

A DECOUPLED STAGGERED SCHEME FOR THE SHALLOW WATER EQUATIONS

Raphaèle Herbin, Jean-Claude Latché, Youssouf Nasserri and Nicolas Therme

Abstract. We present a first order scheme based on a staggered grid for the shallow water equations with topography in two space dimensions, which enjoys several properties: positivity of the water height, preservation of constant states, and weak consistency with the equations of the problem and with the associated entropy inequality.

Keywords: Shallow water, finite volumes, staggered grid.

AMS classification: 65M08, 76B99.

§1. Introduction

The shallow water equations form a hyperbolic system of two conservation equations (mass and momentum) which are obtained when modelling a flow whose vertical height is considered small with respect to the plane scale. The solution of such a system may develop shocks, so that the finite volume method is usually preferred for numerical simulations. Two main approaches are found: one is the collocated approach which is usually based on some approximate Riemann solver, see e.g. [3] and references therein; the other one is based on a staggered arrangement of the unknowns on the grid. Indeed, staggered schemes have been used for some time in the hydraulic and ocean engineering community, see e.g. [1, 2, 12]. They have been recently analysed in the case of one space dimension [5, 8], following the works on the related barotropic Euler equations, see [11] and references therein. In the present work, we obtain a discrete local entropy inequality; furthermore, we extend the consistency analysis of the scheme to the case of two space dimensions, and we weaken the assumptions on the estimates, namely we no longer require a bound on the BV norm of the approximate solutions, at least for the weak formulation (the passage to the limit in the entropy still necessitates a time BV boundedness).

Let Ω be an open bounded domain of \mathbb{R}^2 and let $T > 0$. We consider the shallow water equations with topography over the space and time domain $\Omega \times (0, T)$:

$$\partial_t h + \operatorname{div}(hu) = 0 \quad \text{in } \Omega \times (0, T), \quad (1a)$$

$$\partial_t(hu) + \operatorname{div}(hu \otimes u) + \nabla p + gh\nabla z = 0 \quad \text{in } \Omega \times (0, T), \quad (1b)$$

$$p = \frac{1}{2}gh^2 \quad \text{in } \Omega \times (0, T), \quad (1c)$$

$$\mathbf{u} \cdot \mathbf{n} = 0 \quad \text{on } \partial\Omega \times (0, T), \quad (1d)$$

$$h(\mathbf{x}, 0) = h_0, \quad \mathbf{u}(\mathbf{x}, 0) = \mathbf{u}_0 \quad \text{in } \Omega. \quad (1e)$$

where t stands for the time, g is the standard gravity constant and z the (given) topography, which is supposed to be regular in this paper. These equations solve the water height h and the velocity \mathbf{u} .

Let us recall that if (h, \mathbf{u}) is a regular solution of (1), the following elastic potential energy balance and kinetic energy balance is obtained by manipulations on the mass and momentum equations:

$$\partial_t \left(\frac{1}{2} gh^2 \right) + \operatorname{div} \left(\frac{1}{2} gh^2 \mathbf{u} \right) + \frac{1}{2} gh^2 \operatorname{div} \mathbf{u} = 0 \quad (2)$$

$$\partial_t \left(\frac{1}{2} h |\mathbf{u}|^2 \right) + \operatorname{div} \left(\frac{1}{2} h |\mathbf{u}|^2 \mathbf{u} \right) + \mathbf{u} \cdot \nabla p + gh\mathbf{u} \cdot \nabla z = 0. \quad (3)$$

Summing these equations, we obtain an entropy equality of the form $\partial_t \eta + \operatorname{div} \Phi = 0$, where the entropy-entropy flux pair (η, Φ) is given by:

$$\eta = \frac{1}{2} h |\mathbf{u}|^2 + \frac{1}{2} gh^2 + ghz \text{ and } \Phi = \left(\eta + \frac{1}{2} gh^2 \right) \mathbf{u}. \quad (4)$$

For non regular functions the above manipulations are no longer valid, and the entropy inequality $\partial_t \eta + \operatorname{div} \Phi \leq 0$ is satisfied in a distributional sense.

In this paper, we build a decoupled scheme, involving only explicit steps; the resulting approximate solutions are shown to satisfy some discrete equivalent of (2) and (3); furthermore, under some convergence and boundedness assumptions, the approximate solutions are shown in Section 5 to converge to a weak solution of (1) and to satisfy a weak entropy inequality.

§2. Mesh and space discretizations

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 $(\mathbf{e}^{(1)}, \mathbf{e}^{(2)})$.

Definition 1 (MAC grid). A discretization $(\mathcal{M}, \mathcal{E})$ of Ω with a staggered rectangular grid (or MAC grid), is defined by:

- A primal grid \mathcal{M} which consists in a conforming structured partition of Ω in rectangles, possibly non uniform. A generic cell of this grid is denoted by K , and its mass center by \mathbf{x}_K . The scalar unknowns (water height and pressure) are associated to this mesh.
- The set of all edges of the mesh \mathcal{E} , with $\mathcal{E} = \mathcal{E}_{\text{int}} \cup \mathcal{E}_{\text{ext}}$, where \mathcal{E}_{int} (resp. \mathcal{E}_{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 $\mathbf{e}^{(i)}$ is denoted by $\mathcal{E}^{(i)}$, for $i = 1, 2$. We then have $\mathcal{E}^{(i)} = \mathcal{E}_{\text{int}}^{(i)} \cup \mathcal{E}_{\text{ext}}^{(i)}$, where $\mathcal{E}_{\text{int}}^{(i)}$ (resp. $\mathcal{E}_{\text{ext}}^{(i)}$) are the edges of $\mathcal{E}^{(i)}$ that lie in the interior (resp. on the boundary) of the domain.

For $\sigma \in \mathcal{E}_{\text{int}}$, 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 $\sigma = K|L \in \mathcal{E}_{\text{int}}$ 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);

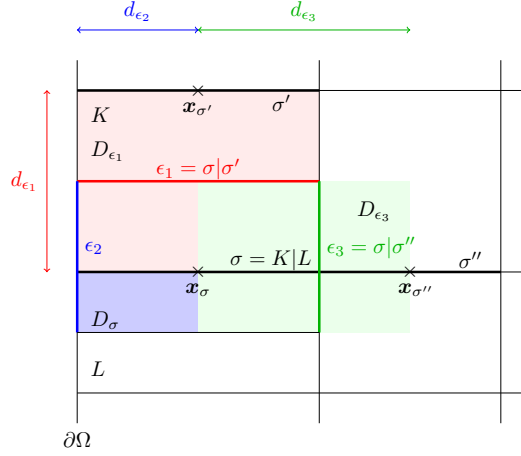


Figure 1: Notations for control volumes and dual cells (in two space dimensions, for the second component of the velocity).

- if $\sigma \in \mathcal{E}_{\text{ext}}$ is adjacent to the cell K , then $D_\sigma = D_{K,\sigma}$.

For each dimension $i = 1, 2$, the domain Ω is partitioned in dual cells: $\Omega = \cup_{\sigma \in \mathcal{E}^{(i)}} D_\sigma$, $i = 1, 2$; the i^{th} partition is referred 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 $\tilde{\mathcal{E}}^{(i)}$ (note that these edges may be orthogonal to any vector of the basis of \mathbb{R}^2 and not only $\mathbf{e}^{(i)}$) and is decomposed into the internal and boundary edges: $\tilde{\mathcal{E}}^{(i)} = \tilde{\mathcal{E}}_{\text{int}}^{(i)} \cup \tilde{\mathcal{E}}_{\text{ext}}^{(i)}$. The dual edge separating two dual cells D_σ and $D_{\sigma'}$ is denoted by $\epsilon = \sigma|\sigma'$. We denote by D_ϵ the dual cell associated to a dual edge $\epsilon \in \tilde{\mathcal{E}}$ defined as follows:

- if $\epsilon = \sigma|\sigma' \in \tilde{\mathcal{E}}_{\text{int}}^{(i)}$ 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 $\epsilon \in \tilde{\mathcal{E}}_{\text{ext}}^{(i)}$ 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 $\mathcal{E}(K)$ and $\tilde{\mathcal{E}}(D_\sigma)$ respectively. For $\sigma \in \mathcal{E}$, we denote by \mathbf{x}_σ the mass center of σ . The vector $\mathbf{n}_{K,\sigma}$ 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 will be denoted $K = \overrightarrow{[\sigma\sigma']}$ if $\sigma, \sigma' \in \mathcal{E}^{(i)}(K)$ for some $i = 1, 2$ are such that $(\mathbf{x}_{\sigma'} - \mathbf{x}_\sigma) \cdot \mathbf{e}^{(i)} > 0$;
- we write $\sigma = \overrightarrow{K|L}$ if $\sigma \in \mathcal{E}^{(i)}$, $\sigma = K|L$ and $\overrightarrow{\mathbf{x}_K \mathbf{x}_L} \cdot \mathbf{e}^{(i)} > 0$ for some $i = 1, 2$;
- the dual edge ϵ separating D_σ and $D_{\sigma'}$ is written $\epsilon = \overrightarrow{\sigma|\sigma'}$ if $\overrightarrow{\mathbf{x}_\sigma \mathbf{x}_{\sigma'}} \cdot \mathbf{e}^{(i)} > 0$ for some $i = 1, 2$.

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

$$\delta_{\mathcal{M}} = \max_{K \in \mathcal{M}} \text{diam}(K), \text{ and } \eta_{\mathcal{M}} = \max \left\{ \frac{|\sigma|}{|\sigma'|}, \sigma \in \mathcal{E}^{(i)}, \sigma' \in \mathcal{E}^{(j)}, i, j = 1, 2, i \neq j \right\}, \quad (5)$$

where $|\cdot|$ stands for the one (or two) dimensional measure of a subset of \mathbb{R} (or \mathbb{R}^2).

The discrete velocity unknowns are associated to the dual cells and denoted by $(u_{i,\sigma})_{\sigma \in \mathcal{E}^{(i)}}$, $i = 1, 2$, while the scalar unknowns (discrete water height and pressure) are associated to the primal cells and are denoted respectively by $(h_K)_{K \in \mathcal{M}}$ and $(p_K)_{K \in \mathcal{M}}$. The scalar unknown space $L_{\mathcal{M}}$ is defined as the set of piecewise constant functions over each grid cell K of \mathcal{M} , and the discrete i^{th} velocity space $H_{\mathcal{E}^{(i)}}$ as the set of piecewise constant functions over each of the grid cells D_{σ} , $\sigma \in \mathcal{E}^{(i)}$. As in the continuous case, the Dirichlet boundary conditions are taken into account by defining the subspaces $H_{\mathcal{E}^{(i)},0} \subset H_{\mathcal{E}^{(i)}}$, $i = 1, 2$ as follows

$$H_{\mathcal{E}^{(i)},0} = \left\{ u_i \in H_{\mathcal{E}^{(i)}}, u_i(\mathbf{x}) = 0, \forall \mathbf{x} \in D_{\sigma}, \sigma \in \mathcal{E}_{\text{ext}}^{(i)} \right\}.$$

We then set $\mathbf{H}_{\mathcal{E},0} = H_{\mathcal{E}^{(1)},0} \times H_{\mathcal{E}^{(2)},0}$. Defining the characteristic function $\mathbb{1}_A$ of any subset $A \subset \Omega$ by $\mathbb{1}_A(\mathbf{x}) = 1$ if $\mathbf{x} \in A$ and $\mathbb{1}_A(\mathbf{x}) = 0$ otherwise, the functions $\mathbf{u} = (u_1, u_2) \in \mathbf{H}_{\mathcal{E},0}$, may then be written:

$$u_i(\mathbf{x}) = \sum_{\sigma \in \mathcal{E}^{(i)}} u_{i,\sigma} \mathbb{1}_{D_{\sigma}}(\mathbf{x}), \quad i = 1, 2. \quad (6)$$

For $\mathbf{u} \in \mathbf{H}_{\mathcal{E},0}$, let $\llbracket u_i \rrbracket_{\epsilon} = |u_{i,\sigma} - u_{i,\sigma'}|$, for $\epsilon = \sigma|\sigma' \in \widetilde{\mathcal{E}}_{\text{int}}^{(i)}$, $i = 1, 2$. In the same way the functions $h \in L_{\mathcal{M}}$ are defined by $h(\mathbf{x}) = \sum_{K \in \mathcal{M}} h_K \mathbb{1}_K(\mathbf{x})$ and the notation $\llbracket h \rrbracket_{\sigma}$ refers to $\llbracket h \rrbracket_{\sigma} = |h_K - h_L|$, for $\sigma = K|L \in \mathcal{E}_{\text{int}}(K)$.

§3. A decoupled explicit scheme

Description of the scheme Let us consider a uniform discretisation $0 = t_0 < t_1 < \dots < t_N = T$ of the time interval $(0, T)$, and let $\delta t = t_{n+1} - t_n$ for $n = 0, 1, \dots, N-1$ be the (constant) time step. The discrete velocity \mathbf{u} and water height h unknowns are defined by:

$$\begin{aligned} \mathbf{u}(\mathbf{x}, t) &= \sum_{n=0}^{N-1} \mathbf{u}^{n+1}(\mathbf{x}) \mathbb{1}_{[t_n, t_{n+1})}(t), \quad \text{with } \mathbf{u}^{n+1} \in \mathbf{H}_{\mathcal{E},0}, \\ h(\mathbf{x}, t) &= \sum_{n=0}^{N-1} h^{n+1}(\mathbf{x}) \mathbb{1}_{[t_n, t_{n+1})}(t), \quad \text{with } h^{n+1} \in L_{\mathcal{M}}, \end{aligned}$$

where $\mathbb{1}_{[t_n, t_{n+1})}$ is the characteristic function of the interval $[t_n, t_{n+1})$ and the space functions \mathbf{u}^n and h^n take the form defined in the previous section. We propose the following decoupled discretisation of the system (1), written in compact form, with the various discrete operators

defined below.

$$\mathbf{Initialisation:} \quad \mathbf{u}^0 = \mathcal{P}_\varepsilon \mathbf{u}_0, \quad h^0 = \mathcal{P}_\mathcal{M} h_0, \quad p^0 = \frac{1}{2} g (h^0)^2. \quad (7a)$$

Iteration n , $0 \leq n \leq N - 1$: solve for $\mathbf{u}^{n+1} \in \mathbf{H}_{\varepsilon,0}$, $h^{n+1} \in L_\mathcal{M}$ and $p^{n+1} \in L_\mathcal{M}$:

$$\delta_t h^{n+1} + \operatorname{div}_\mathcal{M} (h^n \mathbf{u}^n) = 0, \quad (7b)$$

$$p^{n+1} = \frac{1}{2} g (h^{n+1})^2, \quad (7c)$$

$$\delta_t (h\mathbf{u})^{n+1} + \mathbf{C}_\varepsilon (h^n \mathbf{u}^n) \mathbf{u}^n + \nabla_\varepsilon p^{n+1} + g \mathbf{I}_\varepsilon h^{n+1} \nabla_\varepsilon z = 0, \quad (7d)$$

Projection operators - The operators \mathcal{P}_ε and $\mathcal{P}_\mathcal{M}$ used in the initialisation step are defined by $\mathcal{P}_\varepsilon = (\mathcal{P}_{\mathcal{E}^{(i)}})_{i=1,\dots,d}$ with

$$\mathcal{P}_{\mathcal{E}^{(i)}} : \left\{ \begin{array}{l} L^1(\Omega) \longrightarrow H_{\varepsilon^{(i)},0} \\ v \longmapsto \mathcal{P}_{\mathcal{E}^{(i)}} v = \sum_{\sigma \in \mathcal{E}_{\text{int}}^{(i)}} v_\sigma \mathbb{1}_{D_\sigma} \text{ with } v_\sigma = \frac{1}{|D_\sigma|} \int_{D_\sigma} v(\mathbf{x}) \, d\mathbf{x}, \text{ for } \sigma \in \mathcal{E}_{\text{int}}^{(i)}. \end{array} \right. \quad (8)$$

For $q \in L^2(\Omega)$, $\mathcal{P}_\mathcal{M} q \in L_\mathcal{M}$ is defined by:

$$\mathcal{P}_\mathcal{M} q = \sum_{K \in \mathcal{M}} q_K \mathbb{1}_K \text{ with } q_K = \frac{1}{|K|} \int_K q(\mathbf{x}) \, d\mathbf{x} \text{ for } K \in \mathcal{M}. \quad (9)$$

Discrete time derivative - The symbol δ_t denotes the discrete time derivative for both water height and momentum:

$$\delta_t h^{n+1} = \sum_{K \in \mathcal{M}} \delta_t h_K^{n+1} \mathbb{1}_K, \quad \delta_t h_K^{n+1} = \frac{1}{\delta t} (h_K^{n+1} - h_K^n), \text{ and } \delta_t (h\mathbf{u})^{n+1} = (\delta_t (h\mathbf{u}_1)^{n+1}, \delta_t (h\mathbf{u}_2)^{n+1})$$

$$\text{with } \delta_t (h\mathbf{u}_i)^{n+1} = \sum_{\sigma \in \mathcal{E}^{(i)}} \delta_t (h\mathbf{u}_i)_\sigma^{n+1} \mathbb{1}_{D_\sigma}, \text{ and } \delta_t (h\mathbf{u}_i)_\sigma^{n+1} = \frac{1}{\delta t} (h_{D_\sigma}^{n+1} u_{i,\sigma}^{n+1} - h_{D_\sigma}^n u_{i,\sigma}^n), \quad i = 1, 2,$$

where h_{D_σ} is the discrete water height in the dual cell, which is computed from the primal unknowns $(h_K^n)_{n \in \mathbb{N}, K \in \mathcal{M}}$ and defined so as to satisfy a discrete mass balance, see below.

Discrete divergence and gradient operators - The discrete divergence operator $\operatorname{div}_\mathcal{M}$ is defined by:

$$\operatorname{div}_\mathcal{M} : \left\{ \begin{array}{l} \mathbf{H}_{\varepsilon,0} \longrightarrow L_{\mathcal{M},0} \\ \mathbf{u} \longmapsto \operatorname{div}_\mathcal{M} (h\mathbf{u}) = \sum_{K \in \mathcal{M}} \operatorname{div}_K (h\mathbf{u}) \mathbb{1}_K, \text{ with } \operatorname{div}_K (h\mathbf{u}) = \frac{1}{|K|} \sum_{\sigma \in \mathcal{E}(K)} F_{K,\sigma}, \end{array} \right. \quad (10)$$

where $F_{K,\sigma}$ is the (conservative) numerical mass flux, defined by $F_{K,\sigma} = |\sigma| h_\sigma u_{K,\sigma}$ with $u_{K,\sigma} = u_{i,\sigma} \mathbf{n}_{K,\sigma} \cdot \mathbf{e}^{(i)}$ for $\sigma \in \mathcal{E}_{\text{int}}^{(i)}$, $i = 1, 2$, while h_σ is approximated by the first order upwind scheme, namely, for $\sigma = K|L \in \mathcal{E}_{\text{int}}$, $h_\sigma = h_K$ if $u_{K,\sigma} \geq 0$ and $h_\sigma = h_L$ otherwise.

The discrete gradient operator applies to the pressure and the topography and is defined by:

$$\nabla_{\mathcal{E}} : \left\{ \begin{array}{l} L_{\mathcal{M}} \longrightarrow \mathbf{H}_{\mathcal{E},0} \\ p \longmapsto \nabla_{\mathcal{E}} p, \end{array} \right.$$

with for $i = 1, 2$:

$$(\nabla_{\mathcal{E}} p)_i = \sum_{\sigma \in \mathcal{E}_{\text{int}}^{(i)}} \delta_{\sigma} p \mathbb{1}_{D_{\sigma}} \quad \text{with for } \sigma = \overrightarrow{K|L}, \quad \delta_{\sigma} p = \frac{|\sigma|}{|D_{\sigma}|} (p_L - p_K). \quad (11)$$

The above defined discrete divergence and gradient operators satisfy the following div-grad duality relationship [7, Lemma 2.5]:

$$\text{for } p \in L_{\mathcal{M}}, \mathbf{u} \in \mathbf{H}_{\mathcal{E},0}, \quad \int_{\Omega} p \operatorname{div}_{\mathcal{M}}(\mathbf{u}) \, dx + \int_{\Omega} \nabla_{\mathcal{E}} p \cdot \mathbf{u} \, dx = 0.$$

Discrete convection operator – The discrete nonlinear convection operator $\mathbf{C}_{\mathcal{E}}(hu)$ is linked to the discrete divergence operator on the dual mesh by the relation $\mathbf{C}_{\mathcal{E}}(hu)\mathbf{u} = \operatorname{div}_{\mathcal{E}}(hu \otimes \mathbf{u})$, where the full discrete convection operator $\mathbf{C}_{\mathcal{E}}(hu)$ is defined by:

$$\mathbf{C}_{\mathcal{E}}(hu)\mathbf{u} = (\mathbf{C}_{\mathcal{E}^{(1)}}(hu)u_1, \mathbf{C}_{\mathcal{E}^{(2)}}(hu)u_2),$$

and the i -th component $\mathbf{C}_{\mathcal{E}^{(i)}}(hu)$ of the convection operator is defined by:

$$\mathbf{C}_{\mathcal{E}^{(i)}}(hu) : \left\{ \begin{array}{l} H_{\mathcal{E}^{(i)},0} \longrightarrow H_{\mathcal{E}^{(i)},0} \\ u_i \longmapsto \mathbf{C}_{\mathcal{E}^{(i)}}(hu)u_i = \sum_{\sigma \in \mathcal{E}_{\text{int}}^{(i)}} \operatorname{div}_{\mathcal{E}^{(i)}}(hu_i \mathbf{u}) \mathbb{1}_{D_{\sigma}}, \\ \text{with } \operatorname{div}_{\mathcal{E}^{(i)}}(hu_i \mathbf{u}) = \frac{1}{|D_{\sigma}|} \sum_{\epsilon \in \mathcal{E}(D_{\sigma})} F_{\sigma,\epsilon} u_{i,\epsilon}, \end{array} \right. \quad (12)$$

where for $\epsilon = \sigma|\sigma'$, $u_{i,\epsilon}$ is the upwind choice between u_{σ} and $u_{\sigma'}$ with respect to the sign of $F_{\sigma,\epsilon}$. The quantity $F_{\sigma,\epsilon}$ is the numerical mass flux through ϵ outward D_{σ} ; it must be chosen carefully to ensure some stability properties of the scheme as in [7, 11]. Indeed we recall that in order to derive a discrete kinetic energy balance (Lemma 3 below), it is necessary that a discrete equation of the mass balance holds in the dual mesh, namely:

$$\frac{|D_{\sigma}|}{\delta t} (h_{D_{\sigma}}^{n+1} - h_{D_{\sigma}}^n) + \operatorname{div}_{\mathcal{E}}(h^n \mathbf{u}^n) = 0, \quad \text{with } |D_{\sigma}| \operatorname{div}_{\mathcal{E}}(h^n \mathbf{u}^n) = \sum_{\epsilon \in \mathcal{E}(D_{\sigma})} F_{\sigma,\epsilon}^n. \quad (13)$$

The water height $h_{D_{\sigma}}$ and the flux $F_{\sigma,\epsilon}$ are computed from the primal unknowns and fluxes so as to satisfy this latter relation thanks to the discrete mass balance on the primal mesh (7b). For $\sigma = K|L \in \mathcal{E}_{\text{int}}$, the water height $h_{D_{\sigma}}$ is defined as a weighted average between h_K and h_L :

$$|D_{\sigma}| h_{D_{\sigma}} = |D_{K,\sigma}| h_K + |D_{L,\sigma}| h_L, \quad (14)$$

where D_{σ} , $D_{K,\sigma}$ and $D_{L,\sigma}$ are defined in Definition 1. The numerical flux $F_{\sigma,\epsilon}$ on the internal dual edges, is defined according to the location of the edges as follows:

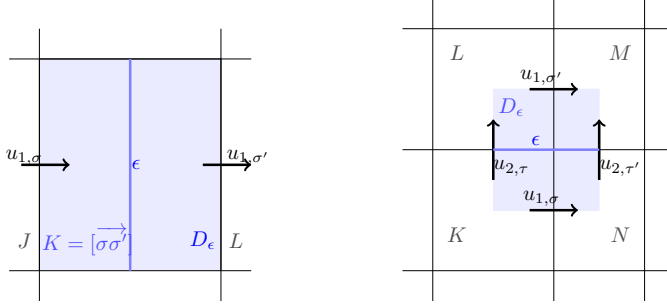


Figure 2: Notations for the definition of the momentum flux on the dual mesh for the first component of the velocity- left: first case - right: second case.

- First case – The vector $e^{(i)}$ is normal to ϵ , and ϵ is included in a primal cell K , with $K = [\overrightarrow{\sigma\sigma'}]$ (see Definition 1 and Figure 2 on the left for $i = 1$). Then for a dual edge $\epsilon \in \widetilde{\mathcal{E}}^{(i)}$ such that $\epsilon = \overrightarrow{\sigma'\sigma}$, the flux $F_{\sigma,\epsilon}$ through the edge ϵ is given by:

$$F_{\sigma,\epsilon} = \frac{1}{2}(F_{K,\sigma'} - F_{K,\sigma}) = \frac{1}{2}|\epsilon| (h_{\sigma}u_{i,\sigma} + h_{\sigma'}u_{i,\sigma'}), \quad (15)$$

since $|\sigma| = |\sigma'| = |\epsilon|$.

- Second case – The vector $e^{(i)}$ is tangent to ϵ , and ϵ is the union of the halves of two primal edges τ and τ' such that $\tau = \overrightarrow{K|L}$, $\tau \in \mathcal{E}(K)$ and $\tau' = \overrightarrow{N|M} \in \mathcal{E}(N)$ (see Definition 1 and Figure 2 on the right for $i = 1$). Let $j \in \{1, 2\}$, $j \neq i$: the numerical flux through ϵ is then given by:

$$F_{\sigma,\epsilon} = \frac{1}{2} (F_{K\tau} + F_{L\tau'}) = \frac{1}{2} (|\tau| h_{\tau}u_{j,\tau} + |\tau'| h_{\tau'}u_{j,\tau'}). \quad (16)$$

Note that the numerical momentum flux on a dual edge is conservative. It is easy to check that the unknowns $h_{D_{\sigma}}^n$ and $F_{\sigma,\epsilon}^n$ thus defined satisfy the discrete dual mass balance (13).

Discrete water height on the dual mesh, for the topography term – In equation (7d) the interpolation operator $\mathcal{I}_{\mathcal{E}}$ is defined as the mean value of the water height:

$$\mathcal{I}_{\mathcal{E}}h = \sum_{\sigma \in \mathcal{E}_{int}} h_{\sigma,c} \mathbb{1}_{D_{\sigma}} \text{ with } h_{\sigma,c} = \begin{cases} \frac{1}{2}(h_K + h_L) & \text{for } \sigma = K|L \in \mathcal{E}_{int}, \\ h_K & \text{for } \sigma \in \mathcal{E}_{ext} \cap \mathcal{E}(K). \end{cases} \quad (17)$$

This choice is important to preserve steady states, see Lemma 2.

§4. Properties of the scheme

The scheme (7) enjoys some interesting properties, which we now state. First of all, thanks to the upwind choice for h^n in (1a), the positivity of the water height is preserved under a CFL like condition.

Lemma 1 (Positivity of the water height). *Let $n \in \llbracket 0, N-1 \rrbracket$, let $(h_K^n, u_{i,\sigma}^n)_{K \in \mathcal{M}, \sigma \in \mathcal{E}^{(i)}}$ be given and such that $h_K^n \geq 0$, for all $K \in \mathcal{M}$, and let h_K^{n+1} be computed by (7b). Then $h_K^{n+1} \geq 0$, for all $K \in \mathcal{M}$ under the following CFL condition,*

$$\forall K \in \mathcal{M}, \delta t \leq \frac{|K|}{\sum_{\sigma \in \mathcal{E}(K)} |\sigma| |u_{K,\sigma}^n|}. \quad (18)$$

Second, thanks to the choice (17) for the reconstruction of the water height, the "lake at rest" steady state is preserved by the scheme.

Lemma 2 (Steady state "lake at rest"). *Let $n \in \llbracket 0, N-1 \rrbracket$, $C \in \mathbb{R}_+$; let $\mathbf{u}^{n+1} \in \mathbf{H}_{\mathcal{E},0}$ and $h^{n+1} \in L_{\mathcal{M}}$ be a solution to (7b)-(7d) with $\mathbf{u}^n = 0$ and $h^n + z = C$, where C is a given real number. Then $\mathbf{u}^{n+1} = 0$ and $h^{n+1} + z = C$.*

As a consequence of the careful discretisation of the convection term, the scheme satisfies a discrete kinetic energy balance, as stated in the following lemma. The proof of this result is an easy adaptation of [10, Lemma 3.2].

Lemma 3 (Discrete kinetic energy balance). *A solution to the scheme (7) satisfies the following equality, for $i = 1, 2$, $\sigma \in \mathcal{E}^{(i)}$ and $0 \leq n \leq N-1$:*

$$\begin{aligned} \frac{1}{2\delta t} (h_{D_\sigma}^{n+1} (u_{i,\sigma}^{n+1})^2 - h_{D_\sigma}^n (u_{i,\sigma}^n)^2) + \frac{1}{2|D_\sigma|} \sum_{\epsilon \in \bar{\mathcal{E}}^{(i)}(D_\sigma)} F_{\sigma,\epsilon}^n (u_{i,\epsilon}^n)^2 \\ + u_{i,\sigma}^{n+1} \delta_\sigma p^{n+1} + g h_{\sigma,c}^{n+1} u_{i,\sigma}^{n+1} \delta_\sigma z = -R_{i,\sigma}^{n+1}, \end{aligned} \quad (19)$$

with $R_{i,\sigma}^{n+1} \geq 0$ under the CFL like restriction:

$$\forall \sigma \in \mathcal{E}^{(i)}, \quad \delta t \leq \frac{|D_\sigma| h_{D_\sigma}^{n+1}}{\sum_{\epsilon \in \bar{\mathcal{E}}(D_\sigma)} (F_{\sigma,\epsilon}^n)^-}. \quad (20)$$

The scheme also satisfies the following potential energy balance [10, Lemma 3.3].

Lemma 4 (Discrete elastic potential balance). *Let, for $K \in \mathcal{M}$ and $0 \leq n \leq N$ the potential energy be defined by $(E_p)_K^n = \frac{1}{2}g (h_K^n)^2$. A solution to the scheme (7) satisfies the following equality, for $K \in \mathcal{M}$ and $0 \leq n \leq N-1$:*

$$(\delta_t E_p)_K^{n+1} + \operatorname{div}_K (E_p^n \mathbf{u}^n) + p_K^n \operatorname{div}_K \mathbf{u}^n = -R_K^{n+1}, \quad (21)$$

with

$$R_K^{n+1} \geq \frac{1}{|K|} g \sum_{\sigma \in \mathcal{E}(K)} |\sigma| u_{K,\sigma}^n h_\sigma^n (h_K^{n+1} - h_K^n). \quad (22)$$

Note that the right-hand side of Equation (22) may be negative, and thus also the quantities R_K^{n+1} . This is specific to explicit schemes (for implicit or pressure-correction schemes, this residual is non-negative [9]) and prevents getting a stability estimate for the scheme.

However, combining the two previous lemmas allows to prove that convergent sequences of solutions to the scheme satisfy an entropy inequality, as depicted in the next section. To this purpose, we will pass to the limit in a discrete entropy balance which is built as follows. Let $K \in \mathcal{M}$ and let us denote by $(E_k)_K^n$ the following quantity, which may be seen as a kinetic energy associated to K :

$$(E_k)_K^n = \frac{1}{4|K|} \sum_{i=1}^2 \sum_{\sigma \in \mathcal{E}(K) \cap \mathcal{E}^{(i)}} |D_\sigma| h_{D_\sigma}^n (u_{i,\sigma}^n)^2.$$

Then, for $\sigma_0 \in \mathcal{E}(K)$, we define a kinetic energy flux, which we denote by G_{K,σ_0}^n , as follows. Let us suppose, for instance, that $\sigma_0 \in \mathcal{E}^{(1)}$. We denote by ϵ the face of D_{σ_0} parallel to σ_0 and included in K and by ϵ' the opposite face of D_{σ_0} . In addition, σ_0 is the union of two half-faces of the dual mesh associated to the second component of the velocity, which we denote by τ and τ' , and we denote by σ and σ' the two faces of K belonging to $\mathcal{E}^{(2)}$ such that $\tau \in \tilde{\mathcal{E}}(D_\sigma)$ and $\tau' \in \tilde{\mathcal{E}}(D_{\sigma'})$. We then set:

$$G_{K,\sigma_0}^n = \frac{1}{4} \left[-F_{\sigma_0,\epsilon}^n (u_{1,\epsilon}^n)^2 + F_{\sigma_0,\epsilon'}^n (u_{1,\epsilon'}^n)^2 + F_{\sigma,\tau}^n (u_{2,\tau}^n)^2 + F_{\sigma',\tau'}^n (u_{2,\tau'}^n)^2 \right]$$

Multiplying the kinetic energy balance equation (19) associated to each face σ of K by $\frac{1}{2} |D_\sigma|$ and summing the four obtained relations with Equation (21) multiplied by $|K|$, we get

$$\begin{aligned} & (\delta_t E_k)_K^{n+1} + (\delta_t E_p)_K^{n+1} + \sum_{\sigma \in \mathcal{E}(K)} [G_{K,\sigma}^n + \frac{1}{2} g h_\sigma^n F_{K,\sigma}^n] + \sum_{\substack{\sigma \in \mathcal{E}(K) \\ \sigma = K|L}} |\sigma| p_K^n u_{K,\sigma}^n \\ & + \sum_{\substack{\sigma \in \mathcal{E}(K) \\ \sigma = K|L}} |\sigma| \frac{1}{2} (p_L^{n+1} - p_K^{n+1}) u_{K,\sigma}^{n+1} + \sum_{\substack{\sigma \in \mathcal{E}(K) \\ \sigma = K|L}} |\sigma| \frac{1}{4} g (h_K^{n+1} + h_L^{n+1}) (z_L - z_K) u_{K,\sigma}^{n+1} = -T_K^{n+1}, \end{aligned}$$

where T_K^{n+1} collects the residual terms in (19) and (21), and thus $T_K^{n+1} \geq R_K^{n+1}$. We now remark that, thanks to the discrete mass balance equation and the fact that the topography does not depend on time,

$$\frac{1}{2} \sum_{\sigma \in \mathcal{E}(K)} F_{K,\sigma}^n (z_L - z_K) = \frac{|K|}{\delta t} (h_K^{n+1} z_K - h_K^n z_K) + \frac{1}{2} \sum_{\sigma \in \mathcal{E}(K), \sigma = K|L} F_{K,\sigma}^n (z_K + z_L),$$

and we finally obtain the following discrete entropy balance:

$$\begin{aligned} & (\delta_t E_k)_K^{n+1} + (\delta_t E_p)_K^{n+1} + g z_K (\delta_t h)_K^{n+1} + \sum_{\sigma \in \mathcal{E}(K)} [G_{K,\sigma}^n + \frac{1}{2} g h_\sigma^n F_{K,\sigma}^n + \frac{1}{2} F_{K,\sigma}^n (z_K + z_L)] \\ & + \sum_{\substack{\sigma \in \mathcal{E}(K) \\ \sigma = K|L}} |\sigma| \frac{1}{2} (p_K^n + p_L^n) u_{K,\sigma}^n = -(R_e)_K^{n+1}, \quad (23) \end{aligned}$$

with

$$\begin{aligned} (R_e)_K^{n+1} &\geq T_K^{n+1} + g \sum_{\sigma \in \mathcal{E}(K)} \left[\frac{1}{2} F_{K,\sigma}^n - \frac{1}{4} |\sigma| (h_K^{n+1} + h_L^{n+1}) u_{K,\sigma}^{n+1} \right] (z_L - z_K) \\ &\quad + \sum_{\sigma \in \mathcal{E}(K), \sigma = KL} |\sigma| \frac{1}{2} (p_K^{n+1} u_{K,\sigma}^{n+1} - p_K^n u_{K,\sigma}^n + p_L^{n+1} u_{L,\sigma}^{n+1} - p_L^n u_{L,\sigma}^n). \end{aligned} \quad (24)$$

Note that the above remainder term R_e cannot be proven to be non negative; however, the relation (24) allows to obtain a weak entropy consistency property when passing to the limit, as stated in the next section.

§5. Consistency analysis

The objective of this section is to show that the schemes are consistent in the Lax-Wendroff sense, namely 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.

A weak solution to the continuous problem satisfies, for any $\varphi \in C_c^\infty(\Omega \times [0, T])$ ($\varphi \in C_c^\infty(\Omega \times [0, T])^2$):

$$\int_0^T \int_\Omega [h \partial_t \varphi + h \mathbf{u} \cdot \nabla \varphi] \, dx \, dt + \int_\Omega h_0(\mathbf{x}) \varphi(\mathbf{x}, 0) \, dx = 0, \quad (25a)$$

$$- \int_0^T \int_\Omega [h \mathbf{u} \cdot \partial_t \boldsymbol{\varphi} + (h \mathbf{u} \otimes \mathbf{u}) : \boldsymbol{\varphi} + \frac{1}{2} g h^2 \operatorname{div}(\boldsymbol{\varphi}) + g h \nabla(z) \boldsymbol{\varphi}] \, dx \, dt \quad (25b)$$

$$- \int_\Omega h_0(\mathbf{x}) \mathbf{u}_0(\mathbf{x}) \cdot \boldsymbol{\varphi}(\mathbf{x}, 0) \, dx = 0.$$

This system is supplemented with a weak entropy inequality, for any nonnegative test functions $\varphi \in C_c^\infty(\Omega \times [0, T], \mathbb{R}_+)$:

$$- \int_0^T \int_\Omega [\eta \partial_t \varphi + \boldsymbol{\Phi} \cdot \nabla \varphi] \, dx \, dt - \int_\Omega \eta_0(\mathbf{x}) \varphi(\mathbf{x}, 0) \, dx \leq 0, \quad (26)$$

with η and $\boldsymbol{\Phi}$ defined by (4).

Before stating the global weak consistency of the scheme (7), some definitions and estimate assumptions are needed.

Let $(\mathcal{M}^{(m)}, \mathcal{E}^{(m)})_{m \in \mathbb{N}}$ be a sequence of meshes in the sense of Definition 1 and let $(h^{(m)} \mathbf{u}^{(m)})_{m \in \mathbb{N}}$ be the associated sequence of solutions of the scheme (7).

Assumed estimates - We need also some a priori estimates on the sequence of discrete solutions $(h^{(m)}, \mathbf{u}^{(m)})_{m \in \mathbb{N}}$ in order to prove the consistency result we are seeking. First of all we assume that $h^{(m)} > 0$, $\forall m \in \mathbb{N}$ which can be obtained under the CFL condition (18). Furthermore:

- The water height $h^{(m)}$ and its inverse are uniformly bounded in $L^\infty(\Omega \times (0, T))$, i.e. there exists some constants $C, C' \in \mathbb{R}_+^*$ such that for $m \in \mathbb{N}$ and $0 \leq n < N^{(m)}$:

$$1/C < (h^{(m)})_K^n \leq C, \quad 1/C' < 1/(h^{(m)})_K^n \leq C' \quad \forall K \in \mathcal{M}^{(m)} \quad (27)$$

- The velocity $\mathbf{u}^{(m)}$ is also uniformly bounded in $L^\infty(\Omega \times (0, T))^2$:

$$|(\mathbf{u}^{(m)})_\sigma^n| \leq C, \quad \forall \sigma \in \mathcal{E}^{(m)}. \quad (28)$$

Finally, the weak consistency to the entropy inequality is only proved under additional assumptions. First we need the following condition on the space and time steps, which is stronger than a CFL condition:

$$\frac{\delta t^{(m)}}{\delta \mathcal{M}^{(m)}} \rightarrow 0 \text{ as } m \rightarrow +\infty \quad (29)$$

Second, the $L^1(\Omega, BV)$ norm of the height is required to be bounded, i.e. there exists one constant C such that, for $m \in \mathbb{N}$,

$$\sum_{n=0}^{N-1} \sum_{K \in \mathcal{M}} |K| |(h^{(m)})_K^{n+1} - (h^{(m)})_K^n| \leq C. \quad (30)$$

We are now in position to state the following consistency result.

Theorem 5 (Weak consistency of the scheme). *Let $(\mathcal{M}^{(m)}, \mathcal{E}^{(m)})_{m \in \mathbb{N}}$ be a sequence of meshes such that $\delta t^{(m)}$ and $\delta \mathcal{M}^{(m)} \rightarrow 0$ as $m \rightarrow +\infty$; assume that there exists $\eta > 0$ such that $\eta_{\mathcal{M}^{(m)}} \leq \eta$ for any $m \in \mathbb{N}$ (with $\eta_{\mathcal{M}^{(m)}}$ defined by (5)); assume moreover that (27) and (28) hold. Let $(h^{(m)}, \mathbf{u}^{(m)})_{m \in \mathbb{N}}$ be a sequence of solutions to the scheme (7) converging to $(\bar{h}, \bar{\mathbf{u}})$ in $L^1(\Omega \times (0, T)) \times L^1(\Omega \times (0, T))^2$. Then $(\bar{h}, \bar{\mathbf{u}})$ satisfies the weak formulation (25) of the shallow water equations.*

If we furthermore assume the space and time steps satisfy (29) and that the sequence of heights is uniformly bounded in $L^1(\Omega, BV)$, i.e. satisfy (30), then $(\bar{h}, \bar{\mathbf{u}})$ satisfies the entropy inequality (26).

Proof. The proof is obtained by passing to the limit in the scheme and in the discrete entropy balance (23), using the tool of [6] (or, more precisely speaking, simplified versions of these tools adapted to Cartesian grids). The additional assumptions required for the entropy condition are used to prove that the residual term appearing in the discrete potential energy balance, given by (22), tends to zero. \square

§6. Numerical results

We now assess the behaviour of the scheme on some numerical experiments. The computations presented here are performed with the CALIF³S free software developed at IRSN [4].

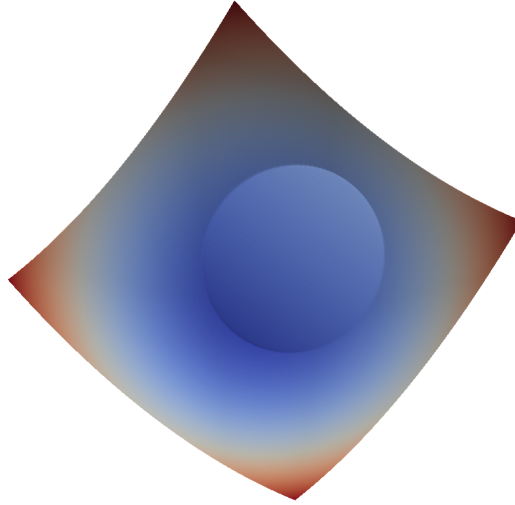


Figure 3: Sloshing of a drop on a parabolic support – State obtained after one revolution (very close to the initial state).

6.1. Rotation in a paraboloid

This first test case consists in calculating the uniform rotation of a circular drop on a support of parabolic shape (see Figure 3). The computational domain is $(0, L) \times (0, L)$ and the elevation of the support is:

$$z = -h_0 \left(1 - \left(x - \frac{L}{2}\right)^2 - \left(y - \frac{L}{2}\right)^2 \right),$$

with $L = 4$ and $h_0 = 0.1$. The fluid height is given by

$$h = h_0 \max\left(0, \left(x - \frac{L}{2}\right) \cos(\omega t) + \left(y - \frac{L}{2}\right) \sin(\omega t) - z - 0.5\right),$$

and the velocity is

$$\mathbf{u} = \frac{1}{2} \omega \begin{bmatrix} -\sin(\omega t) \\ \cos(\omega t) \end{bmatrix}.$$

It is then easy to check that the mass and momentum balance equations are verified provided that $\omega^2 = 2gh_0$. The solution is thus regular, and this test features a regular topography and dry zones (*i.e.* zones where $h = 0$). We compare the numerical and theoretical height obtained after one rotation (*i.e.* $\omega t = 2\pi$), for different uniform grids and with a time step $\delta t = \delta x/8$. (the maximal speed of sound and the maximal velocity are both close to 1); results are gathered in the following table:

grid	error (discrete L^1 norm)
100×100	$3.02 \cdot 10^{-3}$
200×200	$1.54 \cdot 10^{-3}$
400×400	$0.896 \cdot 10^{-3}$
800×800	$0.511 \cdot 10^{-3}$

We observe an order of convergence between 0.8 and 1, which is consistent with a first-order approximation of the fluxes and the time derivative.

6.2. A dam-break problem

In this test, the computational domain is:

$$\Omega = (0, 200) \times (0, 200) \setminus \Omega_w \text{ with } \Omega_w = (95, 105) \times (0, 95) \cup (95, 170) \times (0, 200).$$

The fluid is supposed to be initially at rest, and the initial height is $h = 10$ for $x_1 \leq 100$ and $h = 5$ for $x_1 > 100$. 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 $\delta t = \delta x / 25$ (the maximal speed of sound and the maximal velocity are both close to 10). The obtained fluid height is shown at different times on Figure 4; they confirm the efficiency of the scheme, and its capability to deal with reflexion phenomena very simply (*i.e.* just by setting the normal velocity at the boundary to zero, by contrast with schemes based on Riemann solvers which need to implement fictitious cells techniques).

Acknowledgements

The authors would like to thank Robert Eymard and Thierry Gallouët for several interesting discussions.

References

- [1] ARAKAWA, A., AND LAMB, V. A potential enstrophy and energy conserving scheme for the shallow water equations. *Monthly Weather Review* 109 (1981), 18–36.
- [2] BONAVENTURA, L., AND RINGLER, T. Analysis of discrete shallow-water models on geodesic delaunay grids with c-type staggering. *Monthly Weather Review* 133, 8 (2005), 2351–2373.
- [3] BOUCHUT, F. *Nonlinear Stability of finite volume methods for hyperbolic conservation laws*. Birkhauser, 2004.
- [4] CALIF³S. A software components library for the computation of fluid flows. <https://gforge.irsn.fr/gf/project/califs>.
- [5] DOYEN, D., AND GUNAWAN, H. An explicit staggered finite volume scheme for the shallow water equations. In *Finite volumes for complex applications. VII. Methods and theoretical aspects*, vol. 77 of *Springer Proc. Math. Stat.* Springer, Cham, 2014, pp. 227–235.

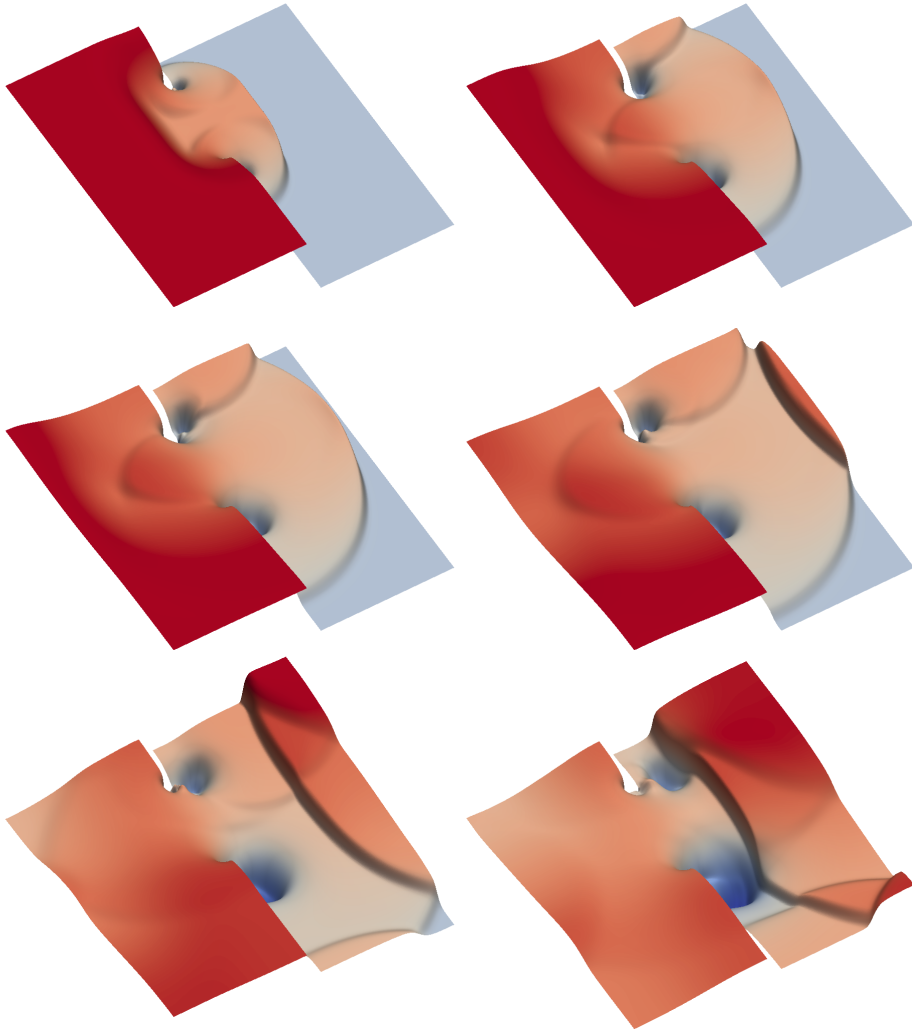


Figure 4: Partial dam break – Height obtained at $t = 4$, $t = 8$, $t = 10$, $t = 12$, $t = 16$ and $t = 20$ with a mesh obtained (by suppression of the zones associated to the obstacles) from a 1000×1000 regular grid. In the last Figure ($t = 20$), the obtained minimal and maximal heights are $h = 2.149$ and $h = 9.306$ respectively.

- [6] GALLOUËT, T., HERBIN, R., AND LATCHÉ. On the weak consistency of finite volumes schemes for conservation laws on general meshes. *under revision* (2019). Available from: <https://hal.archives-ouvertes.fr/hal-02055794>.
- [7] GALLOUËT, T., HERBIN, R., LATCHÉ, J.-C., AND MALLEM, K. Convergence of the marker-and-cell scheme for the incompressible Navier-Stokes equations on non-uniform grids. *Foundations of Computational Mathematics* 18 (2018), 249–289.
- [8] GUNAWAN, H. *Numerical simulation of shallow water equations and related models*. PhD thesis, Université Paris-Est and Institut Teknologi Bandung, 2015.
- [9] HERBIN, R., KHERIJ, W., AND LATCHÉ, J.-C. On some implicit and semi-implicit staggered schemes for the shallow water and Euler equations. *ESAIM: Mathematical Modelling and Numerical Analysis* 48 (2014), 1807–1857.
- [10] HERBIN, R., LATCHÉ, J.-C., AND NGUYEN, T. Explicit staggered schemes for the compressible Euler equations. *ESAIM: Proceedings* 40 (2013), 83–102.
- [11] HERBIN, R., LATCHÉ, J.-C., AND NGUYEN, T. Consistent segregated staggered schemes with explicit steps for the isentropic and full Euler equations. *ESAIM: Mathematical Modelling and Numerical Analysis* 52 (2018), 893–944.
- [12] STELLING, G., AND DUINMEIJER, S. A staggered conservative scheme for every Froude number in rapidly varied shallow water flows. *International Journal for Numerical Methods in Fluids* 43 (2003), 1329–1354.

R. Herbin and Y. Nasser

Aix-Marseille Université, Institut de Mathématiques de Marseille,
39 rue Joliot Curie
13453 Marseille
raphaele.herbin@univ-amu.fr and youssof.nasseri@univ-amu.fr

J.-C. Latché
Institut de Radioprotection et Sûreté Nucléaire,
13115, Saint-Paul-lez-Durance
jean-claude.latche@irsn.fr

N. Therme
CEA/CESTA
33116, Le Barp, France
nicolas.therme@cea.fr