Energy-stable staggered schemes for the Shallow Water equations

Arnaud Duran, Jean-Paul Vila, Rémy Baraille

Energy-stable staggered schemes for the Shallow Water equations

Arnaud Duran [1], Jean-Paul Vila [2,3], Rémy Baraille [2,3,4]

Abstract

In this work we focus on the development and analysis of staggered schemes for the two-dimensional non-linear Shallow Water equations with varying bathymetry. Semi-implicit and fully explicit time-discretizations are proposed. Particular attention is given on non-linear stability results, principally considered here through discrete energy dissipation arguments. To address such an issue, specific convective fluxes are employed, implying diffusive terms relying on the pressure gradient. In addition of providing an explicit control of the discrete energy budget, the method is shown to preserve motionless steady states as well as the positivity of the water height. These properties are still satisfied in a fully explicit context, provided an appropriate discretization of the pressure gradient. These stability results make the approach particularly robust and efficient, for both coastal flows and low Froude number regimes. As a result, in addition of a great ease of implementation, the presented schemes meet the operational requirements attached to the simulation of large and small scale oceanic flows.

Keywords non-linear shallow water equations, energy dissipation, non-linear stability, staggered meshes, well-balanced methods

1 Introduction

We consider in this work numerical discretizations of the two-dimensional Non-linear Shallow Water equations (NSW), formulated below with conservative variables:

$$\left\{ \begin{array}{l} \partial_t h + div(h\mathbf{u}) = 0\,, \\ \partial_t (h\mathbf{u}) + div(h\mathbf{u}\otimes\mathbf{u}) = -h\nabla\Phi\,. \end{array} \right. \tag{1}$$

In the above equations, $h = h(\mathbf{x},t)$ stands for the water depth and $\mathbf{u} = \mathbf{u}(\mathbf{x},t)$ the horizontal velocity, functions of the space and time variables. The pressure force is taken into account through the quantity $\Phi = g(h+z)$, g being the gravity constant and $z = z(\mathbf{x})$ standing for a parametrization of the topography.

Academic and practical interest for approximating weak solutions of the NSW equations has developed substantially over the last few decades, and is still subject to intensive research. Nowadays, most of modern approaches are constrained by well-established non-linear stability properties, relying on the perservation of admissible states (positivity of the water height), preservation of steady solutions and entropy inequalities. These results are still quite complex to achieve simultaneously, especially since applicative issues may imply high order resolutions, complex geometries, stiff source terms or other specific difficulties such as flooding and drying phenomena with occurrence of dry cells. The presented schemes are principally designed to fulfill operational needs related to oceanography or river flows, which are essentially based on these stability criteria.

In light of initial studies carried out to integrate source terms in the equations [6, 24], many researchers have expressed interest in developing *well-balanced* methods, referring to the ability to exactly preserve steady states. Due to crucial practical interests, the focus is usually put on the *lake at rest* configuration:

$$\mathbf{u} = 0 \quad , \quad h + z = cst\,, \tag{2}$$

which, under its appearent simplicity, may already stand for a non-trivial task. We refer to [4, 22, 23, 30, 35, 59] for some reference works on this topic. The list is of course non exhaustive. On the basis of other propositions [7, 33, 47, 55, 61], a quite popular strategy consists in imploying the total free surface elevation $\eta = h + z$ as a new variable in the model, allowing simplified analysis and direct extensions to high order resolutions [16, 31, 36, 39]. The very simple strategy used in this paper, based on a straightforward discretization of Φ, may fit into this category. Note however that most of the above-mentioned methods have been subject to various improvements, with the aim of handling vanishing water depths and occurence of dry cells, high order discretizations and unstructured geometries, or even stiff source terms to account for friction effects for instance (see [12, 18, 21, 41, 44]). In parallel, many other methods have been developed to furnish accurate and efficient resolutions (we refer to [46, 49, 52, 58, 60]

[1] Institut Camille Jordan, Université Claude Bernard Lyon 1, France.
[2] Institut de Mathématiques de Toulouse, Université Paul Sabatier Toulouse 3, France.
[3] INSA - Institut National des Sciences Appliquées, Toulouse, France.
[4] SHOM - Service Hydrographique et Océanographique de la Marine, France.

for some examples). Note finally the recent efforts to propose solutions for moving steady states [10, 13, 14, 40, 43, 57].

Another challenging issue when dealing with the NSW equations is related to the concept of entropy stability. It is well known that in the general context of hyperbolic problems, solutions may develop discontinuities in finite time and the system must be endowed with entropy inequalities. In the particular case of the NSW equations, defining the potential and kinetic energies by

$$\mathcal{E} = g(hz + \frac{1}{2}h^2) \quad \text{and} \quad \mathcal{K} = \frac{1}{2}h\|\mathbf{u}\|^2\,, \tag{3}$$

the total energy $E = \mathcal{E} + \mathcal{K}$ plays the role of an entropy and the stability is often considered through local or global discrete energy estimates. Reaching such results is far from trivial, but mandatory in lots of operational contexts, notably those involving long time integration and/or low-Froude simulations. Historically, most of operational platforms were developed on staggered discretizations, such as HYCOM [8], ROMS [48] or NEMO [38]. These geometric environments are known to provide an appropriate frame with respect to low-velocity flows and non-linear stability, which is fundamental for the targeted applications.

Surprisingly, despite constant advances in numerical modelling during the past decades, the quantity of academic works dedicated to staggered discretizations of the NSW equations is far less plentiful. That being said, an extensive study is proposed in [28] for implicit and semi-implicit pressure-correction schemes, providing global discrete entropy inequalities and consistency results. In [29], similar results are obtained with an explicit time scheme for the barotropic Euler equations. However, the presence of a residual term in the discrete energy balance does not allow to recover a strict total energy decrease. It is shown however that, under some regularity conditions, this do not compromise entropy consistency (see also [26]). The interest for staggered discretizations extends to applications to hydraulics [19] and rotating shallow water models in [2, 27, 50]. In this respect, one can also note [34, 53], where classes of Finite Element and Discontinuous Galerkin methods are investigated, stressing the importance of the unknowns location in the approximation space (notably with respect to the Coriolis force, not accounted for in the theoretical part of this work and subject to work in progress), or refer to [15] for a Discontinuous Galerkin algorithm of arbitrary order.

In this paper we introduce and analyse a class of staggered schemes for the two-dimensional NSW system, on the basis of semi-implicit and explicit time discretizations. The main objective is to guarantee the discrete mechanical energy decrease, together with robustness and well-balanced properties. The cornerstone of the method follows from the strategy used in [25], where a shifted velocity is introduced within the convective fluxes, involving a perturbation expressed in terms of the discrete pressure gradient. During the past two years, the method has been successfully applied to stratified shallow water flows in a collocated frame [11, 45], also with explicit and semi-implicit time discretizations. More recently, a first step in the adaptation to staggered meshes has been proposed in [17], also for the multi-layer system. Following the formalism [3], we hence propose an extension of these works oriented toward operational simulation, with the main objective of maintaining energy stability. Note that the well-balanced property is reached through a simple discretization of the pressure gradient, without the need of any reconstruction step.

The paper is organised as follows. In Section 2 we introduce some notations and the general settings. Section 3 is devoted to the semi-implicit approach. After establishing the kinetic and potential discrete energy balances, we show the non-linear stability of the scheme, understood as the global energy decrease, and then discuss the well-balanced and robustness properties. Following the same steps, a fully explicit time scheme is investigated in Section 4, for which similar results are obtained. Numerical results are finally proposed in Section 5.

2 General framework

To discretize the system (1) we consider a mesh $\mathbb{T}$ of the computational domain $\Omega \subset \mathbb{R}^2$ made of non-overlapping polygonal cells. Fig. 1 gives a focus on the geometry and notations at the level of an edge e shared by two neighbor elements K and K_e of $\mathbb{T}$. On this basis, a dual grid $\mathbb{T}^*$ can be built according to Crouzeix-Raviart (CR), Rannacher-Turek (RT) or Marker And Cell (MAC) discretizations. To illustrate the following developments, we will enter the frame of Crouzeix-Raviart elements, specifying that all the forthcoming results are also valid for RT or MAC geometries. In this context, for each interface e, a dual cell $D \in \mathbb{T}^*$ consists in a quadrilateral admitting e as diagonal and the mass centers of the two neighbors K and K_e as vertices. As indicated in Fig. 1, it results a dual element D made of two triangles D^K,

D^{K_e} separated by the edge e. We will employ the notation $D = \mathbb{T}^*(K, K_e)$ to specify that the dual element D corresponds to the adjoining cells K and K_e of the primal mesh. The length of an edge e will be denoted m_e and m_K, $m_{\partial K}$ will respectively stand for the area and perimeter of an element $K \in \mathbb{T}$. Similar notations are employed for the egdes f and elements D of the dual mesh $\mathbb{T}^*$.

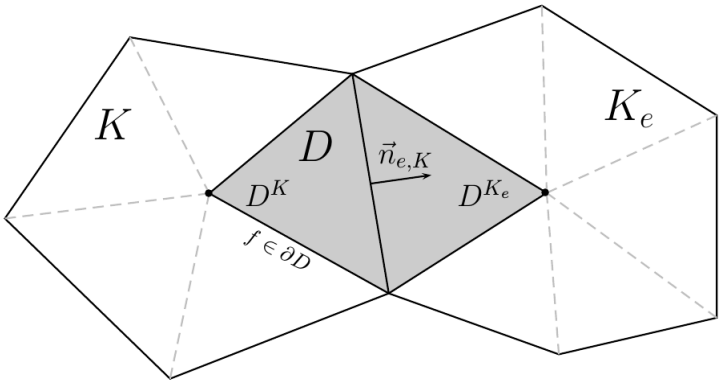

Figure 1: *Geometry associated with the Crouzeix-Raviart discretization. Focus on a dual element* $D = \mathbb{T}^*(K, K_e)$ *at the level of a boundary interface* $e = \partial K \cap \partial K_e$.

3 Semi-implicit upwind scheme

We first consider a semi-implicit time discretization, as an adaptation of [45] to the present formalism. The numerical scheme is the following:

$$h_K^{n+1} = h_K^n - \frac{\Delta t}{m_K} \sum_{e \in \partial K} \mathcal{F}_{e,K}^{n+1} . \vec{n}_{e,K} m_e \quad , \quad K \in \mathbb{T} \tag{4}$$

$$\begin{aligned} h_D^{n+1} \mathbf{u}_D^{n+1} = {} & h_D^n \mathbf{u}_D^n - \frac{\Delta t}{m_D} \sum_{f \in \partial D} \left(\mathbf{u}_D^n \left(\mathcal{F}_f^{n+1} . \vec{n}_{f,D} \right)^+ + \mathbf{u}_{D_f}^n \left(\mathcal{F}_f^{n+1} . \vec{n}_{f,D} \right)^- \right) m_f \\ & - \Delta t h_D^{n+1} \left(\nabla_D \Phi^{n+1} \right) \quad , \quad D \in \mathbb{T}^* \end{aligned} \tag{5}$$

where the notation $w^{\pm} = \dfrac{1}{2}(w \pm |w|)$ is introduced to refer to the positive and negative parts of a generic scalar quantity w. The mass fluxes are given by:

$$\mathcal{F}_{e,K}^{n+1} . \vec{n}_{e,K} = \left(h_K^{n+1} (\mathbf{u}_D^n . \vec{n}_{e,K})^+ + h_{K_e}^{n+1} (\mathbf{u}_D^n . \vec{n}_{e,K})^- \right) - \Pi_{e,K}^{n+1} . \vec{n}_{e,K} \, , \tag{6}$$

where

$$\Pi_{e,K}^{n+1} = \gamma \Delta t \frac{m_e}{m_D} h_D^{n+1} \left(\Phi_{K_e}^{n+1} - \Phi_K^{n+1} \right) \vec{n}_{e,K} \quad , \quad \gamma > 0 \, . \tag{7}$$

As detailed in [11], the quantity (7) can be seen as the discrete counterpart of a diffusive flux correction governed by the pressure gradient, shown to bring energy dissipation at the continuous level for regular solutions. At the level of an interface e (or equivalently a dual cell $D = \mathbb{T}^*(K, K_e)$), following [3], the auxiliary water height h_D^{n+1} is given by:

$$m_D h_D^{n+1} = m_{D^K} h_K^{n+1} + m_{D^{K_e}} h_{K_e}^{n+1} \, , \tag{8}$$

and we assume the following discrete conservation law, in view of derivation of kinetic energy estimates:

$$h_D^{n+1} - h_D^n = - \frac{\Delta t}{m_D} \sum_{f \in \partial D} \mathcal{F}_f^{n+1} . \vec{n}_{f,D} m_f \, . \tag{9}$$

The auxiliary mass fluxes $\mathcal{F}_f^{n+1} . \vec{n}_{f,D}$ are derived from those expressed on the primal mesh, with varying coefficients depending on the staggered discretization (see [3] for details). We employ the following definition of the discrete gradient on $\mathbb{T}^*$, ensuring the *grad/div* duality (see [28]):

$$\nabla_D \Phi^{n+1} = \frac{m_e}{m_D} \left(\Phi_{K_e}^{n+1} - \Phi_K^{n+1} \right) \vec{n}_{e,K} \, . \tag{10}$$

Finally, manipulating the mass and discharge equations (4), (5), we recover the discrete evolution of the velocity:

$$\mathbf{u}_D^{n+1} = \mathbf{u}_D^n - \frac{\Delta t}{m_D} \sum_{f \in \partial D} \left(\frac{\mathbf{u}_{D_f}^n - \mathbf{u}_D^n}{h_D^{n+1}} \right) \left(\mathcal{F}_f^{n+1} . \vec{n}_{f,D} \right)^- m_f - \Delta t \left(\nabla_D \Phi^{n+1} \right) . \tag{11}$$

3.1 Control of the mechanical energy

The objective pursued in the present section is to demonstrate the scheme's stability in the sense of discrete total energy decrease. Due to the staggered geometry, the system's energy at time t^n should be expressed as a combination of local potential and kinetic energy contributions, computed on the primal and dual meshes respectively:

$$E^n = \sum_{K \in \mathbb{T}} m_K \mathcal{E}_K^n + \sum_{D \in \mathbb{T}^*} m_D \mathcal{K}_D^n .$$

The time evolution of kinetik and potential energies is respectively examined in Lemma 1 and Lemma 2. The corresponding local estimates are subsequently exploited to get non-linear stability in Theorem 1. In what follows, we will assume the following time-step restriction:

$$\frac{\Delta t}{m_D} \sum_{f \in \partial D} - \left(\mathcal{F}_f^{n+1} . \vec{n}_{f,D} \right)^- \leq \frac{1}{2} h_D^{n+1} . \tag{12}$$

Lemma 1. *Estimation of the kinetic energy production. We have the following inequality :*

$$\mathcal{K}_D^{n+1} - \mathcal{K}_D^n + \frac{\Delta t}{m_D} \sum_{f \in \partial D} (\mathcal{G}_{K,f}^{n+1} . \vec{n}_{f,D}) m_f \leq \mathcal{R}_{\mathcal{K},D} - \mathcal{Q}_{\mathcal{K},D} , \tag{13}$$

with

$$\begin{aligned} \mathcal{G}_{K,f}^{n+1} . \vec{n}_{f,D} &= \frac{1}{2} \|\mathbf{u}_D^n\|^2 \left(\mathcal{F}_f^{n+1} . \vec{n}_{f,D} \right)^+ + \frac{1}{2} \|\mathbf{u}_{D_f}^n\|^2 \left(\mathcal{F}_f^{n+1} . \vec{n}_{f,D} \right)^- , \\ \mathcal{Q}_{\mathcal{K},D} &= \Delta t \, h_D^{n+1} \mathbf{u}_D^n . \left(\nabla_D \Phi^{n+1} \right) , \\ \mathcal{R}_{\mathcal{K},D} &= (\Delta t)^2 h_D^{n+1} \|\nabla_D \Phi^{n+1}\|^2 . \end{aligned}$$

Proof. The proof is the same as in [17] in the one-layer case and we refer to this paper for further details. We just give here the main steps to help in understanding the links between the constant γ, which governs the scheme's stabillity through (7), and the CFL condition (see Remark 1). After some basic computations, we reach the following equality:

$$K_D^{n+1} - K_D^n + \frac{\Delta t}{m_D} \sum_{f \in \partial D} (\mathcal{G}_{K,f}^{n+1} . \vec{n}_{f,D}) m_f = S_D - \mathcal{Q}_{\mathcal{K},D} ,$$

where

$$S_D = \frac{1}{2} h_D^{n+1} \|\mathbf{u}_D^{n+1} - \mathbf{u}_D^n\|^2 + \frac{\Delta t}{m_D} \sum_{f \in \partial D} \frac{1}{2} \|\mathbf{u}_D^n - \mathbf{u}_{D_f}^n\|^2 \left(\mathcal{F}_f^{n+1} . \vec{n}_{f,D} \right)^- .$$

With the use of Jensen's inequality we have:

$$\begin{aligned} \frac{1}{2} h_D^{n+1} \left\| \mathbf{u}_D^{n+1} - \mathbf{u}_D^n \right\|^2 \leq \quad & h_D^{n+1} (\Delta t)^2 \left\| \nabla_D \Phi^{n+1} \right\|^2 \\ & + \frac{1}{h_D^{n+1}} \left(\frac{\Delta t}{m_D} \right)^2 \left\| \sum_{f \in \partial D} (\mathbf{u}_{D_f}^n - \mathbf{u}_D^n) \left(\mathcal{F}_f^{n+1} . \vec{n}_{f,D} \right)^- m_f \right\|^2 . \end{aligned} \tag{14}$$

Then, Cauchy-Schwarz inequality gives:

$$\begin{aligned} & \left\| \sum_{f \in \partial D} (\mathbf{u}_{D_f}^n - \mathbf{u}_D^n) \left(\mathcal{F}_f^{n+1} . \vec{n}_{f,D} \right)^- m_f \right\|^2 \\ & \qquad \leq \left(\sum_{f \in \partial D} \left\| \mathbf{u}_{D_f}^n - \mathbf{u}_D^n \right\|^2 \left(\mathcal{F}_f^{n+1} . \vec{n}_{f,D} \right)^- m_f \right) \left(\sum_{f \in \partial D} \left(\mathcal{F}_f^{n+1} . \vec{n}_{f,D} \right)^- m_f \right) , \end{aligned}$$

and we get:

$$S_D \leq h_D^{n+1}(\Delta t)^2 \left\|\nabla_D \Phi^{n+1}\right\|^2 + \left(\frac{\Delta t}{m_D}\sum_{f\in\partial D}\left\|\mathbf{u}_{D_f}^n - \mathbf{u}_D^n\right\|^2 \left(\mathcal{F}_f^{n+1}.\vec{n}_{f,D}\right)^- m_f\right)\left(\frac{1}{2} - \frac{\Delta t}{m_D}\sum_{f\in\partial D}\frac{-\left(\mathcal{F}_f^{n+1}.\vec{n}_{f,D}\right)^-}{h_D^{n+1}}m_f\right).$$

Under (12), the second term of the right hand side is negative and we get the annouced result. □

Remark 1. *In the proof above we need to estimate the quantity:*

$$\frac{1}{2}h_D^{n+1}\left\|\mathbf{u}_D^{n+1} - \mathbf{u}_D^n\right\|^2.$$

In (14), we classically used Jensen's inequality with the weights (1/2, 1/2), which at last gave the estimation (13), provided the robustness assumption (12). As a matter of fact, a more general result can be obtained introducing arbitrary weights in this inequality. By doing so we get:

$$\mathcal{K}_D^{n+1} - \mathcal{K}_D^n + \frac{\Delta t}{m_D}\sum_{f\in\partial D}\left(\mathcal{G}_{K,f}^{n+1}.\vec{n}_{f,D}\right)m_f \leq \mathcal{R}_{\mathcal{K},D}^{\lambda} - \mathcal{Q}_{\mathcal{K},D}\,, \tag{15}$$

with this time:

$$\mathcal{R}_{\mathcal{K},D}^{\lambda} = \frac{1}{2\lambda}(\Delta t)^2 h_D^{n+1}\|\nabla_D\Phi^{n+1}\|^2\,. \tag{16}$$

where $\lambda \in]0,1[$. The corresponding time step restriction is

$$\frac{\Delta t}{m_D}\sum_{f\in\partial D} -\left(\mathcal{F}_f^{n+1}.\vec{n}_{f,D}\right)^- m_f \leq \mu h_D^{n+1} \quad , \quad \text{with } \lambda+\mu = 1\,.$$

As will be detailed in Remark 3, this more general formulation can be used to relax the stability condition on γ.

Lemma 2. *Estimation of the potential energy production. We have the following inequality :*

$$\mathcal{E}_K^{n+1} - \mathcal{E}_K^n + \frac{\Delta t}{m_K}\sum_{e\in\partial K}\left(\mathcal{G}_{\mathcal{E},e}^{n+1}.\vec{n}_{e,K}\right)m_e \leq \mathcal{Q}_{\mathcal{E},K} - \mathcal{R}_{\mathcal{E},K}\,, \tag{17}$$

with

$$\begin{aligned}
\mathcal{G}_{\mathcal{E},e}^{n+1}.\vec{n}_{e,K} &= \Phi_K^{n+1}\left(h_K^{n+1}(\mathbf{u}_D^n.\vec{n}_{e,K})^+ - (\Pi_{e,K}^{n+1}.\vec{n}_{e,K})^+\right)\\
&\quad + \Phi_{K_e}^{n+1}\left(h_{K_e}^{n+1}(\mathbf{u}_D^n.\vec{n}_{e,K})^- - (\Pi_{e,K}^{n+1}.\vec{n}_{e,K})^-\right),\\
\mathcal{Q}_{\mathcal{E},K} &= \frac{\Delta t}{m_K}\sum_{e\in\partial K} h_{K_e}^{n+1}(\mathbf{u}_D^n.\vec{n}_{e,K})^-\left(\Phi_{K_e}^{n+1} - \Phi_K^{n+1}\right)m_e\,,\\
\mathcal{R}_{\mathcal{E},K} &= \frac{\Delta t}{m_K}\sum_{e\in\partial K}(\Pi_{e,K}^{n+1}.\vec{n}_{e,K})^-\left(\Phi_{K_e}^{n+1} - \Phi_K^{n+1}\right)m_e\,.
\end{aligned}$$

Proof. Using the equality $\mathcal{E}_K^{n+1} - \mathcal{E}_K^n = \Phi_K^{n+1}\left(h_K^{n+1} - h_K^n\right) - \frac{1}{2}g\left(h_K^{n+1} - h_K^n\right)^2$, we immediately have:

$$\mathcal{E}_K^{n+1} - \mathcal{E}_K^n \leq \Phi_K^{n+1}\left(h_K^{n+1} - h_K^n\right) = -\frac{\Delta t}{m_K}\Phi_K^{n+1}\sum_{e\in\partial K}\mathcal{F}_{e,K}^{n+1}.\vec{n}_{e,K}m_e\,.$$

Then, using the fact that:

$$\Phi_K^{n+1}\mathcal{F}_{e,K}^{n+1}.\vec{n}_{e,K} = \mathcal{G}_{\mathcal{E},e}^{n+1}.\vec{n}_{e,K} - h_{K_e}^{n+1}(\mathbf{u}_D^n.\vec{n}_{e,K})^-\left(\Phi_{K_e}^{n+1} - \Phi_K^{n+1}\right) + (\Pi_{e,K}^{n+1}.\vec{n}_{e,K})^-\left(\Phi_{K_e}^{n+1} - \Phi_K^{n+1}\right),$$

we get the announced result. □

As a remark, one can already note that the estimates (13) and (17) do not involve additionnal antisymmetric terms, as is the case in the collocated frame ([11, 45]). We now combine these two results to establish the main property of the current section:

Theorem 1. *We consider the scheme (4), (5), together with the time-step condition (12). We have:*

$$\begin{aligned} E^{n+1} - E^n \leq & \sum_{D\in\mathbb{T}^*} (\Delta t)^2 h_D^{n+1} \frac{(m_e)^2}{m_D} \left(\Phi_{K_e}^{n+1} - \Phi_K^{n+1}\right)^2 (1-\gamma) \\ & + \sum_{D\in\mathbb{T}^*} \Delta t \frac{m_e}{m_D} (\Phi_{K_e}^{n+1} - \Phi_K^{n+1})(h_{K_e}^{n+1} - h_K^{n+1}) \left[m_{D^K} (\mathbf{u}_D^n.\vec{n}_{e,K})^- - m_{D^{K_e}} (\mathbf{u}_D^n.\vec{n}_{e,K})^+\right] . \end{aligned} \tag{18}$$

Proof. Gathering (13) and (17), we have

$$\begin{aligned} E^{n+1} - E^n &= \sum_{K\in\mathbb{T}} m_K \left(\mathcal{E}_K^{n+1} - \mathcal{E}_K^n\right) + \sum_{D\in\mathbb{T}^*} m_D \left(\mathcal{K}_D^{n+1} - \mathcal{K}_D^n\right) \\ &\leq \sum_{K\in\mathbb{T}} m_K \left(\mathcal{Q}_{\mathcal{E},K} - \mathcal{R}_{\mathcal{E},K}\right) + \sum_{D\in\mathbb{T}^*} m_D \left(\mathcal{R}_{\mathcal{K},D} - \mathcal{Q}_{\mathcal{K},D}\right) . \end{aligned}$$

The previous inequality is then adressed locally, separating the velocity and pressure regularizations as below.

- Terms $\mathcal{Q}_\mathcal{K}$ and $\mathcal{Q}_\mathcal{E}$: At the level of a dual element $D = \mathbb{T}^*(K, K_e)$, the kinetic part brings

$$- m_D \Delta t h_D^{n+1} \mathbf{u}_D^n.(\nabla_D \Phi^{n+1}) = -m_e \Delta t h_D^{n+1} (\Phi_{K_e}^{n+1} - \Phi_K^{n+1}) \mathbf{u}_D^n.\vec{n}_{e,K} . \tag{19}$$

  As concerns the potential part, we get two contributions, coming from each side of the corresponding interface e:

$$\Delta t \big(h_{K_e}^{n+1} (\mathbf{u}_D^n.\vec{n}_{e,K})^- \left(\Phi_{K_e}^{n+1} - \Phi_K^{n+1}\right) m_e + h_K^{n+1} (\mathbf{u}_D^n.\vec{n}_{e,K_e})^- \left(\Phi_K^{n+1} - \Phi_{K_e}^{n+1}\right) m_e\big) . \tag{20}$$

  Summing (19) and (20), and remarking that $(\mathbf{u}_D^n.\vec{n}_{e,K_e})^- = -(\mathbf{u}_D^n.\vec{n}_{e,K})^+$ we get:

$$\Delta t m_e \left(\Phi_{K_e}^{n+1} - \Phi_K^{n+1}\right) \left[(h_{K_e}^{n+1} - h_D^{n+1})(\mathbf{u}_D^n.\vec{n}_{e,K})^- + (h_K^{n+1} - h_D^{n+1})(\mathbf{u}_D^n.\vec{n}_{e,K})^+\right] .$$

  We then use $h_{K_e}^{n+1} - h_D^{n+1} = \dfrac{m_{D^K}}{m_D}(h_{K_e}^{n+1} - h_K^{n+1})$ and $h_K^{n+1} - h_D^{n+1} = \dfrac{m_{D^{K_e}}}{m_D}(h_K^{n+1} - h_{K_e}^{n+1})$ to rewrite the previous term as:

$$\Delta t \frac{m_e}{m_D} \left(\Phi_{K_e}^{n+1} - \Phi_K^{n+1}\right) (h_{K_e}^{n+1} - h_K^{n+1}) \left[m_{D^K} (\mathbf{u}_D^n.\vec{n}_{e,K})^- - m_{D^{K_e}} (\mathbf{u}_D^n.\vec{n}_{e,K})^+\right] . \tag{21}$$

  That way we recover the second term of the right hand side of (18).

- Terms $\mathcal{R}_\mathcal{K}$ and $\mathcal{R}_\mathcal{E}$: Again, we focus on a dual element $D = \mathbb{T}^*(K, K_e)$. The terms issuing from the potential energy estimate (17) are:

$$\begin{aligned} -\Delta t (\Pi_{e,K}^{n+1}.\vec{n}_{e,K})^- \left(\Phi_{K_e}^{n+1} - \Phi_K^{n+1}\right) m_e - \Delta t (\Pi_{e,K}^{n+1}.\vec{n}_{e,K_e})^- \left(\Phi_K^{n+1} - \Phi_{K_e}^{n+1}\right) m_e \\ = -\Delta t (\Pi_{e,K}^{n+1}.\vec{n}_{e,K}) \left(\Phi_{K_e}^{n+1} - \Phi_K^{n+1}\right) m_e . \end{aligned}$$

  Then, adding the kinetic residual $\mathcal{R}_{\mathcal{K},D} = (\Delta t)^2 h_D^{n+1} \|\nabla_D \Phi^{n+1}\|^2$ and using the expression of the discrete gradient (10), we obtain:

$$(\Delta t)^2 h_D^{n+1} \frac{(m_e)^2}{m_D} \left(\Phi_{K_e}^{n+1} - \Phi_K^{n+1}\right)^2 (1-\gamma) , \tag{22}$$

  which is nothing but the local contribution of the first stabilizing term implied in (18).

□

Remark 2. *In the case of flat bottoms, (18) becomes:*

$$\begin{aligned} E^{n+1} - E^n \leq & \sum_{D\in\mathbb{T}^*} (\Delta t)^2 h_D^{n+1} \frac{m_e^2}{m_D} \left(\Phi_{K_e}^{n+1} - \Phi_K^{n+1}\right)^2 (1-\gamma) \\ & + \sum_{D\in\mathbb{T}^*} \Delta t \frac{m_e}{m_D} g \left(h_{K_e}^{n+1} - h_K^{n+1}\right)^2 \left[m_{D^K} (\mathbf{u}_D^n.\vec{n}_{e,K})^- - m_{D^{K_e}} (\mathbf{u}_D^n.\vec{n}_{e,K})^+\right] , \end{aligned} \tag{23}$$

and the condition $\gamma \geq 1$ provides the expected total energy decrease. In the presence of topography, we can introduce the following correction in the mass fluxes (6):

$$\begin{aligned} h_K^{n+1} &\longleftarrow h_K^{n+1} + \frac{m_{D^K}}{m_D}\left(z_K - z_{K_e}\right), \\ h_{K_e}^{n+1} &\longleftarrow h_{K_e}^{n+1} + \frac{m_{D^{K_e}}}{m_D}\left(z_{K_e} - z_K\right), \end{aligned}$$

to write:

$$\begin{aligned} E^{n+1} - E^n \leq & \sum_{D\in\mathbb{T}^*} (\Delta t)^2 h_D^{n+1} \frac{(m_e)^2}{m_D}\left(\Phi_{K_e}^{n+1} - \Phi_K^{n+1}\right)^2 (1-\gamma) \\ & + \sum_{D\in\mathbb{T}^*} \Delta t \frac{m_e}{m_D}\frac{1}{g}\left(\Phi_{K_e}^{n+1} - \Phi_K^{n+1}\right)^2 \left[m_{D^K}(\mathbf{u}_D^n.\vec{n}_{e,K})^- - m_{D^{K_e}}(\mathbf{u}_D^n.\vec{n}_{e,K})^+\right] \leq 0\,. \end{aligned} \tag{24}$$

Remark 3. *In this context, a sufficient condition to get a global energy decrease is given by $\gamma \geq 1$. As shortly discussed in Remark 1, this condition can be relaxed using the general form of the residual (16). Then (22) becomes:*

$$(\Delta t)^2 h_D^{n+1} \frac{(m_e)^2}{m_D}\left(\Phi_{K_e}^{n+1} - \Phi_K^{n+1}\right)^2 \left(\frac{1}{2\lambda} - \gamma\right), \tag{25}$$

In the case of low-Froude number regimes, the first term of the right hand side of (14) is dominant, and the parameter λ can be chosen close to 1 in practise. We thus obtain $\gamma_0 = 1/2$ as a strict stability threshold and the indication that γ can be taken in the vicinity of $1/2$ in these contexts. These conclusions have been confirmed throughout our numerical validations.

Remark 4. *The modification of the mass fluxes (6):*

$$\begin{aligned} h_K^{n+1} &\longleftarrow h_D^{n+1}, \\ h_{K_e}^{n+1} &\longleftarrow h_D^{n+1}, \end{aligned}$$

allows to recover the centred scheme [17], with the following global estimate:

$$E^{n+1} - E^n \leq \sum_{D\in\mathbb{T}^*} (\Delta t)^2 h_D^{n+1} \frac{m_e^2}{m_D}\left(\Phi_{K_e}^{n+1} - \Phi_K^{n+1}\right)^2 (1-\gamma)\,.$$

As can be seen when compared with (23) or (24), the upwind treatment of the mass fluxes brings another stabilizing term in the energy budget, independent of γ. If this quantity is generally insignificant for large scale flows, it is no longer the case for advection-dominated regimes, that are typically encountered in river flows or at the vicinity of coasts for instance. Combined with the robustness property, (see next Section §3.2) this term is expected to help in stabilizing the computations, notably at the level of dry fronts.

3.2 Positivity and well-balanced properties

Here we give a short focus on robustness and stability results. As concerns the positivity preserving features of the scheme, the existence of a space step ensuring (12) is given in [45], based on continuity arguments. Note however that, according to (4) and (6), the upwind formalism leads to the resolution of a linear system involving a monotone matrix with dominant diagonal terms $1 + \dfrac{\Delta t}{m_K}\displaystyle\sum_{e\in\partial K}(\mathbf{u}_D^n.\vec{n}_{e,K})^+ m_e$, allowing to ensure the positivity of h_K^{n+1} without heavy technical difficulties.

Proposition 1. *The numerical scheme (4), (5) preserves the lake at rest configurations:*

$$\mathbf{u}_D^n = 0 \quad , \quad \mathbf{\Phi}_K^n = cst\,. \tag{26}$$

Proof. Assuming (26), the advective term of the fluxes vanishes and (4), (5) becomes:

$$\begin{cases} h_K^{n+1} & = h_K^n + \dfrac{\Delta t}{m_K}\displaystyle\sum_{e\in\partial K}\gamma\Delta t\frac{(m_e)^2}{m_D} h_D^{n+1}\left(\Phi_{K_e}^{n+1} - \Phi_K^{n+1}\right) \quad , \quad K\in\mathbb{T} \\ h_D^{n+1}\mathbf{u}_D^{n+1} & = -\Delta t h_D^{n+1}\left(\nabla_D \Phi^{n+1}\right) \quad , \quad D\in\mathbb{T}^*\,. \end{cases}$$

The first equation implies $h_K^{n+1} = h_K^n$ for all $K\in\mathbb{T}$, from which we immediately deduce $\Phi_K^{n+1} = \Phi_{K_e}^{n+1}$ and then $\mathbf{u}_D^{n+1} = 0$. □

4 Explicit scheme

We now aim at extending the present approach in a fully explicit environment. This task may not be carried out in a straighforward manner, since the introduction of a shifted velocity in the mass fluxes is no longer sufficient to guarantee stability. However, as shown in the collocated frame [11], this lack of regularity can be compensated with a slight modification of the discrete pressure gradient. Following these ideas, and keeping the notations of the previous section, the numerical scheme we consider is expressed as follows:

$$h_K^{n+1} = h_K^n - \frac{\Delta t}{m_K}\sum_{e\in\partial K}\mathcal{F}_{e,K}^n.\vec{n}_{e,K}m_e \quad , \quad K\in\mathbb{T} \tag{27}$$

$$\begin{aligned} h_D^{n+1}\mathbf{u}_D^{n+1} = {} & h_D^n\mathbf{u}_D^n - \frac{\Delta t}{m_D}\sum_{f\in\partial D}\left(\mathbf{u}_D^n\left(\mathcal{F}_f^n.\vec{n}_{f,D}\right)^+ + \mathbf{u}_{D_f}^n\left(\mathcal{F}_f^n.\vec{n}_{f,D}\right)^-\right)m_f \\ & -\Delta t h_D^n\left(\nabla_D\Phi^{n,*}\right) \quad , \quad D\in\mathbb{T}^*\,. \end{aligned} \tag{28}$$

We consider the following mass fluxes:

$$\mathcal{F}_{e,K}^n.\vec{n}_{e,K} = (h\mathbf{u})_D^n.\vec{n}_{e,K} - \Pi_{e,K}^n.\vec{n}_{e,K}\,, \tag{29}$$

where the notation $(h\mathbf{u})_D^n = h_D^n\mathbf{u}_D^n$ is now used for simplification purposes. Following the semi-implicit case, we set:

$$\Pi_{e,K}^n = \gamma\Delta t\frac{m_e}{m_D}h_D^n\left(\Phi_{K_e}^n - \Phi_K^n\right)\vec{n}_{e,K} \quad , \quad \gamma>0\,. \tag{30}$$

As previously discussed, we need the stabilizing virtues of a particular discrete pressure gradient:

$$\nabla_D\Phi^{n,*} = \frac{m_e}{m_D}\left(\Phi_{K_e}^{n,*} - \Phi_K^{n,*}\right)\vec{n}_{e,K}\,, \tag{31}$$

where

$$\Phi_K^{n,*} = \Phi_K^n - \Lambda_{K,D}^n \quad , \quad \Lambda_{K,D}^n = 2\alpha g\Delta t\frac{m_{\partial K}}{m_K}((h\mathbf{u})_D^n - (h\mathbf{u})_K^n).\vec{n}_{e,K}\,, \quad \alpha>0\,. \tag{32}$$

In fact, this choice also finds motivations at the continuous level, since it can be shown that the continuous counterpart of (32), expressed in terms of the discharge divergence, brings energy dissipation. In the expression above, $(h\mathbf{u})_K^n$ stands for an interpolated discharge on the cell K depending on the surrounding elements, designed to obtain entropy stability. This quantity will be defined later on (see proof of Theorem 2, formula (48)). The auxiliary water height is now given by:

$$h_D^{n+1} - h_D^n = -\frac{\Delta t}{m_D}\sum_{f\in\partial D}\mathcal{F}_f^n.\vec{n}_{f,D}m_f\,,$$

and the resulting velocity scheme is:

$$\mathbf{u}_D^{n+1} = \mathbf{u}_D^n - \frac{\Delta t}{m_D}\sum_{f\in\partial D}\left(\frac{\mathbf{u}_{D_f}^n - \mathbf{u}_D^n}{h_D^{n+1}}\right)\left(\mathcal{F}_f^n.\vec{n}_{f,D}\right)^- m_f - \Delta t\frac{h_D^n}{h_D^{n+1}}\left(\nabla_D\Phi^{n,*}\right)\,. \tag{33}$$

Note that the collocated and staggered methods proposed in [5] in the linear case for the Shallow Water equations with Coriolis force share some similarities with the present approach. In particular, regularizing terms of the same nature are used to get linear stability.

4.1 Control of the mechanical energy

This section will follow the same lines as the semi-implicit case. We first derive local inequalities for the kinetic and potential energies (Lemma 3 and Lemma 4), and finally exhibit sufficient conditions for the constants α, γ to get non-linear stability. In what follows we will consider the following advective time-step condition:

$$\frac{\Delta t}{m_D}\sum_{f\in\partial D}\left|\mathcal{F}_f^n.\vec{n}_{f,D}\right| < \beta h_D^{n+1}\,, \tag{34}$$

with $0<\beta\leq 1$. We have the following result:

Lemma 3. *Estimation of the kinetic energy production. We have the following inequality :*

$$\mathcal{K}_D^{n+1} - \mathcal{K}_D^n + \frac{\Delta t}{m_D} \sum_{f \in \partial D} \left(\mathcal{G}_{K,f}^n . \vec{n}_{f,D}\right) m_f + \mathcal{Q}_{\mathcal{K},D} \leq -\mathcal{S}_{\Lambda,D} + \mathcal{R}_{\mathcal{K},D} + \mathcal{A}_{\mathcal{K},D} \,, \tag{35}$$

with

$$\begin{aligned}
\mathcal{G}_{K,f}^n . \vec{n}_{f,D} &= \frac{1}{2} \|\mathbf{u}_D^n\|^2 \left(\mathcal{F}_f^n . \vec{n}_{f,D}\right)^+ + \frac{1}{2} \|\mathbf{u}_{D_f}^n\|^2 \left(\mathcal{F}_f^n . \vec{n}_{f,D}\right)^- \,, \\
\mathcal{Q}_{\mathcal{K},D} &= \Delta t \, (h\mathbf{u})_D^n . \left(\nabla_D \Phi^n\right) \,, \\
\mathcal{S}_{\Lambda,D} &= \Delta t \left(\frac{m_e}{m_D}\right) (h\mathbf{u})_D^n . \vec{n}_{e,K} \left(\Lambda_{K,D}^n - \Lambda_{K_e,D}^n\right) \,, \\
\mathcal{R}_{\mathcal{K},D} &= 2(\Delta t)^2 h_D^n \|\nabla_D \Phi^n\|^2 \,, \\
\mathcal{A}_{\mathcal{K},D} &= 2(\Delta t)^2 \left(\frac{m_e}{m_D}\right)^2 h_D^n \Big[\left(\Lambda_{K,D}^n\right)^2 + \left(\Lambda_{K_e,D}^n\right)^2 \Big] \,.
\end{aligned}$$

Proof. The first steps of the proof are the same as in the previous section. Skipping computation details, we start with:

$$K_D^{n+1} - K_D^n + \frac{\Delta t}{m_D} \sum_{f \in \partial D} \left(\mathcal{G}_{K,f}^n . \vec{n}_{f,D}\right) m_f = S_D - \Delta t (h\mathbf{u})_D^n . \left(\nabla_D \Phi^{n,*}\right) \,, \tag{36}$$

where

$$S_D = \frac{1}{2} h_D^{n+1} \left\|\mathbf{u}_D^{n+1} - \mathbf{u}_D^n\right\|^2 + \frac{\Delta t}{m_D} \sum_{f \in \partial D} \frac{1}{2} \left\|\mathbf{u}_{D_f}^n - \mathbf{u}_D^n\right\|^2 \left(\mathcal{F}_f^n . \vec{n}_{f,D}\right)^- m_f \,.$$

Also, note that under (34), we have:

$$\frac{h_D^n}{h_D^{n+1}} < 1 + \beta \,, \tag{37}$$

For the sake of clarity, we reformulate the velocity scheme (33) as $\mathbf{u}_D^{n+1} - \mathbf{u}_D^n = x\dfrac{\mathbf{a}}{x} + y\dfrac{\mathbf{b}}{y} + z\dfrac{\mathbf{c}}{z}$, with

$$\left\{
\begin{array}{lll}
\mathbf{a} & = -\dfrac{\Delta t}{m_D} \displaystyle\sum_{f \in \partial D} \left(\frac{\mathbf{u}_{D_f}^n - \mathbf{u}_D^n}{h_D^{n+1}}\right) \left(\mathcal{F}_f^n . \vec{n}_{f,D}\right)^- m_f \,, \\
\mathbf{b} & = -\Delta t \dfrac{h_D^n}{h_D^{n+1}} \left(\nabla_D \Phi^n\right) \,, \\
\mathbf{c} & = \Delta t \dfrac{h_D^n}{h_D^{n+1}} \left(\dfrac{m_e}{m_D}\right) \left[\Lambda_{K_e,D}^n - \Lambda_{K,D}^n\right] \vec{n}_{e,K} \,,
\end{array}
\right.$$

with the fixed weights

$$x = \frac{1 - 3\beta}{4} \quad , \quad y = \frac{1 + \beta}{4} \quad , \quad z = \frac{1 + \beta}{2} \,. \tag{38}$$

As in the semi-implicit case, other choice are naturally possible, leading to different sufficient conditions for the constants α and γ (see Section 4.2). Using Jensen's inequality, (37) and chosing β sufficiently small to have $\dfrac{1}{2x} \leq \dfrac{1}{2\beta}$ (in the present case we need $\beta \leq 1/7$) , we get:

$$\begin{aligned}
\frac{1}{2} h_D^{n+1} \left\|\mathbf{u}_D^{n+1} - \mathbf{u}_D^n\right\|^2 \quad &\leq \frac{1}{2\beta} \frac{1}{h_D^{n+1}} \left(\frac{\Delta t}{m_D}\right)^2 \left\| \sum_{f \in \partial D} \left(\mathbf{u}_{D_f}^n - \mathbf{u}_D^n\right) \left(\mathcal{F}_f^n . \vec{n}_{f,D}\right)^- m_f \right\|^2 \\
&\quad + 2(\Delta t)^2 h_D^n \left\|\nabla_D \Phi^n\right\|^2 + (\Delta t)^2 h_D^n \left(\frac{m_e}{m_D}\right)^2 \left[\Lambda_{K_e,D}^n - \Lambda_{K,D}^n\right]^2 .
\end{aligned}$$

With Cauchy Schwarz inequality we have:

$$\left\| \sum_{f \in \partial D} \left(\mathbf{u}_{D_f}^n - \mathbf{u}_D^n\right) \left(\mathcal{F}_f^n . \vec{n}_{f,D}\right)^- m_f \right\|^2 \leq \Big(\sum_{f \in \partial D} \left\|\mathbf{u}_{D_f}^n - \mathbf{u}_D^n\right\|^2 \left(\mathcal{F}_f^n . \vec{n}_{f,D}\right)^- m_f \Big) \Big(\sum_{f \in \partial D} \left(\mathcal{F}_f^n . \vec{n}_{f,D}\right)^- m_f \Big) \,,$$

and write:

$$S_D \leq \left(\frac{\Delta t}{m_D}\right)\Big(\sum_{f\in\partial D}\left\|\mathbf{u}_{D_f}^n-\mathbf{u}_D^n\right\|^2\left(\mathcal{F}_f^n.\vec{n}_{f,D}\right)^-m_f\Big)\left[\frac{1}{2}-\frac{1}{2\beta}\left(\frac{\Delta t}{m_D}\right)\frac{1}{h_D^{n+1}}\sum_{f\in\partial D}-\left(\mathcal{F}_f^n.\vec{n}_{f,D}\right)^-m_f\right]$$
$$+2(\Delta t)^2h_D^n\left\|\nabla_D\Phi^n\right\|^2+(\Delta t)^2h_D^n\left(\frac{m_e}{m_D}\right)^2\left[\Lambda_{K,D}^n-\Lambda_{K_e,D}^n\right]^2.$$

Using Jensen's inequality on the last term of the previous estimation, the time step restriction (34), and returning to (36), we obtain the announced result. □

Lemma 4. *Estimation of the potential energy production. We have the following inequality :*

$$\mathcal{E}_K^{n+1}-\mathcal{E}_K^n+\frac{\Delta t}{m_K}\sum_{e\in\partial K}\left(\mathcal{G}_{\mathcal{E},e}^n.\vec{n}_{e,K}\right)m_e-\mathcal{Q}_{\mathcal{E},K}\leq-\mathcal{S}_{\Phi,K}+\mathcal{R}_{\mathcal{E},K}+\mathcal{A}_{\mathcal{E},K}\,, \tag{39}$$

with

$$\mathcal{G}_{\mathcal{E},e}^n.\vec{n}_{e,K}=\Phi_e^n\mathcal{F}_e^n.\vec{n}_{e,K}\quad,\quad\Phi_e^n=\frac{1}{2}\left(\Phi_K^n+\Phi_{K_e}^n\right),$$
$$\mathcal{Q}_{\mathcal{E},K}=\frac{1}{2}\frac{\Delta t}{m_K}\sum_{e\in\partial K}(h\mathbf{u})_D^n.\left(\Phi_{K_e}^n-\Phi_K^n\right)\vec{n}_{e,K}m_e\,,$$
$$\mathcal{S}_{\Phi,K}=\frac{1}{2}\frac{\Delta t}{m_K}\sum_{e\in\partial K}\Pi_{e,K}^n.\left(\Phi_{K_e}^n-\Phi_K^n\right)\vec{n}_{e,K}m_e\,,$$
$$\mathcal{R}_{\mathcal{E},K}=g\left(\frac{\Delta t}{m_K}\right)^2m_{\partial K}\sum_{e\in\partial K}(\Pi_{e,K}^n.\vec{n}_{e,K})^2m_e\,,$$
$$\mathcal{A}_{\mathcal{E},K}=g\left(\frac{\Delta t}{m_K}\right)^2m_{\partial K}\sum_{e\in\partial K}\Big[\left((h\mathbf{u})_D^n-(h\mathbf{u})_K^n\right).\vec{n}_{e,K}\Big]^2m_e\,.$$

Proof. We write:

$$\mathcal{E}_K^{n+1}-\mathcal{E}_K^n=\Phi_K^n\left(h_K^{n+1}-h_K^n\right)+\frac{1}{2}g\left(h_K^{n+1}-h_K^n\right)^2.$$

The mass equation (27) gives $\Phi_K^n\left(h_K^{n+1}-h_K^n\right)=-\dfrac{\Delta t}{m_K}\sum_{e\in\partial K}\Phi_K^n\mathcal{F}_e^n.\vec{n}_{e,K}m_e$. We then write:

$$\Phi_K^n\mathcal{F}_e^n.\vec{n}_{e,K}=\Phi_e^n\mathcal{F}_e^n.\vec{n}_{e,K}-\frac{1}{2}\left(\Phi_{K_e}^n-\Phi_K^n\right)\mathcal{F}_e^n.\vec{n}_{e,K}$$
$$=\mathcal{G}_{\mathcal{E},e}^n.\vec{n}_{e,K}-\frac{1}{2}(h\mathbf{u})_D^n.\left(\Phi_{K_e}^n-\Phi_K^n\right)\vec{n}_{e,K}+\frac{1}{2}\Pi_{e,K}^n.\left(\Phi_{K_e}^n-\Phi_K^n\right)\vec{n}_{e,K}\,.$$

We thus obtain the potential fluxes and the terms $\mathcal{S}_{\Phi,K}$ and $\mathcal{Q}_{\mathcal{E},K}$. On the other hand, according to discrete Green formula, we can artificially introduce the quantity $(h\mathbf{u})_K^n$ in the mass equation as follows:

$$h_K^{n+1}-h_K^n=-\frac{\Delta t}{m_K}\sum_{e\in\partial K}\left((h\mathbf{u})_D^n-(h\mathbf{u})_K^n\right).\vec{n}_{e,K}m_e+\frac{\Delta t}{m_K}\sum_{e\in\partial K}\Pi_{e,K}^n.\vec{n}_{e,K}m_e\,.$$

Then, Jensen's inequality gives:

$$\frac{1}{2}g\left(h_K^{n+1}-h_K^n\right)^2\leq g\left(\frac{\Delta t}{m_K}\right)^2\Big(\sum_{e\in\partial K}\left((h\mathbf{u})_D^n-(h\mathbf{u})_K^n\right).\vec{n}_{e,K}m_e\Big)^2+g\left(\frac{\Delta t}{m_K}\right)^2\Big(\sum_{e\in\partial K}\Pi_{e,K}^n.\vec{n}_{e,K}m_e\Big)^2$$
$$\leq g\left(\frac{\Delta t}{m_K}\right)^2m_{\partial K}\sum_{e\in\partial K}\Big[\left((h\mathbf{u})_D^n-(h\mathbf{u})_K^n\right).\vec{n}_{e,K}\Big]^2m_e$$
$$+g\left(\frac{\Delta t}{m_K}\right)^2m_{\partial K}\sum_{e\in\partial K}\Big(\Pi_{e,K}^n.\vec{n}_{e,K}\Big)^2m_e\,. \tag{40}$$

This last estimation furnishes the remaining terms $\mathcal{A}_{\mathcal{E},K}$ and $\mathcal{R}_{\mathcal{E},K}$. □

We are now ready to establish the main result of this part, quantifying the total discrete energy production:

Theorem 2. *We consider the scheme (27), (28), together with the time-step condition (34). We have:*

$$\begin{aligned} E^{n+1}-E^n \quad &\leq (\Delta t)^2 \sum_{D\in\mathbb{T}^*} p(\gamma)\left[h_D^n \frac{m_e}{m_D}\left(\Phi_{K_e}^n-\Phi_K^n\right)^2\right] m_e \\ &\quad +(\Delta t)^2 \sum_{K\in\mathbb{T}}\sum_{e\in\partial K} q(\alpha)\Big[g\frac{m_{\partial K}}{m_K}\Big(\left((h\mathbf{u})_D^n-(h\mathbf{u})_K^n\right).\vec{n}_{e,K}\Big)^2\Big] m_e\,, \end{aligned}$$

where

$$p(\gamma)=\left[2\left(\Delta t\right)^2\left(\mu_e\frac{m_e}{m_D}\right)gh_D^n\right]\gamma^2-\gamma+2 \qquad , \qquad 2\mu_e=\frac{m_{\partial K}}{m_K}+\frac{m_{\partial K_e}}{m_{K_e}}\,, \tag{41a}$$

$$q(\alpha)=\left[8(\Delta t)^2\left(\frac{m_{\partial K}}{m_K}\frac{m_e}{m_D}\right)gh_D^n\right]\alpha^2-\alpha+1\,. \tag{41b}$$

Proof. Gathering Lemma 3 and Lemma 4, we have:

$$\begin{aligned} E^{n+1}-E^n \quad &=\sum_{K\in\mathbb{T}} m_K\left(\mathcal{E}_K^{n+1}-\mathcal{E}_K^n\right)+\sum_{D\in\mathbb{T}^*} m_D\left(\mathcal{K}_D^{n+1}-\mathcal{K}_D^n\right) \\ &\leq \sum_{K\in\mathbb{T}} m_K\left(\mathcal{Q}_{\mathcal{E},K}-\mathcal{S}_{\Phi,K}+\mathcal{R}_{\mathcal{E},K}+\mathcal{A}_{\mathcal{E},K}\right)+\sum_{D\in\mathbb{T}^*} m_D\left(\mathcal{A}_{\mathcal{K},D}+\mathcal{R}_{\mathcal{K},D}-\mathcal{Q}_{\mathcal{K},D}-\mathcal{S}_{\Lambda,D}\right)\,. \end{aligned} \tag{42}$$

From this, the proof relies on the following steps:

Lemma 5.

$$\sum_{K\in\mathbb{T}} m_K\mathcal{Q}_{\mathcal{E},K}=\sum_{D\in\mathbb{T}^*} m_D\mathcal{Q}_{\mathcal{K},D}\,. \tag{43}$$

Proof. At the level of a dual element $D=\mathbb{T}^*(K,K_e)$, we have two contributions coming from the terms $\mathcal{Q}_{\mathcal{E},K}$ and $\mathcal{Q}_{\mathcal{E},K_e}$ of the potential part, giving:

$$\Delta t m_e(h\mathbf{u})_D^n.\left(\Phi_{K_e}^n-\Phi_K^n\right)\vec{n}_{e,K}=m_D\left(\Delta t(h\mathbf{u})_D^n.\nabla_D\Phi^n\right)=m_D\mathcal{Q}_{\mathcal{K},D}\,.$$

□

Lemma 6.

$$\sum_{K\in\mathbb{T}} m_K\left(\mathcal{R}_{\mathcal{E},K}-\mathcal{S}_{\Phi,K}\right)+\sum_{D\in\mathbb{T}^*} m_D\mathcal{R}_{\mathcal{K},D}=(\Delta t)^2\sum_{D\in\mathbb{T}^*} p(\gamma)\left[h_D^n\frac{m_e}{m_D}\left(\Phi_{K_e}^n-\Phi_K^n\right)^2\right]m_e\,, \tag{44}$$

where

$$p(\gamma)=\left[2\left(\Delta t\right)^2\left(\mu_e\frac{m_e}{m_D}\right)gh_D^n\right]\gamma^2-\gamma+2 \qquad , \qquad 2\mu_e=\frac{m_{\partial K}}{m_K}+\frac{m_{\partial K_e}}{m_{K_e}}\,.$$

Proof. At the level of a dual element $D=\mathbb{T}^*(K,K_e)$ the terms $\mathcal{R}_{\mathcal{E},K},\mathcal{R}_{\mathcal{E},K_e}$ and $\mathcal{S}_{\Phi,K},\mathcal{S}_{\Phi,K_e}$ furnish the interface terms:

$$g\left(\Delta t\right)^2\left(\frac{m_{\partial K}}{m_K}+\frac{m_{\partial K_e}}{m_{K_e}}\right)\left(\Pi_{e,K}^n.\vec{n}_{e,K}\right)^2 m_e-\Delta t\Pi_{e,K}^n.\left(\Phi_{K_e}^n-\Phi_K^n\right)\vec{n}_{e,K}m_e\,.$$

Adding $m_D\mathcal{R}_{\mathcal{K},D}=2m_D\left(\Delta t\right)^2 h_D^n\left\|\nabla_D\Phi^n\right\|^2$, and using (30), we get the local contribution

$$p(\gamma)\left[h_D^n\left(\Delta t\right)^2\frac{(m_e)^2}{m_D}\left(\Phi_{K_e}^n-\Phi_K^n\right)^2\right]\,,$$

which allows to conclude. □

Lemma 7.

$$\begin{aligned}\sum_{K\in\mathbb{T}} m_K \mathcal{A}_{\mathcal{E},K} + \sum_{D\in\mathbb{T}^*} m_D \left(\mathcal{A}_{\mathcal{K},D} - \mathcal{S}_{\Lambda,D}\right) \\ \leq (\Delta t)^2 \sum_{K\in\mathbb{T}} \sum_{e\in\partial K} q(\alpha)\Big[g\frac{m_{\partial K}}{m_K}\Big(\left((h\mathbf{u})_D^n - (h\mathbf{u})_K^n\right).\vec{n}_{e,K}\Big)^2\Big] m_e\,,\end{aligned} \tag{45}$$

where

$$q(\alpha) = \left[8(\Delta t)^2\left(\frac{m_{\partial K}}{m_K}\frac{m_e}{m_D}\right) g h_D^n\right]\alpha^2 - \alpha + 1\,.$$

Proof. We consider this time the energy budget attached to an element $K \in \mathbb{T}$. We have:

$$\begin{aligned} m_K \mathcal{A}_{\mathcal{E},K} &= g\left(\Delta t\right)^2 \frac{m_{\partial K}}{m_K} \sum_{e\in\partial K}\Big[\left((h\mathbf{u})_D^n - (h\mathbf{u})_K^n\right).\vec{n}_{e,K}\Big]^2 m_e\,,\\ m_D \mathcal{A}_{\mathcal{K},D} &= 2(\Delta t)^2 \frac{(m_e)^2}{m_D} h_D^n\Big[\left(\Lambda_{K,D}^n\right)^2 + \left(\Lambda_{K_e,D}^n\right)^2\Big]\,,\\ m_D \mathcal{S}_{\Lambda,D} &= \Delta t m_e \left(h\mathbf{u}\right)_D^n.\vec{n}_{e,K}\left(\Lambda_{K,D}^n - \Lambda_{K_e,D}^n\right)\,.\end{aligned}$$

In the sequel we will denote $q_{e,D} = (h\mathbf{u})_D^n.\vec{n}_{e,K}$ and $q_{e,K} = (h\mathbf{u})_K^n.\vec{n}_{e,K}$ to alleviate the notations. Considering the dual cells surrounding K, and keeping the terms involving the cell K only, we are left with the control of the following quantity:

$$g\left(\Delta t\right)^2 \frac{m_{\partial K}}{m_K}\sum_{e\in\partial K}\left(q_{e,D} - q_{e,K}\right)^2 m_e + 2\left(\Delta t\right)^2 \sum_{e\in\partial K}\frac{(m_e)^2}{m_D} h_D^n\left(\Lambda_{K,D}^n\right)^2 - \Delta t\sum_{e\in\partial K} q_{e,D}\Lambda_{K,D}^n m_e\,. \tag{46}$$

Using (32), we have

$$q_{e,D}\Lambda_{K,D}^n = 2\alpha g\Delta t\frac{m_{\partial K}}{m_K} q_{e,D}(q_{e,D} - q_{e,K}) = \alpha g\Delta t\frac{m_{\partial K}}{m_K}\left(q_{e,D}^2 - q_{e,K}^2 + (q_{e,D} - q_{e,K})^2\right)\,.$$

It follows that (46) can be expressed as:

$$\begin{aligned}(\Delta t)^2 \sum_{e\in\partial K}\left[8(\Delta t)^2\left(\frac{m_{\partial K}}{m_K}\frac{m_e}{m_D}\right) g h_D^n\alpha^2 - \alpha + 1\right] g\frac{m_{\partial K}}{m_K}(q_{e,D} - q_{e,K})^2 m_e \\ - \left(\Delta t\right)^2 \alpha g\frac{m_{\partial K}}{m_K}\sum_{e\in\partial K}\left(q_{e,D}^2 - q_{e,K}^2\right) m_e\,.\end{aligned}$$

As a consequence, it now remains to define $(h\mathbf{u})_K^n$ in order to guarantee

$$\sum_{e\in\partial K}\left(q_{e,D}^2 - q_{e,K}^2\right) m_e = \sum_{e\in\partial K}\Big(\left((h\mathbf{u})_D^n.\vec{n}_{e,K}\right)^2 - \left((h\mathbf{u})_K^n.\vec{n}_{e,K}\right)^2\Big) m_e \geq 0\,. \tag{47}$$

To this end, one can remark that for any approximation of the mean discharge $(\overline{h\mathbf{u}})_K^n$ over the cell K, it is sufficient to define

$$(h\mathbf{u})_K^n = \lambda_K(\overline{h\mathbf{u}})_K^n\,, \tag{48}$$

with

$$\lambda_K = \begin{cases} 0 & \text{if } \displaystyle\sum_{e\in\partial K}\left((\overline{h\mathbf{u}})_K^n.\vec{n}_{e,K}\right)^2 m_e = 0\,,\\ \left(\displaystyle\sum_{e\in\partial K}\left((h\mathbf{u})_D^n.\vec{n}_{e,K}\right)^2 m_e \Big/ \sum_{e\in\partial K}\left((\overline{h\mathbf{u}})_K^n.\vec{n}_{e,K}\right)^2 m_e\right)^{1/2} & \text{otherwise}\end{cases}$$

to ensure an equality in (47). □

Injecting the estimations (43), (44) and (45) in (42), we get the announced result. □

4.2 Calibration of the stabilization constants

In the following section we discuss how to adjust the constants α, γ adequately, bearing in mind the dual objective of ensuring a global energy decrease while minimizing the diffusive losses in view of applicative issues. As can be seen from Theroem 2, the energy decrease in satisfied under the negativity of the quadratic forms (41a), (41b), which induces conditions linking α and γ to the CFL number. General trends can be extracted from a basic polynomial analysis, illustrated in Fig. 2, where the admissibility range is given with respect to the CFL number. Not surprisingly, it follows that the reduction of the time step allows more flexibility with respect to the choice of α and γ, making the values $\alpha_0 = 1$ and $\gamma_0 = 2$ appear as lower limit values. As done in the semi-implicit case (see Remark 3), it is also possible to

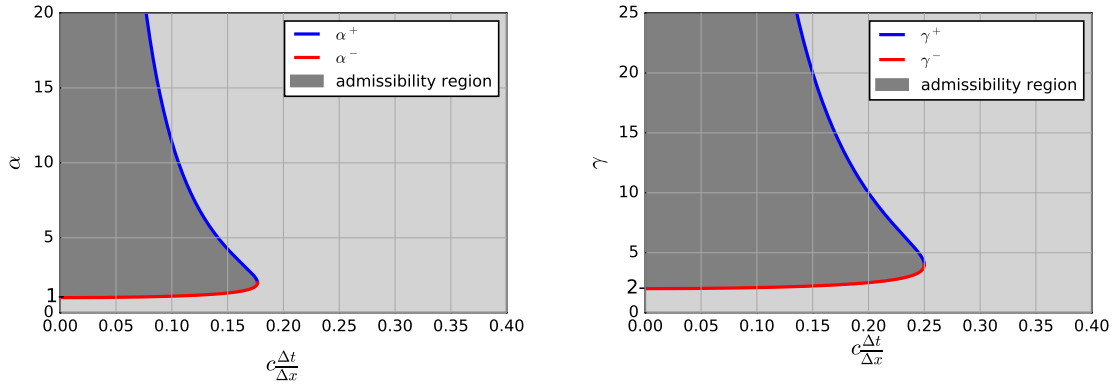

Figure 2: *Admissible values for α and γ issuing from the polynomials (41a), (41b) with respect to the CFL number in the case of a regular grid (we set $c = \sqrt{gh_D^n}$ and $\Delta x = \frac{m_e}{m_D}\frac{m_{\partial K}}{m_K}$). The admissibility range increases as the CFL tends to zero, highligting $\alpha_0 = 1$ and $\gamma_0 = 2$ as treshold values.*

obtain relaxed conditions revisiting some of the Jensen's inequalities implied in the demonstration (more precisely (38) and (40)). Skipping the details of the computations, we fall on the following polynomials:

$$\hat{p}(\gamma) = \left[2\left(\Delta t\right)^2\left(\mu_e \frac{m_e}{m_D}\right) g h_D^n\right]\gamma^2 - \gamma + \frac{1}{2}, \tag{49a}$$

$$\hat{q}(\alpha) = \left[8(\Delta t)^2\left(\frac{m_{\partial K}}{m_K}\frac{m_e}{m_D}\right) g h_D^n\right]\alpha^2 - \alpha + \frac{1}{2}, \tag{49b}$$

leading to the conditions depicted in Fig. 3. Note that we recover the limit value $\gamma_0 = 0.5$ found in the semi-implicit case.

These results are now completed with a linear stability analysis. For the sake of simplicity we detail here the one-dimensional case, which is already sufficient to corroborate the non-linear study. Considering perturbations around the constant state $\bar{h}, \bar{u} = 0$, the linearization procedure gives:

$$\begin{cases} h_i^{n+1} & = h_i^n - \dfrac{\Delta t}{\Delta x}\bar{h}\left(u_{i+1/2}^n - u_{i-1/2}^n\right) + \gamma g\left(\dfrac{\Delta t}{\Delta x}\right)^2\left(h_{i+1}^n - 2h_i^n + h_{i-1}^n\right), \\ u_{i+1/2}^{n+1} & = u_{i+1/2}^n - g\dfrac{\Delta t}{\Delta x}\left(h_{i+1}^n - h_i^n\right) + \alpha g\bar{h}\left(\dfrac{\Delta t}{\Delta x}\right)^2\left(u_{i+3/2}^n - 2u_{i+1/2}^n + u_{i-1/2}^n\right), \end{cases}$$

leading to the following system in the Fourier space:

$$\begin{pmatrix} \hat{h}^{n+1} \\ \\ \hat{u}^{n+1} \end{pmatrix} = \begin{pmatrix} 1 - 4\gamma g\bar{h}\left(\dfrac{\Delta t}{\Delta x}\right)^2 s_j^2 & -2i\bar{h}\left(\dfrac{\Delta t}{\Delta x}\right) s_j \\ -2ig\left(\dfrac{\Delta t}{\Delta x}\right) s_j & 1 - 4\alpha g\bar{h}\left(\dfrac{\Delta t}{\Delta x}\right)^2 s_j^2 \end{pmatrix} \begin{pmatrix} \hat{h}^n \\ \\ \hat{u}^n \end{pmatrix},$$

where $s_j = \sin\left(j\dfrac{\Delta x}{2}\right)$. Studying the eigenvalues we get necessary stability conditions linking γ, α and the CFL number. Skipping computation details, a representation of these conditions is proposed on Fig.

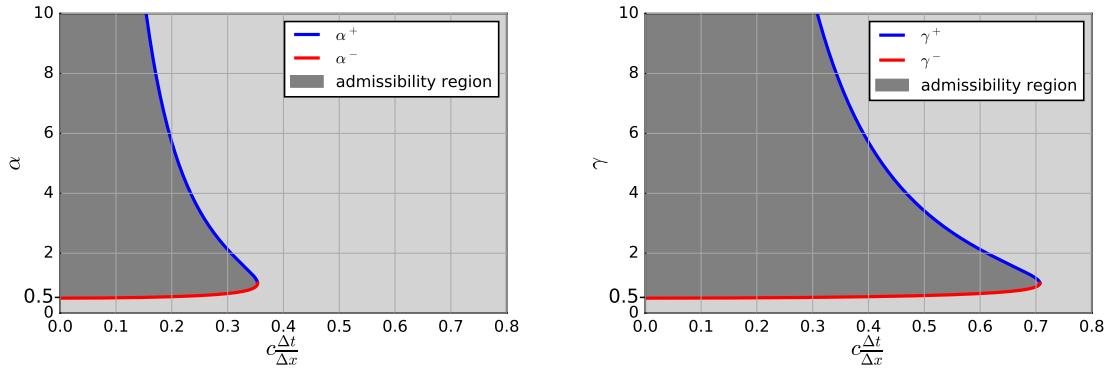

Figure 3: *Optimal admissibility conditions for α and γ issuing from the polynomials (49a), (49b), with respect to the CFL number in the case of a regular grid (we set $c = \sqrt{gh_D^n}$ and $\Delta x = \frac{m_e}{m_D}\frac{m_{\partial K}}{m_K}$). The range of admissibility increases as the CFL tends to zero, highligting $\alpha_0 = \gamma_0 = 0.5$ as threshold values.*

4 (*left*). It can be observed on both pictures the maximum admissible CFL number for given values of the couple (γ, α). A first observation highlights $\alpha + \gamma \geq 1$ as a necessary stability condition. We hence recover the optimized values issuing from the non-linear study $\gamma_0 = \alpha_0 = 0.5$ discussed just before as a particular case. A maximum CFL number is obtained for $(\alpha, \gamma) = (1, 0)$ and $(\alpha, \gamma) = (0, 1)$ which is a notable difference with the collocated case since optimal CFL values were observed all along the line $\alpha + \gamma = 1$. This study has been reproduced considering a Heun time scheme and a standrard MUSCL procedure in space. Looking at Fig. 4 (*right*), the regularizing effects of the RK2 time algorithm are clearly noticeable. More precisely, results indicate that the viscous correction may be reduced by half while preserving an almost optimal CFL condition, or even chosen arbitrarily small. Note finally that a linear analysis turns out to be irrelevant in the semi-implicit case since it simply leads to the necessary condition $\gamma \geq 0$.

4.3 Positivity and well-balanced properties

Since the dual fluxes are built as combinations of the primal ones, the condition (34) can actually be ensured through a collocated time constraint of the form:

$$\frac{\Delta t}{m_D} \sum_{e \in \partial K} |\mathcal{F}_e^n . \vec{n}_{e,K}| < \xi h_K^{n+1} , \tag{50}$$

where ξ is a strictly positive constant depending on β (see (34)) and the choice of the staggered discretization. After some calculations, not reported here, we finally get that the positivity of the water level at time $n+1$ is ensured under the following time step restriction:

$$\Delta t \frac{m_{\partial K}}{m_K} \left(|\mathbf{u}_D^n . \vec{n}_{e,K}| + \gamma \Delta t \frac{m_e}{m_D} |\Phi_{K_e}^n - \Phi_K^n| \right) \leq \left(\frac{\xi}{1+\xi} \right) \frac{h_K^n}{h_D^n} .$$

Note that according to (8) the ratio h_K^n / h_D^n may not lead to a singularity. In practice, this time step condition is much less restrictive than the one ensuring the existence of negative solutions for (41a) and (41b), governed by the speed of gravity waves (displayed in Fig. 2). As regards the well-balanced property, it is straightforward that the lake at rest configuration (26) eliminates the pressure gradient $\nabla_D \Phi^{n,*}$ (formula (31)) and leads to vanishing fluxes for both water height and discharge in (27), (28), giving immediately the steady state preservation.

5 Numerical validations

The present technology has recently been incorporated on an experimental basis within the operational platform of the French Naval Hydrographic and Oceanographic Service (SHOM), referred to as HYCOM

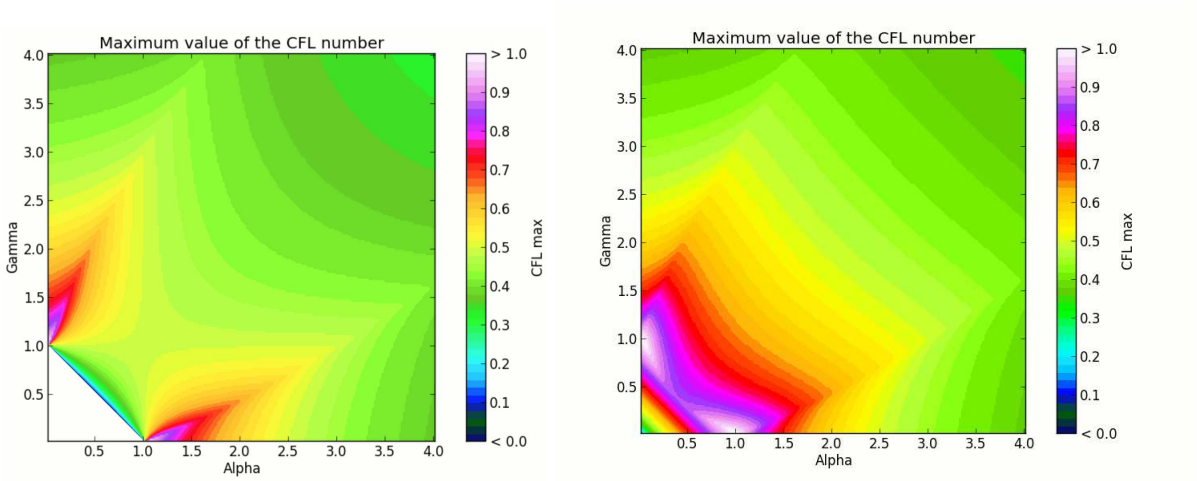

Figure 4: *Linear stability analysis. Representation of admissible values of the CFL number with respect to the stabilization constants α, γ, issuing from (49a), (49b). First order scheme in time and space (left): the threshold values $\alpha = \gamma = 0.5$ obtained with the non-linear analysis are recovered as a particular case. MUSCL scheme with RK2 in time (right) : regularizing effects of the RK2 time scheme.*

[8]. For that purpose, both implicit and explicit schemes have been implemented in a MAC environment, supplemented with a Heun time scheme and a standard MUSCL procedure in space. Although still under development, preliminary results issuing from the validation process are very promising. At this level, it clearly seems that this slight adaptation of the code comes with significant improvements, without any increase of computational time nor technical complexity. In this section, three classical benchmarks are proposed to assess the scheme's performances in various configurations, including propagation around a steady state, description of moving shoreline and displacement of a barotropic vortex. Note that these tests have been already successfully investigated in [11] and may serve as relevant evaluation tool with respect to the collocated methods developed in our previous works.

5.1 Small perturbation of a lake at rest

We first investigate the perturbation of a flow around a steady state, based on the classical test proposed by LeVeque [35]. This benchmark and several variants are widely used in the literature to study the beahaviour of well-balanced methods in the vicinity of steady states, see for instance [33, 37, 43, 56]. This simulation involves a rectangular channel with dimensions $2m \times 1m$, and the topography variations are given by :

$$z(x, y) = 0.8e^{-5(x-0.9)^2-50(y-0.5)^2} .$$

We first consider the so called *lake at rest* steady state $\eta = 1$, $\mathbf{u} = 0$ as initial condition to validate the well-balanced properties of the proposed schemes. As expected, our investigations shown that the rest solution was preserved up to the machine error throughout the computation. Then the steady state is perturbed by imposing the following discontinuity on the free surface:

$$\eta(x, y, t = 0) = \begin{cases} 1.01 \text{ if } 0.05 < x < 0.15\,, \\ 1 \text{ elsewhere } . \end{cases}$$

We can observe on Fig. 5 some snapshots of the solution obtained at time $t = 0.46s$ for two different meshes. The characteristics of the flow propagation are qualitatively well reproduced, even for the coarser one. For a more accurate insight we examine the vertical profiles of the free surface elevation along the centreline $y = 0.5m$, displayed in Fig. 6 (*left*). Regarding the highly resolved solution, the free surface deformations are reproduced with the correct amplitudes, in line with other recent high order approaches [16, 20, 46]. As for the 200×100 grid, we can observe that, notably due to under-resolution, the flow experiences small oscillations juste before the upstream slope. These can be damped by increasing the value of γ, which attests of the regularizing effects of the viscous correction. Note that the constant α has been set to 0 for this test, in accordance with the linear study. As shown in Fig. 6 (*right*), the constant γ is indeed sufficient to govern the dissipation rate. Despite being oustide its application framework,

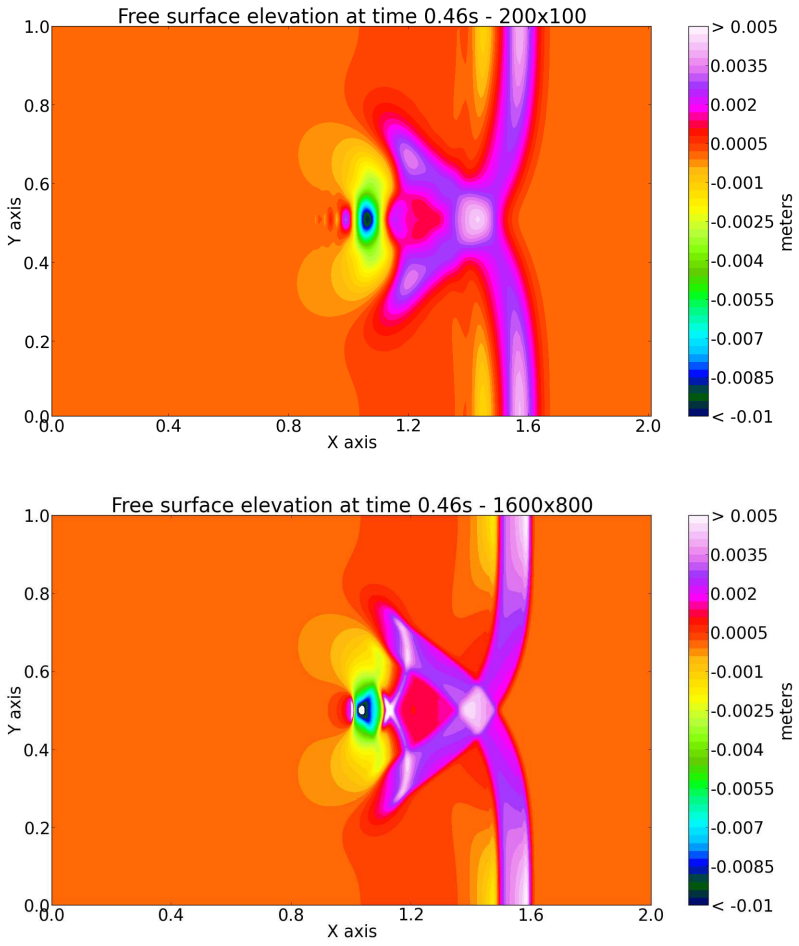

Figure 5: *Small perturbation of a lake at rest: Snapshots of the free surface at time t=0.46s obtained using the explicit scheme. RK2 in time and second order in space, using* $(\alpha, \gamma) = (0, 0.5)$ *as stabilization constants. Top:* 200×100 *grid. Bottom:* 1600×800 *grid.*

results obtained with the semi-implicit scheme are also proposed in Fig. 6 for information purpose; note that the corresponding CFL is ten times larger than the explicit one in this case.

5.2 Oscillatory flow in a parabolic bowl

We consider now the class of periodic solutions derived in [51]. These tests are one of the few for which exact solutions are available and stand for a relevant validation tool in a fully two-dimensional framework, involving dry cells and varying topography. Two classes of exact solutions are available, namely the *planar* and *curved* cases, used for example in [9, 21, 32, 42, 54] and [9, 18, 46] respectively. Based on the COMODO benchmark [1], we consider here the planar case, on a computational domain $200km \times 200km$, in which the bed profile is defined as follows:

$$z(x,y) = D_0 \left(\frac{(x-x_0)^2}{L^2} + \frac{(y-y_0)^2}{L^2} \right) ,$$

with (x_0, y_0) the center of the domain. The exact solution is:

$$\begin{cases} h(x,y,t) = 2a\dfrac{D_0}{L}\left(\dfrac{x-x_0}{L}\cos(\omega t) - \dfrac{y-y_0}{L}\sin(\omega t) - \dfrac{2a}{L} \right) , \\ u(x,y,t) = -a\omega \sin(\omega t) , \\ v(x,y,t) = -a\omega \cos(\omega t) . \end{cases} \tag{51}$$

The amplitude is set to $a = 1m$, and we fix $D_0 = 10km$, $L = 180km$. The frequency is $\omega = f/2 + \sqrt{f^2/4 + 2gD_0/L^2}\, s^{-1}$, with f the Coriolis parameter, set to $10^{-4}s^{-1}$. The total computational time is

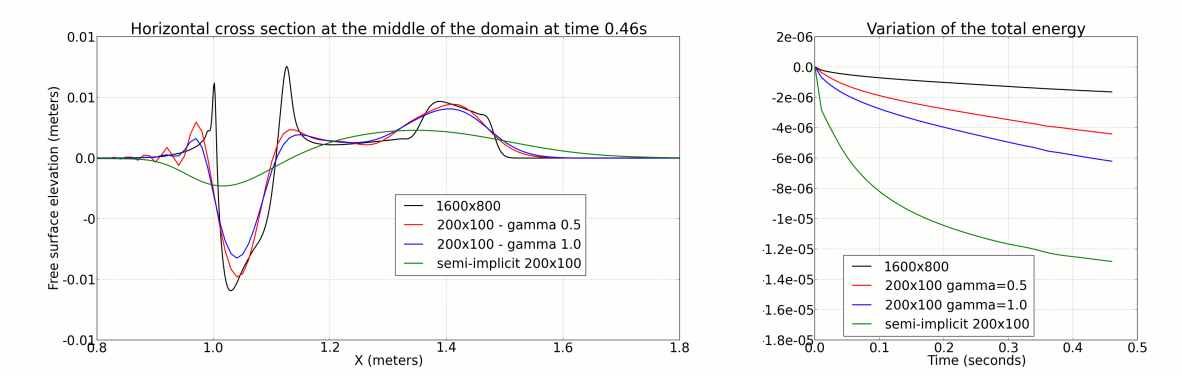

Figure 6: *Small perturbation of a lake at rest: Cross section of the numerical solution at time t=0.46s (left) and total energy variation (right) for* 200×100 *and* 1600×800 *grids at second order in space and time - Comparison with the semi-implicit scheme with a* 200×100 *resolution.*

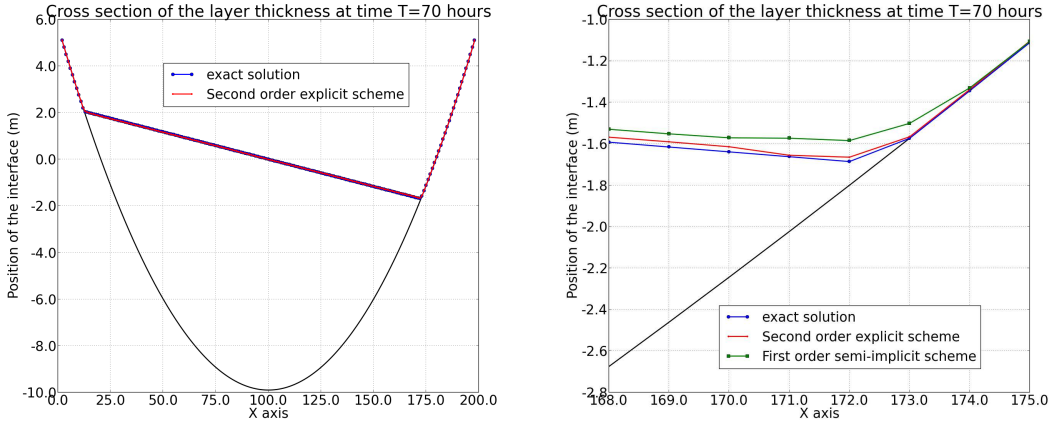

Figure 7: *Oscillatory flow in a parabolic bowl: Comparison between exact solution and second order explicit numerical solution at time* $T = 70$ *hours (left), and focus on the shoreline at first and second order.*

set to 80 hours and the constant γ is fixed to 0.5. In the present case, friction terms are neglected and the initial condition computed from (51) impulses a planar flow expected to follow a periodic motion throughout the whole computation. On can observe on Fig. 7 (*left*) a comparison between analytical and numerical solutions after nine complete oscillations at time $T = 70$ hours on a mesh of 40000 elements, exhibiting an excellent agreement. Fig. 7 (*right*) indicates that dry fronts are also well resolved and do not compromise the efficiency of the second order reconstruction. The scheme's accuracy can also be assessed through Fig. 8, where it is shown the time evolution of the free surface elevation tracked at the points $(100km, 150km)$ and $(100km, 180km)$. Comparisons attest of a very good level of precision; no significant losses of amplitude are observed, even at the level of the shoreline. This can be attributed to the positivity-preserving property and a very low energy dissipation rate, ensured by a proper calibration of the viscous constant γ, as can be observed in Fig. 9.

5.3 Barotropic vortex

We lastly investigate the deviation of a barotropic vortex under the effect of Coriolis force. The computational domain is set $[0, 1800km] \times [0, 1800km]$. The initial condition consists of a centred vortex at geostrophic equilibrium, with free surface elevation

$$\eta(x, y, t = 0) = a \exp\left(-\frac{(x - x_0)^2 + (y - y_0)^2)}{\mu^2}\right),$$

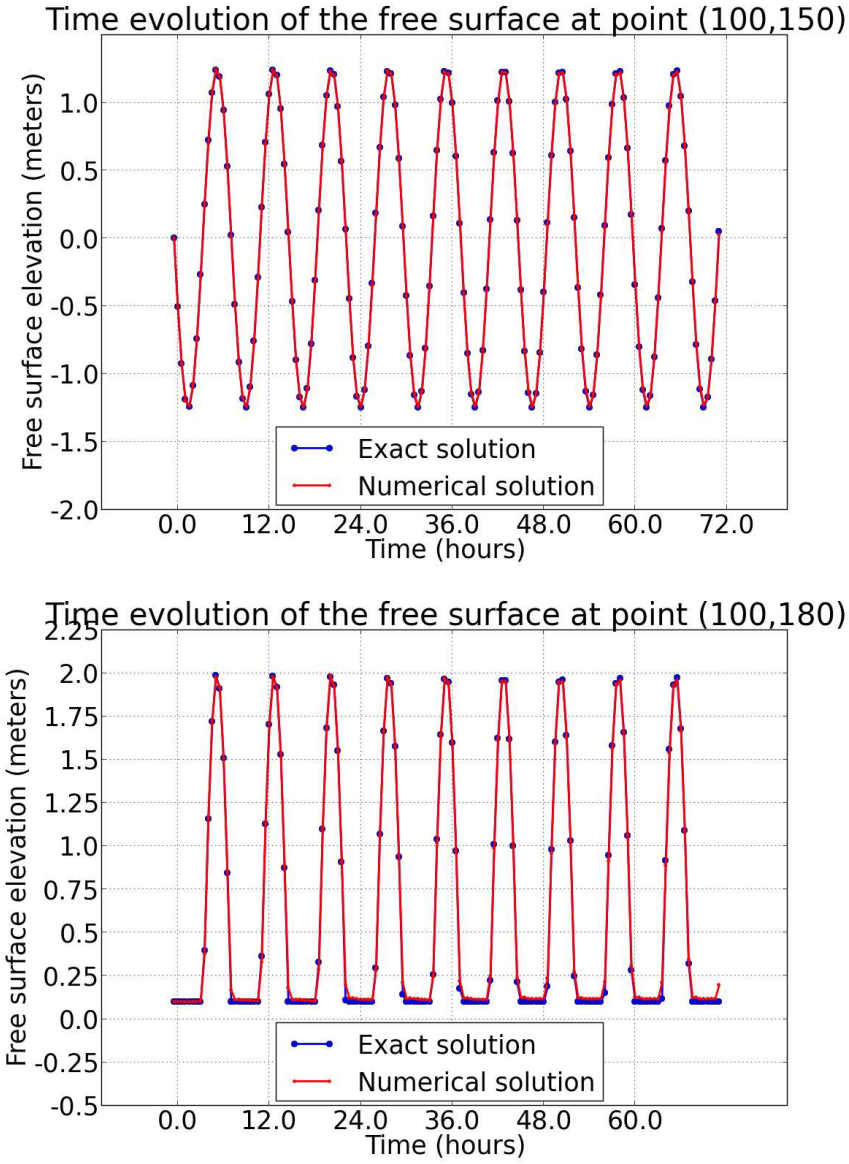

Figure 8: *Oscillatory flow in a parabolic bowl: Time evolution of the free surface at points* $(100km, 150km)$ *(top) and* $(100km, 180km)$ *(bottom), obtained with the second order explicit scheme with* $(\alpha, \gamma) = (0, 0.5)$ *on a regular mesh implying 40000 elements. Comparison with the exact solution.*

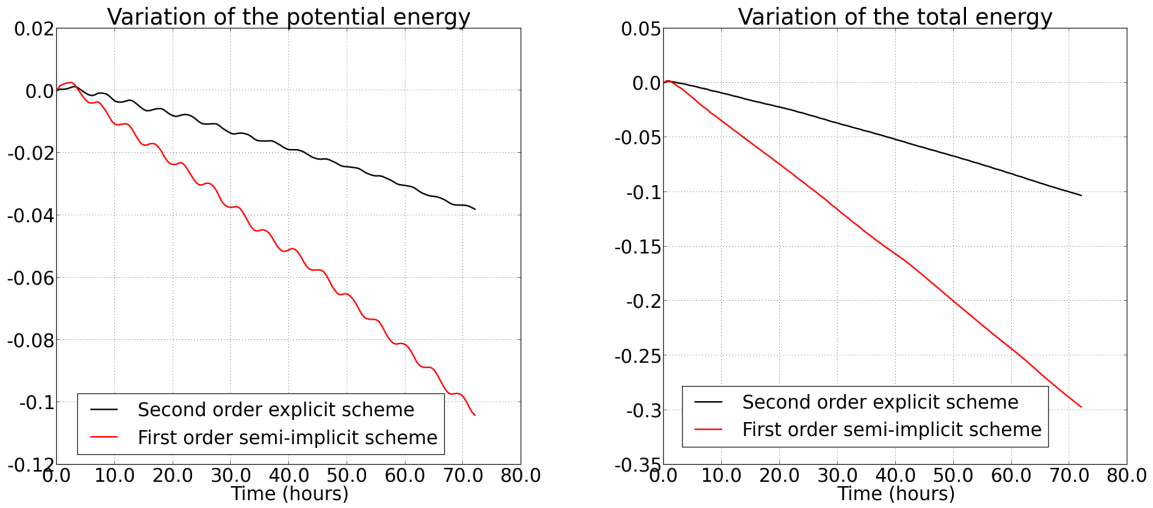

Figure 9: *Oscillatory flow in a parabolic bowl: Variation of potential energy (left) and total energy (right) for the explicit and semi-implicit schemes with* $\gamma = 0.5$.

where $a = 0.1m$, $x_0 = y_0 = 900km$ and $\mu = 60/\sqrt{2}\,km$. According to the original settings [11] we consider a β-plane approximation, setting $f(y) = f_0 + \beta y$ with $f_0 = 9.054.10^{-5}$ and $\beta = 1.788.10^{-11}$. As in the previous case, the resulting source term $(0, -fhv, fhu)^T$ is discretized by means of classical interpolations. On the basis of strong connexions with what has been done recently in the linear case [5] in the presence of rotational effects, specific work is currently being carried out to develop a class of more evolved algorithms for the Coriolis term, able to guarantee the exact geostrophic balance or rigorous discrete energy estimates. Again, the constant α has not proven crucial and has been set to zero for this test. As confirmed by linear study and numerical investigations, and according to the collocated framework [11], the regularizing effects of the RK time scheme allow to go far below the first order condition $\gamma \geq 1$ in practice and we take $\gamma = 0.1$ in the present case. Fig. 10 shows a comparison between numerical results obtained with the current version of HYCOM [8] and its recent improvement relying on the proposed approach. The benefits brought in terms of numerical diffusion are particularly noticeable at this level of resolution. Similar results are displayed in Fig. 11 with an increasing mesh resolution until reaching convergence. We recover the correct position of the vortex and the expected amplitudes [11], which attests to the low-Froude characteristics of the scheme, inherited from collocated approaches. The evolution of several diagnostic quantities related to energy are proposed on Fig. 12. Overall, these results highlight the improvements with respect to the latest uptade of HYCOM in terms of stability and numerical diffusion, in the sense where the proposed method is able to guarantee energy dissipation while minimizing the diffusive losses. Note that the energy variations for the 60×60 grid are not fully relevant here since this level of resolution entails an inaccurate initialisation of the vortex.

6 Conclusion

In this paper we have developed and analysed a numerical approach for the two-dimensional Shallow Water equations, on the basis of staggered discretizations. Inspired from previous works in the collocated frame ([11, 45]), semi-implicit and fully explicit time schemes have been proposed. In the semi-implicit case, the non-linear stability, understood in the sense of discrete mechanical energy decrease, has been achieved introducing appropriate diffusion terms in the advective fluxes. We showed that an additional perturbation term in the pressure gradient brought sufficient regularity to maintain this property in the explicit case. These results were supported by a linear stability analysis. Well-balanced and positivity results have also been obtained for both schemes. This results in a robust approach, easy to implement and satisfying all the fundamental stabiliy criteria attached to operational simulation.

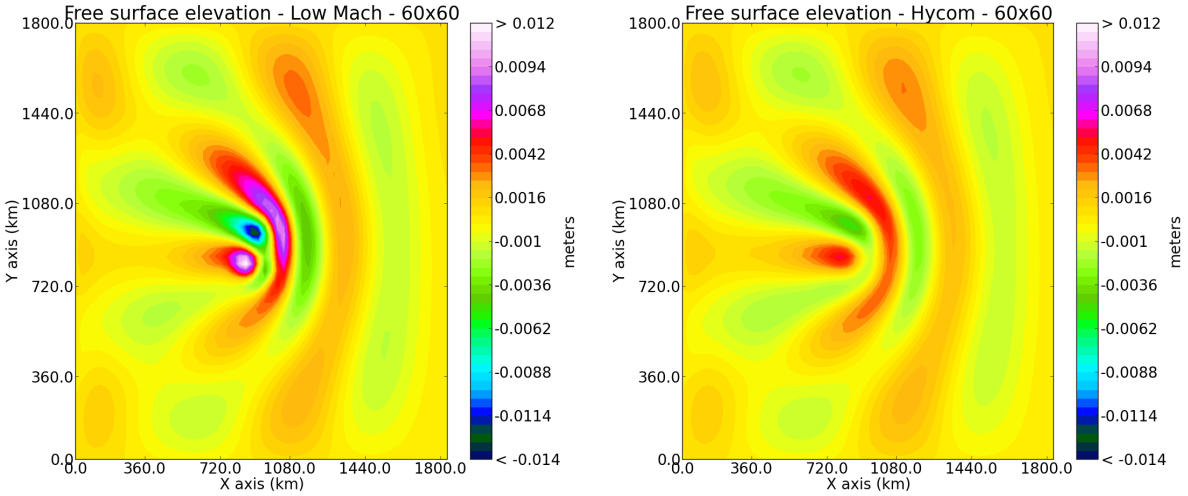

Figure 10: *Barotropic vortex: Representation of the free surface elevation after 60 days of propagation. Comparison between numerical results obtained with the explicit scheme (left) and the current version of HYCOM (right), using a* 60×60 *grid at second order in time and space.*

7 Aknowledgements

This work has been supported by the *Service Hydrographique et Océanographique de la Marine* (SHOM) and the French National Research Agency project NABUCO, grant ANR-17-CE40-0025.

References

[1] COMODO benchmark. `http://indi.imag.fr/wordpress/`.

[2] M. Akbar and S. Aliabadi. Hybrid numerical methods to solve shallow water equations for hurricane induced storm surge modeling. *Environmental Modelling & Software*, 46:118–128, 2013.

[3] G. Ansanay-Alex, T. Babik, J.C. Latché, and D. Vola. An L2 stable approximation of the Navier Stokes convection operator for low order non conforming finite elements. *International Journal for Numerical Methods in Fluids*, 66(5):555–580, 2011.

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

[5] E. Audusse, D.M. Hieu, P. Omnes, and Y. Penel. Analysis of modified Godunov type schemes for the two-dimensional linear wave equation with Coriolis source term on cartesian meshes. *Journal of Computational Physics*, 373:91–129, 2018.

[6] A. Bermúdez and M.E. Vázquez. Upwind methods for hyperbolic conservation laws with source terms. *Comput. & Fluids*, 23:1049–1071, 1994.

[7] C. Berthon and F. Foucher. Efficient well balanced hydrostatic upwind schemes for shallow water equations. *J. Comput. Phys.*, 231:4993–5015, 2012.

[8] R. Bleck. An oceanic general circulation model framed in hybrid isopycnic-cartesian coordinates. *Ocean Model.*, 4:55âĂŞ88, 2002.

[9] A. Bollermann, S. Noelle, and M. Lukácová-Medvidová. Finite volume evolution Galerkin methods for the shallow water equations with dry beds. *Commun. Comput. Phys.*, 10(2):371 – 404, 2011.

[10] L. Cea and M.E. Vázquez-Cendón. Unstructured finite volume discretization of bed friction and convective flux in solute transport models linked to the shallow water equations. *Journal of Computational Physics*, 231:3317 – 3339, 2012.

[11] F. Couderc, A. Duran, and J.P. Vila. An explicit asymptotic preserving low froude scheme for the multilayer shallow water model with density stratification. *Journal of Computational Physics*, 343:235–270, 2017.

[12] M.J. Castro Díaz, J.M. González-Vida, and C. Parès. Numerical treatment of wet/dry fronts in shallow flows with a modified Roe scheme. *Mathematical Models and Methods in Applied Sciences*, 16(6):897–931, 2006.

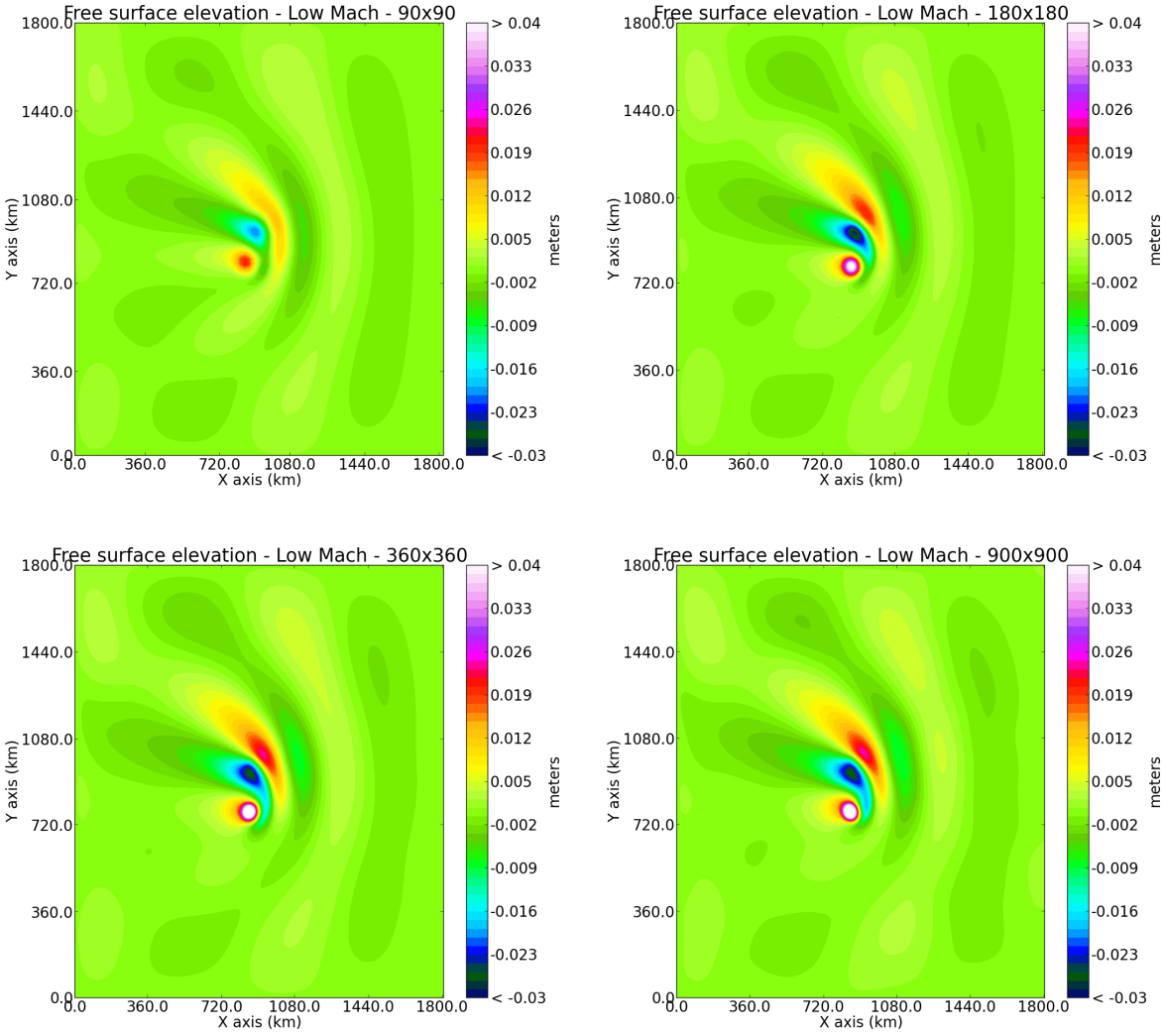

Figure 11: *Barotropic vortex: Representation of the free surface elevation after 60 days of propagation. Numerical results obtained with the explicit scheme at second order in time and space with an increasing mesh resolution.*

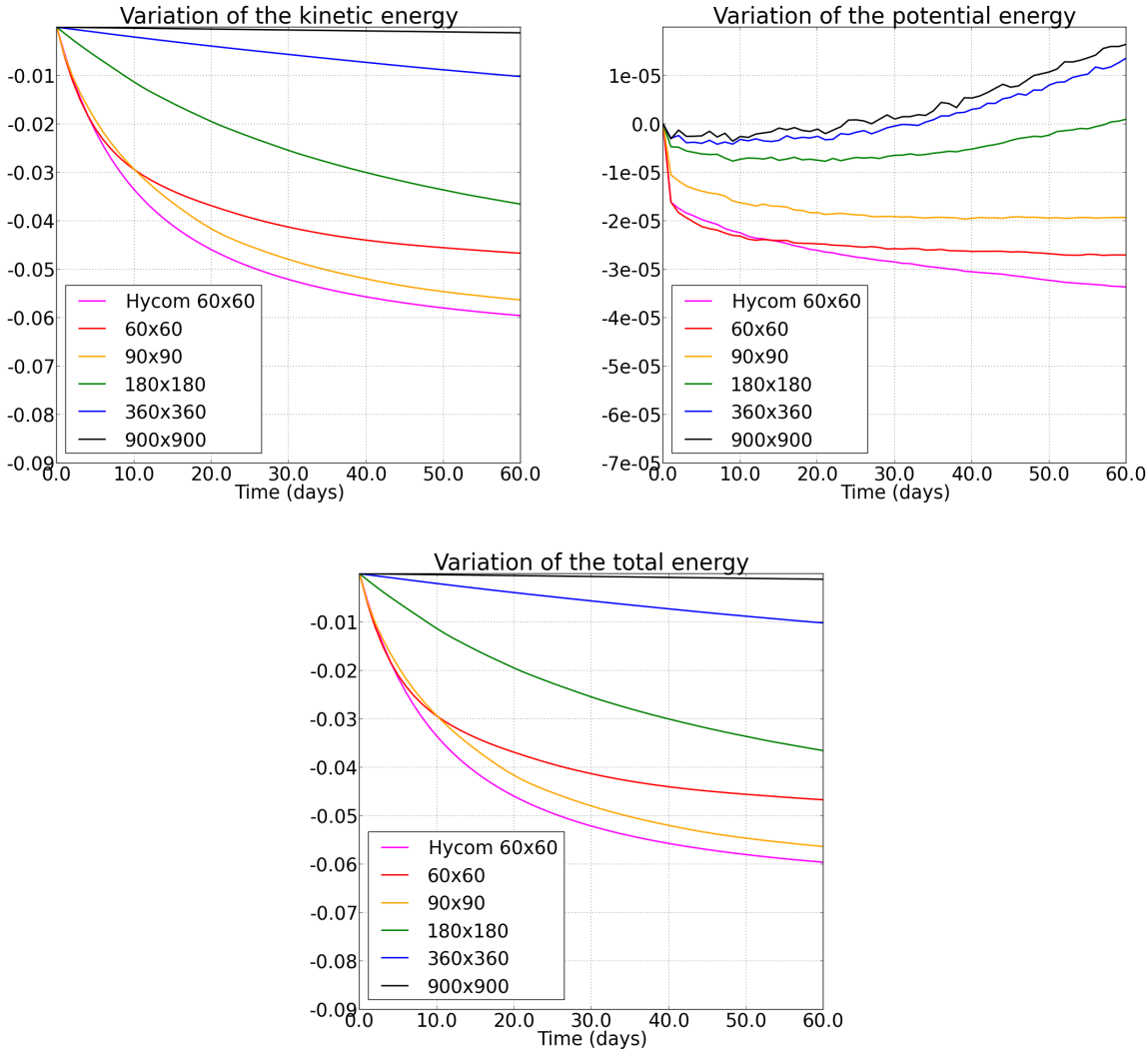

Figure 12: *Barotropic vortex: Evolution of several diagnostic quantities throughout the simulation for an increasing mesh resolution.*

[13] M.J. Castro Díaz, J.A. López-García, and C. Parés. High order exactly well-balanced numerical methods for shallow water systems. *Journal of Computational Physics*, 246:242–264, 2013.

[14] M.J. Castro Díaz, A. Pardo Milanés, and C. Parés. Well-balanced numerical schemes based on a generalized hydrostatic reconstruction technique. *Mathematical Models and Methods in Applied Sciences*, 17:2055–2113, 2007.

[15] M. Dumbser and V. Casulli. A staggered semi-implicit spectral discontinuous Galerkin scheme for the shallow water equations. *Applied Mathematics and Computation*, 219(15):8057–8077, 2013.

[16] A. Duran and F. Marche. Recent advances on the discontinuous Galerkin method for shallow water equations with topography source terms. *Computers & Fluids*, pages 88–104, 2013.

[17] A. Duran, J.P. Vila, and R. Baraille. Semi-implicit staggered mesh scheme for the multi-layer shallow water system. *C. R. Acad. Sci. Paris.*, 355:1298–1306, 2017.

[18] A. Ern, S. Piperno, and K. Djadel. A well-balanced Runge-Kutta discontinuous Galerkin method for the shallow-water equations with flooding and drying. *Int. J. Numer. Meth. Fluids*, 58:1–25, 2008.

[19] R. Fauzi and L.H. Wiryanto. On the staggered scheme for shallow water model down an inclined channel. In *AIP Conference Proceedings, 1867-020002*, 2017.

[20] U.S. Fjordholm, S. Mishra, and E. Tadmor. Well-balanced and energy stable schemes for the shallow water equations with discontinuous topography. *Journal of Computational Physics*, 230(14):5587 – 5609, 2011.

[21] J.M. Gallardo, C. Parés, and M.J. Castro Díaz. On a well-balanced high-order finite volume scheme for shallow water equations with topography and dry areas. *J. Comput. Phys.*, 227:574–601, 2007.

[22] T. Gallouët, J.M. Hérard, and N. Seguin. Some approximate Godunov schemes to compute shallow-water equations with topography. *Computers and Fluids*, 32:479–513, 2003.

[23] L. Gosse. A well-balanced flux-vector splitting scheme designed for hyperbolic systems of conservation laws with source terms. *Comput. Math. Appl.*, 39:135–159, 2000.

[24] J.M. Greenberg and A.Y. Leroux. A well-balanced scheme for the numerical processing of source terms in hyperbolic equations. *SIAM J. Numer. Anal.*, 33:1–16, 1996.

[25] N. Grenier, J.-P. Vila, and P. Villedieu. An accurate low-Mach scheme for a compressible two-fluid model applied to free-surface flows. *Journal of Computational Physics*, 252:1–19, 2013.

[26] P.H. Gunawan. *Numerical simulation of shallow water equations and related models.* PhD thesis, Université Paris Est, 2015.

[27] P.H. Gunawan and S.R. Pudjaprasetya. Explicit staggered grid scheme for rotating shallow water equations on geostrophic flows. *Progress in Computational Fluid Dynamics*, 18(1):46–55, 2018.

[28] R. Herbin, W. Kheriji, and J.C. Latché. On some implicit and semi-implicit staggered schemes for the shallow water and Euler equations. *Mathematical Modelling and Numerical Analysis*, 48:1807–1857, 2014.

[29] R. Herbin, J.C. LatchÃĭ, and T.T. Nguyen. Consistent explicit staggered schemes for compressible flows Part i: the barotropic Euler equations. 2013.

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

[31] G. Kesserwani and Q. Liang. Well-balanced RKDG2 solutions to the shallow water equations over irregular domains with wetting and drying. *Computers and Fluids*, 39:2040–2050, 2010.

[32] G. Kesserwani and Q. Liang. Locally limited and fully conserved RKDG2 shallow water solutions with wetting and drying. *J. Sci. Comput.*, 50:120–144, 2012.

[33] A. Kurganov and D. Levy. Central-upwind schemes for the saint-venant system. *Mathematical Modelling and Numerical Analysis*, 36:397–425, 2002.

[34] D. Le Roux, V. Rostand, and B. Pouliot. Analysis of Numerically Induced Oscillations in 2D Finite Element Shallow Water Models Part I: Inertia Gravity Waves. *SIAM Journal on Scientific Computing*, 29(1):331 – 360, 2007.

[35] R.J. Leveque. Balancing source terms and flux gradients in high-resolution Godunov methods: the quasi-steady wave-propagation algorithm. *J. Comput. Phys.*, 146:346–365, 1998.

[36] G. Li, V. Caleffi, and J. Gao. High-order well-balanced central WENO scheme for pre-balanced shallow water equations. *Computers & Fluids*, 99:182–189, 2014.

[37] M. Lukácová-Medvidová, S. Noelle, and M. Kraft. Well-balanced finite volume evolution Galerkin methods for the shallow water equations. *J. Comput. Phys*, 1:122–147, 2007.

[38] G. Madec. The NEMO team, NEMO ocean engine, Notes PÃťle Model. *(ISSN 1288-1619) 27, Institut Pierre-Simon Laplace (IPSL), France*, 2008.

[39] A. Meister and S. Ortleb. A positivity preserving and well-balanced DG scheme using finite volume subcells in almost dry regions. *Applied Mathematics and Computation*, 272:259 – 273, 2016.

[40] V. Michel-Dansac, C. Berthon, S. Clain, and F. Foucher. A well-balanced scheme for the shallow-water equations with topography. *Comput. Math. Appl.*, 72:568–âĂŞ593, 2016.

[41] J. Murillo and P. García-Navarro. Augmented versions of the HLL and HLLC Riemann solvers including source terms in one and two dimensions for shallow flow applications. *Journal of Computational Physics*, 231:6861 – 6906, 2012.

[42] I.K. Nikolos and A.I. Delis. An unstructured node-centered finite volume scheme for shallow water flows with wet/dry fronts over complex topography. *Comput. Methods Appl. Mech. Engrg*, 198:3723–3750, 2009.

[43] S. Noelle, N. Pankratz, G. Puppo, and J.R. Natvig. Well-balanced finite volume schemes of arbitrary order of accuracy for shallow water flows. *J. Comput. Phys.*, 213:474–499, 2006.

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

[45] M. Parisot and J.-P. Vila. Centered-potential regularization of advection upstream splitting method : Application to the multilayer shallow water model in the low Froude number regime. *SIAM Journal on Numerical Analysis*, 54:3083 – 3104, 2016.

[46] M. Ricchiuto and A. Bollermann. Stabilized residual distribution for shallow water simulations. *J. Comput. Phys.*, 228:1071–1115, 2009.

[47] G. Russo. Central schemes for conservation laws with application to shallow water equations. *Trends and applications of mathematics to mechanics : STAMM 2002, S. Rionero and G. Romano (Editors), Springer-Verlag Italia SRL*, pages 225–246, 2005.

[48] A.F. Shchepetkin and J.C. McWilliams. The regional oceanic modeling system (ROMS): a split-explicit, free-surface, topography-following-coordinate oceanic model. *Ocean Model.*, 9:347âĂŞ404, 2005.

[49] Y.E. Shi, R.K. Ray, and K.D. Nguyen. A projection method-based model with the exact C-property for shallow-water flows over dry and irregular bottom using unstructured finite-volume technique. *Computers & Fluids*, 76:178–195, 2013.

[50] A.L. Stewart and P.J. Dellar. An energy and potential enstrophy conserving numerical scheme for the multilayer shallow water equations with complete Coriolis force. *Journal of Computational Physics*, 313:99 – 120, 2016.

[51] W.C. Thacker. Some exact solutions to the nonlinear shallow water wave equations. *J. Fluid Mech.*, 107:499–508, 1981.

[52] S. Vater, N. Beisiegel, and J. Behrens. A limiter-based well-balanced discontinuous Galerkin method for shallow-water flows with wetting and drying: One-dimensional case. *Advances in Water Resources*, 85:1–13, 2015.

[53] R.A. Walters, E. Hanert, J. Pietrzak, and D. Le Roux. Comparison of unstructured, staggered grid methods for the shallow water equations. *Ocean Modelling*, 28:106–117, 2009.

[54] N. Wintermeyer, A.R. Winters, G.J. Gassner, and T. Warburton. An entropy stable discontinuous Galerkin method for the shallow water equations on curvilinear meshes with wet/dry fronts accelerated by GPUs. *J. Compt. Phys.*, 375:447–480, 2018.

[55] Y. Xing and C.-W. Shu. High order finite difference WENO schemes with the exact conservation property for the shallow water equations. *J. Comput. Phys.*, 208:206–227, 2005.

[56] Y. Xing and C.-W. Shu. A new approach of high order well-balanced finite volume WENO schemes and discontinuous Galerkin methods for a class of hyperbolic systems with source terms. *Commun. Comput. Phys.*, 1:100–134, 2006.

[57] Y. Xing, C.-W. Shu, and S. Noelle. On the advantage of well-balanced schemes for moving-water equilibria of the shallow water equations. *Journal of Scientific Computing*, 48:339–349, 2011.

[58] Y. Xing and X. Zhang. Positivity-preserving well-balanced discontinuous Galerkin methods for the shallow water equations on unstructured triangular meshes. *Journal of Scientific Computing*, 57(1):19–41, 2013.

[59] K. Xu. A well-balanced gas-kinetic scheme for the shallow-water equations with source terms. *Journal of Computational Physics*, 178:533–562, 2002.

[60] L. Zhao, B. Guo, T. Li, E.J. Avital, and J.J.R. Williams. A well-balanced explicit/semi-implicit finite element scheme for shallow water equations in dryingâĂŞwetting areas. *International Journal for Numerical Methods in Fluids*, 75:815–834, 2014.

[61] J.G. Zhou, D.M. Causon, and C. G. Mingham. The surface gradient method for the treatment of source terms in the shallow-water equations. *J. Comput. Phys.*, 168:1–25, 2001.