the shallow water equations with bathymetry. These schemes are based on staggered grids for the space discretization: scalar and vector unknowns are defined on different meshes. MUSCL-like interpolations for the discrete convection operators in the water height and momentum equations are performed in order to improve the precision of the scheme. The time discretization is performed either by a first order segregated forward Euler scheme in time or by the second order Heun scheme. Both schemes are shown to preserve the water height positivity under a CFL condition and an important state equilibrium known as the lake at rest. Using the above mentioned staggered Lax-Wendroff type results, these schemes are shown to be Lax-consistent with the weak formulation of the continuous equations; besides, the forward Euler scheme is shown to be consistent with a weak entropy inequality. Numerical results confirm the efficiency and accuracy of the schemes.

2010 AMS Subject Classification. Primary 65M08, 76N15 ; Secondary 65M12, 76N19.

The dates will be set by the publisher.

Keywords Finite-volume scheme, MAC grid, shallow water flow.

September 15, 2020

Contents

Keywords and phrases: Finite-volume scheme, MAC gird, shallow water flow.

[1] I2M UMR 7373, Aix-Marseille Université, CNRS, Ecole Centrale de Marseille. 39 rue Joliot Curie. 13453 Marseille, France.
(raphaele.herbin@univ-amu.fr)

[2] I2M UMR 7373, Aix-Marseille Université, CNRS, Ecole Centrale de Marseille. 39 rue Joliot Curie. 13453 Marseille, France.
(raphaele.herbin@univ-amu.fr)

[3] IRSN, BP 13115, St-Paul-lez-Durance Cedex, France (jean-claude.latche@irsn.fr)

[4] I2M UMR 7373, Aix-Marseille Université, CNRS, Ecole Centrale de Marseille. 39 rue Joliot Curie. 13453 Marseille, France.
(youssouf.nasseri@univ-amu.fr)

[5] CEA/CESTA 33116, Le Barp, France (nicolas.therme@cea.fr)

1. Introduction

The shallow water equations form a hyperbolic system of two conservation equations (mass and momentum) which models the flow of an incompressible fluid, assuming that the mean vertical height of the fluid is small compared to the plane scale. It is widely used for the simulation of numerous geophysical phenomena, such as flow in rivers and coastal areas. For a fluid occupying the space-time domain $\Omega \times (0,T)$, where Ω is an open bounded subset of $\mathbb{R}^2$ and $T > 0$, the shallow water equations with bathymetry solve the water height h and the (vector) velocity of the fluid $\boldsymbol{u} = (u_1, u_2)$ and read:

$$\partial_t h + \mathrm{div}(h\boldsymbol{u}) = 0 \qquad \text{in } \Omega \times (0,T), \tag{1a}$$
$$\partial_t (h\boldsymbol{u}) + \mathrm{div}(h\boldsymbol{u} \otimes \boldsymbol{u}) + \nabla p + gh\nabla z = 0 \qquad \text{in } \Omega \times (0,T), \tag{1b}$$
$$p = \frac{1}{2} g h^2 \qquad \text{in } \Omega \times (0,T), \tag{1c}$$
$$\boldsymbol{u} \cdot \boldsymbol{n} = 0 \qquad \text{on } \partial\Omega \times (0,T), \tag{1d}$$
$$h(\boldsymbol{x},0) = h_0,\ \boldsymbol{u}(\boldsymbol{x},0) = \boldsymbol{u}_0 \qquad \text{in } \Omega. \tag{1e}$$

where ∂_t is the partial time derivative, div denotes the spatial divergence operator, g is the standard gravity constant and z the (given) bathymetry, which is supposed to be regular in this paper. The initial conditions are $h_0 \in L^\infty(\Omega)$ and $\boldsymbol{u}_0 = (u_{0,1}, u_{0,2}) \in L^\infty(\Omega, \mathbb{R}^2)$ with $h_0 \geq 0$. This system has therefore been intensively studied, both theoretically and numerically, so that it is impossible to give an exhaustive list of references. We refer to the books [6, 29] and to the more recent books or parts of books [3, 8, 30] and the references therein. We recall that it is wellknown that if no dry zone exists, the system is strictly hyperbolic. In all cases, the solution of the system may develop shocks, so that the finite volume method is often preferred for numerical simulations. Two main approaches are found: one is the colocated approach which is usually based on some approximate Riemann solver, see e.g. [6, 8] and references therein; the other one is based on a staggered arrangement of the unknowns on the grid, which is quite classical in the hydraulic and ocean engineering community, see e.g. [2, 5, 28]. These latter staggered schemes have been implemented with an upwind choice for the convection operators and a forward Euler time discretization and analysed in the case of one space dimension [9, 16], following the works on the related barotropic Euler equations, see [22] and references therein. In particular, the weak consistency of the scheme is shown as well as a weak entropy consistency. Let us recall that if $(h, \boldsymbol{u})$ is a regular solution of (1), the following elastic potential energy balance and kinetic energy balance are obtained by manipulations on the mass and momentum equations:

$$\partial_t (\frac{1}{2} g h^2) + \mathrm{div}(\frac{1}{2} g h^2 \boldsymbol{u}) + \frac{1}{2} g h^2 \mathrm{div}\boldsymbol{u} = 0 \tag{2}$$
$$\partial_t (\frac{1}{2} h|\boldsymbol{u}|^2) + \mathrm{div}(\frac{1}{2} h|\boldsymbol{u}|^2 \boldsymbol{u}) + \boldsymbol{u} \cdot \nabla p + gh\boldsymbol{u} \cdot \nabla z = 0. \tag{3}$$

Summing these equations, we obtain an entropy balance equation: $\partial_t E + \mathrm{div}\Phi = 0$, where the entropy-entropy flux pair (E, Φ) is given by:

$$E = \frac{1}{2}h|\boldsymbol{u}|^2 + \frac{1}{2}gh^2 + ghz \text{ and } \Phi = (E + \frac{1}{2}gh^2)\boldsymbol{u}. \tag{4}$$

For non regular functions the above manipulations are no longer valid, and the entropy inequality $\partial_t E + \mathrm{div}\Phi \leq 0$ is satisfied in a distributional sense. The weak entropy consistency consists in showing that any possible limit of the scheme satisfies a weak form of the entropy inequality (4) given in (31) below.

In the case of two space dimensions, the consistency of the upwind scheme with respect to the weak formulation and and to a weak entropy inequality is stated in [20]; a quasi-second order scheme in time and space using the second order Heun method in time dependent and a MUSCL-like interpolation in space was proposed in [14].

Here, we analyse the former schemes both theoretically and numerically. The framework that is developed here includes three schemes : the first order scheme of [20], the same scheme replacing the upwind choice in the numerical convection operator by a MUSCL-like procedure, and the quasi second order scheme proposed in [14]. Generic properties are shown to be preserved, such as the positivity of the water height and the preservation of the "lake at rest" steady state. The weak consistency of the schemes is proven thanks to a generalisation of Lax-Wendroff type result which is given in an appendix; this consistency result is interesting for its own sake and valid for general convection operators on general colocated or staggered grids in any space dimension. Furthermore, the two first schemes are shown to be entropy-weak consistent in the sense that a weak entropy inequality is satisfied by any possible limit of the scheme as the time and space steps tend to 0, under some CFL condition.

The remainder of the paper is organized as follows: In Section 2 we introduce the space and time discretization. The resulting approximate solutions have some discrete stability and well balance properties which are studied in Section 3. Furthermore, under some convergence and boundedness assumptions, the approximate solutions are shown in Section 4 to converge to a weak solution of (1). This proof of these esults heavily relies on the general Lax-Wendroff consistency lemma which is given in the appendix A In Section 5 we consider the first order time discretization and show that any possible limit of the scheme satifies a weak entropy inequality, again using the consistency result of the appendix. Numerical results comparing the first order scheme of [20], the same scheme replacing the upwind choice in the numerical convection operator by a MUSCL-like procedure, and the quasi second order scheme proposed in [14]are presented in Section 6. Finally, the appendix A contains the general consistency result for a nonlinear convection operator on general meshes with a staggered arrangement of the unknowns, which generalizes the result obtained in [12], while the appendix B contains some technical lemmas which were proved formerly and which are recalled for the sake of completeness.

2. Space and time discretization

2.1. Definitions and notations

We concentrate on the MAC discretization in space, see [17, 18] for some seminal papers and [13] for the convergence analysis of the scheme applied to the incompressible Navier-Stokes equations. This scheme is also widely used by the hydrologist and known as the Arakawa scheme [2].

Let Ω be a connected subset of $\mathbb{R}^2$ consisting in a union of rectangles whose edges are assumed to be orthogonal to the canonical basis vectors, denoted by $(\boldsymbol{e}^{(1)}, \boldsymbol{e}^{(2)})$.

Definition 2.1 (MAC discretization)**.** A discretization $(\mathcal{M}, \mathcal{E})$ of Ω with a staggered rectangular grid (or MAC grid), is defined by:

- A primal mesh $\mathcal{M}$ which consists in a conforming structured, possibly non uniform, rectangular grid of Ω. A generic cell of this grid is denoted by K, and its mass center by $\boldsymbol{x}_K$. The scalar unknowns (water height and pressure) are associated to this mesh.
- A set $\mathcal{E}$ of all edges of the mesh, with $\mathcal{E} = \mathcal{E}_{\mathrm{int}} \cup \mathcal{E}_{\mathrm{ext}}$, where $\mathcal{E}_{\mathrm{int}}$ (resp. $\mathcal{E}_{\mathrm{ext}}$) are the edges of $\mathcal{E}$ that lie in the interior (resp. on the boundary) of the domain. The set of edges that are orthogonal to $\boldsymbol{e}^{(i)}$

is denoted by $\mathcal{E}^{(i)}$, for $i \in [\![1,2]\!]$. We then have $\mathcal{E}^{(i)} = \mathcal{E}^{(i)}_{\rm int} \cup \mathcal{E}^{(i)}_{\rm ext}$, where $\mathcal{E}^{(i)}_{\rm int}$ (resp. $\mathcal{E}^{(i)}_{\rm ext}$) are the edges of $\mathcal{E}^{(i)}$ that lie in the interior (resp. on the boundary) of the domain.

For $\sigma \in \mathcal{E}_{\rm int}$, we write $\sigma = K|L$ if $\sigma = \partial K \cap \partial L$. A dual cell D_σ associated to an edge $\sigma \in \mathcal{E}$ is defined as follows:

- if $\sigma = K|L \in \mathcal{E}_{\rm int}$ then $D_\sigma = D_{K,\sigma} \cup D_{L,\sigma}$, where $D_{K,\sigma}$ (resp. $D_{L,\sigma}$) is the half-part of K (resp. L) adjacent to σ (see Fig. 1);
- if $\sigma \in \mathcal{E}_{\rm ext}$ is adjacent to the cell K, then $D_\sigma = D_{K,\sigma}$.

For each dimension $i = 1, 2$, the domain Ω can also be split up in dual cells: $\Omega = \cup_{\sigma \in \mathcal{E}^{(i)}} \overline{D_\sigma}$, $i \in [\![1,2]\!]$; the i^{th} grid is refered to as the i^{th} dual mesh; it is associated to the i^{th} velocity component, in a sense which is clarified below. The set of the edges of the i^{th} dual mesh is denoted by $\widetilde{\mathcal{E}}_i$ (note that these edges may be non-orthogonal to $\boldsymbol{e}^{(i)}$); the set $\widetilde{\mathcal{E}}_i$ is decomposed into the internal and boundary edges: $\widetilde{\mathcal{E}}_i = \widetilde{\mathcal{E}}^{(i)}_{\rm int} \cup \widetilde{\mathcal{E}}^{(i)}_{\rm ext}$. The dual edge separating two duals cells D_σ and $D_{\sigma'}$ is denoted by $\epsilon = \sigma|\sigma'$. We denote by D_ϵ the cell associated to a dual edge $\epsilon \in \widetilde{\mathcal{E}}$ defined as follows:

- if $\epsilon = \sigma|\sigma' \in \widetilde{\mathcal{E}}_{\rm int}$ then $D_\epsilon = D_{\sigma,\epsilon} \cup D_{\sigma',\epsilon}$, where $D_{\sigma,\epsilon}$ (resp. $D_{\sigma',\epsilon}$) is the half-part of D_σ (resp. $D_{\sigma'}$) adjacent to ϵ (see Fig. 1);
- if $\epsilon \in \widetilde{\mathcal{E}}^{\rm ext}$ is adjacent to the cell D_σ, then $D_\epsilon = D_{\sigma,\epsilon}$.

In order to define the scheme, we need some additional notations. The set of edges of a primal cell K and of a dual cell D_σ are denoted by $\mathcal{E}(K) \subset \mathcal{E}$ and $\widetilde{\mathcal{E}}(D_\sigma)$ respectively; note that $\widetilde{\mathcal{E}}(D_\sigma) \subset \widetilde{\mathcal{E}}_i$ if $\sigma \in \mathcal{E}^{(i)}$. For $\sigma \in \mathcal{E}$, we denote by $\boldsymbol{x}_\sigma$ the mass center of σ. The vector $\boldsymbol{n}_{K,\sigma}$ stands for the unit normal vector to σ outward K. In some cases, we need to specify the orientation of various geometrical entities with respect to the axis:

- a primal cell K is denoted $K = [\overrightarrow{\sigma\sigma'}]$ if $\sigma, \sigma' \in \mathcal{E}^{(i)}(K)$ for some $i \in [\![1,2]\!]$ are such that $(\boldsymbol{x}_{\sigma'} - \boldsymbol{x}_\sigma) \cdot \boldsymbol{e}^{(i)} > 0$;
- we write $\sigma = \overrightarrow{K|L}$ if $\sigma \in \mathcal{E}^{(i)}$, $\sigma = K|L$ and $\overrightarrow{\boldsymbol{x}_K \boldsymbol{x}_L} \cdot \boldsymbol{e}^{(i)} > 0$ for some $i \in [\![1,2]\!]$;
- the dual edge ϵ separating D_σ and $D_{\sigma'}$ is written $\epsilon = \overrightarrow{\sigma|\sigma'}$ if $\overrightarrow{\boldsymbol{x}_\sigma \boldsymbol{x}_{\sigma'}} \cdot \boldsymbol{e}^{(i)} > 0$ for some $i \in [\![1,2]\!]$.

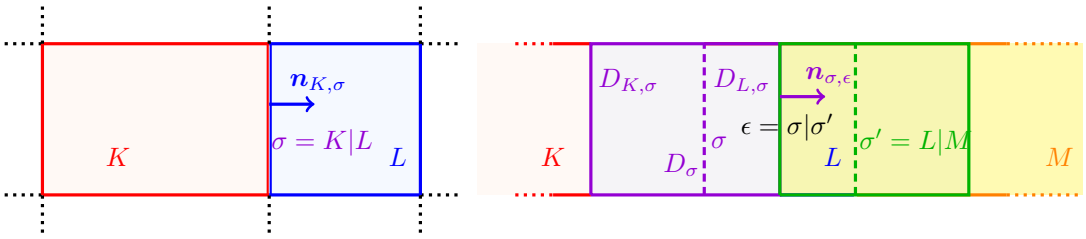

FIGURE 1. Notations for the prima and dual meshes (in two space dimensions, for the first component of the velocity).

The size $\delta_{\mathcal{M}}$ of the mesh and its regularity $\theta_{\mathcal{M}}$ are defined by:

$$\delta_{\mathcal{M}} = \max_{K \in \mathcal{M}} \operatorname{diam}(K), \text{ and } \theta_{\mathcal{M}} = \max_{K \in \mathcal{M}} \max_{\sigma \in \mathcal{E}_K} \frac{|D_\sigma|}{|K|}. \tag{5}$$

where $|\cdot|$ stands for the one (or two) dimensional measure of a subset of $\mathbb{R}$ (or $\mathbb{R}^2$). Note that in the rectangular case that is considered here, the regularity parameter $\theta_{\mathcal{M}}$ is also equal to:

$$\theta_{\mathcal{M}} = \frac{1}{2}(1 + \max\Big\{\frac{|\sigma|}{|\sigma'|},\ (\sigma,\sigma') \in \mathcal{E}^{(i)^2},\ \ i = 1,2\Big\}.$$

The discrete velocity unknowns are associated to the velocity cells and are denoted by $(u_{i,\sigma})_{\sigma \in \mathcal{E}^{(i)}}$, $i \in [\![1,2]\!]$, while the discrete scalar unknowns (water height and pressure) are associated to the primal cells and are denoted respectively by $(h_K)_{K \in \mathcal{M}}$ and $(p_K)_{K \in \mathcal{M}}$.

Let us consider a uniform discretisation $0 = t_0 < t_1 < \cdots < t_N = T$ of the time interval $(0, T)$, and let $\delta t = t_{n+1} - t_n$ for $n = 0, 1, \cdots, N-1$ be the (constant, for the sake of simplicity) time step.

Here we present two schemes: a first order in time segregated scheme using the forward Euler scheme and the second order in time Heun scheme. Both schemes use a MUSCL-like technique for the computation of the numerical flux, see [27], so that they are quasi second-order in space.

2.2. The segregated forward Euler scheme

We propose here a first order in time segregated discretisation and MAC discretization in space of the system (1); the scheme is written in compact form as follows:

$$\textbf{Initialisation}: \qquad u_\sigma^0 = \frac{1}{|D_\sigma|}\int_{D_\sigma} u_{i,0}(\boldsymbol{x})\ \mathrm{d}\boldsymbol{x},\ h^0 = \frac{1}{|K|}\int_K h_0(\boldsymbol{x})\ \mathrm{d}\boldsymbol{x},\ p^0 = \frac{1}{2}g(h^0)^2. \tag{6a}$$

$$\textbf{For}\ \ 0 \le n \le N-1 : \text{solve for } h^{n+1},\ p^{n+1} \text{ and } \boldsymbol{u}^{n+1} = (u_i^{n+1})_{i=1,2} :$$

$$\eth_t h_K^{n+1} + \mathrm{div}_K\,(h^n \boldsymbol{u}^n) = 0, \qquad \forall K \in \mathcal{M} \tag{6b}$$

$$p^{n+1} = \frac{1}{2}g(h^{n+1})^2, \tag{6c}$$

$$\eth_t(h\ u_i)_\sigma^{n+1} + \mathrm{div}_{D_\sigma}(h^n \boldsymbol{u}^n u_i^n) + \eth_\sigma p^{n+1} + g\, h_{\sigma,c}^{n+1}\ \eth_\sigma z = 0, \forall \sigma \in \mathcal{E}_{\mathrm{int}}^{(i)}, i \in [\![1,2]\!], \tag{6d}$$

where the different discrete terms and operators introduced here are now defined.

Discrete time derivative - In the sequel, we shall denote by $\eth_t v^{n+1}$ the discrete forward time derivative of a given discrete function of time v, *i.e.*:

$$\eth_t v^{n+1} = \frac{v^{n+1} - v^n}{\delta t} \tag{7}$$

Discrete divergence and gradient operators - The discrete divergence operator on the primal mesh denoted by div_K is defined as follows:

$$\mathrm{div}_K\,(h\boldsymbol{u}) = \frac{1}{|K|}\sum_{\sigma\in\mathcal{E}(K)} |\sigma|\ \boldsymbol{F}_\sigma \cdot \boldsymbol{n}_{K,\sigma},\ \text{with } \boldsymbol{F}_\sigma = h_\sigma\ \boldsymbol{u}_\sigma,\ \text{with } \boldsymbol{u}_\sigma = u_{i,\sigma}\ \boldsymbol{e}^{(i)} \text{ for } \sigma \in \mathcal{E}^{(i)},\ i \in [\![1,2]\!], \tag{8}$$

and h_σ is approximated by the MUSCL-like interpolation technique with respect to $\boldsymbol{u}_\sigma$; in the subsequent analysis, we do not need to have an explicit formula for h_σ, but we need the following conditions to be satisfied:

$$\forall\ K \in \mathcal{M},\ \forall \sigma = K|L \in \mathcal{E}_{\mathrm{int}}(K),$$

$$- \exists \lambda_{K,\sigma} \in [0,1] : h_\sigma = \lambda_{K,\sigma} h_K + (1-\lambda_{K,\sigma}) h_L \text{ if } \boldsymbol{F}_\sigma \cdot \boldsymbol{n}_{K,\sigma} \ge 0. \tag{9}$$

$$- \exists\ \alpha_\sigma^K \in [0,1] \text{ and } M_\sigma^K \in \mathcal{M}\ :\ h_\sigma - h_K = \begin{cases} \alpha_\sigma^K (h_K - h_{M_\sigma^K}) & \text{if } \boldsymbol{u}_\sigma \cdot \boldsymbol{n}_{K,\sigma} \ge 0, \\ \alpha_\sigma^K (h_{M_\sigma^K} - h_K) & \text{otherwise.} \end{cases} \tag{10}$$

By (9), h_σ is a convex combination of h_K and h_L, and if $\boldsymbol{u}_\sigma \cdot \boldsymbol{n}_{K,\sigma} < 0$, the cell M_σ^K in (10) can be chosen as L and $\alpha_{K,\sigma}$ as $1 - \lambda_{K,\sigma}$. In the case of a discrete divergence free velocity field $\boldsymbol{u}$, this assumption ensures that h_K^{n+1} is a convex combination of the values h_K^n and $(h_M^n)_{M\in\mathcal{N}_m((K)}$, where $\mathcal{N}_m(K)$ denotes the set of cells M_σ^K satisfying (10), see [27, Lemma 3.1], for any structured or unstructured mesh.

Note that if $K = [\sigma'\sigma]$ with $\sigma' = J|K$ and $\sigma = K|L$ and $\boldsymbol{u}_\sigma \cdot \boldsymbol{n}_{K,\sigma} \geq 0$, the cell M_σ^K in Relation (10) can be chosen as the cell J and the value h_σ computed using the following limitation procedure:

$$h_\sigma - h_K = \frac{1}{2}\,\psi\big(h_L, h_K, h_J\big), \text{ where}$$

$$\psi\big(h_L, h_K, h_J\big) = \begin{cases} \text{minmod}\big(\dfrac{h_L - h_J}{2}, \zeta^+(h_L - h_K), \zeta^-(h_K - h_J)\big), & \text{if } (h_L - h_K)(h_K - h_J) > 0, \\ 0, & \text{otherwise,} \end{cases}$$

where the limitation parameters ζ^+, ζ^- are such that $\zeta^+, \zeta^- \in [0, 2]$. Observe that if $\zeta^+ = \zeta^- = 1$, the classical minmod limiter $\big(\text{minmod}\big(h_L - h_K, h_K - h_J\big)\big)$ is recovered.

A local discrete derivative applied to a discrete scalar field ξ (with $\xi = p, h$ or z) is defined by:

$$\eth_\sigma \xi = \frac{|\sigma|}{|D_\sigma|}\,(\xi_L - \xi_K) \text{ for } \sigma = \overrightarrow{K|L} \in \mathcal{E}_{\text{int}}. \tag{11}$$

The above defined discrete divergence and discrete derivatives satisfy the following div-grad duality relationship [13, Lemma 2.4]:

$$\sum_{K\in\mathcal{M}} |K|\xi_K \text{div}_K(h\boldsymbol{u}) + \sum_{i=1}^{2} \sum_{\sigma\in\mathcal{E}_{\text{int}}^{(i)}} |D_\sigma| h_\sigma u_{i,\sigma}\; \eth_\sigma \xi = 0. \tag{12}$$

Discrete water height for the bathymetry term – In equation (6d) the term $\eth_\sigma z$ denotes the discrete derivative (in the sense of (11)) of the piecewise constant function $z_\mathcal{M} = \sum_{K\in\mathcal{M}} z(\boldsymbol{x}_K)\mathbb{1}_K$, that is:

$$\eth_\sigma z = \frac{|\sigma|}{|D_\sigma|}\,(z(\boldsymbol{x}_L) - z(\boldsymbol{x}_K)) \text{ for } \sigma = \overrightarrow{K|L} \in \mathcal{E}_{\text{int}}. \tag{13}$$

The value $h_{\sigma,c}$ of the water height is defined so as to satisfy:

$$\eth_\sigma p + g\, h_{\sigma,c}\; \eth_\sigma z = 0 \text{ if } \eth_\sigma(h + z) = 0,\ \forall i = 1, 2. \tag{14}$$

This requirement is fulfilled if $h_{\sigma,c}$ is centered:

$$h_{\sigma,c} = \begin{cases} \frac{1}{2}(h_K + h_L) & \text{for } \sigma = K|L \in \mathcal{E}_{\text{int}}, \\ h_K & \text{for } \sigma \in \mathcal{E}_{\text{ext}} \cap \mathcal{E}(K). \end{cases} \tag{15}$$

Indeed, if $h_{\sigma,c}$ is defined by (15), since $p = \frac{1}{2}gh^2$, one has from the definition of the discrete gradient (11), for $\sigma = K|L$,

$$\eth_\sigma p + g\, h_{\sigma,c}\; \eth_\sigma z = \frac{1}{2} g \frac{|\sigma|}{|D_\sigma|}(h_K + h_L)\eth_\sigma(h + z)$$

and therefore (14) holds, so that the "lake at rest" steady state is preserved, see Lemma 3.2 below.

Discrete convection operator – The term $(h\ u_i)_\sigma^{n+1}$ in the discrete time derivative in (6d) is defined by

$$(h\ u_i)_\sigma^{n+1} = h_{D_\sigma}^{n+1}\ u_{i,\sigma}^{n+1}, \tag{16a}$$

$$h_{D_\sigma} = \frac{1}{|D_\sigma|}\Big(|D_{K,\sigma}|\ h_K + |D_{L,\sigma}|\ h_L\Big), \text{ with } \sigma = K|L \in \mathcal{E}_{\text{int}}, \tag{16b}$$

where D_σ, $D_{K,\sigma}$ and $D_{L,\sigma}$ are defined in Definition 2.1.

The discrete divergence operator on the dual mesh div_{D_σ} is given by:

$$\text{div}_{D_\sigma}(hu_i\boldsymbol{u}) = \frac{1}{|D_\sigma|} \sum_{\epsilon\in\widetilde{\mathcal{E}}^{(i)}(D_\sigma)} |\epsilon|\ \boldsymbol{G}_\epsilon \cdot \boldsymbol{n}_{\sigma,\epsilon}, \text{ with } \boldsymbol{G}_\epsilon = \boldsymbol{F}_\epsilon u_{i,\epsilon}, \tag{17}$$

where

- the flux $\boldsymbol{F}_\epsilon$ is computed from the primal numerical mass fluxes; following [19] (see also [21], and [1] for an extension to triangular or quandrangular meshes using low order non-conforming finite element), it is defined as follows:

$$\text{for } \epsilon = \sigma|\sigma',\ \epsilon \subset K, \qquad \boldsymbol{F}_\epsilon = \frac{1}{2}\big(\boldsymbol{F}_\sigma + \boldsymbol{F}_{\sigma'}\big), \quad \epsilon \subset K, \quad \text{(left on Figure 2)} \tag{18a}$$

$$\text{for } \epsilon = \sigma|\sigma', \epsilon \not\subset K, \epsilon \subset \tau \cup \tau', \qquad \boldsymbol{F}_\epsilon = \frac{1}{|\epsilon|}\ \big(\frac{1}{2}|\tau|\ \boldsymbol{F}_\tau + \frac{1}{2}|\tau'|\ \boldsymbol{F}_{\tau'}\big), \quad \text{(right on Figure 2)}, \tag{18b}$$

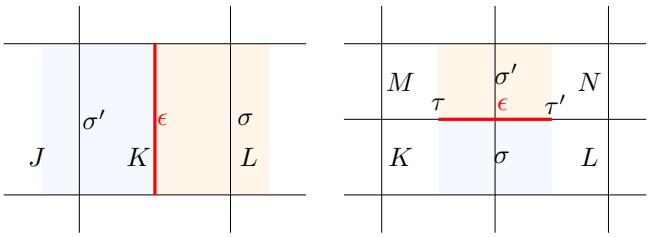

FIGURE 2. Notation for the definition of the momentum flux on the dual mesh for the first component of the velocity- left: $\epsilon \subset K$ - right: $\epsilon \subset \tau \cup \tau'$.

- the value $u_{i,\epsilon}$ is expressed in terms of the unknowns $u_{i,\sigma}$, for $\sigma \in \mathcal{E}^{(i)}$ by a second order MUSCL-like interpolation scheme with respect to $\boldsymbol{F}_\epsilon \cdot \boldsymbol{n}_{\sigma,\epsilon}$ [27]; the values $u_{i,\sigma}$ satisfy the following property:

$$\forall\ \sigma \in \mathcal{E}_{\rm int}^{(i)}, i = 1,2,\ \forall\ \epsilon = \sigma|\sigma' \in \widetilde{\mathcal{E}}(D_\sigma),$$

$$u_{i,\epsilon} \text{ is a convex combination of } u_{i,\sigma} \text{ and } u_{i,\sigma'} : \exists \mu_{\sigma,\epsilon} \in [0,1] : u_{i,\epsilon} = \mu_{\sigma,\epsilon} u_{i,\sigma} + (1-\mu_{\sigma,\epsilon}) u_{i,\sigma'} \tag{19}$$

$$\exists\ \alpha_\epsilon^\sigma \in [0,1] \text{ and } \tau_\epsilon^\sigma \in \mathcal{E}_{\rm int}^{(i)}\ : u_{i,\epsilon} - u_{i,\sigma} = \begin{cases} \alpha_\epsilon^\sigma (u_{i,\sigma} - u_{i,\tau_\epsilon^\sigma}) & \text{if } \boldsymbol{F}_\epsilon \cdot \boldsymbol{n}_{\sigma,\epsilon} \geq 0, \\ \alpha_\epsilon^\sigma (u_{i,\tau_\epsilon^\sigma} - u_{i,\sigma}) & \text{otherwise.} \end{cases} \tag{20}$$

Again note that in the case $\boldsymbol{F}_\epsilon \cdot \boldsymbol{n}_{\sigma,\epsilon} < 0$, the edge τ_ϵ^σ may be chosen as σ'.

Let us emphasize that thanks to the definitions (16b) and (18) the following discrete mass balance version on the dual mesh holds:

$$\frac{|D_\sigma|}{\delta t}(h_{D_\sigma}^{n+1} - h_{D_\sigma}^n) + \sum_{\epsilon \in \widetilde{\mathcal{E}}(D_\sigma)} |\epsilon|\ \boldsymbol{F}_\epsilon^n \cdot \boldsymbol{n}_{\sigma,\epsilon} = 0. \tag{21}$$

2.3. **A second order in time Heun scheme**

We retain here the quasi-second order space discretization which we just set up, but consider now a second order time discretization using the Heun (or Runge Kutta 2) scheme.

The initialization of the scheme is the same as that of the forward Euler scheme, see (6a), but the n-th step now reads:

Step n : For h^n and $\boldsymbol{u}^n = (u_i^n)_{i=1,2}$ known,

$$\widehat{h}_K^{n+1} = h_K^n - \delta t\, \mathrm{div}_K(h^n \boldsymbol{u}^n), \qquad \forall K \in \mathcal{M} \tag{22a}$$

$$\widehat{h}_{D_\sigma}^{n+1}\ \widehat{u}_{i,\sigma}^{n+1} = h_{D_\sigma}^n u_{i,\sigma}^n - \delta t\, \mathcal{F}_{D_\sigma}(h^n, u_i^n), \qquad \forall \sigma \in \mathcal{E}_{\rm int}^{(i)} \tag{22b}$$

$$\tilde{h}_K^{n+1} = \widehat{h}_K^{n+1} - \delta t\, \mathrm{div}_K(\widehat{h}^{n+1} \widehat{\boldsymbol{u}}^{n+1}), \qquad \forall K \in \mathcal{M} \tag{22c}$$

$$\tilde{h}_{D_\sigma}^{n+1}\ \tilde{u}_{i,\sigma}^{n+1} = \widehat{h}_{D_\sigma}^{n+1} \widehat{u}_{i,\sigma}^{n+1} - \delta t\, \mathcal{F}_{D_\sigma}(\widehat{h}^{n+1}, \widehat{u}_i^{n+1}), \qquad \forall \sigma \in \mathcal{E}_{\rm int}^{(i)} \tag{22d}$$

$$h_K^{n+1} = \frac{1}{2}\,(h_K^n + \tilde{h}_K^{n+1}), \qquad \forall K \in \mathcal{M} \tag{22e}$$

$$h_{D_\sigma}^{n+1}\ u_{i,\sigma}^{n+1} = \frac{1}{2}\big(h_{D_\sigma}^n u_{i,\sigma}^n + \tilde{h}_{D_\sigma}^{n+1}\ \tilde{u}_{i,\sigma}^{n+1}\big), \qquad \forall \sigma \in \mathcal{E}_{\rm int}^{(i)} \tag{22f}$$

where

$$\mathcal{F}_{D_\sigma}(h^n, u_i^n) = {\rm div}_{D_\sigma}(h^n \boldsymbol{u}^n u_i^n) + g\, h_{\sigma,c}^n\, \big((\eth_\sigma h^n) + (\eth_\sigma z)\big) \tag{23}$$

and the dual cell values $\widehat{h}_{D_\sigma}^{n+1}$, $\tilde{h}_{D_\sigma}^{n+1}$ and $h_{D_\sigma}^{n+1}$ are computed from the corresponding cell values by the analogue of the formula (16b), so that they satisfy a dual mass balance of the type (21).

The steps (22c)-(22f) of the above scheme (22) may be replaced by the more compact form

$$\eth_t h_K^{n+1} = -\frac{1}{2}\,\Big({\rm div}_K(h^n \boldsymbol{u}^n) + {\rm div}_K(\widehat{h}^{n+1}\widehat{\boldsymbol{u}}^{n+1})\Big), \qquad \forall K \in \mathcal{M} \tag{24a}$$

$$\eth_t(h_{D_\sigma}\ u_{i,\sigma})^{n+1} = -\frac{1}{2}\,\Big(\mathcal{F}_{D_\sigma}(h^n, u_i^n) + \mathcal{F}_{D_\sigma}(\widehat{h}^{n+1}, \widehat{u}_i^{n+1})\Big), \qquad \forall \sigma \in \mathcal{E}^{(i)}, \tag{24b}$$

where the dual cell value $h_{D_\sigma}^{n+1}$ is computed by the formula (16b) and hence satisfies a dual mass balance of the type (21).

3. Stability of the schemes

The positivity of the water height under a CFL like condition is ensured by both the schemes (6) and (22); it is a consequence of the property (10) of the MUSCL choice for the interface values. Indeed, the proof of the positivity in [27, Lemma 3.1] remains valid even if the discrete velocity field is not divergence free, as is the case here.

Lemma 3.1 (Positivity of the water height)**.** *Let $n \in [\![0, N-1]\!]$, let $(h_K^n)_{K\in\mathcal{M}} \subset \mathbb{R}_+^*$ and $(\boldsymbol{u}_\sigma^n)_{\sigma\in\mathcal{E}} \subset \mathbb{R}^d$ be given, and let h_K^{n+1} be computed by the forward Euler scheme, step* (6b)*. Then $h_K^{n+1} > 0$, for all $K \in \mathcal{M}$ under the following CFL condition,*

$$2\ \delta t \le \frac{|K|}{\displaystyle\sum_{\sigma\in\mathcal{E}(K)} |\sigma|\,|\boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{K,\sigma}|}. \tag{25}$$

If (25) *is fullfilled and if furthermore*

$$2\ \delta t \le \frac{|K|}{\displaystyle\sum_{\sigma\in\mathcal{E}(K)} |\sigma|\,|\widehat{\boldsymbol{u}}_\sigma^{n+1} \cdot \boldsymbol{n}_{K,\sigma}|}, \tag{26}$$

then h_K^{n+1} computed by the Heun scheme (22) *is positive.*

Secondly, thanks to the choice (15) for the reconstruction of the water height, the property (14) holds, so that the co-called "lake at rest" steady state is preserved by both schemes.

Lemma 3.2 (Steady state "lake at rest")**.** *Let $n \in [\![0, N-1]\!]$, $C \in \mathbb{R}_+$; let $(h_K^n)_{K\in\mathcal{M}} \subset \mathbb{R}$ such that $h_K^n + z_K = C$ for all $K \in \mathcal{M}$ and $\boldsymbol{u}_\sigma^n = 0$ for $\sigma \in \mathcal{E}$. Then the solution $(h_K^{n+1})_{K\in\mathcal{M}}$, $(\boldsymbol{u}_\sigma^{n+1})_{\sigma\in\mathcal{E}}$ of the forward Euler scheme* (6) *(resp. Heun scheme* (22)*) satisfies $h_K^{n+1} + z = C$ for all $K \in \mathcal{M}$ and $\boldsymbol{u}_\sigma^{n+1} = 0$ for $\sigma \in \mathcal{E}$.*

As a consequence of the careful discretisation of the convection term, the segregated forward Euler scheme satisfies a discrete kinetic energy balance, as stated in the following lemma. The proof of this result is an easy adaptation of [22, Lemma 3.2].

Lemma 3.3 (Discrete kinetic energy balance, forward Euler scheme)**.** *A solution to the scheme* (6) *satisfies the following equality, for* $i = 1, 2$, $\sigma \in \mathcal{E}^{(i)}$ *and* $0 \le n \le N-1$*:*

$$\frac{|D_\sigma|}{2\,\delta t}(h_{D_\sigma}^{n+1}(u_{i,\sigma}^{n+1})^2 - h_{D_\sigma}^{n}(u_{i,\sigma}^{n})^2) + \frac{1}{2}\sum_{\epsilon\in\widetilde{\mathcal{E}}^{(i)}(D_\sigma)} |\epsilon|\ (u_{i,\epsilon}^n)^2\ \boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon} + |D_\sigma|\ u_{i,\sigma}^{n+1}(\eth_\sigma p^{n+1}) + |D_\sigma|\ g\, h_{\sigma,c}^{n+1}\, u_{i,\sigma}^{n+1}(\eth_\sigma z) = -R_{i,\sigma}^{n+1}, \quad (27)$$

with

$$R_{i,\sigma}^{n+1} = \frac{1}{2\,\delta t}\ |D_\sigma|\ h_{D_\sigma}^{n+1}\ \left(u_{i,\sigma}^{n+1} - u_{i,\sigma}^{n}\right)^2 - \frac{1}{2}\sum_{\epsilon\in\widetilde{\mathcal{E}}^{(i)}(D_\sigma)} |\epsilon|\ \boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}\ \left(u_{i,\epsilon}^n - u_{i,\sigma}^n\right)^2 + \sum_{\epsilon\in\widetilde{\mathcal{E}}^{(i)}(D_\sigma)} |\epsilon|\ \boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}\ \left(u_{i,\epsilon}^n - u_{i,\sigma}^n\right)\left(u_{i,\sigma}^{n+1} - u_{i,\sigma}^n\right).$$

The scheme also satisfies the following potential energy balance.

Lemma 3.4 (Discrete potential balance, forward Euler scheme)**.** *Let, for* $K \in \mathcal{M}$ *and* $0 \le n \le N$ *the potential energy be defined by* $(E_p)_K^n = \frac{1}{2}g\,(h_K^n)^2 + g h_K^n z_K$*. A solution to the scheme* (6) *satisfies the following equality, for* $K \in \mathcal{M}$ *and* $0 \le n \le N-1$*:*

$$\eth_t(E_p)_K^{n+1} + \mathrm{div}_K(\frac{1}{2}g(h^n)^2\boldsymbol{u}^n) + g z_K \mathrm{div}_K(h^n\boldsymbol{u}^n) + p_K^n \mathrm{div}_K(\boldsymbol{u}^n) = -r_K^{n+1}, \quad (28)$$

with $\eth_t(E_p)_K^{n+1} = \frac{1}{\delta t}((E_p)_K^{n+1} - (E_p)_K^n)$, $\mathrm{div}_K(\frac{1}{2}g(h^n)^2\boldsymbol{u}^n) = \sum_{\sigma\in\mathcal{E}(K)} |\sigma|(\frac{1}{2}g(h_\sigma^n)^2)\boldsymbol{u}_\sigma\cdot\boldsymbol{n}_{K,\sigma}$ *and*

$$|K|\ r_K^{n+1} = \frac{1}{2}\frac{|K|}{\delta t}g(h_K^{n+1} - h_K^n)^2 - \frac{1}{2}\sum_{\sigma\in\mathcal{E}(K)} |\sigma|\ g(h_\sigma^n - h_K^n)^2\ \boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma} + \sum_{\sigma\in\mathcal{E}(K)} |\sigma| g(h_K^{n+1} - h_K^n)\ h_\sigma^n\ \boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma}. \quad (29)$$

Proof. Applying [22, Lemma A1], (re-stated in Lemma B.1 below for the sake of completeness), with $P = K$, $\psi : x \mapsto \frac{1}{2}gx^2$, $\rho_P = h_K^{n+1}$, $\rho_P^* = h_K^n$, $\eta = \sigma$, $\rho_\eta^* = h_\sigma^n$ and $V_\eta^* = |\sigma|\boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma}$, and $R_K^{n+1} = |K|\ r_K^{n+1}$, we get that

$$\frac{g}{2}\eth_t(h_K^{n+1})^2 + \mathrm{div}_K(\frac{g}{2}(h^n)^2\boldsymbol{u}^n) + p_K^n\mathrm{div}_K(\boldsymbol{u}^n) = -\frac{g}{2\delta t}(h_K^{n+1} - h_K^n)^2 + \frac{g}{2}\frac{1}{|K|}\sum_{\sigma\in\mathcal{E}(K)} |\sigma|\ (h_\sigma^n - h_K^n)^2\ \boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma} - \frac{1}{|K|}\sum_{\sigma\in\mathcal{E}(K)} |\sigma| g(h_K^{n+1} - h_K^n)\ h_\sigma^n\ \boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma},$$

Then, multiplying the discrete mass balance equation (6b) by $g z_K$ yields

$$\eth_t(ghz)_K^{n+1} + \mathrm{div}_K(h^n z\boldsymbol{u}) + \frac{1}{|K|}\sum_{\sigma\in\mathcal{E}(K)^{(m)}} |\sigma| g(z_K - z_\sigma)h_\sigma^n\boldsymbol{u}^n\cdot\boldsymbol{n}_{K,\sigma} = 0$$

Summing the two above equations yields (29). □

Since the discrete kinetic and potential energies are computed on the dual and primal meshes respectively, the obtention of a discrete entropy inequality is not straightforward. In [20], a kinetic energy inequality on

the primal cell is obtained from the inequality (1d) to get a discrete local entropy inequality. Here, however we proceed otherwise, thanks to a general Lax-Wendroff Lemma for staggered grids (Lemma A.1 in the appendix), which allows to handle each energy inequality on its respective mesh, without any reconstruction, see Section 5 below.

4. Weak consistency of the schemes

We now wish to prove the weak consistency of the scheme in the Lax-Wendroff sense, namely to prove that if a sequence of solutions is controlled in suitable norms and converges to a limit, this latter necessarily satisfies a weak formulation of the continuous problem.

The pair of functions $(\bar{h}, \bar{\boldsymbol{u}}) \in L^1(\Omega\times[0,T)) \times L^1(\Omega\times[0,T))^2$ is a weak solution to the continuous problem if it satisfies, for any $\varphi \in C_c^\infty\big(\Omega \times [0,T)\big)$ $(\boldsymbol{\varphi} \in C_c^\infty\big(\Omega \times [0,T)\big)^2)$:

$$\int_0^T \int_\Omega \Big[\bar{h}\,\partial_t\varphi + \bar{h}\,\bar{u}\,\cdot\nabla\varphi\Big]\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t + \int_\Omega h_0(\boldsymbol{x})\,\varphi(\boldsymbol{x},0)\,\mathrm{d}\boldsymbol{x} = 0, \tag{30a}$$

$$\int_0^T \int_\Omega \Big[\bar{h}\,\bar{\boldsymbol{u}}\cdot\partial_t\boldsymbol{\varphi} + (\bar{h}\bar{\boldsymbol{u}}\otimes\bar{\boldsymbol{u}}) : \nabla\boldsymbol{\varphi} + \frac{1}{2}\,g\,\bar{h}^2\mathrm{div}\boldsymbol{\varphi} + g\,\bar{h}\,\nabla z\boldsymbol{\varphi}\Big]\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \tag{30b}$$
$$+ \int_\Omega h_0(\boldsymbol{x})\,\boldsymbol{u}_0(\boldsymbol{x})\cdot\varphi(\boldsymbol{x},0)\,\mathrm{d}\boldsymbol{x} = 0.$$

A weak solution of (30) is an entropy weak solution if for any nonnegative test function $\varphi \in C_c^\infty\big(\Omega \times [0,T), \mathbb{R}_+\big)$:

$$\int_0^T \int_\Omega \Big[\bar{E}\,\partial_t\varphi + \bar{\boldsymbol{\Phi}}\cdot\nabla\varphi\Big]\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t + \int_\Omega E_0(\boldsymbol{x})\,\varphi(\boldsymbol{x},0)\,\mathrm{d}\boldsymbol{x} \geq 0, \tag{31}$$

with

$$\bar{E} = \frac{1}{2}\bar{h}|\bar{\boldsymbol{u}}|^2 + \frac{1}{2}g\bar{h}^2 + g\bar{h}z \text{ and } \bar{\Phi} = (\bar{E} + \frac{1}{2}g\bar{h}^2)\bar{\boldsymbol{u}}.$$

Before stating the global weak consistency of the schemes (6) and (22), some definitions and assumptions are needed.

Let $(\mathcal{M}^{(m)}, \mathcal{E}^{(m)})_{m\in\mathbb{N}}$ be a sequence of meshes in the sense of Definition 2.1 and let $(h^{(m)},\, \boldsymbol{u}^{(m)})_{m\in\mathbb{N}}$ be the associated sequence of solutions of the scheme (6)) defined almost everywhere on $(\Omega\times[0,T)$ by:

$$\begin{aligned} u_i^{(m)}(\boldsymbol{x},t) &= \sum_{n=0}^{N-1} \sum_{\sigma\in(\mathcal{E}^{(i)})^{(m)}} (u_i^{(m)})_\sigma^{n+1} \mathbb{1}_{D_\sigma}(\boldsymbol{x}) \mathbb{1}_{[t_n,t_{n+1})}(t), \text{ for } i \in [\![1,2]\!] \\ h^{(m)}(\boldsymbol{x},t) &= \sum_{n=0}^{N-1} \sum_{K\in\mathcal{M}^{(m)}} (h^{(m)})_K^{n+1} \mathbb{1}_K(\boldsymbol{x}) \mathbb{1}_{[t_n,t_{n+1})}(t), \end{aligned}$$

where $\mathbb{1}_A$ is the characteristic function of a given set A, that is $\mathbb{1}_A(y) = 1$ if $y \in A$, $\mathbb{1}_A(y) = 0$ otherwise.

Assumed estimates - Some boundedness and compactness assumptions on the sequence of discrete solutions $(h^{(m)},\, \boldsymbol{u}^{(m)})_{m\in\mathbb{N}}$ are needed in order to prove the Lax-Wendroff type consistency result. First of all we assume that $h^{(m)} > 0$, $\forall m \in \mathbb{N}$ which can be obtained under uniform versions of the CFL conditions (25) and (26), thanks to Lemma 3.1. Furthermore, we assume that:

– the water height $h^{(m)}$ and its inverse are uniformly bounded in $\mathrm{L}^\infty(\Omega\times(0,T))$, *i.e.* there exists $C_{\mathcal{M}}^h \in \mathbb{R}_+^*$ such that for $m \in \mathbb{N}$ and $0 \leq n < N^{(m)}$:

$$\frac{1}{C^h} < (h^{(m)})_K^n \leq C^h, \quad \forall K \in \mathcal{M}^{(m)}, \tag{32}$$

– the velocity $\boldsymbol{u}^{(m)}$ is also uniformly bounded in $\mathrm{L}^\infty(\Omega\times(0,T))^2$, *i.e.* there exists $C^u\in\mathbb{R}_+^*$ such that

$$|(\boldsymbol{u}^{(m)})_\sigma^n|\le C^u,\quad \forall\sigma\in\mathcal{E}^{(m)}. \tag{33}$$

Theorem 4.1 (Weak consistency of the schemes)**.** *Let $(\mathcal{M}^{(m)},\mathcal{E}^{(m)})_{m\in\mathbb{N}}$ be a sequence of meshes such that $\delta t^{(m)}$ and $\delta_{\mathcal{M}^{(m)}}\to 0$ as $m\to+\infty$; assume that there exists $\theta>0$ such that $\theta_{\mathcal{M}^{(m)}}\le\theta$ for any $m\in\mathbb{N}$ (with $\theta_{\mathcal{M}^{(m)}}$ defined by (5)).*

Let $(h^{(m)},\boldsymbol{u}^{(m)})_{m\in\mathbb{N}}$ be a sequence of solutions to the scheme (6) satisfying (32) and (33) converging to $(\bar h,\bar{\boldsymbol{u}})$ in $L^1(\Omega\times(0,T))\times L^1(\Omega\times(0,T))^2$. Then $(\bar h,\bar{\boldsymbol{u}})$ satisfies the weak formulation (30) of the shallow water equations.

Similarly, if $(h^{(m)},\boldsymbol{u}^{(m)})_{m\in\mathbb{N}},(\widehat h^{(m)},\widehat{\boldsymbol{u}}^{(m)})_{m\in\mathbb{N}}$ are sequences of solutions to the scheme (22) both uniformly bounded in the sense of (32) and (33) and converging to $(\bar h,\bar{\boldsymbol{u}})$ in $L^1(\Omega\times(0,T))\times L^1(\Omega\times(0,T))^2$, then the limit $(\bar h,\bar{\boldsymbol{u}})$ satisfies (30).

The proof of this theorem is the object of the following paragraphs; it relies on some general consistency lemmas which generalize the results of [12] to staggered meshes; these results are independent of the problem at hand and are given in the Appendix A. The proof of the consistency of the schemes is given in Section 4.1 for the forward Euler time discretization and in Section 4.2 for the Heun time discretization.

Note that because the convergence and boundedness of the approximate solutions are assumed, no CFL condition is required in Theorem 4.1. However, recall that a CFL condition is for instance already needed to show the positivity of the water height, see Lemma 3.1.

Finally, in Section 4.3, we give some conditions that imply the boundedness and convergence of the sequence $(\widehat h^{(m)},\widehat{\boldsymbol{u}}^{(m)})_{m\in\mathbb{N}}$ if the boundedness and convergence of the sequence $(h^{(m)},\boldsymbol{u}^{(m)})_{m\in\mathbb{N}}$ is assumed. One of this condition is a rather strong CFL-like condition.

4.1. **Proof of consistency of the forward Euler MAC scheme**

4.1.1. *Consistency, mass equation*

Under the assumptions of Theorem 4.1, the aim here is to prove that the limit $(\bar h,\bar{\boldsymbol{u}})$ of the scheme (6) satisfies the weak form of the mass equation (30a). In order to do so, we apply the consistency result of Lemma A.1 in the appendix A, with $U=(h,\boldsymbol{u})$, $\beta(U)=h$, $\boldsymbol{f}(U)=h\boldsymbol{u}$, $\mathcal{P}^{(m)}=\mathcal{M}^{(m)},\mathfrak{F}^{(m)}=\mathcal{E}^{(m)}$, and

$$\begin{aligned}\mathcal{C}^{(m)}_{\mathrm{MASS}}(U^{(m)}):\quad &\Omega\times(0,T)\to\mathbb{R},\\ &(\boldsymbol{x},t)\mapsto \eth_t(h^{(m)})_K^{n+1}+\ \mathrm{div}_K\left((h^{(m)})^n\boldsymbol{u}^n\right)\text{ for }\boldsymbol{x}\in K\text{ and }t\in(t_n,t_{n+1})\end{aligned} \tag{34}$$

We first note that the assumptions (32) and (33) imply that (77) holds. Furthermore, the assumption of Theorem 4.1 that $(h^{(m)},\boldsymbol{u}^{(m)})_{m\in\mathbb{N}}$ is a sequence of solutions to the scheme (6) converging to $(\bar h,\bar{\boldsymbol{u}})$ in $L^1(\Omega\times(0,T))\times L^1(\Omega\times(0,T))^2$ implies that (78) holds.

By the initialisation (6a) of the scheme, it is clear that

$$\sum_{K\in\mathcal{M}^{(m)}}\int_K|(h^{(m)})_K^0-h_0(\boldsymbol{x}))|d\boldsymbol{x}=0,$$

so that the assumption (80) is satisfied.

Since for any $n\in[\![0,N_m-1]\!]$ and $K\in\mathcal{M}$, one has $\beta(U^{(m)}(\boldsymbol{x},t))=h_K^n$ for any $(\boldsymbol{x},t)\in K\times[t_nt_{n+1})$ and $(\beta^{(m)})_K^n=h_K^n$,

$$\sum_{n=0}^{N_m-1}\sum_{K\in\mathcal{M}^{(m)}}\int_{t_n}^{t_{n+1}}\int_K|(h^{(m)})_K^n-h^{(m)}(\boldsymbol{x},t))|d\boldsymbol{x}\ dt=0,$$

and therefore the assumption (81) is also clearly satisfied. Now $(\boldsymbol{F}^{(m)})^n_\sigma = h^n_\sigma \boldsymbol{u}^n_\sigma$ and, because the velocity components are piecewise constant on different grids,

$$\boldsymbol{f}(U^m(\boldsymbol{x},t)) = (f_1(U^m(\boldsymbol{x},t)), f_2(U^m(\boldsymbol{x},t))), \text{ with}$$
$$f_i(U^m(\boldsymbol{x},t)) = \begin{cases} h^n_K u^n_{i,\sigma} \text{ if } \boldsymbol{x} \in D_{K,\sigma} \\ h^n_K u^n_{i,\sigma'} \text{ if } \boldsymbol{x} \in D_{K,\sigma'}, \end{cases} \quad \text{with } K = [\sigma\sigma'] \text{ and where } \sigma \text{ and } \sigma' \perp \boldsymbol{e}^{(i)}.$$

For $\boldsymbol{x} \in K$, $\sigma = K|L$ and $t \in [t_n t_{n+1})$,

$$\begin{aligned} \Big|\Big((\boldsymbol{F}^{(m)})^n_\sigma - \boldsymbol{f}(U^m(\boldsymbol{x},t)\Big) \cdot \boldsymbol{n}_{K,\sigma}\Big| &= \Big|\Big(h^n_\sigma \boldsymbol{u}^n_\sigma - h^n_K \boldsymbol{u}^n_\sigma + h^n_K \boldsymbol{u}^n_\sigma - h^n_K \boldsymbol{u}(\boldsymbol{x},t)\Big) \cdot \boldsymbol{n}_{K,\sigma} \\ &\le C^u \Big|h_K - h_L\Big| + C^h \Big|\boldsymbol{u}_\sigma - \boldsymbol{u}_{\sigma'}\Big|. \end{aligned}$$

Thanks to Lemma B.2 ([12, Lemma 4.2], recalled in Lemma B.2 in the appendix below) we have

$$\sum_{n=0}^{N_m-1} \sum_{K \in \mathcal{M}^{(m)}} \int_{t_n}^{t_{n+1}} \frac{\operatorname{diam}(K)}{|K|} \int_K |\sigma| \Big|\Big(h^n_\sigma \boldsymbol{u}^n_\sigma - h(\boldsymbol{x},t)\boldsymbol{u}(\boldsymbol{x},t)\Big) \cdot \boldsymbol{n}_{K,\sigma}\Big| d\boldsymbol{x}\, dt \to 0 \text{ as } m \to +\infty.$$

so that the assumption (82) is also satisfied.

Hence, by Lemma A.1,

$$\begin{aligned} &\forall \varphi \in C^\infty_c\big(\Omega \times [0,T)\big), \int_0^T \int_\Omega \mathfrak{C}^{(m)}_{\mathrm{MASS}}(U^{(m)})\varphi(\boldsymbol{x},t)\, d\boldsymbol{x}\, dt \to \\ &\quad - \int_\Omega h_0(\boldsymbol{x})\, \varphi(\boldsymbol{x},0)\, \mathrm{d}\boldsymbol{x} - \int_0^T \int_\Omega \Big[\bar{h}(\boldsymbol{x},t)\, \partial_t \varphi(\boldsymbol{x},t) + \bar{h}(\boldsymbol{x},t)\bar{\boldsymbol{u}}(\boldsymbol{x},t)\, \cdot \nabla\varphi(\boldsymbol{x},t)\Big] d\boldsymbol{x}\, dt \text{ as } m \to +\infty. \quad (35) \end{aligned}$$

From (6b) and (35), we conclude that the limit $(\bar{h}, \bar{\boldsymbol{u}})$ of the approximate solutions defined by the forward Euler scheme (6) satisfies (30a).

4.1.2. *Consistency, momentum equation*

Let $\boldsymbol{\varphi} = (\varphi_1, \cdots, \varphi_d) \in (C^\infty_c(\Omega \times [0,T)))^d$ be a test function and let $\varphi^{n+1}_{i,\sigma}$ denote the mean value of φ_i over $\sigma \times (t_n, t_{n+1})$. Multiplying the equation (6d) by $|D_\sigma| \varphi^{n+1}_{i,\sigma}$, summing the result over $\sigma \in \mathcal{E}^{(i)}$ and then summing over $n \in [\![0, N-1]\!]$ and $i = 1, 2$ yields:

$$\sum_{i=1}^{2} Q^{(m)}_{1,i} + Q^{(m)}_{2,i} + Q^{(m)}_{3,i} + Q^{(m)}_{4,i} = 0, \quad (36)$$

with (dropping the exponents (m) in the summations for the sake of simplicity)

$$Q^{(m)}_{1,i} = \sum_{n=0}^{N_m-1} \delta t^{(m)} \sum_{\sigma \in (\mathcal{E}^{(m)})^{(i)}} |D_\sigma|\, \eth_t (h\, u_i)^{n+1}_\sigma, \quad (37)$$

$$Q^{(m)}_{2,i} = \sum_{n=0}^{N_m-1} \delta t^{(m)} \sum_{\sigma \in (\mathcal{E}^{(m)})^{(i)}} |D_\sigma|\, \mathrm{div}_{D_\sigma}(h^n \boldsymbol{u}^n u^n_i) \varphi^{n+1}_{i,\sigma}, \quad (38)$$

$$Q^{(m)}_{3,i} = \sum_{n=0}^{N_m-1} \delta t^{(m)} \sum_{\sigma \in (\mathcal{E}^{(m)})^{(i)}} |D_\sigma|\, \eth_\sigma p^{n+1}\, \varphi^{n+1}_{i,\sigma}, \quad (39)$$

$$Q^{(m)}_{4,i} = \sum_{n=0}^{N_m-1} \delta t^{(m)} \sum_{\sigma \in (\mathcal{E}^{(m)})^{(i)}} |D_\sigma| g\, h^{n+1}_{\sigma,c}\, \eth_\sigma z\, \varphi^{n+1}_{i,\sigma}. \quad (40)$$

The nonlinear convection operator. In order to study the limit of the discrete non linear convection operator defined by $Q_i^{(m)} = Q_{1,i}^{(m)} + Q_{2,i}^{(m)}$, we apply Lemma A.1 with $U = (h, \boldsymbol{u})$, $\beta(U) = hu_i$, $\boldsymbol{f}(U) = h\boldsymbol{u}u_i$, with $\mathcal{P}^{(m)}$ the set of dual cells associated with u_i (that is with the cells corresponding to the vertical edges for $i = 1$ and the horizontal edges for $i = 2$), with $\mathfrak{F} = \widetilde{\mathcal{E}}_i^{(m)}$ and with the dual fluxes $(\boldsymbol{G})_\epsilon^n$ defined by (18). The discrete non linear convection operator thus reads

$$[\mathcal{C}_{\mathrm{MOM}}^{(m)}(U^{(m)})]_i : \quad \Omega \times (0,T) \to \mathbb{R}, \\ (\boldsymbol{x}, t) \mapsto \eth_t(hu_i)_\sigma^{n+1} - \ \mathrm{div}_{D_\sigma}\left(h^n u_i \boldsymbol{u}^n\right) \text{ for } \boldsymbol{x} \in D_\sigma \text{ and } t \in (t_n, t_{n+1}) \tag{41}$$

(again dropping the exponents $^{(m)}$ for the sake of simplicity).

Again, by the initialisation of the scheme (6a) and by the definition of $(h\boldsymbol{u})_{i,\sigma}^0$ (see (16)), it is clear that

$$\sum_{\sigma \in \widetilde{\mathcal{E}}_i^{(m)}} \int_{D_\sigma} |(h\boldsymbol{u})_{i,\sigma}^0 - h_0(\boldsymbol{x}) u_{i,0}(\boldsymbol{x}))| d\boldsymbol{x} = 0 \text{ and } \sum_{n=0}^{N_m-1} \sum_{\sigma \in \mathcal{E}^{(m)}} \int_{t_n}^{t_{n+1}} \int_{D_\sigma} |(h\boldsymbol{u})_{i,\sigma}^n - h(\boldsymbol{x},t) u_i(\boldsymbol{x},t)| d\boldsymbol{x}\, dt = 0,\ i = 1, 2.$$

so that the assumptions (80) and (81) are satisfied.

In order to show that the assumption (82) is satisfied, we need to show that

$$\sum_{n=0}^{N_m-1} \sum_{\sigma \in \mathcal{E}^{(m)}} \int_{t_n}^{t_{n+1}} \frac{\mathrm{diam}(D_\sigma)}{|D_\sigma|} \int_{D_\sigma} |\epsilon| \Big| \sum_{\epsilon \in \widetilde{\mathcal{E}}_i^{(m)}} \Big((\boldsymbol{G}^{(m)})_\epsilon^n - h(\boldsymbol{x},t) u_i(\boldsymbol{x},t) \boldsymbol{u}(\boldsymbol{x},t) \Big) \cdot \boldsymbol{n}_{\sigma,\epsilon} \Big| d\boldsymbol{x}\, dt \to 0 \\ \text{as } m \to +\infty. \tag{42}$$

Let us then estimate, for any $\epsilon \in \widetilde{\mathcal{E}}_i^{(m)}$, $n \in [\![0, N_m - 1]\!]$ and $\boldsymbol{x} \in D_\sigma$ the quantity Y_ϵ^n defined by:

$$Y_\epsilon^n(\boldsymbol{x}) = \Big| \Big((\boldsymbol{G}^{(m)})_\epsilon^n - h(\boldsymbol{x},t) u_i(\boldsymbol{x},t) \boldsymbol{u}(\boldsymbol{x},t) \Big) \cdot \boldsymbol{n}_{\sigma,\epsilon} \Big|.$$

Let L be the (primal) cell such that $\sigma = K|L$.

(1) If $\epsilon = \sigma'|\sigma \subset K$, then $(\boldsymbol{G}^{(m)})_\epsilon^n$ is defined by (18a). By the triangular inequality and thanks to the assumptions (9), (19),(32), and (33), we get that

$$Y_\epsilon^n(\boldsymbol{x}) \le \frac{1}{2}(C^u)^2 |h_K - h_L| + \frac{1}{2}(C^u)^2 |h_K - h_J| + C^h C^u |u_{\sigma,i} - u_{\sigma',i}|,\ \forall \boldsymbol{x} \in D_\sigma,$$

where J is the (primal) cell such that $\sigma' = J|K$, see Figure 2, left.

(2) If $\epsilon \subset K$, then $(\boldsymbol{G}^{(m)})_\epsilon^n$ is defined by (18b). Again by the triangular inequality and thanks to the assumptions (9), (19),(32), and (33), we get that

$$Y_\epsilon^n(\boldsymbol{x}) \le \frac{1}{2}(C^u)^2 |h_K - h_M| + \frac{1}{2}(C^u)^2 |h_K - h_N| + C^h C^u |u_{\sigma,i} - u_{\sigma',i}|,\ \forall \boldsymbol{x} \in D_\sigma,$$

where M and N are the two (primal) cells such that $\tau = K|M$ and $\tau' = L|N$, as depicted on Figure 2, right.

Now recall that the sequence of meshes is assumed to be regular in the sense that $\theta^{(m)} \le \theta$ with $\theta^{(m)}$ defined by (5); therefore, since the sequences $h^{(m)}$ and $\boldsymbol{u}^{(m)}$ converge in L^1 as m tends to $+\infty$, we may again apply

Lemma B.2 below, to get that (42) holds. Hence, owing to Lemma A.1, we get that

$$Q_i^{(m)} = Q_{1,i}^{(m)} + Q_{2,i}^{(m)} = \int_0^T \int_\Omega \mathcal{C}_{\rm MOM}^{(m)}(U^{(m)})\ \varphi(\boldsymbol{x},t)\ d\boldsymbol{x}\ dt \to$$
$$\int_0^T \int_\Omega \Big[\bar{h}(\boldsymbol{x},t)\bar{u}_i(\boldsymbol{x},t)\,\partial_t\varphi(\boldsymbol{x},t) + \bar{h}(\boldsymbol{x},t)\bar{u}_i(\boldsymbol{x},t)\bar{\boldsymbol{u}}(\boldsymbol{x},t)\cdot\nabla\varphi(\boldsymbol{x},t)\Big] d\boldsymbol{x}\ dt$$
$$+ \int_\Omega h_0(\boldsymbol{x})u_{i,0}(\boldsymbol{x})\,\varphi(\boldsymbol{x},0)\ \mathrm{d}\boldsymbol{x} \text{ as } m\to+\infty. \quad (43)$$

Pressure gradient and bathymetry. Let us now study the terms $Q_{3,i}^{(m)}$ and $Q_{4,i}^{(m)}$ defined by (39) and (40). By the definition (11) of $\eth_\sigma p$ and by conservativity, we have (again dropping the exponents (m))

$$\sum_{i=1}^2 Q_{3,i}^{(m)} = -\sum_{n=0}^{N_m-1}\delta t^{(m)} \sum_{K\in\mathcal{M}^{(m)}}\sum_{\sigma\in\mathcal{E}(K)} p_K^{n+1}\int_\sigma \boldsymbol{\varphi}_\sigma^{n+1}\cdot\boldsymbol{n}_{K,\sigma}$$
$$= -\int_0^T\int_\Omega p^{(m)}(\boldsymbol{x},t)\ \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x},t)\ \mathrm{d}\boldsymbol{x}\ \mathrm{d}t$$

Since the sequence $(h^{(m)})m\in\mathbb{N}$ is bounded in $L^\infty(\Omega\times(0,T))$ and converges to $\bar{h}$ in $L^1(\Omega\times(0,T))$, the sequence $(p^{(m)})_{m\in\mathbb{N}}$ converges to $\bar{p}=\frac{1}{2}g\bar{h}^2$ in $L^1(\Omega\times(0,T))$ as $m\to+\infty$. Hence we get

$$\sum_{i=1}^2 Q_{3,i}^{(m)} \to \int_0^T\int_\Omega \bar{p}(\boldsymbol{x},t)\ \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x},t)\ \mathrm{d}\boldsymbol{x}\ \mathrm{d}t \text{ as } m\to+\infty. \quad (44)$$

Let us now turn to the bathymetry term $Q_{4,i}^{(m)}$, which may be written

$$Q_{4,i}^{(m)} = \int_0^T\int_\Omega \widetilde{h}^{(m)}(\boldsymbol{x},t)\eth_i^{(m)}z(\boldsymbol{x})\widetilde{\varphi}_i^{(m)}(\boldsymbol{x},t)\ \mathrm{d}\boldsymbol{x}\ \mathrm{d}t,$$

where

- the function $\widetilde{h}^{(m)}:\Omega\times(0,T)\to\mathbb{R}$ is defined by $\widetilde{h}(\boldsymbol{x},t)=h_{\sigma,c}^{n+1}=\frac{1}{2}(h_K^{n+1}+h_L^{n+1})$ for $\boldsymbol{x}\in D_\sigma$ and $t\in(t_n,t_{n+1})$; the sequence $(\widetilde{h}^{(m)})_{m\in\mathbb{N}}$ is therefore bounded in $L^\infty(\Omega\times(0,T))$ and converges to $\bar{h}$ in $L^1(\Omega\times(0,T))$;
- the function $\widetilde{\varphi}_i^{(m)}:\Omega\times(0,T)\to\mathbb{R}$ is defined by $\widetilde{\varphi}_i^{(m)}(\boldsymbol{x},t)=\varphi_\sigma^{n+1}$ for $\boldsymbol{x}\in D_\sigma$ and $t\in(t_n,t_{n+1})$; by the regularity of $\boldsymbol{\varphi}$, the sequence $(\widetilde{\varphi}_i^{(m)})_{m\in\mathbb{N}}$ converges to φ_i uniformly.
- by (13), the function $\eth_i^{(m)}z:\Omega\to\mathbb{R}$ is defined by $\eth_i^{(m)}z=\sum_{\substack{\sigma\in\mathcal{E}_{\rm int}\\ \sigma=K|L}}\frac{|\sigma|}{|D_\sigma|}(z(\boldsymbol{x}_L)-z(\boldsymbol{x}_K))\mathbb{1}_{D_\sigma}$. Since z is a regular function, the sequence of functions $(\eth_i^{(m)}z)_{m\in\mathbb{N}}$ converges uniformly to the derivative $\partial_i z$ of z with respect to the i-th variable as $m\to+\infty$.

Hence

$$Q_{4,i}^{(m)} \to \int_0^T\int_\Omega \bar{h}(\boldsymbol{x},t)\partial_i z(\boldsymbol{x})\varphi_i(\boldsymbol{x},t)\ \mathrm{d}\boldsymbol{x}\ \mathrm{d}t \text{ as } m\to\infty. \quad (45)$$

Limit of the momentum equation. Passing to the limit in (36) as $m\to+\infty$, using (43), (44) and (45), we get that the limit $(\bar{h},\bar{\boldsymbol{u}})$ of the approximate solutions defined by the forward Euler scheme (6) satisfies (30b), which concludes first part of the proof of Theorem 4.1.

4.2. **Proof of the weak consistency of the Heun scheme**

4.2.1. *Mass balance*

Under the assumptions of Theorem 4.1, the aim here is to prove that the limit $(\bar{h},\bar{\boldsymbol{u}})$ of the scheme (22a)-(22f) satisfies the weak form of the mass equation (30a). In order to do so, we consider the equivalent mass

equation (24a). Because of the structure of the scheme, we cannot use here Lemma A.1 straightforwardly as in the case of the forward Euler scheme. We apply Lemma A.2 with $U = (h, \boldsymbol{u})$, $\beta(U) = h$, $\boldsymbol{f}(U) = h\boldsymbol{u}$, $\mathcal{P}^{(m)} = \mathcal{M}^{(m)}, \mathfrak{F}^{(m)} = \mathcal{E}^{(m)}$ and then Lemma A.3 twice: once with $U = (h, \boldsymbol{u})$, $\boldsymbol{f}(U) = h\boldsymbol{u}$, $\mathcal{P}^{(m)} = \mathcal{M}^{(m)}, \mathfrak{F}^{(m)} = \mathcal{E}^{(m)}$, and then with $U = (\widehat{h}, \widehat{\boldsymbol{u}})$, $\boldsymbol{f}(U) = \widehat{h}\widehat{\boldsymbol{u}}$. Thanks to the arguments developed in Section 4.1.1, it is easy to check that in each case, the assumptions of the lemmas are satisfied, so that we can conclude that $(\bar{h}, \bar{\boldsymbol{u}})$ satisfies (30a).

4.2.2. *Momentum balance*

Still under the assumptions of Theorem 4.1, we now prove that the limit $(\bar{h}, \bar{\boldsymbol{u}})$ of the scheme (22a)-(22f) satisfies the weak form of the mass equation (30b). Again we consider the equivalent momentum equation (24b). Multiplying the equation (24b) by $|D_\sigma|\varphi_{i,\sigma}^{n+1}$, summing the result over $\sigma \in \mathcal{E}^{(i)}$ and then summing over $n \in [\![0, N-1]\!]$ and $i = 1, 2$ yields:

$$\sum_{i=1}^{2} \Big[Q_{1,i}^{(m)} + \frac{1}{2}(Q_{2,i}^{(m)} + \widehat{Q}_{2,i}^{(m)} + Q_{3,i}^{(m)} + \widehat{Q}_{3,i}^{(m)}) + Q_{4,i}^{(m)} + \widehat{Q}_{4,i}^{(m)}) \Big] = 0, \tag{46}$$

where $Q_{1,i}^{(m)}, \dots, Q_{4,i}^{(m)}$ are defined by (37)-(40), and $\widehat{Q}_{2,i}^{(m)}, \widehat{Q}_{3,i}^{(m)}, \widehat{Q}_{4,i}^{(m)}$ are defined by (38)-(40), replacing the unknowns h, p, u by $\widehat{h}, \widehat{p}, \widehat{u}$.

Again, because of the structure of the scheme, we cannot use Lemma A.1 directly: we use Lemma A.2 for the time derivative term $Q_{1,i}^{(m)}$ and Lemma A.3 for the terms $Q_{2,i}^{(m)}$ and $\widehat{Q}_{2,i}^{(m)}$, with $\mathcal{P}^{(m)}$ the set of dual cells associated with u_i (that is with the vertical edges for $i = 1$ and the horizontal edges for $i = 2$), with $\mathfrak{F} = \widetilde{\mathcal{E}}_i^{(m)}$ and with the dual fluxes $(\boldsymbol{G})_\epsilon^n$ defined by (18). We first apply Lemma A.2 with $U = (h, \boldsymbol{u})$, $\beta(U) = hu_i$, $\boldsymbol{f}(U) = h\boldsymbol{u}u_i$, and then Lemma A.3, once with $U = (h, \boldsymbol{u})$, $\beta(U) = hu_i$, $\boldsymbol{f}(U) = h\boldsymbol{u}u_i$ and then with $U = (\widehat{h}, \widehat{\boldsymbol{u}})$, $\beta(U) = \widehat{h}\widehat{u}_i$, $\boldsymbol{f}(U) = \widehat{h}\widehat{\boldsymbol{u}}\widehat{u}_i$. Thanks to the arguments developed in Section 4.1.2, it is easy to check that in each case, the assumptions of the lemmas are satisfied, so that

$$Q_{1,i}^{(m)} + \frac{1}{2}(Q_{2,i}^{(m)} + \widehat{Q}_{2,i}^{(m)}) \to \int_0^T \int_\Omega \Big[\bar{h}(\boldsymbol{x},t)\bar{u}_i(\boldsymbol{x},t)\, \partial_t\varphi(\boldsymbol{x},t) + \bar{h}(\boldsymbol{x},t)\bar{u}_i(\boldsymbol{x},t)\bar{\boldsymbol{u}}(\boldsymbol{x},t) \cdot \nabla\varphi(\boldsymbol{x},t) \Big] d\boldsymbol{x}\, dt + \int_\Omega h_0(\boldsymbol{x})u_{i,0}(\boldsymbol{x})\, \varphi(\boldsymbol{x},0)\, \mathrm{d}\boldsymbol{x} \text{ as } m \to +\infty. \tag{47}$$

The proof of convergence of the pressure gradient and bathymetry terms $Q_{3,i}^{(m)}$), $Q_{4,i}^{(m)}$, $\widehat{Q}_{3,i}^{(m)}$) and $\widehat{Q}_{4,i}^{(m)}$ follow the exact same lines as that of the terms $Q_{3,i}^{(m)}$) and $Q_{4,i}$ in Section 4.1.2. Hence

$$\sum_{i=1}^{2} \frac{1}{2} \Big(Q_{3,i}^{(m)} + \widehat{Q}_{3,i}^{(m)} + Q_{4,i}^{(m)} + \widehat{Q}_{4,i}^{(m)} \Big) \to \int_0^T \int_\Omega \Big(\bar{p}(\boldsymbol{x},t)\, \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x},t) + \bar{h}(\boldsymbol{x},t)\nabla z(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x},t) \Big)\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}t \quad \text{as } m \to +\infty. \tag{48}$$

Therefore, owing to (47) and (48), we may pass to the limit in (46) and conclude that $(\bar{h}, \bar{\boldsymbol{u}})$ satisfies (30b). This concludes the proof of Theorem 4.1.

4.3. **A sufficient condition for the convergence of the intermediate solutions**

In Theorem 4.1, we assumed the boundedness and convergence of both sequences $(h^{(m)}, u^{(m)})$ and $(\widehat{h}^{(m)}, \widehat{u}^{(m)})$. In fact, under a restricted CFL condition, we may prove that the convergence and boundedness of the sequence $(h^{(m)}, u^{(m)})$ implies the convergence and boundedness of the sequence $(\widehat{h}^{(m)}, \widehat{u}^{(m)})$.

Lemma 4.2 (Bound on the intermediate step, Heun scheme)**.** *Let $n \in [\![0, N-1]\!]$, let $(h_K^n)_{K\in\mathcal{M}} \subset \mathbb{R}_+^*$ and $(\boldsymbol{u}_\sigma^n)_{\sigma\in\mathcal{E}} \subset \mathbb{R}^d$ be given. Assume that there exists $\zeta \in (0,1)$ such that the following restricted CFL-like*

condition holds (note that it slightly more restrictive than (25)*):*

$$2\ \delta t \le \zeta \frac{|K|}{\sum_{\sigma\in\mathcal{E}(K)} |\sigma|\,|\boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma}|} \text{ for all } K\in\mathcal{M}. \tag{49}$$

Let $C_{\mathcal{M}}^{\delta t}$, $C_{\mathcal{M}}^{h}$ *and* $C_{\mathcal{M}}^{u}\in\mathbb{R}_+^*$ *be such that*

$$\delta t \le C_{\mathcal{M}}^{\delta t}\min_{\sigma\in\mathcal{E}}|\sigma|, \tag{50a}$$

$$\frac{1}{C_{\mathcal{M}}^{h}} \le h_K^n \le C_{\mathcal{M}}^{h}, \forall n\in[\![0,N-1]\!], \forall K\in\mathcal{M}, \tag{50b}$$

$$\max_{\sigma\in\mathcal{E}}|\boldsymbol{u}_\sigma^n| \le C_{\mathcal{M}}^{u}, \forall n\in[\![0,N-1]\!]. \tag{50c}$$

Then the solutions $(\widehat{h}_K^{n+1})_{K\in\mathcal{M}}$ $(\widehat{\boldsymbol{u}}_\sigma^n)_{\sigma\in\mathcal{E}}$ *of the Heun steps* (22a)-(22b) *satisfy:*

$$\frac{1-\zeta}{C_{\mathcal{M}}^{h}} \le \widehat{h}_K^{n+1} \le 2C_{\mathcal{M}}^{h}\ \forall K\in\mathcal{M}, \tag{51a}$$

$$|\widehat{\boldsymbol{u}}_\sigma^{n+1}| \le C_{\mathcal{M}}^{u} + C_{\mathcal{M}}^{\delta t}\frac{(C_{\mathcal{M}}^{h})^2}{1-\zeta}\Big(4(C_{\mathcal{M}}^{u})^2 + g(C_{\mathcal{M}}^{h}+||z||_\infty)\Big).\ \forall\sigma\in\mathcal{E}, \tag{51b}$$

Proof. From (22a) and by the definition (8) of the discrete divergence, we have

$$\widehat{h}_K^{n+1} = h_K^n - \sum_{\sigma\in\mathcal{E}(K)}(\omega_{K,\sigma}^n)^+h_\sigma^n + \sum_{\sigma\in\mathcal{E}(K)}(\omega_{K,\sigma}^n)^-h_\sigma^n \text{ with } \omega_{K,\sigma}^n = \delta t\frac{|\sigma|}{|K|}\,\boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma}.$$

Thanks to the condition (10), $\exists\,\alpha_\sigma^K\in[0,1]$ and $M_\sigma^K\in\mathcal{M}$ such that $h_\sigma^n - h_K^n = \alpha_\sigma^K(h_K^n - h_{M_\sigma^K}^n)$ if $\omega_{K,\sigma}^n\ge 0$, and therefore

$$h_\sigma^n = h_K^n(1+\alpha_\sigma^K) - \alpha_\sigma^K h_{M_\sigma^K}^n.$$

Hence

$$\widehat{h}_K^{n+1} = \Big(1 - \sum_{\sigma\in\mathcal{E}(K)}(\omega_{K,\sigma}^n)^+(1+\alpha_\sigma^K)\Big)h_K^n + \sum_{\sigma\in\mathcal{E}(K)}(\omega_{K,\sigma}^n)^-h_\sigma^n + \sum_{\sigma\in\mathcal{E}(K)}\alpha_\sigma^K(\omega_{K,\sigma}^n)^+h_{M_\sigma^K}^n.$$

Therefore, thanks to the condition (49), we get (51a).

Let us now prove (51b); from (22b) we have

$$\begin{aligned}
\widehat{u}_{i,\sigma}^{n+1} &= \frac{1}{\widehat{h}_{D_\sigma}^{n+1}}\Big(h_{D_\sigma}^n u_{i,\sigma}^n - \frac{\delta t}{|D_\sigma|}\sum_{\epsilon\in\widetilde{\mathcal{E}}(D_\sigma)}|\epsilon|\boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}u_{i,\epsilon}^n\Big) - \frac{\delta t|\sigma|}{|D_\sigma|}\frac{gh_{\sigma,c}^n}{\widehat{h}_{D_\sigma}^{n+1}}\big(h_L^n - h_K^n + z_L - z_K\big)\\
&= \frac{1}{\widehat{h}_{D_\sigma}^{n+1}}\left[\Big(h_{D_\sigma}^n - \frac{\delta t}{|D_\sigma|}\sum_{\epsilon\in\widetilde{\mathcal{E}}(D_\sigma)}|\epsilon|\boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}\Big)u_{i,\sigma}^n - \frac{\delta t}{|D_\sigma|}\sum_{\epsilon\in\widetilde{\mathcal{E}}(D_\sigma)}|\epsilon|\boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}\,\big(u_{i,\epsilon}^n - u_{i,\sigma}^n\big)\right]\\
&\qquad - \frac{\delta t|\sigma|}{|D_\sigma|}\frac{gh_{\sigma,c}^n}{\widehat{h}_{D_\sigma}^{n+1}}\big(h_L^n - h_K^n + z_L - z_K\big).
\end{aligned}$$

Since the values $\widehat{h}_{D_\sigma}^{n+1}$ and $\widehat{h}_{D_\sigma}^{n}$ are computed by an equivalent formula to (17), they satisfy a discrete dual mass balance of the type (21), and therefore:

$$\widehat{u}_{i,\sigma}^{n+1} = u_{i,\sigma}^n - \frac{1}{\widehat{h}_{D_\sigma}^{n+1}}\left[\frac{\delta t}{|D_\sigma|}\sum_{\epsilon\in\widetilde{\mathcal{E}}(D_\sigma)}|\epsilon|\boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}\,\big(u_{i,\epsilon}^n - u_{i,\sigma}^n\big)\right] - \frac{\delta t|\sigma|}{|D_\sigma|}\frac{gh_{\sigma,c}^n}{\widehat{h}_{D_\sigma}^{n+1}}\big(h_L^n - h_K^n + z_L - z_K\big).$$

Thanks to the CFL condition (50a) and to the bounds on $\boldsymbol{u}_\sigma^n$ and $\widehat{h}_{D_\sigma}^{n+1}$ for all σ (recall that for $\sigma = K|L$, $\widehat{h}_{D_\sigma}^{n+1}$ is a convex combination of $\widehat{h}_K^{n+1}$ and $\widehat{h}_L^{n+1}$),

$$\frac{1}{\widehat{h}_{D_\sigma}^{n+1}}\left|\frac{\delta t}{|D_\sigma|}\sum_{\epsilon\in\widetilde{\mathcal{E}}(D_\sigma)}|\epsilon|\boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}\,\left(u_{i,\epsilon}^n-u_{i,\sigma}^n\right)\right|\le\frac{4C_{\mathcal{M}}^{\delta}t(C_{\mathcal{M}}^u)^2(C_{\mathcal{M}}^h)^2}{1-\zeta}.$$

Furthermore, since $2h_{\sigma,c}^n = h_K^n + h_L^n$ and agina owing to (50a),

$$\frac{\delta t|\sigma|}{|D_\sigma|}\frac{gh_{\sigma,c}^n}{\widehat{h}_{D_\sigma}^{n+1}}\big(h_L^n-h_K^n+z_L-z_K\big)\le C_{\mathcal{M}}^{\delta t}g\frac{(C_{\mathcal{M}}^h)^2}{1-\zeta}\Big(\max_{K\in\mathcal{M}}(h_K^n)+\max_{K\in\mathcal{M}}(z_K)\Big)\le\frac{C_{\mathcal{M}}^{\delta t}(C_{\mathcal{M}}^h)^2g}{1-\zeta}(C_{\mathcal{M}}^h+||z||_\infty).$$

Therefore,

$$|\widehat{u}_{i,\sigma}^{n+1}|\le C_{\mathcal{M}}^u+\frac{4C_{\mathcal{M}}^{\delta}t(C_{\mathcal{M}}^u)^2(C_{\mathcal{M}}^h)^2}{1-\zeta}+\frac{C_{\mathcal{M}}^{\delta t}(C_{\mathcal{M}}^h)^2g}{1-\zeta}(C_{\mathcal{M}}^h+||z||_\infty),$$

which concludes the proof that (51b) holds. □

Lemma 4.3 (L^1 convergence of the intermediate step, Heun scheme)**.** *Let $(\mathcal{M}^{(m)},\mathcal{E}^{(m)})_{m\in\mathbb{N}}$ be a sequence of meshes such that $\delta t^{(m)}$ and $\delta_{\mathcal{M}^{(m)}}\to 0$ as $m\to+\infty$; assume that $(\mathcal{M}^{(m)},\mathcal{E}^{(m)})_{m\in\mathbb{N}}$ is uniformly regular, in the sense that there exists $\theta>0$ such that $\theta_{\mathcal{M}^{(m)}}\le\theta$ for any $m\in\mathbb{N}$ (with $\theta_{\mathcal{M}^{(m)}}$ defined by (5)).*

Let $(h^{(m)},\boldsymbol{u}^{(m)})_{m\in\mathbb{N}}$ be a sequence of solutions to the scheme (6) satisfying (32) and (33) converging to $(\bar{h},\bar{\boldsymbol{u}})$ in $L^1(\Omega\times(0,T))\times L^1(\Omega\times(0,T))^2$.

Assume that there exists $\zeta\in(0,1)$ such that the following restricted CFL-like condition holds:

$$2\ \delta t^{(m)}\le\zeta\frac{|K|}{\sum_{\sigma\in\mathcal{E}(K)}|\sigma|\,|(\boldsymbol{u}^{(m)})_\sigma^n\cdot\boldsymbol{n}_{K,\sigma}|},\ \forall K\in\mathcal{M}^{(m)},\ \forall m\in\mathbb{N},\tag{52}$$

and assume that there exists $C^{\delta t}\in\mathbb{R}_+^*$ not depending on m such that

$$\delta t^{(m)}\le C^{\delta t}\min_{\sigma\in\mathcal{E}^{(m)}}|\sigma|,\forall m\in\mathbb{N}.\tag{53}$$

Then there exists $\widehat{C}^u,\widehat{C}^h\in\mathbb{R}_+^*$ such that

$$\frac{1}{\widehat{C}^h}<(\widetilde{h}^{(m)})_K^n\le\widehat{C}^h,\quad\forall K\in\mathcal{M}^{(m)},\tag{54a}$$

$$|(\widetilde{\boldsymbol{u}}^{(m)})_\sigma^n|\le\widehat{C}^u,\quad\forall\sigma\in\mathcal{E}^{(m)}.\tag{54b}$$

Furthermore, the sequence $\left(\widehat{h}^{(m)},\widehat{\boldsymbol{u}}^{(m)}\right)_{m\in\mathbb{N}}$ converges to $(\bar{h},\bar{\boldsymbol{u}})$ in $L^1(\Omega\times(0,T))\times L^1(\Omega\times(0,T))^2$.

Proof. Under the above assumptions, the hypotheses (49) and (50) hold uniformly with respect to m, so that the bounds (54) are a direct consequence of Lemma 4.2.

Now, from equation (22a), we get that

$$\widehat{h}_K^{n+1}-h_K^n=-\frac{\delta t}{|K|}\sum_{\sigma\in\mathcal{E}(K)}|\sigma|\,\boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma}\,(h_\sigma^n-h_K^n)-\frac{\delta t}{|K|}\sum_{\sigma\in\mathcal{E}(K)}|\sigma|\,h_K^n\ \boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma},\ \forall K\in\mathcal{M}^{(m)}.$$

For $K\in\mathcal{M}^{(m)}$, let us denote by $\sigma_{K,i}$ and $\sigma'_{K,i}$ the edges of K in the direction $i\in[\![1,2]\!]$, so that $K=[\overrightarrow{\sigma_{K,i}\sigma'_{K,i}}]$ for $i\in[\![1,2]\!]$; noting that $\boldsymbol{n}_{K,\sigma_{K,i}}=-\boldsymbol{n}_{K,\sigma'_{K,i}}$ and that $|\sigma_{K,i}|=|\sigma'_{K,i}|$, and owing to (32), we get that

$$\Big|\sum_{\sigma\in\mathcal{E}(K)}|\sigma|\,h_K^n\ \boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma}\Big|\le C^h\sum_{i=1}^d|\sigma_{K,i}||u_{i,\sigma_{K,i}}^n-u_{i,\sigma'_{K,i}}^n|,\ \forall K\in\mathcal{M}^{(m)}.$$

Since h_σ is a convex combination of h_K and h_L, with K and L such that $\sigma = K|L$, we get:

$$|\widehat{h}_K^{n+1} - h_K^n| \leq \sum_{\substack{\sigma\in\mathcal{E}(K)\\ \sigma=K|L}} \frac{\delta t^{(m)}}{|K|}\,|\sigma|\,\big|\boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma}\big|\,|h_L^n - h_K^n| + C^h\sum_{i=1}^{2}\delta t\frac{|\sigma|}{|K|}|u_{i,\sigma_{K,i}}^n - u_{i,\sigma'_{K,i}}^n|,\ \forall K\in\mathcal{M}^{(m)}.$$

Noting that (53) implies that $\dfrac{\delta t^{(m)}}{|K|}\,|\sigma| \leq 1$ and thanks to the condition (33), we thus get that there exists $C\in\mathbb{R}_+$ depending on C^h, C^u, $C^{\delta t}$ such that

$$|\widehat{h}_K^{n+1} - h_K^n| \leq C\Big[\sum_{\substack{\sigma\in\mathcal{E}(K)\\ \sigma=K|L}} |h_L^n - h_K^n| + \sum_{i=1}^{d}|u_{i,\sigma_{K,i}}^n - u_{i,\sigma'_{K,i}}^n|\Big],\ \forall K\in\mathcal{M}^{(m)}.$$

Multiplying this latter inequality by $|K|\delta t^{(m)}$ and summing over $K\in\mathcal{M}^{(m)}$ and $n\in[\![0,N]\!]$, using the uniform regularity of the mesh and owing again to the convergence result on the space translates given in Lemma B.2, we conclude that

$$\int_0^T\int_\Omega |\widehat{h}^{(m)} - h^{(m)}|\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \to 0 \text{ as } m\to+\infty.$$

Let us now turn to the intermediate velocities. Owing to (22b), (23) and since $\widehat{u}$ satisfies a dual mass balance of the form (21), we have for $\sigma = K|L\in(\mathcal{E}_{\rm int}^{(i)})^{(m)}$, $i\in[\![1,2]\!]$:

$$\begin{aligned}\widehat{h}_{D_\sigma}^{n+1}(\widehat{u}_{i,\sigma}^{n+1} - u_{i,\sigma}^n) &= -(\widehat{h}_{D_\sigma}^{n+1} - h_{D_\sigma}^n)u_{i,\sigma}^n - \frac{\delta t}{|D_\sigma|}\sum_{\epsilon\in\widetilde{\mathcal{E}}(D_\sigma)}\boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}u_{i,\epsilon}^n - \delta t\ gh_{\sigma,c}^n\big((\eth_\sigma h^n) + (\eth_\sigma z)\big)\\ &= -\sum_{\epsilon\in\widetilde{\mathcal{E}}(D_\sigma)}\Big[\frac{\delta t\ |\epsilon|}{|D_\sigma|}\,\boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}(u_{i,\epsilon}^n - u_{i,\sigma}^n) + \frac{\delta t|\sigma|}{|D_\sigma|}gh_{\sigma,c}^n\big(h_L^n - h_K^n + z_L - z_K\big)\Big].\end{aligned}$$

Hence, owing to (32), (33), (53) and to the fact that for $\epsilon = \sigma|\sigma'$, $u_{i,\epsilon}^n$ is a convex combination of $u_{i,\sigma}^n$ and $u_{i,\sigma'}^n$, there exists $C\in\mathbb{R}_+$ depending only on C^h, C^u, $C^{\delta t}$ and g such that

$$|\widehat{u}_{i,\sigma}^{n+1} - u_{i,\sigma}^n| \leq C\Big[\sum_{\substack{\epsilon\in\widetilde{\mathcal{E}}(D_\sigma)\\ \epsilon=\sigma|\sigma'}}|u_{i,\sigma'}^n - u_{i,\sigma}^n| + |h_L^n - h_K^n| + |z_L - z_K|\Big],\ \text{for}\ \ i=1,2.$$

Multiplying this latter inequality by $|D_\sigma|\delta t^{(m)}$ and summing over $\sigma\in\mathcal{M}^{(m)}$ and $n\in[\![0,N]\!]$, using the uniform regularity of the mesh and again thanks to Lemma B.2 we conclude that

$$\int_0^T\int_\Omega |\widehat{u}_i^{(m)} - u_i^{(m)}|\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \to 0 \text{ as } m\to+\infty,\ \text{for}\ \ i=1,2.$$

□

5. Weak entropy consistency of the forward Euler- MAC scheme

Theorem 5.1 (Weak entropy consistency of the forward Euler MAC scheme)**.** *Let $(\mathcal{M}^{(m)},\mathcal{E}^{(m)})_{m\in\mathbb{N}}$ be a sequence of meshes such that $\delta t^{(m)}$ and $\delta_{\mathcal{M}^{(m)}}\to 0$ as $m\to+\infty$; assume that there exists $\theta>0$ such that $\theta_{\mathcal{M}^{(m)}}\leq\theta$ for any $m\in\mathbb{N}$ (with $\theta_{\mathcal{M}^{(m)}}$ defined by (5)). Let $(h^{(m)},\boldsymbol{u}^{(m)})_{m\in\mathbb{N}}$ be a sequence of solutions to the*

scheme (6) *converging to* $(\bar{h}, \bar{\boldsymbol{u}})$ *in* $L^1(\Omega \times (0,T)) \times L^1(\Omega \times (0,T))^2$*, such that* (32), (33) *hold. Assume the following CFL-like condition:*

$$\delta t^{(m)} \leq \frac{|D_\sigma| h_{D_\sigma}^{n+1}}{\sum\limits_{\substack{\epsilon \in \mathcal{E} D_\sigma \\ \boldsymbol{F}_\epsilon^n \boldsymbol{n}_{\sigma,\epsilon} > 0}} |\epsilon| \ \boldsymbol{F}_\epsilon^n \cdot \boldsymbol{n}_{\sigma,\epsilon}}. \tag{55}$$

Assume furthermore that

$$\exists C_{BVt} \in \mathbb{R}_+ \ : \ \sum_{K \in \mathcal{M}^{(m)}} |K| |(h^{(m)})_K^{n+1} - (h^{(m)})_K^n| \sum_{K \in \mathcal{M}^{(m)}} |K| |(h^{(m)})_K^{n+1} - (h^{(m)})_K^n| \leq C_{BVt} \ \forall m \in \mathbb{N}, \tag{56a}$$

$$\frac{\delta t^{(m)}}{\inf_{K \in \mathcal{M}^{(m)}} \operatorname{diam}(K)} \sum_{n=0}^{N_m - 1} \to 0 \ \textit{as} \ m \to +\infty., \tag{56b}$$

and that the coefficients $\lambda_{K,\sigma}$ *and* $\mu_{\sigma,\epsilon}$ *in* (9) *and* (19) *satisfy:*

$$\lambda_{K,\sigma} \in [\frac{1}{2}, 1] : \ \textit{if} \ \boldsymbol{F}_\sigma \cdot \boldsymbol{n}_{K,\sigma} \geq 0, \tag{57}$$

$$\mu_{\sigma,\epsilon} \in [\frac{1}{2}, 1] : \ \textit{if} \ \boldsymbol{F}_\sigma \cdot \boldsymbol{n}_{K,\sigma} \geq 0. \tag{58}$$

Then $(\bar{h}, \bar{\boldsymbol{u}})$ *satisfies the entropy inequality* (31)*.*

Note also that the condition (9) implies that

$$\forall \ K \in \mathcal{M}, \ \forall \sigma = K|L \in \mathcal{E}_{\rm int}(K) \text{ with } K \text{ such that } \boldsymbol{F}_\sigma \cdot \boldsymbol{n}_{K,\sigma} \geq 0,$$
$$h_\sigma = \frac{1}{2}(h_K + h_L) + (\lambda_{K,\sigma} - \frac{1}{2})(h_K - h_L), \text{ with } \lambda_{K,\sigma} - \frac{1}{2} \geq 0. \tag{59}$$

Also Note that the condition (57) is rather restrictive. Indeed, it is satisfied by the usual two slopes minmod limiter [15] only in the case of a uniform Cartesian mesh [27], and it is not satisfied by the three slopes minmod limiter.

Proof. Let $\varphi \in C_c^\infty(\Omega \times [0,T), \mathbb{R}_+)$, and for a given discretization $(\mathcal{M}^{(m)}, \mathcal{E}^{(m)})$ let φ_K^n (resp. φ_σ^n) denote the mean value of φ on $K \times (t_n, t_{n+1})$ (resp. $D_\sigma \times (t_n, t_{n+1})$), for any $K \in \mathcal{M}^{(m)}$ (resp. $\sigma \in \mathcal{E}^{(m)}$) and $n \in [\![0, N_m - 1]\!]$. Let us multiply the discrete kinetic energy balance (27) by $\delta t \varphi_\sigma^{n+1}$ and sum over $\sigma \in \mathcal{E}^{(m)}$ and $i \in \{1,2\}$; let us then multiply the discrete potential energy balance (28) by $\delta t |K| \varphi_K^n$ and sum over $K \in \mathcal{M}^{(m)}$. Summing the two resulting equations and summing over $n \in [\![0, N_m - 1]\!]$, we get, owing to lemmas 3.3 and 3.4,

$$\int_0^T \int_\Omega \mathfrak{C}_{\rm KIN}^{(m)}(U^{(m)}) \varphi(\boldsymbol{x}, t) \ d\boldsymbol{x} \ dt + \int_0^T \int_\Omega \mathfrak{C}_{\rm POT}^{(m)}(U^{(m)}) \varphi(\boldsymbol{x}, t) \ d\boldsymbol{x} \ dt + \mathcal{P}^{(m)} + \mathcal{Z}^{(m)} = -\mathcal{R}_k^{(m)} - \mathcal{R}_p^{(m)}, \tag{60}$$

with

$$\mathcal{C}^{(m)}_{\rm KIN}(U^{(m)})|_{D_\sigma} = \sum_{i=1}^{2} \mathcal{C}^{(m,i)}_{\rm KIN}(U^{(m)})|_{D_\sigma} \text{ with } \mathcal{C}^{(m,i)}_{\rm KIN}(U^{(m)})|_{D_\sigma} = (\eth_t E_{k,i})^{n+1}_\sigma + \sum_{\epsilon\in\widetilde{\mathcal{E}}(D_\sigma)} |\epsilon| \frac{(u^n_{i,\epsilon})^2}{2} \boldsymbol{F}^n_\epsilon \cdot \boldsymbol{n}_{\sigma,\epsilon},$$

$$\text{and } (E_{k,i})^n_\sigma = \frac{1}{2} h^n_{D_\sigma} (u^n_{i,\sigma})^2,$$

$$\mathcal{C}^{(m)}_{\rm POT}(U^{(m)})|_K = \frac{1}{2}(\eth_t h^2_K)^n + \mathrm{div}_K\Big(\frac{1}{2} g (h^n)^2 \boldsymbol{u}^n\Big),$$

$$\mathcal{P}^{(m)} = \sum_{n=0}^{N_m-1} \delta t^{(m)} \Big[\sum_{\sigma\in\mathcal{E}^{(m)}} |D_\sigma| u^{n+1}_\sigma \eth_\sigma p^{n+1} \varphi^{n+1}_\sigma + \sum_{K\in\mathcal{M}^{(m)}} |K| p^n_K \mathrm{div}_K \boldsymbol{u}^n \varphi^n_K \Big],$$

$$\mathcal{Z}^{(m)} = + \sum_{n=0}^{N_m-1} \delta t^{(m)} \Big[\sum_{\sigma\in\mathcal{E}^{(m)}} |D_\sigma| h^{n+1}_{\sigma,c} u^{n+1}_\sigma \eth_\sigma z \varphi^{n+1}_\sigma + \sum_{K\in\mathcal{M}^{(m)}} g\Big(z_K (\eth_t h_K)^n + g z_K \mathrm{div}_K (h^n \boldsymbol{u}^n)\Big) \varphi^n_K \Big],$$

$$\mathcal{R}^{(m)}_k = \sum_{i=1}^{2} \sum_{n=0}^{N_m-1} \delta t^{(m)} \sum_{\sigma\in\mathcal{E}^{(i,m)}_{\rm int}} \Big[\frac{1}{2} \frac{|D_\sigma|}{\delta t^{(m)}} h^{n+1}_{D_\sigma} (u^{n+1}_\sigma - u^n_\sigma)^2$$

$$+ \sum_{\epsilon\in\widetilde{\mathcal{E}}^{(i)}(D_\sigma)} |\epsilon| \ \boldsymbol{F}^n_\epsilon \cdot \boldsymbol{n}_{\sigma,\epsilon} \Big(-\frac{1}{2}(u^n_{i,\epsilon} - u^n_{i,\sigma})^2 + (u^n_{i,\epsilon} - u^n_{i,\sigma})(u^{n+1}_{i,\sigma} - u^n_{i,\sigma}) \Big) \Big] \varphi^{n+1}_\sigma$$

$$\mathcal{R}^{(m)}_p \geq \sum_{n=0}^{N_m-1} \delta t^{(m)} \sum_{K\in\mathcal{M}^{(m)}} \Big[-\frac{1}{2} g \sum_{\sigma\in\mathcal{E}(K)} |\sigma| \ (h^n_\sigma - h^n_K)^2 \ \boldsymbol{u}^n_\sigma \cdot \boldsymbol{n}_{K,\sigma} + \sum_{\sigma\in\mathcal{E}(K)} |\sigma| g (h^{n+1}_K - h^n_K) \ h^n_\sigma \ \boldsymbol{u}^n_\sigma \cdot \boldsymbol{n}_{K,\sigma} \Big] \varphi_K.$$

Kinetic energy convection term

Let us check that the above defined convection operator $\mathcal{C}^{(m,i)}_{\rm KIN}$ satisfies the hypotheses (80)–(82) of Lax-Wendroff type consistency Lemma A.1 given in the appendix which we apply here with $d = 2$, $\mathcal{P}^{(m)}$ and $\mathfrak{F}^{(m)}$ the i-th dual mesh and its set of edges, $U = (h, \boldsymbol{u})$, $\beta(U) = E_{k,i}(U) = \frac{1}{2} h u_i^2$, for $i = 1, 2$.

Let us start with the assumption (80). For a given function $\psi \in L^1(\Omega)$, and any subset A of Ω we denote by $\langle\psi\rangle_A$ the mean value of ψ on A. By definition of the kinetic energy, we have $(E_{k,i})^0_\sigma = \frac{1}{2} h^0_{D_\sigma} |u^0_{i,\sigma}|^2 = \frac{1}{2}\langle h_0\rangle_{D_\sigma} (|\langle u_{i,0}\rangle_{D_\sigma}|)^2$ and $E_{k,i}(U_0) = E_{k,i}(h_0, \boldsymbol{u}_0) = \frac{1}{2} h_0 u^2_{i,0}$. Therefore, owing to the assumptions (32)-(33) on the functions $h^{(m)}$ and $\boldsymbol{u}^{(m)}$ and to the fact that these sequences converge in L^1

$$\begin{aligned}
\sum_{P\in\mathcal{P}^{(m)}} \int_P |(\beta^{(m)})^0_P - \beta(U_0(\boldsymbol{x}))| d\boldsymbol{x} &= \sum_{\sigma\in\mathcal{E}^{(m)}} \int_{D_\sigma} |(E_{k,i})^0_\sigma - E_{k,i}(h_0, \boldsymbol{u}_0)| d\boldsymbol{x} \\
&= \frac{1}{2} \sum_{\sigma\in\mathcal{E}^{(m)}} |D_\sigma| \Big| \langle h_0\rangle_{D_\sigma} \langle u_{i,0}\rangle^2_{D_\sigma} - \langle h_0 u^2_{i,0}\rangle_{D_\sigma} \Big| \\
&\to 0 \text{ as } m \to +\infty.
\end{aligned}$$

The assumption (80) is thus satisfied.

Let us then note that the assumption (81), which reads

$$\sum_{n=0}^{N_m-1} \sum_{\sigma\in\mathcal{E}^{(m)}} \int_{t_n}^{t_{n+1}} \int_{D_\sigma} |E^{(m)}_{k,i})^n_\sigma - E_{k,i}(U^{(m)}(\boldsymbol{x}, t))| d\boldsymbol{x} \ dt \to 0 \text{ as } m \to +\infty,$$

is satisfied, again thanks to the assumptions (32)-(33) on the functions $h^{(m)}$ and $\boldsymbol{u}^{(m)}$ and to the fact that these sequences converge in L^1.

Let us now turn to the assumption (82), which reads

$$\sum_{n=0}^{N_m-1}\sum_{\sigma\in\mathcal{E}^{(m)}}\int_{t_n}^{t_{n+1}}\frac{\mathrm{diam}(D_\sigma)}{|D_\sigma|}\int_{D_\sigma}\Big|\sum_{\epsilon\in\widetilde{\mathcal{E}}^{(m)}}|\epsilon|\Big((\boldsymbol{F}^{(m)})_\epsilon^n-\boldsymbol{f}(U^m(\boldsymbol{x},t)\Big)\cdot\boldsymbol{n}_{\sigma,\epsilon}\Big|d\boldsymbol{x}\;dt\to 0\text{ as }m\to+\infty,$$

with $(\boldsymbol{F}^{(m)})_\epsilon^n=\frac{1}{2}(u_{i,\epsilon^n})^2\boldsymbol{F}_\epsilon^n$ and $\boldsymbol{f}(U)=\frac{1}{2}h|u|^2u_i$. This assumption is indeed satisfied since $\boldsymbol{F}_\epsilon^n$ is a convex combination of $h_\sigma\boldsymbol{u}_\sigma$ and $h_{\sigma'}\boldsymbol{u}_{\sigma'}$ for $\epsilon=\sigma|\sigma'$, and thanks to the boundedness and convergence assumptions on the sequences $(h^{(m)})_{m\in\mathbb{N}}$ and $(\boldsymbol{u}^{(m)})_{m\in\mathbb{N}}$.

By Lemma A.1, we thus get that

$$\int_0^T\int_\Omega[C_{\mathrm{KIN}}^{(m)}(U^{(m)})]_i\varphi(\boldsymbol{x},t)\;d\boldsymbol{x},\;dt\to-\int_\Omega E_{k,i}(U^{(0)}\varphi(x,0)\;\mathrm{d}\boldsymbol{x}-\int_0^T\int_\Omega E_{k,i}(\bar{U})\partial_t\varphi+\frac{1}{2}E_{k,i}(\bar{U})\bar{u}_i\partial_i\varphi\;\mathrm{d}\boldsymbol{x}\;\mathrm{d}t$$

with $E_{k,i}(\bar{U})=\frac{1}{2}g\bar{h}\bar{u}_i^2$; summing over $i=1,2$, we get that

$$\int_0^T\int_\Omega\mathcal{C}_{\mathrm{KIN}}^{(m)}(U^{(m)})\varphi(\boldsymbol{x},t)\;d\boldsymbol{x}\;dt\to$$
$$-\int_\Omega E_k(U^{(0)}\varphi(x,0)\;\mathrm{d}\boldsymbol{x}-\int_0^T\int_\Omega\Big[E_k(\bar{U})\partial_t\varphi+\frac{1}{2}E_k(\bar{U})\boldsymbol{u}\cdot\nabla\varphi\Big]\;\mathrm{d}\boldsymbol{x}\;\mathrm{d}t\text{ as }m\to+\infty.\quad(61)$$

with $E_k(\bar{U})=\frac{1}{2}g\bar{|}\bar{\boldsymbol{u}}|^2$.

Potential energy convection terms

Let us now check that the above defined convection operator $\mathcal{C}_{\mathrm{POT}}^{(m)}$ satisfies the hypotheses (80)–(82) of Lemma A.1 which we now apply with $d=2$, $\mathcal{P}^{(m)}$ and $\mathfrak{F}^{(m)}$ the primal mesh and its set of edges, $U=(h,\boldsymbol{u})$, $\beta(U)=\frac{1}{2}gh^2$ and $\boldsymbol{f}(U)=\frac{1}{2}gh^2\boldsymbol{u}$.

Indeed,

$$\sum_{K\in\mathcal{M}}\int_K\Big|\langle h(\cdot,0)^2\rangle_K-h(x,0)^2\Big|\;\mathrm{d}\boldsymbol{x}\to 0\text{ as }m\to+\infty,$$

so that the hypothesis (80) is satisfied. Next,

$$\sum_{n=0}^{N_m-1}\int_{t_n}^{t_{n+1}}\sum_{K\in\mathcal{M}}\int_K\Big|(h_K^n)^2-h^2(\boldsymbol{x},t)\Big|\;\mathrm{d}\boldsymbol{x}\;\mathrm{d}t\to 0\text{ as }m\to+\infty,$$

thanks to the boundedness and convergence assumptions on the sequence $(h^{(m)})_{m\in\mathbb{N}}$. so that the hypothesis (81) is satisfied. Finally, the left hand side of (82) reads

$$\begin{aligned}X_F&=\sum_{n=0}^{N_m-1}\int_{t_n}^{t_{n+1}}\sum_{K\in\mathcal{M}}\frac{\mathrm{diam}(K)}{|K|}\int_K\Big|\sum_{\sigma\in\mathcal{E}(K)}|\sigma|\Big(\frac{1}{2}g(h_\sigma^n)^2\boldsymbol{u}_\sigma^n-\frac{1}{2}gh^2(\boldsymbol{x},t)\boldsymbol{u}(\boldsymbol{x},t)\Big)\cdot\boldsymbol{n}_{K,\sigma}\Big|\;\mathrm{d}\boldsymbol{x}\;\mathrm{d}t\\&=\sum_{n=0}^{N_m-1}\int_{t_n}^{t_{n+1}}\sum_{K\in\mathcal{M}}\frac{\mathrm{diam}(K)}{|K|}\Big|\sum_{\sigma\in\mathcal{E}(K)}|\sigma|\int_{D_{K,\sigma}}\Big(\frac{1}{2}g(h_\sigma^n)^2-\frac{1}{2}g(h_K^n)^2\Big)\boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma}\Big|\;\mathrm{d}\boldsymbol{x}\;\mathrm{d}t\\&\to 0\text{ as }m\to+\infty\end{aligned}$$

thanks to the fact that h_σ^n is a convex combination of h_K^n and h_L^n for $\sigma=K|L$, and thanks to the boundedness and convergence assumptions on the sequences $(h^{(m)})_{m\in\mathbb{N}}$ and $(\boldsymbol{u}^{(m)})_{m\in\mathbb{N}}$. Therefore, the assumption (82) is also satisfied.

Hence by Lemma A.1,

$$\int_0^T \int_\Omega \mathcal{C}_{\mathrm{POT}}^{(m)}(U^{(m)})\varphi(\boldsymbol{x},t)\ d\boldsymbol{x}\ dt \to -\frac{1}{2}\int_\Omega gh^2(\boldsymbol{x},0)\varphi(\boldsymbol{x},0)\ \mathrm{d}\boldsymbol{x} - \int_0^T \int_\Omega \Big[E_p(\bar{U})\partial_t\varphi + E_p(\bar{U})\bar{\boldsymbol{u}}\cdot\nabla\varphi\Big]\ \mathrm{d}\boldsymbol{x}\ \mathrm{d}t, \tag{62}$$

with $E_p(\bar{U}) = \frac{1}{2}g\bar{h}^2$.

Pressure terms

Let us rewrite $\mathcal{P}^{(m)}$ as

$$\mathcal{P}^{(m)} = \sum_{n=0}^{N_m-1} \delta t^{(m)}(A^{n+1} + B^{n+1}) \ - \delta t^{(m)} B^0,$$
$$\text{with } A^n = \sum_{\substack{\sigma\in\mathcal{E}_{\mathrm{int}}^{(m)}\\ \sigma=K|L}} |D_\sigma| u_\sigma^n \eth_\sigma p^n \varphi_\sigma^n, \text{ and } B^n = \sum_{K\in\mathcal{M}^{(m)}} |K| p_K^n \mathrm{div}_K(\boldsymbol{u}^n)\varphi_K^n.$$

By Lemma 5.2 below,

$$A^{n+1} + B^{n+1} = \sum_{K\in\mathcal{M}^{(m)}} \sum_{\sigma\in\mathcal{E}(K)} |D_{K,\sigma}| p_K^{n+1} \boldsymbol{u}_\sigma^{n+1} \cdot \frac{|\sigma|(\varphi_K^{n+1} - \varphi_\sigma^{n+1})}{|D_{K,\sigma}|}\ \boldsymbol{n}_{K,\sigma}$$

On each subcell $D_{K,\sigma}$ the quantity $\dfrac{|\sigma|(\varphi_K^{n+1} - \varphi_\sigma^{n+1})}{|D_{K,\sigma}|}\ \boldsymbol{n}_{K,\sigma}$ is, up to higher order terms, a discrete differential quotient of φ between $\boldsymbol{x}_K$ and $\boldsymbol{x}_\sigma$, in the direction i if $\sigma \in \mathcal{E}^{(i)}$, which uniformly converges to $\partial_i\varphi \boldsymbol{e}_i$ in the case of a rectangular grid, and therefore,

$$\sum_{n=0}^{N_m-1} \delta t^{(m)}(A^{n+1} + B^{n+1}) \to -\int_0^T \int_\Omega \bar{p}(\boldsymbol{x},t)\ \bar{\boldsymbol{u}}(\boldsymbol{x},t)\cdot\nabla\varphi(\boldsymbol{x},t)d\boldsymbol{x}\ dt \text{ as } m \to +\infty.$$

Now, since we assume $u_0 \in W^{1,1}(\Omega)$,

$$\begin{aligned}
\delta t^{(m)}|B^0| &= \delta t^{(m)}\Big|\sum_{K\in\mathcal{M}^{(m)}} |K| p_K^0 \mathrm{div}_K(\boldsymbol{u}^0)\varphi_K^0\Big| \\
&\le g\delta t^{(m)}\|h_0\|_\infty^2\|\varphi\|_\infty \sum_{K\in\mathcal{M}^{(m)}} |K||\mathrm{div}_K(\boldsymbol{u}^0)| \\
&\le 2g\frac{\delta t^{(m)}}{\inf_{K\in\mathcal{M}^{(m)}}\mathrm{diam}(K)}\|h_0\|_\infty^2\|\varphi\|_\infty \sum_{\sigma\in\mathcal{E}(K)} |\sigma| d_\sigma \|\boldsymbol{u}_0\|_\infty,
\end{aligned}$$

so that, by the assumption (56b), $\displaystyle\sum_{n=0}^{N_m-1} \delta t^{(m)} B^0 \to 0$ as $m \to +\infty$; and therefore,

$$\mathcal{P}^{(m)} \to -\int_0^T \int_\Omega \bar{p}(\boldsymbol{x},t)\ \bar{\boldsymbol{u}}(\boldsymbol{x},t)\cdot\nabla\varphi(\boldsymbol{x},t)d\boldsymbol{x}\ dt \text{ as } m \to +\infty. \tag{63}$$

In the above bound, we used the assumption (56b); this could be avoided if we assume $u_0 \in W^{1,1}(\Omega)$)or $u_0 \in L^1(0,T;BV(\Omega))$); indeed, in this case we have

$$|B^0| \le g\|h_0\|_\infty^2\|\varphi\|_\infty\|\boldsymbol{u}_0\|_{W^{1,1}(\Omega)}.$$

However, the assumption (56) seems unavoidable to deal with the remainder term appearing in the discrete potential energy, see below.

Bathymetry terms

Let us introduce the following piecewise constant functions:

- $\widetilde{h}^{(m)}$ is the piecewise constant function equal to $h_{\sigma,c}^{n+1} = \frac{1}{2}(h_K^{n+1} + h_L^{n+1})$ on each set $D_\sigma \times (t_n, t_{n+1})$, for $\sigma = K|L \in \mathcal{E}_{\rm int}^{(m)}$ and $n \in\in [\![0, N_m - 1]\!]$;
- $\nabla^{(m)} z^{(m)}$ is the piecewise constant function equal to $\dfrac{|\sigma|}{|D_\sigma|}(z_L - z_K)$ on each set D_σ, for $\sigma = K|L \in \mathcal{E}_{\rm int}^{(m)}$;
- $\widetilde{\varphi}^{(m)}$ is the piecewise constant function equal to φ_σ on each set on each set $D_\sigma \times (t_n, t_{n+1})$, for $\sigma = K|L \in \mathcal{E}_{\rm int}^{(m)}$ and $n \in\in [\![0, N_m - 1]\!]$;

With these notations, we get that

$$\sum_{n=0}^{N_m-1} \delta t^{(m)} \sum_{\sigma\in\mathcal{E}^{(m)}} |D_\sigma| h_{\sigma,c}^{n+1} u_\sigma^{n+1} \eth_\sigma z \varphi_\sigma^{n+1} = \int_\Omega \widetilde{h}^{(m)}(\boldsymbol{x},t)\boldsymbol{u}^{(m)}(\boldsymbol{x},t)\cdot\nabla z^{(m)}(\boldsymbol{x})\ \widetilde{\varphi}^{(m)}(\boldsymbol{x},t)\ \mathrm{d}\boldsymbol{x}\ \mathrm{d}t \text{ as } n\to+\infty$$
$$\to \int_\Omega h(\boldsymbol{x},t)\boldsymbol{u}(\boldsymbol{x},t)\cdot\nabla z(\boldsymbol{x})\ \varphi(\boldsymbol{x},t)\ \mathrm{d}\boldsymbol{x}\ \mathrm{d}t \text{ as } m\to+\infty, \tag{64}$$

thanks to the convergence assumptions on $h^{(m)}$ and $\boldsymbol{u}^{(m)}$ and owing to the strong convergence of the discrete gradient $\nabla^{(m)}$ (which would be only a weak convergence in the case of a non rectangular mesh, see [12, Lemma 3.1]).

Now let $T_K^n = g\ \eth_t h_K^{n+1}\ z_K$ and $Z_K^n = \dfrac{1}{|K|} g \sum_{\sigma\in\mathcal{E}(K)} |\sigma| h_\sigma^n u_{K,\sigma}^n z_K$. Using a discrete summation by parts in time and thanks to the convergence assumption on $h^{(m)}$, we get that

$$\sum_{n=0}^{N_m-1} \delta t^{(m)} \sum_{K\in\mathcal{M}^{(m)}} |K| T_K^n \to -\int_\Omega g z(\boldsymbol{x}) h(\boldsymbol{x},0)\varphi(\boldsymbol{x},0)\ \mathrm{d}\boldsymbol{x} - \int_0^T\int_\Omega g z(\boldsymbol{x}) h(\boldsymbol{x},t)\partial_t\varphi(\boldsymbol{x},t)\ \mathrm{d}\boldsymbol{x}\ \mathrm{d}t. \tag{65}$$

Using next a discrete summation by parts in space, we get

$$\begin{aligned}\sum_{K\in\mathcal{M}^{(m)}} |K| Z_K^n &= \sum_{K\in\mathcal{M}^{(m)}} g z_K \varphi_K^{n+1} \sum_{\sigma\in\mathcal{E}(K)} |\sigma| h_\sigma^n u_{K,\sigma}^n z_K \\ &= \sum_{\substack{\sigma\in\mathcal{E}_{\rm int}^{(m)}\\ \sigma=K|L}} |\sigma| h_\sigma^n u_{K,\sigma}^n (z_K\varphi_K^{n+1} - z_L\varphi_L^{n+1}) \\ &= -\sum_{\substack{\sigma\in\mathcal{E}_{\rm int}^{(m)}\\ \sigma=K|L}} |D_\sigma| h_\sigma^n \boldsymbol{u}_\sigma^n \cdot (\nabla^{(m)}(z\varphi))_\sigma^{n+1},\end{aligned}$$

where $\nabla^{(m)}(z\varphi)$ is the piecewise constant discrete gradient defined by:

$$\forall \sigma = K|L \in \mathcal{E}_{\rm int}^{(m)},\ \forall n \in [\![0, N_m-1]\!],\ \forall (\boldsymbol{x},t)\in D_\sigma\times[t_n,t_{n+1}),$$
$$\nabla^{(m)}(z\varphi)^{n+1}(\boldsymbol{x},t) = (\nabla^{(m)}(z\varphi))_\sigma =^{n+1} \frac{|\sigma|}{|D_\sigma|}(z_K\varphi_K^{n+1} - z_L\varphi_L^{n+1})\boldsymbol{n}_{K,\sigma},$$

which converges to $\nabla(z\varphi)$ uniformly in the case of a rectangular mesh, and weakly in the case of a general mesh, see [12, Lemma 3.1].

Therefore, thanks to the convergence assumptions on h and $\boldsymbol{u}$,

$$\sum_{n=0}^{N_m-1} \delta t^{(m)} \sum_{K\in\mathcal{M}^{(m)}} |K| Z_K^n \to -\int_0^T\int_\Omega g\bar{h}(\boldsymbol{x},t)\bar{\boldsymbol{u}}(\boldsymbol{x},t)\cdot\nabla(z\varphi)(\boldsymbol{x},t)\ \mathrm{d}\boldsymbol{x}\ \mathrm{d}t \text{ as } m\to+\infty. \tag{66}$$

Owing to (64), (65) and (66), we thus get that

$$\mathcal{Z}^{(m)} \to -\int_\Omega gz(\boldsymbol{x})h(\boldsymbol{x},0)\varphi(\boldsymbol{x},0)\,\mathrm{d}\boldsymbol{x} - \int_0^T\int_\Omega gz(\boldsymbol{x})\bar{h}(\boldsymbol{x},t)\partial_t\varphi(\boldsymbol{x},t)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \\ - \int_0^T\int_\Omega g\bar{h}(\boldsymbol{x},t)z(\boldsymbol{x})\bar{\boldsymbol{u}}(\boldsymbol{x},t)\cdot\nabla\varphi(\boldsymbol{x},t)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \text{ as } m\to+\infty. \quad (67)$$

Remainder terms
The remainder term $\mathcal{R}_k^{(m)}$ in (60) satisfies

$$\mathcal{R}_k^{(m)} = \mathcal{R}_{k,1}^{(m)} + \mathcal{R}_{k,2}^{(m)} + \mathcal{R}_{k,3}^{(m)} \quad (68)$$

with

$$\mathcal{R}_{k,1}^{(m)} = \frac{1}{2}\sum_{i=1}^{2}\sum_{n=0}^{N_m-1}\delta t^{(m)}\sum_{\sigma\in\mathcal{E}_{\mathrm{int}}^{(i)}}\frac{1}{\delta t}\,|D_\sigma|\,h_{D_\sigma}^{n+1}\,\left(u_{i,\sigma}^{n+1}-u_{i,\sigma}^{n}\right)^2\varphi_\sigma^{n+1},$$

$$\mathcal{R}_{k,2}^{(m)} = -\frac{1}{2}\sum_{i=1}^{2}\sum_{n=0}^{N_m-1}\delta t^{(m)}\sum_{\sigma\in\mathcal{E}_{\mathrm{int}}^{(i)}}\sum_{\epsilon\in\widetilde{\mathcal{E}}^{(i)}(D_\sigma)}|\epsilon|\,\boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}\,\left(u_{i,\epsilon}^{n}-u_{i,\sigma}^{n}\right)^2\varphi_\sigma^{n+1}$$

$$\mathcal{R}_{k,3}^{(m)} = \sum_{i=1}^{2}\sum_{n=0}^{N_m-1}\delta t^{(m)}\sum_{\sigma\in\mathcal{E}_{\mathrm{int}}^{(i)}}\sum_{\epsilon\in\widetilde{\mathcal{E}}^{(i)}(D_\sigma)}|\epsilon|\,\boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}\big(u_{i,\epsilon}^{n}-u_{i,\sigma}^{n}\big)\big(u_{i,\sigma}^{n+1}-u_{i,\sigma}^{n}\big)\varphi_\sigma^{n+1}.$$

The term $\mathcal{R}_{k,3}^{(m)}$ satisfies

$$\mathcal{R}_{k,3}^{(m)} \geq \mathcal{R}_{k,3,1}^{(m)} + \mathcal{R}_{k,3,2}^{(m)} \quad (69)$$

with

$$\mathcal{R}_{k,3,1}^{(m)} = -\frac{1}{2}\sum_{i=1}^{2}\sum_{n=0}^{N_m-1}\delta t^{(m)}\sum_{\sigma\in\mathcal{E}_{\mathrm{int}}^{(i)}}\sum_{\substack{\epsilon\in\widetilde{\mathcal{E}}^{(i)}(D_\sigma)\\ \boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}>0}}|\epsilon|\,\boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}\big(u_{i,\sigma}^{n+1}-u_{i,\sigma}^{n}\big)^2\varphi_\sigma^{n+1},$$

$$\mathcal{R}_{k,3,2}^{(m)} = -\frac{1}{2}\sum_{i=1}^{2}\sum_{n=0}^{N_m-1}\delta t^{(m)}\sum_{\sigma\in\mathcal{E}_{\mathrm{int}}^{(i)}}\sum_{\epsilon\in\widetilde{\mathcal{E}}^{(i)}(D_\sigma)}|\epsilon|\,\boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}\big(u_{i,\epsilon}^{n}-u_{i,\sigma}^{n}\big)^2\varphi_\sigma^{n+1}.$$

Thanks to the CFL condition (55), we get that

$$\mathcal{R}_{k,1}^{(m)} + \mathcal{R}_{k,3,1}^{(m)} \geq 0. \quad (70)$$

Let us now study the term

$$\widetilde{\mathcal{R}}_{k,2}^{(m)} = \mathcal{R}_{k,3,2}^{(m)} + \mathcal{R}_{k,2}^{(m)} = -\sum_{i=1}^{2}\sum_{n=0}^{N_m-1}\delta t^{(m)}\sum_{\sigma\in\mathcal{E}_{\mathrm{int}}^{(i)}}\sum_{\epsilon\in\widetilde{\mathcal{E}}^{(i)}(D_\sigma)}|\epsilon|\,\boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}\,\left(u_{i,\epsilon}^{n}-u_{i,\sigma}^{n}\right)^2\varphi_\sigma^{n+1},$$

which we decompose as: $\widetilde{\mathcal{R}}_{k,2}^{(m)} \geq \widetilde{\mathcal{R}}_{k,2,1}^{(m)} + \widetilde{\mathcal{R}}_{k,2,2}^{(m)}$, with

$$\widetilde{\mathcal{R}}_{k,2,1}^{(m)} = -\sum_{i=1}^{2}\sum_{n=0}^{N_m-1} \delta t^{(m)} \sum_{\sigma\in\mathcal{E}_{\rm int}^{(i)}} \sum_{\substack{\epsilon\in\widetilde{\mathcal{E}}^{(i)}(D_\sigma)\\ \boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}>0}} |\epsilon|\ \boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}\ \big(u_{i,\epsilon}^n - u_{i,\sigma}^n\big)^2\varphi_\epsilon^{n+1},$$

$$\widetilde{\mathcal{R}}_{k,2,2}^{(m)} = -\sum_{i=1}^{2}\sum_{n=0}^{N_m-1} \delta t^{(m)} \sum_{\sigma\in\mathcal{E}_{\rm int}^{(i)}} \sum_{\epsilon\in\widetilde{\mathcal{E}}^{(i)}(D_\sigma)} |\epsilon|\ \boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}\ \big(u_{i,\epsilon}^n - u_{i,\sigma}^n\big)^2(\varphi_\epsilon^{n+1} - \varphi_\sigma^{n+1}),$$

and, by conservativity,

$$\begin{aligned}\widetilde{\mathcal{R}}_{k,2,1}^{(m)} &\geq \sum_{i=1}^{2}\sum_{n=0}^{N_m-1} \delta t^{(m)} \sum_{\substack{\epsilon=\sigma|\sigma'\in\widetilde{\mathcal{E}}_{\rm int}^{(i)}\\ \boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}>0}} |\epsilon|\ \boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}\ \Big[\big(u_{i,\epsilon}^n - u_{i,\sigma}^n\big)^2 - \big(u_{i,\epsilon}^n - u_{i,\sigma'}^n\big)^2\Big]\varphi_\epsilon^{n+1}\\ &\geq \sum_{i=1}^{2}\sum_{n=0}^{N_m-1} \delta t^{(m)}\|u_i\|_\infty \sum_{\substack{\epsilon=\sigma|\sigma'\in\widetilde{\mathcal{E}}_{\rm int}^{(i)}\\ \boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}>0}} |D_\epsilon|(2u_{i,\epsilon} - u_{i,\sigma} - u_{i,\sigma'})(u_{i,\sigma'} - u_{i,\sigma'}).\end{aligned}$$

Therefore, thanks to (19) and (58),

$$\widetilde{\mathcal{R}}_{k,2,1}^{(m)} \geq \sum_{i=1}^{2}\sum_{n=0}^{N_m-1} \delta t^{(m)} \sum_{\substack{\epsilon=\sigma|\sigma'\in\widetilde{\mathcal{E}}_{\rm int}^{(i)}\\ \boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}>0}} |\epsilon|\ \boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}(2\mu_{\sigma,\epsilon}-1)\big(u_{i,\sigma}^n - u_{i,\sigma'}^n\big)^2\varphi_\epsilon^{n+1} \geq 0. \tag{71}$$

Let us then write that, thanks to the regularity of φ,

$$\begin{aligned}|\widetilde{\mathcal{R}}_{k,2,2}^{(m)}| &\leq C_\varphi \sum_{i=1}^{2}\sum_{n=0}^{N_m-1} \delta t^{(m)} \sum_{\sigma\in\mathcal{E}_{\rm int}^{(i)}} \sum_{\epsilon\in\widetilde{\mathcal{E}}^{(i)}(D_\sigma)} |D_\epsilon|\ |\boldsymbol{F}_\epsilon^n\cdot\boldsymbol{n}_{\sigma,\epsilon}|\ \big(u_{i,\epsilon}^n - u_{i,\sigma}^n\big)^2\\ &\leq C_\varphi\|h\|_\infty\|\boldsymbol{u}\|_\infty \sum_{i=1}^{2}\sum_{n=0}^{N_m-1} \delta t^{(m)} \sum_{\sigma\in\mathcal{E}_{\rm int}^{(i)}} \sum_{\epsilon\in\widetilde{\mathcal{E}}^{(i)}(D_\sigma)} |D_\epsilon|\ u_{i,\epsilon}^n - u_{i,\sigma}^n|\end{aligned}$$

so that, thanks to the L^1 convergence of $\boldsymbol{u}^{(m)}$ and to the regularity of the mesh, we may again apply Lemma B.2 to obtain

$$|\widetilde{\mathcal{R}}_{k,2,2}^{(m)}| \to 0 \text{ as } m \to +\infty \tag{72}$$

Thefore, owing to (68)-(72)

$$\lim_{m\to+\infty} \mathcal{R}_k^{(m)} \geq 0. \tag{73}$$

Let us now turn to the remainder $\mathcal{R}_p^{(m)}$. We have $\mathcal{R}_p^{(m)} \geq \mathcal{R}_{p,1}^{(m)} + \mathcal{R}_{p,2}^{(m)}$, with

$$\mathcal{R}_{p,1}^{(m)} = -\sum_{n=0}^{N_m-1} \frac{\delta t^{(m)}}{2} \sum_{K\in\mathcal{M}^{(m)}} g \sum_{\sigma\in\mathcal{E}(K)} |\sigma|\ (h_\sigma^n - h_K^n)^2\ \boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma}\varphi_K^n,$$

$$\mathcal{R}_{p,2}^{(m)} = \sum_{n=0}^{N_m-1} \delta t^{(m)} \sum_{K\in\mathcal{M}^{(m)}} \sum_{\sigma\in\mathcal{E}(K)} |\sigma| g(h_K^{n+1} - h_K^n)\ h_\sigma^n\ \boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma}\varphi_K^n.$$

Note that if h_σ^n is the upwind choice for any $\sigma \in \mathcal{E}^{(m)}$, then $\mathcal{R}_{p,1}^{(m)} \geq 0$. In the general case, we may write that

$$\mathcal{R}_{p,1}^{(m)} = \mathcal{R}_{p,1,1}^{(m)} + \mathcal{R}_{p,1,2}^{(m)}$$

with

$$\mathcal{R}_{p,1,1}^{(m)} = -\sum_{n=0}^{N_m-1} \frac{\delta t^{(m)}}{2} \sum_{K\in\mathcal{M}^{(m)}} g \sum_{\sigma\in\mathcal{E}(K)} |\sigma|\ (h_\sigma^n - h_K^n)^2\ \boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{K,\sigma}\varphi_\sigma^n$$

$$\mathcal{R}_{p,1,2}^{(m)} = -\sum_{n=0}^{N_m-1} \frac{\delta t^{(m)}}{2} \sum_{K\in\mathcal{M}^{(m)}} g \sum_{\sigma\in\mathcal{E}(K)} |\sigma|\ (h_\sigma^n - h_K^n)^2\ \boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{K,\sigma}(\varphi_K^n - \varphi_\sigma^n)$$

By conservativity,

$$\begin{aligned}\mathcal{R}_{p,1,1}^{(m)} &\geq -g \sum_{n=0}^{N_m-1} \frac{\delta t^{(m)}}{2} \sum_{\substack{\sigma=K|L\in\mathcal{E}\\ \boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma}>0}} |\sigma|\ \left[(h_\sigma^n - h_K^n)^2 - (h_\sigma^n - h_L^n)^2\right]\ \boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{K,\sigma}\varphi_\sigma^n \\ &= -g \sum_{n=0}^{N_m-1} \frac{\delta t^{(m)}}{2} \sum_{\substack{\sigma=K|L\in\mathcal{E}\\ \boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma}>0}} |\sigma|\ (h_L^n - h_K^n))(2h_\sigma^n - h_K^n - h_L^n)\ \boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{K,\sigma}\varphi_\sigma^n\end{aligned}$$

Owing to the assumption (9), one has

$$(h_L^n - h_K^n))(2h_\sigma^n - h_K^n - h_L^n) = -2\lambda_{K,\sigma}(h_K^n - h_L^n)$$

and since by (57), $\lambda_{K,\sigma} \geq \frac{1}{2}$ if $\boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{K,\sigma} > 0$,

$$\mathcal{R}_{p,1,1}^{(m)} = 2g \sum_{n=0}^{N_m-1} \delta t^{(m)} \sum_{\substack{\sigma=K|L\in\mathcal{E}\\ \boldsymbol{u}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma}>0}} |\sigma|(\lambda_{K,\sigma} - \frac{1}{2})\ (h_L^n - h_K^n)^2\ \boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{K,\sigma}\varphi_\sigma^n \geq 0.$$

Now

$$|\mathcal{R}_{p,1,2}^{(m)}| \leq \sum_{n=0}^{N_m-1} \frac{\delta t^{(m)}}{2} \sum_{K\in\mathcal{M}^{(m)}} g\|h\|_\infty \|\boldsymbol{u}\|_\infty C_\varphi \sum_{\sigma\in\mathcal{E}(K)} |D_{K,\sigma}|\ |h_L^n - h_K^n| \to 0 \text{ as } m \to +\infty$$

so that

$$\lim_{m\to+\infty} \mathcal{R}_{p,1}^{(m)} \geq 0.$$

Let us now turn to $\mathcal{R}_{p,2}^{(m)}$. Since for all $K \in \mathcal{M}$ and $\sigma \in \mathcal{E}(K)$ we have $|\sigma| \leq \dfrac{|K|}{\inf_{K\in\mathcal{M}} \operatorname{diam}(K)}$, we have

$$\begin{aligned}|\mathcal{R}_{p,2}^{(m)}| &\leq g\|h\|_\infty\|\boldsymbol{u}\|_\infty\|\varphi\|_\infty \sum_{n=0}^{N_m-1} \frac{\delta t^{(m)}}{\inf_{K\in\mathcal{M}^{(m)}} \operatorname{diam}(K)} \sum_{n=0}^{N_m-1} \sum_{K\in\mathcal{M}^{(m)}} |K||(h^{(m)})_K^{n+1} - (h^{(m)})_K^n| \\ &\to 0 \text{ as } m \to +\infty,\end{aligned}$$

thanks to the assumption (56). Hence

$$\lim_{m\to+\infty} \mathcal{R}_p^{(m)} \geq 0. \tag{74}$$

Conclusion of the proof

Owing to (73) and (74), passing to the limit in (60) as $m \to +\infty$ yields, together with (61), (62), (63) and (67), that the limit $(\bar{h}, \bar{u})$ satisfies the weak entropy inequality (31). □

The next lemma, used to pass to the limit in the pressure terms of the entropy is the discrete equivalent, on a staggered grid, of the formal equality $\displaystyle\int_\Omega (\boldsymbol{u}\cdot\nabla p\ \varphi + p\ \mathrm{div}\boldsymbol{u}\ \varphi)\ \mathrm{d}\boldsymbol{x} = -\int_\Omega p\ \boldsymbol{u}\cdot\nabla\varphi\ d\boldsymbol{x}$.

Lemma 5.2 (Pressure terms)**.** *Let $(\mathcal{M},\mathcal{E})$ be a MAC discretization of Ω in the sense of Definition 2.1 ; Let $(p_K)_{K\in\mathcal{M}}\subset\mathbb{R}$ and $(\boldsymbol{u}_\sigma)_{\sigma\in\mathcal{E}}\subset\mathbb{R}^d$ be some discrete unknowns associated to $\mathcal{M}$ and $\mathcal{E}$ respectively. Let $\varphi\in C_c^\infty(\Omega)$, and let φ_K (resp. φ_σ) denote the mean value of φ on K (resp. D_σ), for any $K\in\mathcal{M}$ (resp. $\sigma\in\mathcal{E}^{(m)}$). Then*

$$\sum_{\substack{\sigma\in\mathcal{E}_{\rm int}\\ \sigma=K|L}} |D_\sigma|\ u_\sigma\ \eth_\sigma p\ \varphi_\sigma + \sum_{K\in\mathcal{M}} |K|\ p_K\ \mathrm{div}_K\boldsymbol{u}\ \varphi_K = \sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}(K)} |D_{K,\sigma}|\ p_K\ \boldsymbol{u}_\sigma\cdot\frac{|\sigma|(\varphi_K-\varphi_\sigma)}{|D_{K,\sigma}|}\boldsymbol{n}_{K,\sigma}.$$

Proof. Let us denote by A and B the first and second terms of the right hand side. Then, with the notations of Definition 2.1,

$$\begin{aligned} A = \sum_{K\in\mathcal{M}^{(m)}}\sum_{\sigma\in\mathcal{E}(K)} |D_{K,\sigma}| u_\sigma\ \eth_\sigma p\ \varphi_\sigma) &= \sum_{K\in\mathcal{M}^{(m)}}\sum_{\substack{\sigma\in\mathcal{E}(K)\\ \sigma=K|L}} |D_{K,\sigma}|\ u_{K,\sigma}\frac{p_L-p_K}{|D_\sigma|}|\sigma|\ \varphi_\sigma \\ &= \sum_{K\in\mathcal{M}^{(m)}}\sum_{\substack{\sigma\in\mathcal{E}(K)\\ \sigma=K|L}} |D_{K,\sigma}|\ u_{K,\sigma}\frac{p_\sigma-p_K}{|D_{K,\sigma}|}|\sigma|\ \varphi_\sigma, \end{aligned}$$

where p_σ is defined by $\dfrac{p_\sigma-p_K}{|D_{K,\sigma}|}=\dfrac{p_L-p_K}{|D_\sigma|}$. By conservativity, $\displaystyle\sum_{K\in\mathcal{M}^{(m)}}\sum_{\substack{\sigma\in\mathcal{E}(K)\\ \sigma=K|L}} u_{K,\sigma}p_\sigma|\sigma|\varphi_\sigma=0$, so that

$$A = -\sum_{K\in\mathcal{M}^{(m)}}\sum_{\substack{\sigma\in\mathcal{E}(K)\\ \sigma=K|L}} |D_{K,\sigma}|\ u_{K,\sigma}\frac{p_K}{|D_{K,\sigma}|}|\sigma|\ \varphi_\sigma$$

Now

$$B = \sum_{K\in\mathcal{M}^{(m)}} |K|\ p_K\ \mathrm{div}_K\boldsymbol{u}\ \varphi_K = \sum_{K\in\mathcal{M}^{(m)}}\sum_{\sigma\in\mathcal{E}(K)} |\sigma|\ p_K\ |D_{K,\sigma}|\ u_\sigma\ \varphi_K.$$

Adding the results for A and B concludes the proof. □

6. Numerical results

This section is devoted to numerical tests: we first check the order of convergence of the proposed scheme on a two-dimensional regular solution (Section 6.1); then we turn to one-dimensional and two-dimensional shock solutions on a plane topography (Sections 6.2 and 6.3); in Section 6.4, we address a two-dimensional dam-break problem in a closed computational domain with a variable topography, which, in particular, shows tha ability of staggered scheme to "natively" cope with reflection boundary conditions; finally, we compute the motion of a liquid slug over a partly dry support (6.5).

In this section, we compare three schemes: the second-order scheme developed here, the scheme referred to in Section 2.2 as the segregated forward Euler scheme (combining a segregated forward Euler scheme in time and the proposed MUSCL-like discretization of the convection fluxes) and a first order scheme which still features the segregated forward Euler scheme in time but with first-order upwind convection fluxes. These schemes are referred to in the following as the *second-order*, *segregated* and *first-order* scheme respectively.

The schemes have been implemented within the CALIF^{3}S open-source software [7] of the French Institut de Sûreté et de Radioprotection Nucléaire (IRSN); this software is used for the following tests.

6.1. **A smooth solution**

We begin here by checking the accuracy of the scheme on a known regular solution consisting in a travelling vortex. This solution is obtained through the following steps: we first derive a compact-support H^2 solution consisting in a standing vortex which becomes time-dependent by adding a constant velocity motion. The velocity field of the standing vortex and the pressure are sought under the form:

$$\hat{\boldsymbol{u}} = f(\xi) \begin{bmatrix} -x_2 \\ x_1 \end{bmatrix}, \quad \hat{p} = \wp(\xi),$$

with $\xi = x_1^2 + x_2^2$. A simple derivation of these expressions yields:

$$\hat{\boldsymbol{u}} \cdot \boldsymbol{\nabla} \hat{\boldsymbol{u}} = -f(\xi)^2 \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

and

$$\boldsymbol{\nabla} \hat{p} = 2\, \wp'(\xi) \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

Using the relation $p = \frac{1}{2} g h^2$, we thus obtain a stationary solution of the shallow water equations (1) with a topography $z = 0$ if $\wp$ satisfies $8\, g\, \wp = (F + c)^2$, where F is such that $F' = f^2$, $F(0) = 0$ and c is a positive real number. For the present numerical study, we choose $f(\xi) = 10\, \xi^2 (1 - \xi)^2$ if $\xi \in (0, 1)$, $f = 0$ otherwise, which indeed yields an $H^2(\mathbb{R}^2)$ velocity field (note that as a consequence, the pressure and the water height are also regular), and $c = 1$. The problem is made unsteady by a time translation: given a constant vector field $\boldsymbol{a}$, the pressure p and the velocity $\boldsymbol{u}$ are deduced from the steady state solution $\hat{p}$ and $\hat{\boldsymbol{u}}$:

$$h(\boldsymbol{x}, t) = \hat{h}(\boldsymbol{x} - \boldsymbol{a}t), \qquad \boldsymbol{u}(\boldsymbol{x}, t) = \hat{\boldsymbol{u}}(\boldsymbol{x} - \boldsymbol{a}t) + \boldsymbol{a}.$$

The center of the vortex is initially located at $\boldsymbol{x}_0 = (0, 0)^t$, the translation velocity $\boldsymbol{a}$ is set to $\boldsymbol{a} = (1, 1)^t$, the computational domain is $\Omega = (-1.2,\, 2.)^2$ and the computation is run on the time interval $(0, 0.8)$.

Computations are performed with successively refined meshes with square cells, and the time step is $\delta t = \delta_{\mathcal{M}}/8$, and corresponds to a Courant (or CFL) number with respect to the celerity of the fastest waves close to $1/3$. The discrete L^1-norm of the difference between the exact solution and the solution obtained by the second-order scheme is given in Table 1. The observed order of convergence over the whole sequence is 2 for the water height and 1.5 for the velocity. Results with the first-order scheme are given in Table 2; one observes that the second-order scheme is much more accurate. Finally, the segregated scheme yields good results on coarse meshes (it is the most accurate scheme on the 32×32 mesh); unfortunately, when refining the mesh, oscillations appear, and the convergence is lost. This results confirms a behaviour already observed for the transport operator in [27]: for multi-dimensional problems, the smoothing produced by the Heun time-stepping seems to be necessary to compensate the oscillatory character of the MUSCL scheme (which, for the transport operator, does not lead, of course, to violate the local maximum principle warranted by construction of the limitation process).

mesh	error(h)	ord(h)	error(u)	ord(u)
32×32	$3.61\,10^{-3}$	/	$2.93\,10^{-1}$	/
64×64	$1.15\,10^{-3}$	1.65	$1.14\,10^{-1}$	1.36
128×128	$2.58\,10^{-4}$	2.16	$4.06\,10^{-2}$	1.49
256×256	$5.85\,10^{-5}$	2.14	$1.49\,10^{-2}$	1.45
512×512	$1.53\,10^{-5}$	1.93	$4.67\,10^{-3}$	1.68

TABLE 1. Measured numerical errors for the travelling vortex – Discrete L^1-norm of the difference between the numerical and exact solution at $t = 0.8$, for the height and the velocity, and corresponding order of convergence.

mesh	error(h)	ord(h)	error(u)	ord(u)
32×32	$8.04\,10^{-3}$	/	$6.55\,10^{-1}$	/
64×64	$5.56\,10^{-3}$	0.53	$4.84\,10^{-1}$	0.44
128×128	$3.53\,10^{-3}$	0.66	$3.22\,10^{-1}$	0.59
256×256	$2.08\,10^{-3}$	0.76	$1.96\,10^{-1}$	0.72
512×512	$1.15\,10^{-3}$	0.85	$1.16\,10^{-1}$	0.76

TABLE 2. Measured numerical errors for the travelling vortex with the first order scheme - Discrete L^1-norm of the difference between the numerical and exact solution at $t = 0.8$, for the height and the velocity, and corresponding order of convergence.

mesh	error(h)	error(u)
32×32	$2.06\,10^{-3}$	$2.33\,10^{-1}$
64×64	$1.37\,10^{-3}$	$1.18\,10^{-1}$
128×128	$1.24\,10^{-3}$	$8.50\,10^{-2}$
256×256	$1.26\,10^{-3}$	$6.16\,10^{-2}$
512×512	$1.56\,10^{-3}$	$4.85\,10^{-2}$

TABLE 3. Measured numerical errors for the travelling vortex with the segregated scheme - Discrete L^1-norm of the difference between the numerical and exact solution at $t = 0.8$, for the height and the velocity.

6.2. **A Riemann problem**

We now turn to a one-dimensional shock solution, corresponding to a Riemann problem posed over $\Omega = (0, 1)$. The initial height is $h = 1$ if $x < 0.5$ and $h = 0.2$ otherwise, and the topography z is set to zero over the computational domain; the fluid is initially at rest. The solution consists in a 1-rarefaction wave and a 2-shock.

We plot on Figure 3 and Figure 4 the results obtained a $t = 0.1$ with the second-order scheme, the segregated scheme and the first-order scheme. The space step is $\delta x = 1/200$ and the time step is chosen as $\delta t = \delta x/10$, which corresponds to a CFL number lower than 0.5 with respect to the waves celerity (the maximal speed of sound is close to 3 and the maximal velocity is close to 2). As expected, the first order scheme is more diffusive than the other ones. As in the previous test, the segregated forward Euler scheme (with MUSCL fluxes) exhibits some oscillations, which are damped by the Heun time discretization (see the Figure 4). In this test case, for both the second-order and the segregated scheme, the shock is captured with only one intermediate cell between the left and the right state.

6.3. **A circular dam break problem**

The objective of this test-case is to check the capability of the scheme to capture a multi-dimensional shock solution. The fluid is initially at rest and the height is given by:

$$h = 2.5 \text{ if } r < 2.5,\ h = 0.5 \text{ otherwise, with } r^2 = x_1^2 + x_2^2.$$

The computational domain is $\Omega = (-20, 20) \times (-20, 20)$ and the final time is $T = 4.7$.

We plot on Figure 5 the results obtained with a 800×800 uniform mesh, with the second-order scheme. The time-step is $\delta t = h_{\mathcal{M}}/10$ (with a maximal velocity in the range of 3.5 and a maximal speed of sound in the range of 5). In addition, to cure some oscillations (see Figure 7), we add a slight stabilization in the momentum balance equation which consists in adding to the dicrete momentum equation associated to an

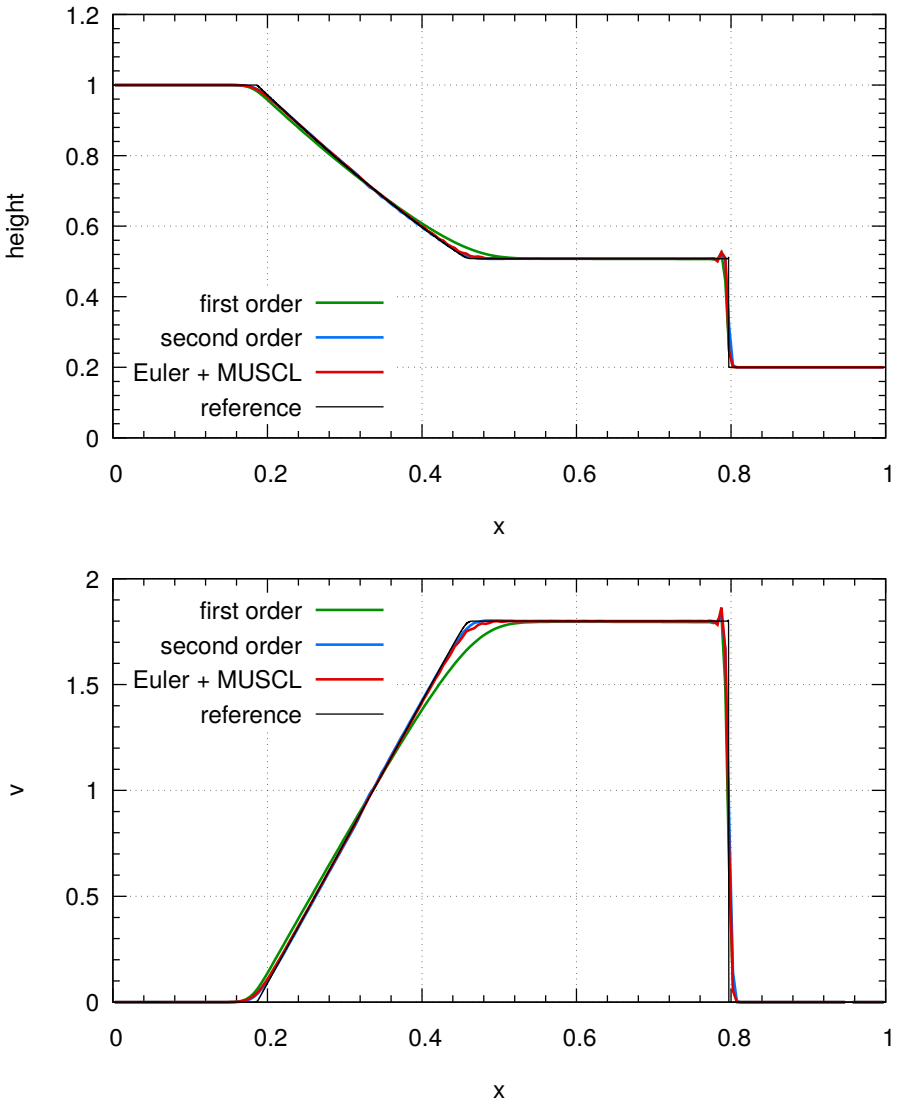

FIGURE 3. Riemann problem. Top: flow height – Bottom: velocity.

edge σ included in a cell K the following flux through a dual edge $\epsilon = D_\sigma | D_{\sigma'}$:

$$F_{\mathrm{stab},\sigma,\epsilon} = \zeta \; h_K \; \mathrm{diam}(K)^{d-1} \; (u_\sigma - u'_\sigma),$$

where ζ is a user-defined parameter. Here, $\zeta = 0.1$, which is significantly lower than the diffusion generated by the use of an upwind scheme in the momentum balance equation; indeed, the upwind scheme may be seen as the centered one complemented by a diffusion taking the same expression as $F_{\mathrm{stab},\sigma,\epsilon}$ with $\zeta \; h_K$ replaced by $|F_{\sigma,\epsilon}|/2$. The interest of this stabilization stems from the fact that the numerical diffusion introduced in the present family of schemes depends on the material velocity (and not on the waves celerity as, for instance, in colocated schemes based on Riemann solvers), and is sometimes too low in the zones where the fluid is almost at rest [22]. Note that, as a counterpart, the scheme does not become overdiffusive for low-Mach number flows. For the same computation, we give on Figure 6 the height and the radial velocity along the axis $x_2 = 0$ (*i.e.* the first component of the velocity) at different times.

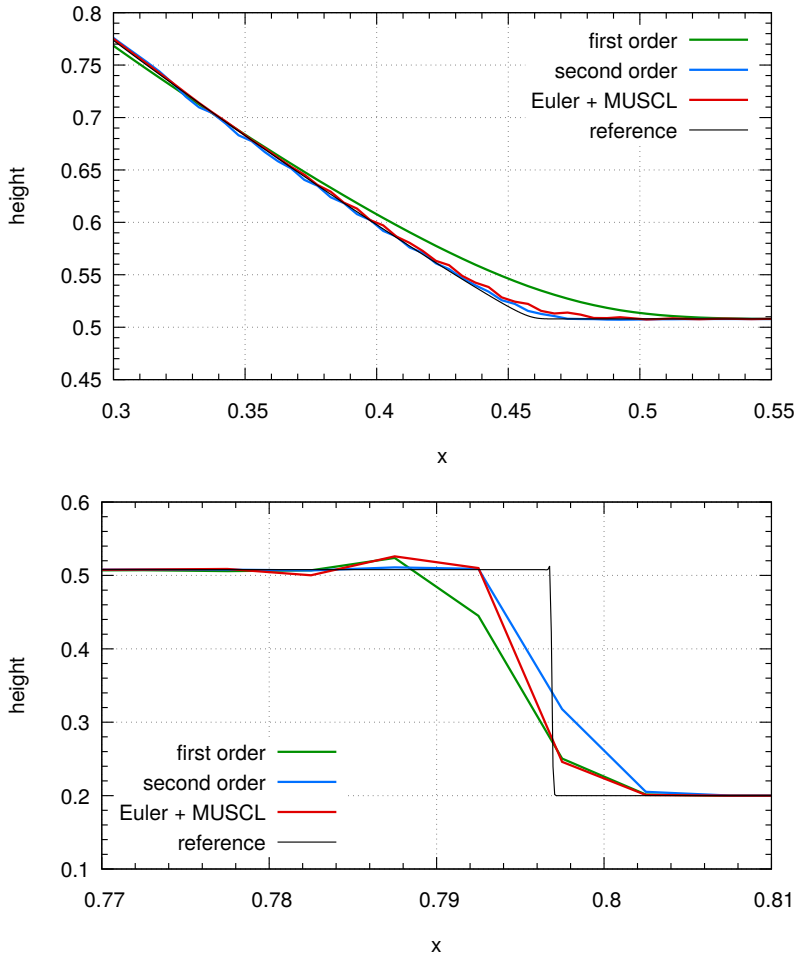

FIGURE 4. Riemann problem. Details of the flow height.

This computation is also used as "reference computation" on Figure 7, where we compare the results obtained at $t = 3T/5$ with a 200×200 mesh with the second-order scheme, the second-order scheme with stabilization and the first-order scheme. This latter is significantly more diffusive, and we observe how the stabilization (even if added to the momentum balance only and not on the mass balance) damps the oscillations obtained with the second-order scheme for both the flow height and the velocity.

6.4. **A so-called partial dam-break problem**

We now turn to a test consisting in a partial dam-break problem with reflection phenomena, and with a non-flat bathymetry. In this test, the computational domain is $\Omega = (0, 200) \times (0, 200) \setminus \Omega_w$ with $\Omega_w = (95, 105) \times (0, 95) \cup (95, 105) \times (170, 200)$. The fluid is supposed to be initially at rest, the initial water height is $h = 10$ for $x_1 \leq 100$ and $h = 5 - 0.04\,(x_1 - 100)$ otherwise, and the bathymetry is $z = 0$ if $x_1 \leq 100$ and $z = 0.04\,(x_1 - 100)$ otherwise. A zero normal velocity is prescribed at all the boundaries of the computational domain. The computation is performed with a mesh obtained from a 1000×1000 regular grid by removing the cells included in Ω_w. The time step is $\delta t = \delta_{\mathrm{M}}/40$ (the maximal speed of sound and the maximal velocity are both close to 10). A stabilization with $\zeta = 0.25$ (so two orders of magnitude lower than the artificial viscosity generated by the upwind scheme in high momentum zones) is added to damp oscillations appearing in the zones at rest, where no numerical diffusion is generated by our schemes. Results obtained at $t = 20$ with the first order in time and space and the present scheme are compared on Figure 8. One can observe that the second-order scheme is clearly less diffusive. In addition, these results illustrate the capacity of the

staggered scheme to deal with reflection conditions by simply imposing the normal velocity to the boundary at zero.

6.5. **Uniform circular motion in a paraboloid**

We address in this section a classical test which admits a closed-form solution and corresponds to the uniform rotation of a drop of liquid on a paraboloid-shaped support. The solution is very regular (at a given time, the velocity field is constant and $h+z$ is affine outside the dry zones), and the essential interest of this test is to check whether the scheme is able to cope with dry zones, *i.e.* zones where the height is zero (in the continuous setting) or very close to zero, as we shall use numerically. The computational domain is $\Omega = (0, L) \times (0, L)$ and the topography is given by

$$z = -\frac{h_0}{a^2}\Big(a^2 - (x - \frac{L}{2})^2 - (y - \frac{L}{2})^2\Big),$$

with h_0 and a parameters which are given below. The height is:

$$h = \max(0, \bar{h}) \text{ with } \bar{h} = \eta\frac{h_0}{a^2}\Big(2\,(x - \frac{L}{2})\,\cos(\omega t) + 2(y - \frac{L}{2})\,\sin(\omega t) - \eta\Big) - z,$$

with η a parameter and ω (the angular rotation velocity of the drop) given by

$$\omega = \frac{(2gh_0)^{1/2}}{a}.$$

Finally, the velocity is

$$\boldsymbol{u} = \eta\,\omega\begin{bmatrix} -\sin(\omega t) \\ \cos(\omega) t \end{bmatrix}.$$

The computation is run up to $T = 6\,\pi/\omega$, so the drop is supposed to perform 3 turns and to lie at the final time at its initial position. The parameters are fixed here to $L = 4$, $h_0 = 0.1$, $a = 1$ and $\eta = 0.5$.

For numerical tests, we bound h from below by 10^{-8}, *i.e.* we set $h = \max(10^{-8}, \bar{h})$, in particular to avoid divisions by zero in the averaging steps of the Heun scheme (Equations (22e) and (22f)). The computation are performed with a uniform 100×100 mesh, with $\delta t = \delta_{\mathcal{M}}/16$, without changing anything to the numerical fluxes to cope with dry zones. This is clearly dangerous, since a non-upwind approximation of the water height at a face separating two cells with a large ratio of water height may lead to a huge outflow mass flux in view of the cell mass inventory (or, in other words, a very large CFL number). This probably explains the rather small time step used here (the CFL number with repect to the celerity of the fastest waves is in the range of 1/8); the first-order scheme, which uses upwind fluxes, works with time steps four times larger. This problem would be probably cured by a more careful limitation of the mass fluxes outward an almost dry cell.

Results obtained with the first order, the segregated and the second order scheme at $t = 6\,\pi/\omega$ are plotted on Figure 9. All schemes give good results, which, for the first-order scheme, is probably due to the regularity of the solution. For the momentum, one observes that the second-order scheme is less accurate than the other ones; this seems to be due to the time-stepping procedure, which perhaps generates some diffusion at the interface between dry and wet zones, especially in the last averaging step, since the segregated scheme is the most accurate one (and superimposed to the exact solution on Figure 9).

Appendix A. Consistency of numerical non linear convection fluxes on staggered meshes

We give here some general lemmas which generalise the Lax-Wendroff theorem to multidimensional staggered meshes, and which we state for any space dimension $d = 1, 2$ or 3. The well-known Lax-Wendroff theorem [25] states that, on uniform 1D grids, a flux-consistent and conservative cell-centered finite-volume scheme for a system of conservation laws is weakly consistent, in the sense that the limit of any a.e. convergent sequence of L^∞-bounded numerical solutions, obtained with a sequence of grids with mesh and time

steps tending to zero, is a weak solution of the conservation law; it is also stated in a different form [26, Section 12.10], with a BV bound assumption on the scheme It is generalised to non uniform 1D or Cartesian meshes in [11, Theorem 21.2]. In a recent work [4], the Lax-Wendroff theorem is extended to obtain some error estimates for higher order schemes on uniform 1D meshes. The case of general (and, in particular, unstructured) discretizations has been also been tackled over the past decades: [24], [15, Section 4.2.2] [10], [12]. In this latter work, the quasi-uniformity assumption that is required in [10] is relaxed, but while in [10] the flux is only required to be continuous, it is supposed to be Lipschitz continuous or at least "lip-diag". In all these works, the scheme is supposed to be colocated, in the sense that the discrete unknowns are associated to the cells of the mesh; these results may not be used directly on staggered meshes, and for instance, in [23], the consistency of an explicit staggered scheme for the full compressible Euler equations is proven recovering the kinetic energy inequality on the primal mesh.

The consistency result that we give here is valid for general polygonal or polyhedral grids with a colocated or staggered arrangement of the unknowns. The main new idea is that in the proof of consistency, rather than using a convergence result for the discrete gradient, which is only weak and demands some regularity on the mesh, we use the actual mean value of the gradient of the test function on each cell, which converges strongly to the gradient, and does not require any regularity of the mesh. As in [12], the proof also relies on the control of some residual terms, involving the difference between the numerical solution and a space or time translate of this latter, and we use the estimate on the translates given [12, Lemma 4.2] to this purpose, which we recall in the appendix B for the sake of completeness.

Let us suppose that:

$$\Omega \subset \mathbb{R}^d,\ d = 1, 2, 3,\ T \in (0, +\infty), \tag{75a}$$

$$p \in \mathbb{N}^*,\ \beta \in C^1(\mathbb{R}^p, \mathbb{R}),\ \boldsymbol{f} \in C^1(\mathbb{R}^p, \mathbb{R}^d),\ U \in L^\infty(\Omega \times (0,T), \mathbb{R}^p), \tag{75b}$$

and consider the conservative convection operator defined (in the distributional sense) by:

$$\begin{aligned} \mathcal{C}(U): \quad & \Omega \times (0,T) \to \mathbb{R}, \\ & (\boldsymbol{x}, t) \mapsto \partial_t(\beta(U))(\boldsymbol{x}, t) + \operatorname{div}(\boldsymbol{f}(U))(\boldsymbol{x}, t). \end{aligned} \tag{76}$$

Lemma A.1 (Weak consistency for a multi-dimensional conservative convection operator)**.** *Under the assumptions* (75)*, let* $(U^{(m)})_{m\in\mathbb{N}} \subset L^\infty(\Omega \times (0,T), \mathbb{R}^p)$ *be a sequence of functions such that:*

$$\exists\ C^u \in \mathbb{R}_+^*\ :\ \|U^{(m)}\|_\infty \le C^u\ \forall m \in \mathbb{N}, \tag{77}$$

$$\exists\ \bar{U} \in L^\infty(\Omega \times (0,T), \mathbb{R}^p) : \|U^{(m)} - \bar{U}\|_{L^1(\Omega\times(0,T),\mathbb{R}^p)} \to 0 \text{ as } m \to +\infty. \tag{78}$$

Let $(\mathcal{P}_m)_{m\in\mathbb{N}}$ *be a sequence of polygonal or polyhedral conforming mesh of* Ω *such that*

$$\delta(\mathcal{P}_m) = \max_{P\in\mathcal{P}_m} \operatorname{diam}(P) \to 0 \text{ as } m \to +\infty.$$

Let $\mathfrak{F}^{(m)}$ *denote the set of faces (or edges) of the mesh, and for a given polyhedron (or polygon)* $P \in \mathcal{P}^{(m)}$*, let* $\mathfrak{F}^{(m)}(P)$ *be the set of faces (or edges) of* P*. For* $m \in \mathbb{N}$*, let* $t_0^{(m)} = 0 < t_1^{(m)} < \ldots < t_{N_m}^{(m)} = T$ *be a discretization of* $(0,T)$ *with* $\delta t^{(m)} = t_{k+1}^{(m)} - t_k^{(m)} \to 0$ *as* $m \to +\infty$*, and consider the discrete convection operator*

$$\begin{aligned} \mathcal{C}^{(m)}(U^{(m)}): \quad & \Omega \times (0,T) \to \mathbb{R}, \\ & (\boldsymbol{x}, t) \mapsto \eth_t(\beta^{(m)})_P^n + \frac{1}{|P|} \sum_{\zeta\in\mathfrak{F}^{(m)}(P)} |\zeta| (\boldsymbol{F}^{(m)})_\zeta^n \cdot \boldsymbol{n}_{P,\zeta} \text{ for } \boldsymbol{x} \in P \text{ and } t \in (t_n, t_{n+1}) \end{aligned} \tag{79}$$

with $\eth_t(\beta^{(m)})_P^n = \frac{1}{\delta t}((\beta^{(m)})_P^{n+1} - (\beta^{(m)})_P^n)$ *and where the families* $\{(\beta^{(m)})_P^n, P \in \mathcal{P}^{(m)},\ n \in [\![0, N_m - 1]\!]\}$ *of real numbers and* $\{(\boldsymbol{F}^{(m)})_\zeta^n, \zeta \in \mathfrak{F}^{(m)},\ n \in [\![0, N_m - 1]\!]\}$ *of real vectors are such that*

$$\sum_{P\in\mathcal{P}^{(m)}} \int_P |(\beta^{(m)})_P^0 - \beta(U_0(\boldsymbol{x}))| d\boldsymbol{x} \to 0 \text{ as } m \to +\infty, \text{ with } U_0 \in L^\infty(\Omega, \mathbb{R}^p), \tag{80}$$

$$\sum_{n=0}^{N_m-1} \sum_{P\in\mathcal{P}^{(m)}} \int_{t_n}^{t_{n+1}} \int_P |(\beta^{(m)})_P^n - \beta(U^{(m)}(\boldsymbol{x},t))| d\boldsymbol{x}\ dt \to 0 \text{ as } m \to +\infty, \tag{81}$$

$$\sum_{n=0}^{N_m-1} \sum_{P\in\mathcal{P}^{(m)}} \int_{t_n}^{t_{n+1}} \frac{\mathrm{diam}(P)}{|P|} \int_P \Big| \sum_{\zeta\in\mathfrak{F}^{(m)}} |\zeta| \Big((\boldsymbol{F}^{(m)})_\zeta^n - \boldsymbol{f}(U^m(\boldsymbol{x},t)\Big) \cdot \boldsymbol{n}_{P,\zeta} \Big| d\boldsymbol{x}\ dt \to 0 \text{ as } m \to +\infty. \tag{82}$$

Let $\varphi \in C_c^\infty(\Omega \times [0,t))$*, then*

$$\begin{aligned}\int_0^T \int_\Omega \mathcal{C}^{(m)}(U^{(m)})(\boldsymbol{x},t)\varphi(\boldsymbol{x},t)\ d\boldsymbol{x}\ dt \to &- \int_\Omega \beta(U_0(\boldsymbol{x}))\varphi(\boldsymbol{x},0)\ d\boldsymbol{x} \\ &- \int_0^T \int_\Omega \Big(\beta(\bar{U})(\boldsymbol{x},t)\partial_t\varphi(\boldsymbol{x},t) + \ \boldsymbol{f}(\bar{U})(\boldsymbol{x},t)\cdot\nabla\varphi(\boldsymbol{x},t)\Big) d\boldsymbol{x}\ dt \text{ as } m \to +\infty. \end{aligned} \tag{83}$$

Proof. The result of this lemma is the consequence of the two following lemmas, which prove respectively the convergence of the time derivative part and the space derivative part. Indeed, let us decompose

$$\int_0^T \int_\Omega \mathcal{C}^{(m)}(U^{(m)})(\boldsymbol{x},t)\varphi(\boldsymbol{x},t)\ d\boldsymbol{x}\ dt = X_1^{(m)} + X_2^{(m)}, \text{ with} \tag{84}$$

$$X_1^{(m)} = \sum_{n=0}^{N_m-1} \delta t \sum_{P\in\mathcal{P}^{(m)}} |P| \eth_t^n \beta_P^{(m)}\ \varphi_P^n \tag{85}$$

$$X_2^{(m)} = \sum_{n=0}^{N_m-1} \delta t \sum_{P\in\mathcal{P}^{(m)}} \sum_{\zeta\in\mathfrak{F}^{(m)}(P)} |\zeta| (\boldsymbol{F}^{(m)})_\zeta^n \cdot \boldsymbol{n}_{P,\zeta} \varphi_P^n \tag{86}$$

where φ_P^n denotes the mean value of φ on $P \times (t_n, t_{n+1})$. Then, by Lemma A.2 below,

$$X_1^{(m)} \to -\int_\Omega \beta(U_0(\boldsymbol{x}))\ d\boldsymbol{x} - \int_0^T \int_\Omega \beta(\bar{U})(\boldsymbol{x},t)\partial_t\varphi(\boldsymbol{x},t) d\boldsymbol{x}\ dt \text{ as } m \to +\infty,$$

and by Lemma A.3 below,

$$X_2^{(m)} \to -\int_0^T \int_\Omega \boldsymbol{f}(\bar{U})(\boldsymbol{x},t)\cdot\nabla\varphi(\boldsymbol{x},t) d\boldsymbol{x}\ dt \text{ as } m \to +\infty,$$

which concludes the proof □

Lemma A.2 (Weak consistency, time derivative)**.** *Under the assumptions and notations of Lemma A.1,*

$$\int_0^T \int_\Omega \eth_t(\beta^{(m)})_P^n \varphi(\boldsymbol{x},t)\ d\boldsymbol{x}\ dt \to -\int_\Omega \beta(U_0(\boldsymbol{x}))\varphi(\boldsymbol{x},0)\ d\boldsymbol{x} - \int_0^T \int_\Omega \beta(\bar{U})(\boldsymbol{x},t)\partial_t\varphi(\boldsymbol{x},t) d\boldsymbol{x}\ dt$$

Proof. By definition of $\eth_t^n \beta_P^{(m)}(\boldsymbol{x},t)$ and thanks to a discrete integration by parts,

$$\begin{aligned} X_1^{(m)} &= \int_0^T \int_\Omega \eth_t(\beta^{(m)})_P^n \varphi(\boldsymbol{x},t)\ d\boldsymbol{x}\ dt \\ &= -\sum_{P\in\mathcal{P}^{(m)}} |P|\ (\beta^{(m)})_P^0 \varphi_P^0 - \sum_{n=1}^{N_m} \delta t \sum_{P\in\mathcal{P}^{(m)}} |P| \beta_P^{(m)}(\boldsymbol{x},t) \frac{1}{\delta t}\Big(\varphi_P^n - \varphi_P^{n-1}\Big). \end{aligned}$$

Thanks to the assumptions (77), (78) (80) and (81), we get that

$$\lim_{m\to+\infty} X_1^{(m)} = -\int_\Omega \beta(U_0)(\boldsymbol{x})\varphi(\boldsymbol{x},0)\ d\boldsymbol{x} - \int_0^T \int_\Omega \Big(\beta(\bar{U})(\boldsymbol{x},t)\partial_t\varphi(\boldsymbol{x},t)\ dt\ d\boldsymbol{x}. \tag{87}$$

□

Lemma A.3 (Weak consistency, space derivative)**.** *Under the assumptions and notations of Lemma A.1,*

$$\int_0^T \int_\Omega \frac{1}{|P|} \sum_{\zeta\in\mathfrak{F}^{(m)}(P)} |\zeta| (\boldsymbol{F}^{(m)})_\zeta^n \cdot \boldsymbol{n}_{P,\zeta} \varphi(\boldsymbol{x},t)\ d\boldsymbol{x}\ dt \to -\int_0^T \int_\Omega \boldsymbol{f}(\bar{U})(\boldsymbol{x},t)\cdot\nabla\varphi(\boldsymbol{x},t) d\boldsymbol{x}\ dt \text{ as } m\to+\infty.$$

Proof. Let $X_2^{(m)}$ denote the left-hand-side of the above assertion. Since for a face ζ separating P and P', one has $\boldsymbol{n}_{P,\zeta} = -\boldsymbol{n}_{P',\zeta}$, we may rewrite $X_2^{(m)}$ as

$$\begin{aligned} X_2^{(m)} &= \int_0^T \int_\Omega \frac{1}{|P|} \sum_{\zeta\in\mathfrak{F}^{(m)}(P)} |\zeta| (\boldsymbol{F}^{(m)})_\zeta^n \cdot \boldsymbol{n}_{P,\zeta} \varphi(\boldsymbol{x},t)\ d\boldsymbol{x}\ dt \\ &= \sum_{n=0}^{N_m-1} \delta t^{(m)} \sum_{P\in\mathcal{P}^{(m)}} A_P^n \text{ with } A_P^n = \sum_{\zeta\in\mathfrak{F}^{(m)}(P)} |\zeta| (\boldsymbol{F}^{(m)})_\zeta^n \cdot \boldsymbol{n}_{P,\zeta}\Big(\varphi_P^n - \varphi_\zeta^n\Big), \end{aligned}$$

where φ_P^n (resp. φ_ζ^n) denotes the mean value of φ over $P\times(t_n,t_{n+1})$ (resp. $\zeta\times(t_n,t_{n+1})$). Now for any $\boldsymbol{x}\in P$, $t\in[t_n,t_{n+1})$, we can decompose A_P^n as

$$\begin{aligned} &A_P^n = B_P^n(\boldsymbol{x},t) + R_P^n(\boldsymbol{x},t), \text{ with } B_P^n(\boldsymbol{x}) = \sum_{\zeta\in\mathfrak{F}^{(m)}(P)} |\zeta| \boldsymbol{f}(U^{(m)}(\boldsymbol{x},t))\cdot\boldsymbol{n}_{P,\zeta}\Big(\varphi_P^n - \varphi_\zeta^n\Big), \text{ and} \\ &R_P^n(\boldsymbol{x},t) = \sum_{\zeta\in\mathfrak{F}^{(m)}(P)} |\zeta| (\boldsymbol{F}_\zeta^n - \boldsymbol{f}(U^{(m)}(\boldsymbol{x},t)))\cdot\boldsymbol{n}_{P,\zeta}\Big(\varphi_P^n - \varphi_\zeta^n\Big)\ d\boldsymbol{x}. \end{aligned}$$

Since $\sum_{\zeta\in\mathfrak{F}^{(m)}(P)} |\zeta|\boldsymbol{n}_{P,\zeta} = 0$, we have

$$\begin{aligned} B_P^n(\boldsymbol{x},t) &= -\sum_{\zeta\in\mathfrak{F}^{(m)}(P)} |\zeta| \boldsymbol{f}(U^{(m)}(\boldsymbol{x},t)\cdot\boldsymbol{n}_{P,\zeta}\varphi_\zeta^n = -|P|\boldsymbol{f}(U^{(m)}(\boldsymbol{x},t)\cdot(\nabla\varphi)_P^n, \\ &\text{with } (\nabla\varphi)_P^n = \frac{1}{|P|}\sum_{\zeta\in\mathfrak{F}^{(m)}(P)} |\zeta|\varphi_\zeta^n \boldsymbol{n}_{P,\zeta} = \frac{1}{|P|}\nabla\varphi(\boldsymbol{x})\ \mathrm{d}\boldsymbol{x}. \end{aligned} \tag{88}$$

Note that the piecewise function $\nabla^{(m)}\varphi : \Omega\times(0,T)\to\mathbb{R}^d$ defined by $\nabla^{(m)}\varphi(\boldsymbol{x},t) = (\nabla\varphi)_P^n$ for $(\boldsymbol{x},t)\in\in P\times(t_n,t_{n+1})$ converges uniformly to $\nabla\varphi$ in $L^\infty(\Omega\times(0,T))^d$. Integrating (88) over $\boldsymbol{x}\in P$,

$$\int_P B_P^n(\boldsymbol{x},t)d\boldsymbol{x} = |P|\int_P \boldsymbol{f}(U^{(m)}(\boldsymbol{x},t))\cdot\nabla^{(m)}\varphi(\boldsymbol{x},t)\ d\boldsymbol{x},$$

Since $A_P^n = \dfrac{1}{\delta t^{(m)}|P|}\Big(\displaystyle\int_{t_n}^{t_{n+1}}\int_P B_P^n(\boldsymbol{x},t)d\boldsymbol{x}dt + \int_{t_n}^{t_{n+1}}\int_P R_P^n(\boldsymbol{x},t)d\boldsymbol{x}dt\Big)$, we get

$$X_2^{(m)} = \sum_{n=0}^{N_m-1}\sum_{P\in\mathcal{P}^{(m)}}\Big(\int_{t_n}^{t_{n+1}}\int_P B_P^n(\boldsymbol{x},t)d\boldsymbol{x}dt + \int_{t_n}^{t_{n+1}}\frac{1}{|P|}\int_P R_P^n(\boldsymbol{x},t)d\boldsymbol{x}dt\Big)$$
$$= -\int_0^T\int_\Omega \boldsymbol{f}(U^{(m)}(\boldsymbol{x},t))\nabla^{(m)}\varphi(\boldsymbol{x},t)\ d\boldsymbol{x}dt + \sum_{n=0}^{N_m-1}\sum_{P\in\mathcal{P}^{(m)}}\int_{t_n}^{t_{n+1}}\frac{1}{|P|}\int_P R_P^n(\boldsymbol{x},t)d\boldsymbol{x}dt.$$

Owing to the boundedness and convergence assumptions on $U^{(m)}$ and to the uniform convergence of $\nabla^{(m)}\varphi$ to $\nabla\varphi$, the first term tends to $-\displaystyle\int_0^T\int_\Omega \boldsymbol{f}(U(\boldsymbol{x},t))\nabla\varphi(\boldsymbol{x},t)\ d\boldsymbol{x}dt$ as $m\to+\infty$. Since $|\varphi_\zeta^n-\varphi_P^n|\le C_\varphi \mathrm{diam}(P)$, with C_φ depending only on φ, the second term tends to 0 thanks to the assumption (82). Therefore

$$\lim_{m\to+\infty} X_2^{(m)} \to -\int_0^T\int_\Omega \boldsymbol{f}(U(\boldsymbol{x},t))\cdot\nabla\varphi(\boldsymbol{x},t)\ d\boldsymbol{x}dt. \tag{89}$$

□

Appendix B. Former lemmas

B.1. **A result on a finite volume convection operator**

We begin with a property of the convection operator $\mathcal{C}:\rho\mapsto\partial_t(\rho)+\mathrm{div}(\rho\boldsymbol{u})$; at the continuous level, this property may be formally obtained as follows (see [22] for the detailed derivation). Let ψ be a regular function from $(0,+\infty)$ to $\mathbb{R}$; then:

$$\psi'(\rho)\ \mathcal{C}(\rho) = \partial_t\big(\psi(\rho)\big) + \mathrm{div}\big(\psi(\rho)\boldsymbol{u}\big) + \big(\rho\psi'(\rho)-\psi(\rho)\big)\ \mathrm{div}\boldsymbol{u}. \tag{90}$$

This computation is of course completely formal and only valid for regular functions ρ and $\boldsymbol{u}$. The following lemma states a discrete analogue to (90), and its proof follows the formal computation which we just described.

Lemma B.1. *[On the discrete convection operator, [21, Lemma A1]] Let P be a polygonal (resp. polyhedral) bounded set of $\mathbb{R}^2$ (resp. $\mathbb{R}^3$), and let $\mathcal{E}(P)$ be the set of its edges (resp. faces). Let ψ be a twice continuously differentiable function defined over $(0,+\infty)$. Let $\rho_P^*>0$, $\rho_P>0$, $\delta t>0$; consider three families $(\rho_\eta^*)_{\eta\in\mathcal{E}(P)}\subset\mathbb{R}_+\setminus\{0\}$, $(V_\eta^*)_{\eta\in\mathcal{E}(P)}\subset\mathbb{R}$ and $(F_\eta^*)_{\eta\in\mathcal{E}(P)}\subset\mathbb{R}$ such that*

$$\forall\eta\in\mathcal{E}(P),\qquad F_\eta^* = \rho_\eta^*\ V_\eta^*.$$

Let $R_{P,\delta t}$ be defined by:

$$R_{P,\delta t} = \Big[\frac{|P|}{\delta t}\ (\rho_P-\rho_P^*) + \sum_{\eta\in\mathcal{E}(P)} F_\eta^*\Big]\ \psi'(\rho_P)$$
$$-\Big[\frac{|P|}{\delta t}\ [\psi(\rho_P)-\psi(\rho_P^*)] + \sum_{\eta\in\mathcal{E}(P)}\psi(\rho_\eta^*)V_\eta^* + [\rho_P^*\psi'(\rho_P^*)-\psi(\rho_P^*)]\sum_{\eta\in\mathcal{E}(P)} V_\eta^*\Big].$$

Then this quantity may be expressed as follows:

$$R_{P,\delta t} = \frac{1}{2}\frac{|P|}{\delta t}(\rho_P-\rho_P^*)^2\,\psi''(\overline{\rho}_P^{(1)}) - \frac{1}{2}\sum_{\eta\in\mathcal{E}(P)} V_\eta^*\,(\rho_P^*-\rho_\eta^*)^2\,\psi''(\overline{\rho}_\eta^*) + \sum_{\eta\in\mathcal{E}(P)} V_\eta^*\rho_\eta^*\,(\rho_P-\rho_P^*)\,\psi''(\overline{\rho}_P^{(2)}),$$

where $\overline{\rho}_P^{(1)},\ \overline{\rho}_P^{(2)}\in[\![\rho_P,\rho_P^*]\!]$ and $\forall\eta\in\mathcal{E}(P)$, $\overline{\rho}_\eta^*\in[\![\rho_P^*,\rho_\eta^*]\!]$. We recall that, for $a,\ b\in\mathbb{R}$, we denote by $[\![a,b]\!]$ the interval $[\![a,b]\!]=\{\theta a+(1-\theta)b,\ \theta\in[0,1]\}$.

B.2. **A result on the space translates**

Lemma B.2 (Convergence of the space translates [12, Lemma 4.2])**.** *For a given mesh* $\mathcal{M}$*, let*

$$\theta_{\mathcal{M}} = \max_{K\in\mathcal{M}} \max_{\sigma\in\mathcal{E}_K} \frac{|D_\sigma|}{|K|}.$$

Let $\theta > 0$ *and* $(\mathcal{M}^{(m)})_{m\in\mathbb{N}}$ *be a sequence of meshes such that* $\theta_{\mathcal{M}^{(m)}} \le \theta$ *for all* $m \in \mathbb{N}$ *and* $\lim_{m\to+\infty} h_{\mathcal{M}^{(m)}} = 0$*. We suppose that the number of faces of a cell* $K \in \mathcal{M}^{(m)}$ *is bounded by* $\mathcal{N}_{\mathcal{E}}$*, for any* $m \in \mathbb{N}$*. Let* $\psi \in L^1(\Omega)$*, let* $\langle\psi\rangle_K$ *denote the mean value of* ψ *on a cell* K*. Then,*

$$\lim_{m\to+\infty} \sum_{\substack{\sigma\in\mathcal{E}_{\rm int}\\ \sigma=K|L}} |D_\sigma|\ |\langle\psi\rangle_K - \langle\psi\rangle_L| = 0. \tag{91}$$

References

[1] G. Ansanay-Alex, F. Babik, J.-C. Latché, and D. Vola. An L^2-stable approximation of the Navier-Stokes convection operator for low-order non-conforming finite elements. *International Journal for Numerical Methods in Fluids*, 66:555–580, 2011.

[2] A. Arakawa and V. Lamb. A potential enstrophy and energy conserving scheme for the shallow water equations. *Monthly Weather Review*, 109:18–36, 1981.

[3] E. Audusse. *Autour du système de Saint-Venant : Méthodes numériques pour le transport sédimentaire, les fluides en rotation et les équations primitives*. Habilitation à diriger des recherches, Universite Paris 13 Villeraneuse, Nov. 2018.

[4] M. Ben-Artzi and J. Li. Consistency and convergence of finite volume approximations to nonlinear hyperbolic balance laws, 2019.

[5] L. Bonaventura and T. Ringler. Analysis of discrete shallow-water models on geodesic delaunay grids with c-type staggering. *Monthly Weather Review*, 133(8):2351–2373, 2005.

[6] F. Bouchut. *Nonlinear Stability of finite volume methods for hyperbolic conservation laws*. Birkhauser, 2004.

[7] CALIF^{3}S. A software components library for the computation of fluid flows. `https://gforge.irsn.fr/gf/project/califs`.

[8] M. J. Castro, T. Morales de Luna, and C. Parés. Well-balanced schemes and path-conservative numerical methods. In *Handbook of numerical methods for hyperbolic problems*, volume 18 of *Handb. Numer. Anal.*, pages 131–175. Elsevier/North-Holland, Amsterdam, 2017.

[9] D. Doyen and H. Gunawan. An explicit staggered finite volume scheme for the shallow water equations. In *Finite volumes for complex applications. VII. Methods and theoretical aspects*, volume 77 of *Springer Proc. Math. Stat.*, pages 227–235. Springer, Cham, 2014.

[10] V. Elling. A Lax-Wendroff type theorem for unstructured quasi-uniform grids. *Mathematics of Computation*, 76:251–272, 2007.

[11] R. Eymard, T. Gallouët, and R. Herbin. Finite volume methods. In P. Ciarlet and J. Lions, editors, *Handbook of Numerical Analysis, Volume VII*, pages 713–1020. North Holland, 2000, https://hal.archives-ouvertes.fr/.

[12] T. Gallouët, R. Herbin, and J.-C. Latché. On the weak consistency of finite volumes schemes for conservation laws on general meshes. *SeMA J.*, 76(4):581–594, 2019.

[13] T. Gallouët, R. Herbin, J.-C. Latché, and K. Mallem. Convergence of the marker-and-cell scheme for the incompressible Navier-Stokes equations on non-uniform grids. *Foundations of Computational Mathematics*, 18:249–289, 2018.

[14] T. Gallouët, R. Herbin, J.-C. Latché, and Y. Nasseri. A second order consistent MAC scheme for the shallow water equations on non uniform grids. In *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, pages 123–131. Springer, 6 2020.

[15] E. Godlewski and P.-A. Raviart. Numerical approximation of hyperbolic systems of conservation laws. In *Springer*, page 118. Applied Mathematical Sciences, New York, 1996.

[16] H. Gunawan. *Numerical simulation of shallow water equations and related models*. PhD thesis, Université Paris-Est and Institut Teknologi Bandung, 2015.

[17] F. Harlow and A. Amsden. A numerical fluid dynamics calculation method for all flow speeds. *Journal of Computational Physics*, 8:197–213, 1971.

[18] F. Harlow and J. Welsh. Numerical calculation of time-dependent viscous incompressible flow of fluid with free surface. *Physics of Fluids*, 8:2182–2189, 1965.

[19] R. Herbin and J.-C. Latché. Kinetic energy control in the MAC discretization of the compressible Navier-Stokes equations. *International Journal on Finite Volumes*, 7, 2010.

[20] R. Herbin, J.-C. Latché, Y. Nasseri, and N. Therme. A decoupled staggered scheme for the shallow water equations. *Monografías Matemáticas García de Galdeano*, 52:1–16, 2019.

[21] R. Herbin, J.-C. Latché, and T. Nguyen. Explicit staggered schemes for the compressible Euler equations. *ESAIM: Proceedings*, 40:83–102, 2013.

[22] R. Herbin, J.-C. Latché, and T. Nguyen. Consistent segregated staggered schemes with explicit steps for the isentropic and full Euler equations. *ESAIM: Mathematical Modelling and Numerical Analysis*, 52:893–944, 2018.

[23] R. Herbin, J.-C. Latch, S. Minjeaud, and N. Therme. Conservativity and weak consistency of a class of staggered finite volume methods for the euler equations. *Mathematics of Computation*, on line 2020.

[24] D. Kroner, M. Rokyta, and M. Wierse. A Lax-Wendroff type theorem for upwind finite volume schemes in 2-D. *East-West Journal of Numerical Mathematics*, 4:279–292, 1996.

[25] P. Lax and B. Wendroff. Systems of conservation laws. *Communications in Pure and Applied Mathematics*, 13:217–237, 1960.

[26] R. Leveque. *Finite Volume Methods for Hyperbolic Problems*. Cambridge texts in applied mathematics. Cambridge University Press, 2002.

[27] L. Piar, F. Babik, R. Herbin, and J.-C. Latché. A formally second-order cell centred scheme for convection-diffusion equations on general grids. *Internat. J. Numer. Methods Fluids*, 71(7):873–890, 2013.

[28] G. Stelling and S. Duinmeijer. A staggered conservative scheme for every Froude number in rapidly varied shallow water flows. *International Journal for Numerical Methods in Fluids*, 43:1329–1354, 2003.

[29] W.-Y. Tan. *Shallow water hydrodynamics: Mathematical theory and numerical solution for a two-dimensional system of shallow-water equations*. Elsevier, 1992.

[30] Y. Xing. Numerical methods for the nonlinear shallow water equations. In *Handbook of numerical methods for hyperbolic problems*, volume 18 of *Handb. Numer. Anal.*, pages 361–384. Elsevier/North-Holland, Amsterdam, 2017.

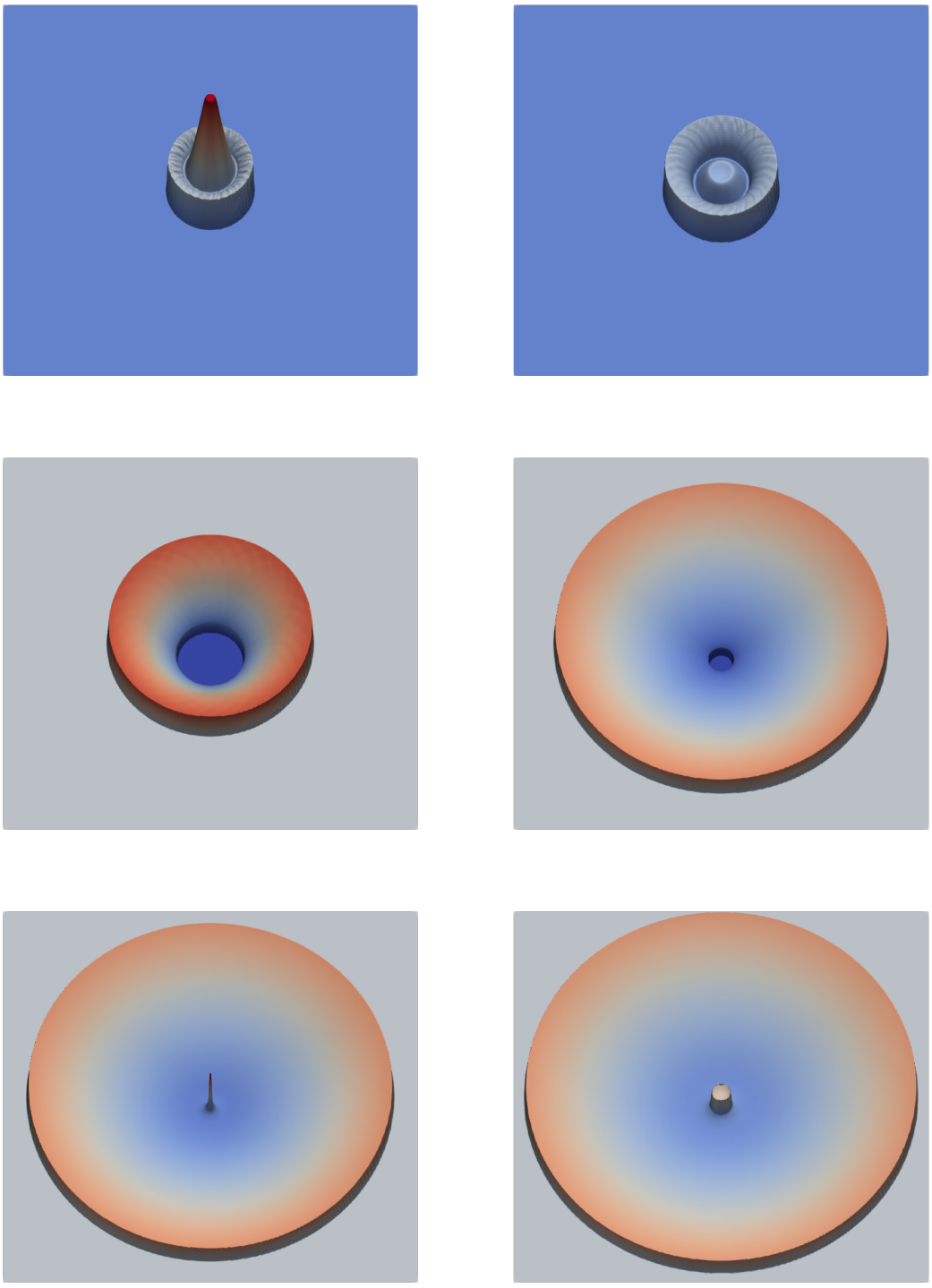

FIGURE 5. Circular dam-break problem. Height obtained at $t = 0.38$, $t = 0.705$, $t = 1.88$, $t = 3.76$, $t = 4.28$ and $t = T = 4.7$ with the stabilized second-order scheme and a 800×800 mesh. The color range corresponds to the $(0.1, 2.5)$ interval for the first two plots, and to the $(0.1, 1)$ interval for the last four ones.

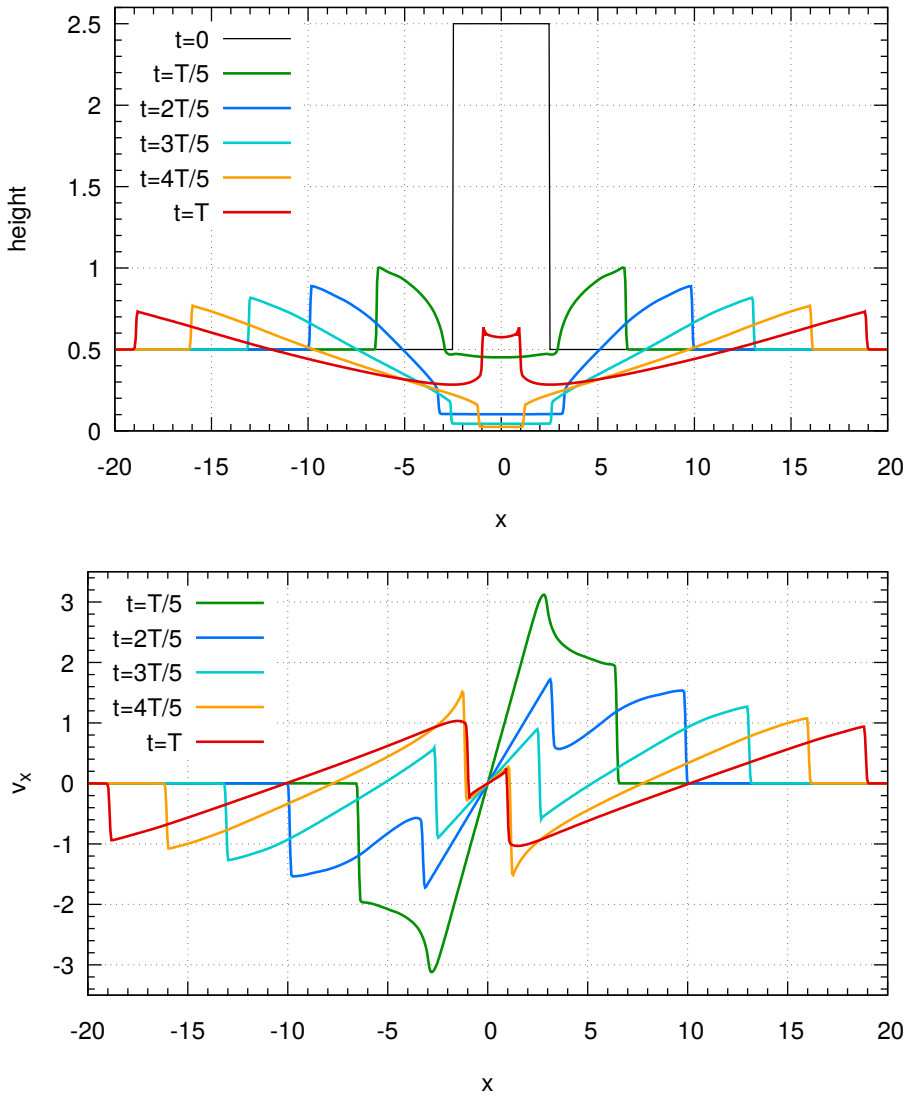

FIGURE 6. Circular dam-break problem. Height and radial velocity obtained at different times along the line $x_2 = 0$ with the stabilized second-order scheme and a 800×800 mesh.

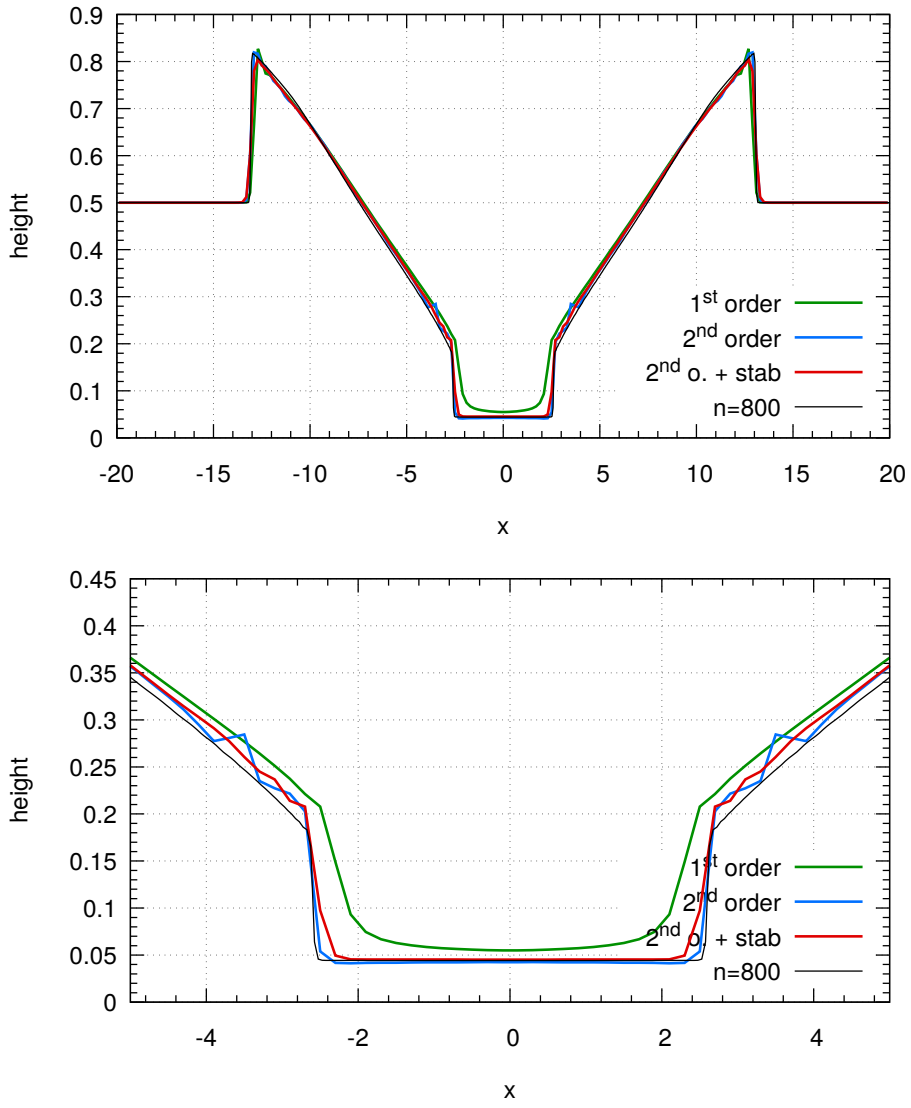

FIGURE 7. Circular dam-break problem. Height obtained at $t = 3T/5$ with the first-order scheme and the second-order scheme with and without stabilization, with a 200×200 mesh.

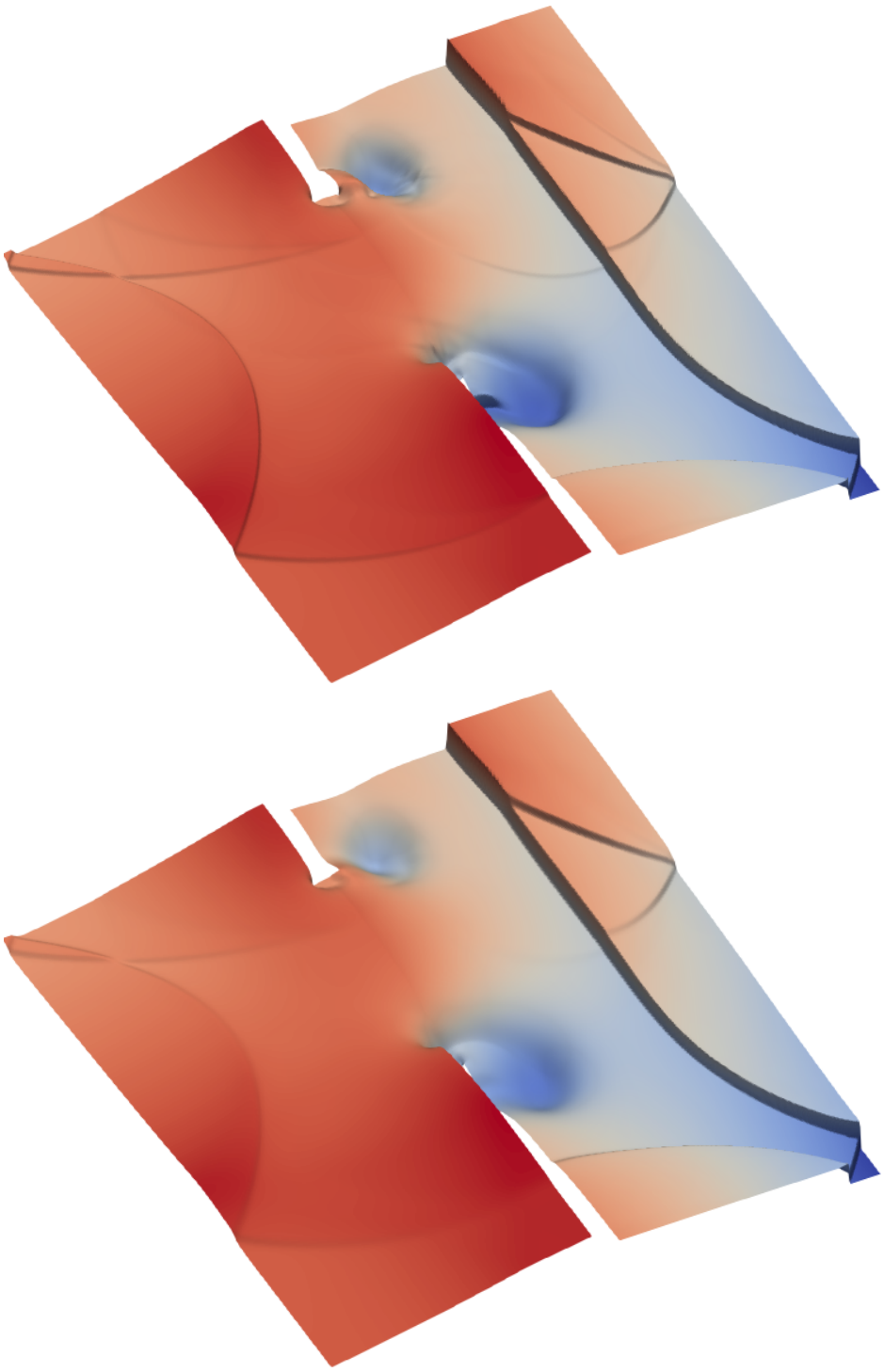

FIGURE 8. Partial dam-break flow. Top: MUSCL scheme – Bottom: upwind scheme.

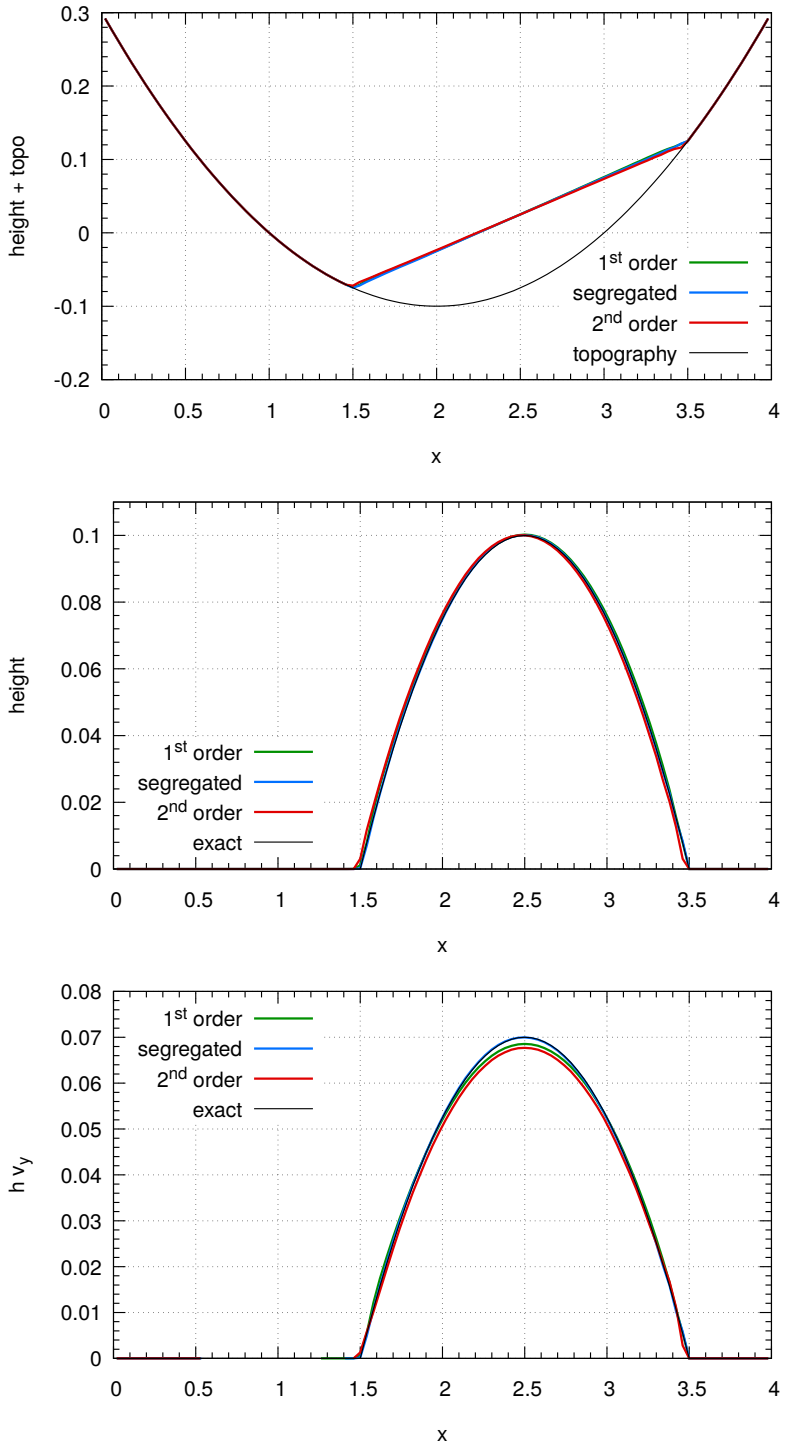

FIGURE 9. Circular motion of a drop over a paraboloid-shaped topography. Sum of the height and the topography, height alone and second component of the momentum along the $y = L/2$ line at $t = 6\,\pi$.