Novel well-balanced continuous interior penalty stabilizations

Lorenzo Micalizzi, Mario Ricchiuto, Rémi Abgrall

Novel well-balanced continuous interior penalty stabilizations

L. Micalizzi*, M. Ricchiuto† and R. Abgrall‡

July 20, 2023

Abstract

In this work, in a monodimensional setting, the high order accuracy and the well-balanced (WB) properties of some novel continuous interior penalty (CIP) stabilizations for the Shallow Water (SW) equations are investigated. The underlying arbitrary high order numerical framework is given by a Residual Distribution (RD)/continuous Galerkin (CG) finite element method (FEM) setting for the space discretization coupled with a Deferred Correction (DeC) time integration, to have a fully-explicit scheme. If, on the one hand, the introduced CIP stabilizations are all specifically designed to guarantee the exact preservation of the lake at rest steady state, on the other hand, some of them make use of general structures to tackle the preservation of general steady states, whose explicit analytical expression is not known. Several basis functions have been considered in the numerical experiments and, in all cases, the numerical results confirm the high order accuracy and the ability of the novel stabilizations to exactly preserve the lake at rest steady state and to capture small perturbations of such equilibrium. Moreover, some of them, based on the notions of space residual and global flux, have shown very good performances and superconvergences in the context of general steady solutions not known in closed-form. Despite the simulations addressing the monodimensional SW equations only, many elements can be extended to other general hyperbolic systems and to a multidimensional setting.

*Affiliation: Institut für Mathematik, Universität Zürich, Winterthurerstrasse 190, Zurich, 8057, Switzerland. Email: lorenzo.micalizzi@math.uzh.ch

†Affiliation: Team CARDAMOM, INRIA, University of Bordeaux, CNRS, Bordeaux INP, IMB, UMR 5251, France. Email: mario.ricchiuto@inria.fr

‡Affiliation: Institut für Mathematik, Universität Zürich, Winterthurerstrasse 190, Zurich, 8057, Switzerland. Email: remi.abgrall@math.uzh.ch

Contents

1 Introduction

In the context of the numerical resolution of hyperbolic partial differential equations (PDEs), one has to deal with several challenges, among which: the presence of instabilities and the exact preservation of some analytical solutions at the discrete level, namely well-balancing.

The instability issues are usually solved through an upwinding in the Finite Volume/discontinuous Galerkin FEM setting, through stabilization techniques in the RD/CG FEM setting. In particular, in this last context, the existing literature offers many possible options, for example: Streamline-Upwind Petrov-Galerkin, orthogonal subscale stabilization and CIP, respectively introduced in [9], [18] and [21]. For more details, the reader is referred to [36] and [37] in which a complete Fourier analysis and numerical investigation of the mentioned stabilizations, with different basis functions and time discretizations, has been performed.

We refer to well-balancing or C-property as the ability of a numerical scheme to exactly preserve a particular analytical solution or to be superconvergent, toward such solution, with respect to the general accuracy of the underlying discretization. In many applications, one is interested in embedding such feature in the adopted numerical method. This happens, for example, in the context of the study of physical systems admitting nontrivial stationary equilibria. In fact, such systems can stay for very long time in a neighborhood of a steady state. For this reason, researchers are interested in studying the evolution of small perturbations of steady solutions and, in this regard, it is desirable not to confuse the evolution of the perturbations with the natural noise arising from the numerical discretization. In such context, there are essentially two possibilities: using very refined meshes, with consequent increase in the computational cost, or modifying the numerical scheme in such a way that it preserves exactly the analytical solution of interest, without wasting the accuracy toward any other general solution. The latter option looks indeed very appealing, however, it is also very challenging since the steady states are usually not available in closed-form and, more in general, we rely on numerics because we do not have the analytical solutions to the systems of PDEs that we are trying to numerically solve.

Several strategies have been introduced to achieve well-balancing. The interested reader is referred to [13, 8, 43, 20, 23, 28, 41] and references therein. In particular, a successful approach is the one introduced in [25] and is based on the definition of a global flux, i.e., a new flux which keeps into account the source term, allowing to recast the initial problem into an equivalent one which is homogeneous.

In this work, we introduce some novel arbitrary high order WB CIP stabilizations for the SW equations in an RD/CG setting. They all are designed in such a way to exactly preserve the lake at rest steady state, however, some of them address the challenge of the preservation of general steady equilibria not known in closed-form.

The time discretization is achieved via the bDeCu method, introduced in [34] as an efficient modification of the DeC for hyperbolic problems designed in [7] to get arbitrary high order fully explicit schemes avoiding the issues associated with the mass matrix.

The structure of this work is the following. We will start by introducing the SW equations and their steady solutions in Section 2. Then, we will introduce, in Section 3, the CG FEM and explain how this can be put into an RD formulation. The issue of achieving well-balancing in the presented formulation is addressed in Section 4, which is the main section of this work. There, we will present two WB space discretizations and the novel CIP stabilizations. In Section 5, we will describe the time-stepping strategy, the bDeCu. In Section 6, we will present the numerical results. Finally, Section 7 is dedicated to conclusions and to future perspectives.

2 Shallow water equations

The SW equations are a system of hyperbolic PDEs used to model water flows, e.g., flows in sees, rivers, lakes or channels. Their monodimensional formulation, without rain and assuming a bottom topography fixed in time, reads

$$\frac{\partial}{\partial t}\boldsymbol{u} + \frac{\partial}{\partial x}\boldsymbol{F}(\boldsymbol{u}) = \boldsymbol{S}(x, \boldsymbol{u}), \quad (x, t) \in \Omega \times \mathbb{R}_0^+, \tag{1}$$

where $\Omega \subseteq \mathbb{R}$ is the space domain and the vector of the conserved variables, the flux and the source term are respectively defined as

$$\boldsymbol{u} := \begin{pmatrix} H \\ q \end{pmatrix}, \quad \boldsymbol{F}(\boldsymbol{u}) := \begin{pmatrix} q \\ \frac{q^2}{H} + g\frac{H^2}{2} \end{pmatrix}, \quad \boldsymbol{S}(x, \boldsymbol{u}) := -\begin{pmatrix} 0 \\ gH\frac{\partial}{\partial x}B(x) + g\frac{n_M^2|q|}{H^{\frac{7}{3}}}q \end{pmatrix}, \tag{2}$$

where H is water height, $q := Hv$ is the momentum of the flow, with v being the water speed averaged in the vertical direction, g is the gravitational constant, B is the bathymetry (or bottom topography) and n_M the Manning friction coefficient. Further, we introduce the total height $\eta := H + B$ and the sound speed $c := \sqrt{gH}$.

The Jacobian of the flux with respect to the conserved variables is given by

$$\boldsymbol{J}(\boldsymbol{u}) := \frac{\partial \boldsymbol{F}}{\partial \boldsymbol{u}} = \begin{pmatrix} 0 & 1 \\ -\frac{q^2}{H^2} + gH & 2\frac{q}{H} \end{pmatrix}, \tag{3}$$

with the two real eigenvalues given by $\lambda_{1,2} = v \pm c$.

When no friction is present, the SW system is also endowed with an entropy pair (s, F_s), with entropy s and entropy flux F_s respectively given by [40, 26]

$$s := \frac{1}{2}\frac{q^2}{H} + \frac{1}{2}gH^2 + gHB\,, \quad F_s := q\left(\frac{1}{2}\frac{q^2}{H^2} + g\eta\right), \tag{4}$$

with associated entropy variables

$$\boldsymbol{w} := \left(g\eta - \frac{1}{2}\frac{q^2}{H^2}, \frac{q}{H}\right)^T. \tag{5}$$

2.1 Steady states

The SW equations are well known to be characterized by nontrivial stationary solutions satisfying, in the weak sense, the ODE

$$\frac{\partial}{\partial x}\boldsymbol{F}(\boldsymbol{u}) = \boldsymbol{S}(x, \boldsymbol{u}). \tag{6}$$

The simplest stationary solution is the so-called "lake at rest" steady state given by constant total height and zero velocity

$$\eta = H + B \equiv \overline{\eta} \in \mathbb{R}_0^+, \quad v \equiv 0. \tag{7}$$

When no friction is present, through basic analysis, from (6) one can show that smooth steady states are characterized by constant momentum and energy

$$q(x,t) \equiv \overline{q} \in \mathbb{R}, \quad E = \frac{1}{2}\frac{\overline{q}^2}{H^2} + g(H + B) \equiv \overline{E} \in \mathbb{R}_0^+. \tag{8}$$

When also the friction is present, one can easily prove [32, 31] that smooth steady solutions satisfy

$$q(x,t) \equiv \overline{q} \in \mathbb{R}, \quad \left(-\frac{\overline{q}^2}{H^3} + g\right)\frac{\partial}{\partial x}H = -g\frac{\partial}{\partial x}B - g\frac{n_M^2|\overline{q}|}{H^{\frac{10}{3}}}\overline{q}. \tag{9}$$

In general, steady states are not available in closed-form and are obtained by solving (6). The interested reader is referred to [19], in which a wide collection of analytical solutions (not only steady) is provided.

3 Continuous Galerkin FEM and Residual Distribution

We will introduce in this section the CG FEM for hyperbolic problems and show how such method can be put in an RD formalism. For more information, the interested reader is referred to [4].

3.1 CG

We would like to numerically solve a hyperbolic system of balance laws in the form (1) over the bounded domain $\Omega := (x_L, x_R)$ in the time interval $[0, T_f]$, with some initial and boundary conditions. The main ingredients of the CG method are

- a tessellation $\mathcal{T}_h$ of the space domain made by non-overlapping closed elements K, segments in this case since we are in a monodimensional setting, covering its closure exactly;
- the finite dimensional space $W_M := \{\varphi \in C^0(\overline{\Omega}) \ : \ \varphi|_K \in \mathbb{P}_M(K), \ \forall K \in \mathcal{T}_h\}$ of the continuous functions φ which are such that their restriction to each element K of the tessellation is a polynomial of degree M;
- a basis $\{\varphi_i\}_{i=1,\dots,I}$ of W_M normalized in such a way that $\sum_{i=1}^{I} \varphi \equiv 1$ and which is such that each basis function φ_i can be associated to a spatial node $x_i \in \overline{\Omega}$, usually referred to as "degree of freedom" (DoF).

Further, the adopted bases are such that each basis function φ_i has support in the union of the elements $K \in K_i$, with $K_i := \{K \in \mathcal{T}_h \ : \ x_i \in K\}$ being the set of the elements containing the DoF x_i to which the function is associated. Let us notice that the previous assumptions imply

$$\sum_{x_i \in K} \varphi_i(x) \equiv 1, \quad \forall x \in K. \tag{10}$$

Examples of such bases, considered in the numerical tests, are the Bernstein polynomials and the Lagrange polynomials associated to equispaced nodes or to Gauss–Lobatto (GL) ones.

Finally, the CG method is given by a projection of the weak formulation in space of (1) over W_M, i.e., we look for an approximated solution $\boldsymbol{u}_h(x,t) := \sum_{i=1}^{I} \boldsymbol{c}_i(t)\varphi_i(x)$, linear combination of the basis functions through unknown coefficients which depend on time, such that it satisfies the following system of equations

$$\int_\Omega \left(\frac{\partial}{\partial t}\boldsymbol{u}_h + \left[\frac{\partial}{\partial x}\boldsymbol{F}(\boldsymbol{u}_h) - \boldsymbol{S}(x, \boldsymbol{u}_h) \right]_h \right) \varphi_i(x) dx + \mathbf{ST}_i(\boldsymbol{u}_h) = \mathbf{0}, \quad \forall i = 1, \dots, I, \tag{11}$$

where the term in square brackets is a consistent discretization of the spatial part of our initial PDE and $\mathbf{ST}_i$ is a stabilization term introduced at the discrete level to prevent the instabilities of central schemes.

Equation (11) is the semidiscretization of the CG method and consists of a nonlinear system of ODEs in all the coefficients $\boldsymbol{c}_i(t)$, collected in a single vector $\boldsymbol{c}(t)$, characterized by a mass matrix which is big and sparse. By numerically solving such system, one gets the evolution in time of the approximated solution $\boldsymbol{u}_h$.

Let us leave aside for one moment the problem of the time-stepping, which is not central in the context of this work, and let us observe that, if the discretization of the spatial part of the equation and the stabilization term in (11) are defined in such a way to be exactly zero for a particular

steady state, then the resulting numerical scheme will be WB with respect to such steady state. This will be the main topic of Section 4 but, before addressing the problem of well-balancing, we introduce here a short subsection, in which we show how the CG method can be easily embedded in an RD framework.

3.2 RD and link with with CG

We assume a classical CG FEM setting, i.e., a tessellation $\mathcal{T}_h$ of the space domain, the space W_M of continuous piecewise polynomial functions and a basis $\{\varphi_i\}_{i=1,\dots,I}$ of such space satisfying the properties previously mentioned. This allows us to consider the continuous approximation $\boldsymbol{u}_h(x,t) := \sum_{i=1}^{I} \boldsymbol{c}_i(t)\varphi_i(x)$ of the analytical solution. Then, the RD approach can be summarized in three main steps

i) **Definition of the element residuals**
For each element K of the tessellation $\mathcal{T}_h$, we define the element residual

$$\boldsymbol{\Phi}^K(\boldsymbol{u}_h) := \int_K \left(\frac{\partial}{\partial t}\boldsymbol{u}_h + \left[\frac{\partial}{\partial x}\boldsymbol{F}(\boldsymbol{u}_h) - \boldsymbol{S}(x,\boldsymbol{u}_h) \right]_h \right) dx, \quad \forall K \in \mathcal{T}_h, \tag{12}$$

which represents an integral balance at the considered element;

ii) **Definition of the node residuals**
For each element K, we consider the DoFs belonging to it, $x_i \in K$, and define the node residuals $\boldsymbol{\Phi}_i^K(\boldsymbol{u}_h)$ satisfying the following conservation relation

$$\sum_{x_i \in K} \boldsymbol{\Phi}_i^K(\boldsymbol{u}_h) = \boldsymbol{\Phi}^K(\boldsymbol{u}_h), \quad \forall K \in \mathcal{T}_h, \tag{13}$$

which corresponds to isolating the contribution of each DoF $x_i \in K$ to the integral balance introduced in the previous step;

iii) **Imposition of the balance at the nodes**
For each DoF x_i, we impose an equilibrium between all the node residuals $\boldsymbol{\Phi}_i^K(\boldsymbol{u}_h)$ of the elements that contain that DoF

$$\sum_{K \in K_i} \boldsymbol{\Phi}_i^K(\boldsymbol{u}_h) = \boldsymbol{0}, \quad \forall i = 1, \dots, I, \tag{14}$$

where we recall that K_i is the set of the elements of the tessellation containing the node x_i. This amounts to imposing that the global contribution of each node to all the balances of all the elements that share it is 0, which is indeed a reasonable constraint: nothing is created or destroyed at the nodes.

Equation (14) is a system of ODEs in the coefficients $\boldsymbol{c}_i(t)$, which must be solved in time. The recipe is quite general, as we did not specify how to choose the node residuals $\boldsymbol{\Phi}_i^K(\boldsymbol{u}_h)$. In fact, under this point of view, there are plenty of possibilities and the properties of the resulting scheme depend on this choice. In particular, due to (10), one can easily verify that the following definition of the node residuals

$$\boldsymbol{\Phi}_i^K(\boldsymbol{u}_h) := \int_K \left(\frac{\partial}{\partial t}\boldsymbol{u}_h + \left[\frac{\partial}{\partial x}\boldsymbol{F}(\boldsymbol{u}_h) - \boldsymbol{S}(x,\boldsymbol{u}_h) \right]_h \right) \varphi_i(x) dx + \mathbf{ST}_i^K(\boldsymbol{u}_h) = \boldsymbol{0} \tag{15}$$

fulfills the conservation relation (13), provided that the terms $\mathbf{ST}_i^K(\boldsymbol{u}_h)$ are defined in such a way that $\sum_{x_i \in K} \mathbf{ST}_i^K(\boldsymbol{u}_h) = \mathbf{0}$. Moreover, the resulting RD scheme given by system (14), for such choice of the node residuals, is formally equivalent to the CG semidiscretization (11) if $\sum_{K \in K_i} \mathbf{ST}_i^K(\boldsymbol{u}_h) = \mathbf{ST}_i(\boldsymbol{u}_h)$. The equivalence is essentially based on the fact that the support of the basis function φ_i is in the union of the elements $K \in K_i$.

Let us remark that the presented formulations, as well as the equivalence between them, extend in a natural way to a multidimensional framework. The interested reader is referred to [4], where several possible choices of the node residuals are presented and the link between RD and several classical approaches, e.g., discontinuous Galerkin and Finite Volume, are analyzed in depth.

4 Well-balancing

The evolution in time of the numerical solution $\boldsymbol{u}_h$ is given by the CG/RD formulation (11), which is recalled here for clarity

$$\int_\Omega \left(\frac{\partial}{\partial t} \boldsymbol{u}_h + \left[\frac{\partial}{\partial x} \boldsymbol{F}(\boldsymbol{u}_h) - \boldsymbol{S}(x, \boldsymbol{u}_h) \right]_h \right) \varphi_i(x) dx + \mathbf{ST}_i(\boldsymbol{u}_h) = \mathbf{0}, \quad \forall i = 1, \dots, I. \tag{16}$$

As anticipated, if we are able to design the discretization of the spatial part of the equation and the stabilization term in such a way that they are exactly zero for a particular steady state, then, we will get an exact well-balancing with respect to such steady state for any general time-stepping method. The goal of this section is to do precisely this.

Generally speaking, there are two main possibilities to achieve well-balancing:

- choosing a particular steady equilibrium and define ad hoc the mentioned ingredients of the scheme to be zero with respect to it;
- introducing some general structures aiming at preserving (6) at the discrete level.

The second strategy is the most desirable since, as already specified, the analytical expression of the steady states is almost never known in closed-form. In accordance with the first approach, all the WB elements that will be presented in this section are designed to be exactly zero with respect to the lake at rest steady state (7); nevertheless, some of them address the problem of the preservation of general stationary solutions not known in closed-form.

We will start by presenting a basic non-WB reference framework and, afterwards, we will continue with the definition of some WB alternatives. In order to light the notation, in this section we drop the dependence on time, which is not central in this context, being clear that all the space discretizations are performed for a given $\boldsymbol{u}_h$ and a fixed time t.

4.1 A reference non-well-balanced framework

A consistent space discretization is given by a simple interpolation of the flux and the source onto the functional space W_M

$$\left[\frac{\partial}{\partial x} \boldsymbol{F}(\boldsymbol{u}_h) - \boldsymbol{S}(x, \boldsymbol{u}_h) \right]_h = \frac{\partial}{\partial x} \boldsymbol{F}_h - \boldsymbol{S}_h, \tag{17}$$

$$\boldsymbol{F}_h := \sum_{i=1}^{I} \boldsymbol{F}_i \varphi_i(x), \quad \boldsymbol{S}_h := \sum_{i=1}^{I} \boldsymbol{S}_i \varphi_i(x), \tag{18}$$

with $\boldsymbol{F}_i$ and $\boldsymbol{S}_i$ interpolation coefficients, coinciding with the evaluations at the DoFs, respectively $\boldsymbol{F}(\boldsymbol{u}_h(x_i))$ and $\boldsymbol{S}(x_i, \boldsymbol{u}_h(x_i))$, if one assumes a Lagrange basis for W_M.

For what concerns the stabilization term, we adopt the CIP stabilization, firstly introduced in [21] in an elliptic-parabolic setting by Douglas and Dupont and then applied to the hyperbolic framework in [10] by Burman and Hansbo. Such stabilization is based on the introduction of a penalization term based on the jump of the normal derivatives of the numerical solution $\boldsymbol{u}_h$ across the faces of the tessellation, reading in general

$$\boldsymbol{ST}_i(\boldsymbol{u}_h) := \sum_{f \in \mathcal{F}_h} \sum_{r=1}^{R} \alpha_{f,r} \int_f \left[\!\left[\nabla^r_{\boldsymbol{\nu}_f} \varphi_i \right]\!\right] \left[\!\left[\nabla^r_{\boldsymbol{\nu}_f} \boldsymbol{u}_h \right]\!\right] d\sigma(\boldsymbol{x}), \quad \alpha_{f,r} = \delta_r \bar{\rho}_f h_f^{2r}, \tag{19}$$

where $\mathcal{F}_h$ denotes the set of the faces shared by two elements of the tessellation, $[\![\cdot]\!]$ is the jump across the face f, $\nabla^r_{\boldsymbol{\nu}_f}$ is the r-th partial derivative in the direction $\boldsymbol{\nu}_f$ normal to the face f, $\bar{\rho}_f$ is a local reference value for the spectral radius of the normal Jacobian of the flux, h_f is a reference characteristic size of the elements containing f and δ_r are constant parameters to be tuned. The orientation of the normal $\boldsymbol{\nu}_f$ and the direction of evaluation for the jump can be chosen freely. Originally, only the jump of the first derivative was taken into account, the stabilization on higher order derivatives has been introduced in [30].

Clearly, in a monodimensional context, the faces between the elements are just points and the integrals reduce to point-evaluations. Hence, (19) reduces to

$$\boldsymbol{ST}_i(\boldsymbol{u}_h) := \sum_{f \in \mathcal{F}_h} \sum_{r=1}^{R} \alpha_{f,r} \left[\!\left[\frac{\partial^r}{\partial x^r} \varphi_i \right]\!\right] \left[\!\left[\frac{\partial^r}{\partial x^r} \boldsymbol{u}_h \right]\!\right]. \tag{20}$$

The CG/RD formulation (11), along with the space discretization (17)-(18) coupled with the jump stabilization (20), properly solved in time through a suitable ODE integrator, provides an arbitrary high order framework for the numerical solution of the PDE (1). Nevertheless, as no particular attention has been paid to design the space discretization and the stabilization in such a way to achieve well-balancing, the resulting formulation is not WB.

Actually, neither the space discretization nor the jump stabilization, taken individually, are zero with respect to any particular steady state. In fact, the naive interpolation (18) of the flux and the source leads to a natural mismatch preventing any possibility of well-balancing, as $\frac{\partial}{\partial x}\boldsymbol{F}_h$ and $\boldsymbol{S}_h$ belong to two different polynomial spaces and their difference can be zero only in very trivial cases. Further, in the context of the lake a rest steady state, the jump of the derivatives of H_h across the interfaces, in the first component of (20), leads to a lack of well-balancing.

In the following, we will introduce some possible WB substitutes. We conclude this section with some final remarks.

Remark 4.1. *The CIP stabilizations can be naturally put in an RD formalism, even in a general multidimensional setting, as shown in the next proposition.*

Proposition 4.1. *Under the assumption of a conformal tessellation, if we define*

$$\boldsymbol{ST}_i^K(\boldsymbol{u}_h) := \sum_{\substack{f \subset \partial K \\ f \in \mathcal{F}_h}} \sum_{r=1}^{R} \alpha_{f,r} \int_f \nabla^r_{\boldsymbol{\nu}_f} \varphi_i|_K \left[\!\left[\nabla^r_{\boldsymbol{\nu}_f} \boldsymbol{u}_h \right]\!\right] d\sigma(\boldsymbol{x}), \tag{21}$$

where here the jump is evaluated from the inside of K to the neighboring element K' sharing f, $[\![z]\!] := z|_K - z|_{K'}$, then we have that

- $\sum_{\boldsymbol{x}_i \in K} \boldsymbol{ST}_i^K(\boldsymbol{u}_h) = \boldsymbol{0}$;
- *the stabilization term* (19) *is given by* $\sum_{K \in K_i} \boldsymbol{ST}_i^K(\boldsymbol{u}_h) = \boldsymbol{ST}_i(\boldsymbol{u}_h)$.

In the previous proposition, the bold font has been used for the generic DoF $\boldsymbol{x}_i \in K$ in order to emphasize the fact that the result holds in a general multidimensional setting. The proof can be found in Appendix A.

Remark 4.2 (Arbitrary high order stabilizations)**.** *Not all the stabilizations allow to reach arbitrary high order. For example, the Lax-Friedrichs stabilization presented in [1], in the context of the local Lax–Friedrichs node residuals in an RD setting, is at most first order accurate.*

4.2 Well-balanced space discretizations

We start by recalling, for clarity, the main focus of this section

$$\left[\frac{\partial}{\partial x}\boldsymbol{F} - \boldsymbol{S}\right]_h, \tag{22}$$

$$\boldsymbol{F}(\boldsymbol{u}) = \begin{pmatrix} q \\ \frac{q^2}{H} + g\frac{H^2}{2} \end{pmatrix}, \quad \boldsymbol{S}(x, \boldsymbol{u}) = -\begin{pmatrix} 0 \\ gH\frac{\partial}{\partial x}B + g\frac{n_M^2 |q|}{H^{\frac{7}{3}}}q \end{pmatrix}. \tag{23}$$

We will introduce here two WB discretizations of the spatial part of equation (1) with respect to the lake at rest steady state (7). While the second one is strongly based on the assumption of a monodimensional framework, the first one can be easily generalized to a multidimensional setting. Before starting, it is useful to introduce here the following splitting of the flux and the source

$$\boldsymbol{F} = \boldsymbol{F}^V + \boldsymbol{F}^{HS}, \quad \boldsymbol{F}^V = \begin{pmatrix} q \\ \frac{q^2}{H} \end{pmatrix}, \qquad \boldsymbol{F}^{HS} = \begin{pmatrix} 0 \\ g\frac{H^2}{2} \end{pmatrix}, \tag{24}$$

$$\boldsymbol{S} = \boldsymbol{S}^V + \boldsymbol{S}^{HS}, \quad \boldsymbol{S}^V = -\begin{pmatrix} 0 \\ g\frac{n_M^2 |q|}{H^{\frac{7}{3}}}q \end{pmatrix}, \quad \boldsymbol{S}^{HS} = -\begin{pmatrix} 0 \\ gH\frac{\partial}{\partial x}B \end{pmatrix}, \tag{25}$$

where the superscripts "V" and "HS" are used in order to identify respectively the velocity and the hydrostatic parts.

4.2.1 WB-HS

This WB discretization, presented in [42] and here denoted by "WB-HS", relies on a particular treatment of the terms $\frac{\partial}{\partial x}\left(g\frac{H^2}{2}\right)$ and $gH\frac{\partial}{\partial x}B$. Rather than simply interpolating $\boldsymbol{F}$ and $\boldsymbol{S}$, we consider the splitting (24)-(25). In the context of a lake at rest steady state, the velocity parts of the flux and of the source are identically zero as $q \equiv 0$, therefore, one can easily discretize such terms with a simple interpolation

$$\boldsymbol{F}_h^V = \sum_{i=1}^{I} \boldsymbol{F}_i^V \varphi_i(x), \quad \boldsymbol{S}_h^V = \sum_{i=1}^{I} \boldsymbol{S}_i^V \varphi_i(x). \tag{26}$$

A WB treatment of the hydrostatic part is less trivial. The mentioned approach consists in interpolating separately the water height and the bathymetry, thus getting

$$\left[\frac{\partial}{\partial x}\boldsymbol{F}^{HS} - \boldsymbol{S}^{HS}\right]_h = \begin{pmatrix} 0 \\ gH_h \frac{\partial}{\partial x} H_h \end{pmatrix} + \begin{pmatrix} 0 \\ gH_h \frac{\partial}{\partial x} B_h \end{pmatrix} = \begin{pmatrix} 0 \\ gH_h \frac{\partial}{\partial x}(H_h + B_h) \end{pmatrix}, \tag{27}$$

where, by linearity of the interpolation, $(H_h + B_h) = (H + B)_h = \eta_h$ which is constant in the context of the lake at rest steady state, leading to an exact well-balancing. To sum up, the final WB discretization reads

$$\left[\frac{\partial}{\partial x}\boldsymbol{F} - \boldsymbol{S}\right]_h := \frac{\partial}{\partial x}\boldsymbol{F}_h^V - \boldsymbol{S}_h^V + \begin{pmatrix} 0 \\ gH_h \frac{\partial}{\partial x}(H_h + B_h) \end{pmatrix}, \tag{28}$$

where the subscript h at the right-hand side indicates a simple interpolation.

4.2.2 WB-GF

The global flux approach has been firstly introduced in [25] and has already been employed in many works [14, 17, 33, 15, 12, 11, 29] to design WB methods. In particular, in [5], it has been shown how the notion of global flux can be naturally embedded in RD formulations.

The underlying idea is to define a new flux $\boldsymbol{G}$ keeping into account the source term in order to rephrase the original PDE (1) into an equivalent homogeneus formulation

$$\frac{\partial}{\partial t}\boldsymbol{u} + \frac{\partial}{\partial x}\boldsymbol{G} = \boldsymbol{0}. \tag{29}$$

Despite being absolutely non-trivial in a multidimensional setting, in the monodimensional case one can easily define the global flux through a simple integration of the source term

$$\boldsymbol{R} := -\int_{x_L}^{x} \boldsymbol{S}(s, \boldsymbol{u}(s))ds, \quad \boldsymbol{G} = \boldsymbol{F} + \boldsymbol{R}. \tag{30}$$

At the discrete level, $\boldsymbol{G}_h$ is got by interpolation, providing an approximation at each DoF, and the simplest idea that one could have is to set for any i

$$\boldsymbol{G}_h(x_i) := \boldsymbol{F}_h(x_i) + \boldsymbol{R}_h(x_i), \tag{31}$$

$$\boldsymbol{F}_h(x_i) := \sum_{j=1}^{I} \boldsymbol{F}_j \varphi_j(x_i), \quad \boldsymbol{R}_h(x_i) := -\int_{x_L}^{x_i} \sum_{j=1}^{I} \boldsymbol{S}_j \varphi_j(s)ds, \tag{32}$$

with $\boldsymbol{F}_h$ being the interpolation of the flux and $\boldsymbol{R}_h$ the integral of the interpolation of the source. Again, we remark that the dependence on time is dropped in order to light the notation.

Unfortunately, despite this choice providing a consistent discretization of the spatial part of our PDE, given by $\frac{\partial}{\partial x}\boldsymbol{G}_h$, this formulation is not WB with respect to the lake at rest steady state ($\boldsymbol{G}_h \neq$ const), as no special care has been taken under this point of view. In fact, in such a case, the flux and the source reduce to their hydrostatic part

$$\boldsymbol{F} = \boldsymbol{F}^{HS} = \begin{pmatrix} 0 \\ g\frac{H^2}{2} \end{pmatrix}, \quad \boldsymbol{S} = \boldsymbol{S}^{HS} = -\begin{pmatrix} 0 \\ gH \frac{\partial}{\partial x} B \end{pmatrix}, \tag{33}$$

and there is no reason why the integral of the interpolation of the second component of the source should match the second component of the flux. A WB alternative is the one presented in [45, 17] and consists in adopting, in each element K, the following discretization for the second component of $\boldsymbol{S}^{HS}$

$$\left[gH\frac{\partial}{\partial x}B\right]_h := \left[g(H_h + B_h)\frac{\partial}{\partial x}B_h\right]_h - \frac{\partial}{\partial x}\left[\frac{gB^2}{2}\right]_h, \quad \forall x \in K, \quad \forall K \in \mathcal{T}_h, \tag{34}$$

where again the subscripts at the right-hand side stand for simple interpolations. More formally, we can state the following proposition.

Proposition 4.2. *By adopting the discretization* (34) *for the second component of the hydrostatic part of the source and a simple interpolation of the velocity part, the resulting global flux got by interpolating its values at the DoFs*

$$\boldsymbol{G}_h(x_i) := \boldsymbol{F}_h(x_i) + \boldsymbol{R}_h(x_i), \tag{35}$$

$$\boldsymbol{F}_h(x_i) := \sum_{j=1}^{I} \boldsymbol{F}_j \varphi_j(x_i), \quad \boldsymbol{R}_h(x_i) := -\int_{x_L}^{x_i} \left[-\begin{pmatrix} 0 \\ \left[gH\frac{\partial}{\partial x}B\right]_h(s) \end{pmatrix} + \sum_{j=1}^{I} \boldsymbol{S}_j^V \varphi_j(s) \right] ds, \tag{36}$$

is constant for a lake at rest steady state.

The proof can be found in Appendix B. Summarizing, the WB discretization based on the notion of global flux, here denoted as "WB-GF", reads

$$\left[\frac{\partial}{\partial x}\boldsymbol{F} - \boldsymbol{S}\right]_h := \frac{\partial}{\partial x}\boldsymbol{G}_h, \tag{37}$$

with $\boldsymbol{G}_h$ defined by interpolating its values at the DoFs given by (35)-(36).

Remark 4.3 (Local interpolation)**.** *We remark that the discretization* (34) *is meant to be performed separately in each element, as the term* $\frac{\partial}{\partial x}B_h$ *is in general discontinuous across the interfaces between the elements.*

Remark 4.4 (Global flux quadrature)**.** *The global flux can be also interpreted as a suitable modification of the quadrature formula adopted in order to evaluate the source integral in* (11)*. For more information, the reader is referred to [33].*

4.3 Well-balanced continuous interior penalty stabilizations

We start by recalling the original non-WB CIP stabilization (20) in a monodimensional setting

$$\boldsymbol{ST}_i(\boldsymbol{u}_h) := \sum_{f \in \mathcal{F}_h} \sum_{r=1}^{R} \alpha_{f,r} \left[\!\left[\frac{\partial^r}{\partial x^r}\varphi_i \right]\!\right] \left[\!\left[\frac{\partial^r}{\partial x^r}\boldsymbol{u}_h \right]\!\right] \tag{38}$$

$$= \sum_{f \in \mathcal{F}_h} \sum_{r=1}^{R} \alpha_{f,r} \left[\!\left[\frac{\partial^r}{\partial x^r}\varphi_i \right]\!\right] \left[\!\left[\frac{\partial^r}{\partial x^r}\begin{pmatrix} H_h \\ q_h \end{pmatrix} \right]\!\right]. \tag{39}$$

This stabilization is based on the jump of the derivatives of the conserved variables, this is why we will refer to it as "**jc**".

We propose here some novel WB alternative CIP stabilizations. The main idea is to change the object of the stabilization in such a way to achieve well-balancing. In the following definitions the subscript h denotes an interpolation.

- **Total height (jt)**

$$\boldsymbol{ST}_i(\boldsymbol{u}_h) := \sum_{f\in\mathcal{F}_h}\sum_{r=1}^{R}\alpha_{f,r}\left[\!\left[\frac{\partial^r}{\partial x^r}\varphi_i\right]\!\right]\left[\!\left[\frac{\partial^r}{\partial x^r}\begin{pmatrix}\eta_h\\ q_h\end{pmatrix}\right]\!\right], \quad \eta_h := H_h + B_h. \tag{40}$$

- **Entropy variables (je)**

$$\boldsymbol{ST}_i(\boldsymbol{u}_h) := \sum_{f\in\mathcal{F}_h}\alpha_f\left[\!\left[\frac{\partial}{\partial x}\varphi_i\right]\!\right]\mathcal{A}_f\left[\!\left[\frac{\partial}{\partial x}\boldsymbol{w}_h\right]\!\right], \quad \boldsymbol{w} := \begin{pmatrix} g\eta - \frac{v^2}{2}\\ v\end{pmatrix}, \tag{41}$$

where $\mathcal{A}_f \in \mathbb{R}^{2\times 2}$ is a matrix which is used to make the stabilization dimensionally consistent. One possible choice for it, the one assumed here, is given by the Jacobian of the transformation from entropy to conserved variables

$$\mathcal{A}_f := \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{w}} = \begin{pmatrix}\frac{1}{g} & \frac{v}{g}\\ \frac{v}{g} & H + \frac{v^2}{g}\end{pmatrix}, \tag{42}$$

evaluated at the interface f and computed assuming a flat bathymetry in such a way to have a proper bijective map between $\boldsymbol{w}$ and $\boldsymbol{u}$.

- **Space residual (jr)**

$$\boldsymbol{ST}_i(\boldsymbol{u}_h) := \sum_{f\in\mathcal{F}_h}\alpha_f\left[\!\left[\boldsymbol{J}\frac{\partial}{\partial x}\varphi_i\right]\!\right]\mathcal{B}_f\left[\!\left[\boldsymbol{J}\frac{\partial}{\partial x}\boldsymbol{u}_h - \boldsymbol{S}^*\right]\!\right], \quad \boldsymbol{S}^* := -\begin{pmatrix}0\\ gH_h\frac{\partial}{\partial x}B_h\end{pmatrix}, \tag{43}$$

where $\boldsymbol{J} := \frac{\partial \boldsymbol{F}}{\partial \boldsymbol{u}}$ is the Jacobian of the flux (3) at the interface f and the matrix $\mathcal{B}_f \in \mathbb{R}^{2\times 2}$, just like $\mathcal{A}_f$, is used for consistency purposes. In this work, we assume $\mathcal{B}_f := |\boldsymbol{J}|^{-1}$ with $|\boldsymbol{J}|$ absolute value of the Jacobian, defined as $|\boldsymbol{J}| := R|\Lambda|R^{-1}$, where R is the matrix of the right eigenvectors of $\boldsymbol{J}$ and $|\Lambda|$ a diagonal matrix having as entries the absolute values of the eigenvalues of $\boldsymbol{J}$. More explicitly, for the sake of completeness, R and $|\Lambda|$ are respectively given by

$$R := \begin{pmatrix}1 & 1\\ v-c & v+c\end{pmatrix}, \quad |\Lambda| := \begin{pmatrix}|v-c| & 0\\ 0 & |v+c|\end{pmatrix}. \tag{44}$$

- **Global flux (jg)**

$$\boldsymbol{ST}_i(\boldsymbol{u}_h) := \sum_{f\in\mathcal{F}_h}\alpha_f\left[\!\left[\boldsymbol{J}\frac{\partial}{\partial x}\varphi_i\right]\!\right]\mathcal{B}_f\left[\!\left[\frac{\partial}{\partial x}\boldsymbol{G}_h\right]\!\right], \tag{45}$$

where again the presence of $\mathcal{B}_f := |\boldsymbol{J}|^{-1}$ allows to make the stabilization consistent with the other elements in (11).

The abbreviations are based on the objects of the stabilizations. Before going to the numerical results, we make some useful remarks.

All the new CIP stabilizations are WB with respect to the lake at rest steady state, as shown in the next proposition.

Proposition 4.3. *The stabilization terms* (40)*,* (41)*,* (43) *and* (45) *are exactly zero in the context of the lake at rest steady state* (7)*.*

Proof. The claim is trivial for jt (40) and je (41): the terms dependent on q and v cancel and we are left with the terms $\overline{\eta}$ and $g\overline{\eta}$, which, in such a case, are constant and, therefore, have zero derivatives and related jump. Clearly, the same holds for jg (45) because, as shown in Proposition 4.2, the global flux $\boldsymbol{G}_h$ has been specifically designed to be constant for a lake at rest steady state. For what concerns instead jr (43), we have that the argument of the second jump, $\boldsymbol{J}\frac{\partial}{\partial x}\boldsymbol{u}_h - \boldsymbol{S}^*$, for a lake at rest steady state reduces to

$$\boldsymbol{J}\frac{\partial}{\partial x}\boldsymbol{u}_h - \boldsymbol{S}^* = \begin{pmatrix} 0 & 1 \\ gH_h & 0 \end{pmatrix}\begin{pmatrix} \frac{\partial}{\partial x}H_h \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ gH_h\frac{\partial}{\partial x}B_h \end{pmatrix} = \begin{pmatrix} 0 \\ gH_h\frac{\partial}{\partial x}\left(H_h + B_h\right) \end{pmatrix}, \tag{46}$$

which is indeed zero because $(H_h + B_h) = \eta_h \equiv \overline{\eta}$ is constant. In practice, the definition of $\boldsymbol{S}^*$ is given in such a way to mimic the trick of the first WB space discretization, WB-HS, presented in Section 4.2.1. □

The last two stabilizations, (43) and (45), are particularly interesting, as in such cases the stabilization is based on discretizations of $\frac{\partial}{\partial x}\boldsymbol{F} - \boldsymbol{S}$, a quantity which is supposed to be zero in the context of a general steady state, not only in the context of the lake at rest. In fact, as we are going to see in the numerical experiments, they have special properties in terms of superconvergence and capturing of small perturbations of general stationary solutions.

Remark 4.5 (About non-differential terms)**.** *One could wonder why, in the context of jr* (43)*, the term $\boldsymbol{S}^*$ seems to take into account only the hydrostatic part of the source $\boldsymbol{S}^{HS}$, defined in* (25)*, and not $\boldsymbol{S}^V$. The point is that the remaining velocity part $\boldsymbol{S}^V$ has no differential terms and, thus, it would cancel due to the assumption of a continuous representation of the numerical solution. Under this point of view, it is worth underlying that the terms $\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{w}}$, $\boldsymbol{J}$ and $|\boldsymbol{J}|^{-1}$ are non-differential and therefore well-defined at each interface f.*

Remark 4.6 (Another possible choice for $\mathcal{B}_f$)**.** *As already pointed out, the matrices $\mathcal{A}_f$ and $\mathcal{B}_f$ are used in order to achieve dimensional consistency. Other choices, with respect to the ones presented, are possible. In particular, another valid option for $\mathcal{B}_f$ is given by $\rho_f^{-1}\mathbb{I}$, with ρ_f being the spectral radius of $\boldsymbol{J}$ at the interface f and $\mathbb{I} \in \mathbb{R}^{2\times 2}$ the identity matrix. The numerical results got with the two definitions of $\mathcal{B}_f$ are qualitatively similar but slightly better with $\mathcal{B}_f := |\boldsymbol{J}|^{-1}$. Therefore, for the sake of compactness, we will only present the ones obtained for such definition.*

Remark 4.7. *The corresponding terms $\boldsymbol{ST}_i^K$ for the definitions of the new stabilizations in an RD*

setting are respectively given by

$$\boldsymbol{ST}_i^K(\boldsymbol{u}_h) := \sum_{\substack{f\subset\partial K\\ f\in\mathcal{F}_h}} \sum_{r=1}^{R} \alpha_{f,r} \frac{\partial^r}{\partial x^r} \varphi_i|_K \left[\!\!\left[\frac{\partial^r}{\partial x^r} \begin{pmatrix} \eta_h \\ q_h \end{pmatrix} \right]\!\!\right], \tag{47}$$

$$\boldsymbol{ST}_i^K(\boldsymbol{u}_h) := \sum_{\substack{f\subset\partial K\\ f\in\mathcal{F}_h}} \alpha_f \frac{\partial}{\partial x} \varphi_i|_K \mathcal{A}_f \left[\!\!\left[\frac{\partial}{\partial x} \boldsymbol{w}_h \right]\!\!\right], \tag{48}$$

$$\boldsymbol{ST}_i^K(\boldsymbol{u}_h) := \sum_{\substack{f\subset\partial K\\ f\in\mathcal{F}_h}} \alpha_f \boldsymbol{J} \frac{\partial}{\partial x} \varphi_i|_K \mathcal{B}_f \left[\!\!\left[\boldsymbol{J} \frac{\partial}{\partial x} \boldsymbol{u}_h - \boldsymbol{S}^* \right]\!\!\right], \tag{49}$$

$$\boldsymbol{ST}_i^K(\boldsymbol{u}_h) := \sum_{\substack{f\subset\partial K\\ f\in\mathcal{F}_h}} \alpha_f \boldsymbol{J} \frac{\partial}{\partial x} \varphi_i|_K \mathcal{B}_f \left[\!\!\left[\frac{\partial}{\partial x} \boldsymbol{G}_h \right]\!\!\right], \tag{50}$$

where the convention on the direction of evaluation of the jump is the one adopted in the context of Proposition 4.1, $[\![z]\!] := z|_K - z|_{K'}$.

Remark 4.8 (Extension to a multidimensional setting)**.** *Apart from jg* (45) *which requires the notion of global flux and, hence, is strictly related to a monodimensional setting, the other jump stabilizations can be easily generalized to a multidimensional context.*

5 Deferred Correction

Aiming at an arbitrary high order framework, once the discretization in space has been fixed, we need to select a suitable arbitrary high order time integration technique for the numerical resolution of the CG/RD semidiscretization (11). For this purpose, we adopt a DeC time discretization.

Originally introduced in [24], the DeC approach has been extensively developed over the years. Several formulations have been proposed [22, 38, 7], with applications to many different fields, for example [35, 16, 44, 39, 3, 27, 6], ranging from adaptivity to structure–preserving.

In particular, here we consider the bDeCu method, introduced in [34] as an efficient modification of the DeC formulation presented in [7]. The advantage of such formulation is that it allows to get rid of the burden associated with the big and sparse mass matrix, typical of CG/RD discretizations. In particular, its size and sparse structure make its inversion unfeasible in concrete applications, determining, in general, high computational costs related to the resolution, at each time iteration, of several linear systems for standard time integration methods. Moreover, the adoption of some particular stabilization terms $\boldsymbol{ST}_i(\boldsymbol{u}_h)$, dependent on the time derivative of the approximated solution, may determine contributions to the mass matrix, leading to a new solution-dependent mass matrix, which implies heavy complications. In fact, the mass matrix should be recomputed multiple times at each time iteration and, in general, no warranties hold concerning its invertibility (and so concerning the well–posedness of the resulting method), see for example [2] in a Lagrangian framework. All these problems do not exist with the adopted approach.

We will introduce now the DeC formulation for hyperbolic problems defined in [7] and, afterwards, we will describe the efficient modification introduced in [34]. The idea behind the approach is based on having two operators dependent on a same discretization parameter Δ and associated

to different discretizations of the same problem, $\mathcal{L}^1_\Delta, \mathcal{L}^2_\Delta : X \longrightarrow Y$, with X and Y normed vector spaces. The operator $\mathcal{L}^2_\Delta$ corresponds to a high order and implicit discretization and is, hence, difficult to solve, while, $\mathcal{L}^1_\Delta$ corresponds to a low order and explicit discretization and, in particular, we assume that it is easy to solve problems of the type $\mathcal{L}^1_\Delta(\underline{\boldsymbol{u}}) = \underline{\boldsymbol{z}}$ for $\underline{\boldsymbol{z}} \in Y$ given. Due to its simplicity, we would prefer solving $\mathcal{L}^1_\Delta$, rather than $\mathcal{L}^2_\Delta$, however, the solution to such operator would not be accurate enough. Under some assumptions on the operators, we can consider $\underline{\boldsymbol{u}}^{(p)}$ given by the following iterative procedure

$$\mathcal{L}^1_\Delta(\underline{\boldsymbol{u}}^{(p)}) = \mathcal{L}^1_\Delta(\underline{\boldsymbol{u}}^{(p-1)}) - \mathcal{L}^2_\Delta(\underline{\boldsymbol{u}}^{(p-1)}), \quad p \geq 1, \tag{51}$$

which is subjected to the following accuracy estimate with respect to the solution $\underline{\boldsymbol{u}}_\Delta$ to the high order operator $\mathcal{L}^2_\Delta$

$$\left\| \underline{\boldsymbol{u}}^{(p)} - \underline{\boldsymbol{u}}_\Delta \right\|_X \leq \left(\Delta \frac{\alpha_2}{\alpha_1} \right)^p \left\| \underline{\boldsymbol{u}}^{(0)} - \underline{\boldsymbol{u}}_\Delta \right\|_X . \tag{52}$$

Let us observe that the updating formula (51) is explicit as a result of the assumptions made on the operator $\mathcal{L}^1_\Delta$. Moreover, thanks to the accuracy estimate (52), the convergence to the solution $\underline{\boldsymbol{u}}_\Delta$ of the operator $\mathcal{L}^2_\Delta$ is ensured for Δt small enough, independently of the chosen $\underline{\boldsymbol{u}}^{(0)}$, and the number of iterations to achieve a given accuracy with respect to it is controlled.

In our case, the reference problem is the semidiscrete formulation (11), which can be rephrased more compactly as

$$\sum_{j=1}^{I} \left(\int_\Omega \varphi_i \varphi_j dx \right) \frac{d}{dt} \boldsymbol{c}_j(t) + \boldsymbol{\phi}_i(\boldsymbol{c}(t)) = \boldsymbol{0}, \quad \forall i = 1, \ldots, I, \tag{53}$$

with $\boldsymbol{\phi}_i(\boldsymbol{c}(t))$ containing the terms not related to the time derivative. As in the context of a classical one-step method, we assume to know an approximation $\boldsymbol{c}_n \approx \boldsymbol{c}(t_n)$ of the solution to system (11) at the generic time t_n and we look for an approximation $\boldsymbol{c}_{n+1} \approx \boldsymbol{c}(t_{n+1})$ at $t_{n+1} := t_n + \Delta t$.

In order to define the operators $\mathcal{L}^1_\Delta$ and $\mathcal{L}^2_\Delta$, we introduce $M+1$ equispaced subtimenodes in the interval $[t_n, t_{n+1}]$, such that $t_n = t^0 < t^1 < \cdots < t^M = t_{n+1}$. Then, the operator $\mathcal{L}^2_\Delta : \mathbb{R}^{(I \times Q \times M)} \to \mathbb{R}^{(I \times Q \times M)}$ is given by

$$\mathcal{L}^2_\Delta(\underline{\boldsymbol{c}}) = \left(\mathcal{L}^2_{\Delta,1}(\underline{\boldsymbol{c}}), \mathcal{L}^2_{\Delta,2}(\underline{\boldsymbol{c}}), \ldots, \mathcal{L}^2_{\Delta,I}(\underline{\boldsymbol{c}}) \right), \tag{54}$$

$$\mathcal{L}^2_{\Delta,i}(\underline{\boldsymbol{c}}) = \begin{pmatrix} \sum_{j=1}^{I} \left(\int_\Omega \varphi_i \varphi_j dx \right) \left(\boldsymbol{c}_j^1 - \boldsymbol{c}_j^0 \right) + \Delta t \sum_{\ell=0}^{M} \theta_\ell^1 \boldsymbol{\phi}_i(\boldsymbol{c}^\ell) \\ \vdots \\ \sum_{j=1}^{I} \left(\int_\Omega \varphi_i \varphi_j dx \right) \left(\boldsymbol{c}_j^M - \boldsymbol{c}_j^0 \right) + \Delta t \sum_{\ell=0}^{M} \theta_\ell^M \boldsymbol{\phi}_i(\boldsymbol{c}^\ell) \end{pmatrix}, \quad \underline{\boldsymbol{c}} = \begin{pmatrix} \boldsymbol{c}^1 \\ \vdots \\ \boldsymbol{c}^M \end{pmatrix}, \tag{55}$$

with $\boldsymbol{c}_j^0$ known components of $\boldsymbol{c}^0 := \boldsymbol{c}_n$ and Q number of scalar equations in the original hyperbolic system, 2 in this case. Its definition is based on replacing the function $\boldsymbol{\phi}_i(\boldsymbol{c}(t))$ by a high order interpolation in time with the Lagrange polynomials ψ^m associated to the subtimenodes t^m. The generic coefficient θ_ℓ^m is, in fact, the normalized integral of the function ψ^m over $[t^0, t^m]$. The solution $\underline{\boldsymbol{c}}_\Delta$, such that $\mathcal{L}^2_\Delta(\underline{\boldsymbol{c}}_\Delta) = \boldsymbol{0}$, contains M components $\boldsymbol{c}_\Delta^m$ $m = 1, \ldots, M$, which are $(M+1)$-th order accurate approximations of the exact solution $\boldsymbol{c}(t^m)$. Trying to solve such operator directly is indeed very complicated as it amounts to solving a huge nonlinear system of algebraic equations.

The low order explicit operator $\mathcal{L}^1_\Delta : \mathbb{R}^{(I\times Q\times M)} \to \mathbb{R}^{(I\times Q\times M)}$ is, instead, given by

$$\mathcal{L}^1_\Delta(\underline{\boldsymbol{c}}) = \left(\mathcal{L}^1_{\Delta,1}(\underline{\boldsymbol{c}}), \mathcal{L}^1_{\Delta,2}(\underline{\boldsymbol{c}}), \dots, \mathcal{L}^1_{\Delta,I}(\underline{\boldsymbol{c}})\right), \tag{56}$$

$$\mathcal{L}^1_{\Delta,i}(\underline{\boldsymbol{c}}) = \begin{pmatrix} C_i\left(\boldsymbol{c}_i^1 - \boldsymbol{c}_i^0\right) + \Delta t \beta^1 \boldsymbol{\phi}_i(\boldsymbol{c}^0) \\ \vdots \\ C_i\left(\boldsymbol{c}_i^M - \boldsymbol{c}_i^0\right) + \Delta t \beta^M \boldsymbol{\phi}_i(\boldsymbol{c}^0) \end{pmatrix}, \quad \underline{\boldsymbol{c}} = \begin{pmatrix} \boldsymbol{c}^1 \\ \vdots \\ \boldsymbol{c}^M \end{pmatrix}, \tag{57}$$

with $\beta^m := \frac{t^m - t^0}{\Delta t}$ and $C_i := \int_\Omega \varphi_i dx$. Such definition is based on an Euler approximation in time and a first order mass lumping in space. The components of the solution to $\mathcal{L}^1_\Delta$ are, in fact, first order accurate approximations of the exact solution to (11) in the subtimendodes t^m $m = 1, \dots, M$.

By a direct computation, the updating formula (51), in this case reduces to

$$\begin{pmatrix} \vdots \\ \boldsymbol{c}_i^{m,(p)} \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ \boldsymbol{c}_i^{m,(p-1)} \\ \vdots \end{pmatrix} - \frac{1}{C_i} \begin{pmatrix} \vdots \\ \sum_{j=1}^I \left(\int_K \varphi_i \varphi_j dx\right)\left(\boldsymbol{c}_j^{m,(p-1)} - \boldsymbol{c}_j^0\right) + \Delta t \sum_{\ell=0}^M \theta_\ell^m \boldsymbol{\phi}_i(\boldsymbol{c}^{\ell,(p-1)}) \\ \vdots \end{pmatrix}, \tag{58}$$

for any i.

The vectors $\boldsymbol{c}_i^{m,(p)}$ are the components of $\boldsymbol{c}^{m,(p)}$ which is, itself, a component of $\underline{\boldsymbol{c}}^{(p)}$, the output vector at the p-th iteration. For what concerns the initial vector $\underline{\boldsymbol{c}}^{(0)}$, the most reasonable choice is to set for any subtimenode $\underline{\boldsymbol{c}}^{m,(0)} = \boldsymbol{c}^0 := \boldsymbol{c}_n$. The updating does not involve the solution at the subtimenode t^0, therefore, we also have $\underline{\boldsymbol{c}}^{0,(p)} = \boldsymbol{c}^0 := \boldsymbol{c}_n$ for any p. In the end, we set $\boldsymbol{c}_{n+1} := \boldsymbol{c}^{M,(P)}$, where P is the final number of iterations performed. The optimal number of iterations is given by $P := M+1$, as the accuracy of the approximation with respect to the solution of $\mathcal{L}^2_\Delta$ increases by one at each iteration but we are not interested in approximating it with accuracy higher than the one of the underlying discretization.

One can see that the well-posedness of the explicit update (58) is strongly related to the fact that the coefficients C_i in (57) are not zero. This is not always the case for any choice of the polynomial basis. A safe option is given by the Bernstein polynomials, for which we always have $C_i > 0\ \forall i$. Another possibility, particularly convenient, is to choose a basis of Lagrange polynomials associated to points defining a quadrature formula sufficiently accurate to guarantee a high order mass lumping, e.g., the GL points, and to adopt such quadrature for the integrals.

The modification presented in [34] consists in introducing interpolation processes between the iterations to increase the number of subtimenodes according to the order of accuracy achieved in the specific iterations. In particular, at the iteration p, the bDeCu method involves an interpolation in time of $\underline{\boldsymbol{c}}^{(p-1)}$, associated to p subtimenodes $t^{m,(p-1)}$, to get $\underline{\boldsymbol{c}}^{*(p-1)}$, associated to $p+1$ subtimenodes $t^{m,(p)}$. Such vector is then used to perform the iteration via (58). The updating formula stays formally identical, up to the fact that the coefficients θ_ℓ^m, at each iteration p, are the ones associated to the considered subtimenodes $t^{m,(p)}$. For efficiency reasons, the interpolation is not performed at the first and at the last iteration. Useful sketches of the original method and of the modification are shown in Figure 1, in particular, the crosses indicate the location in time of the quantities of interest. For further details, the reader is referred to [34].

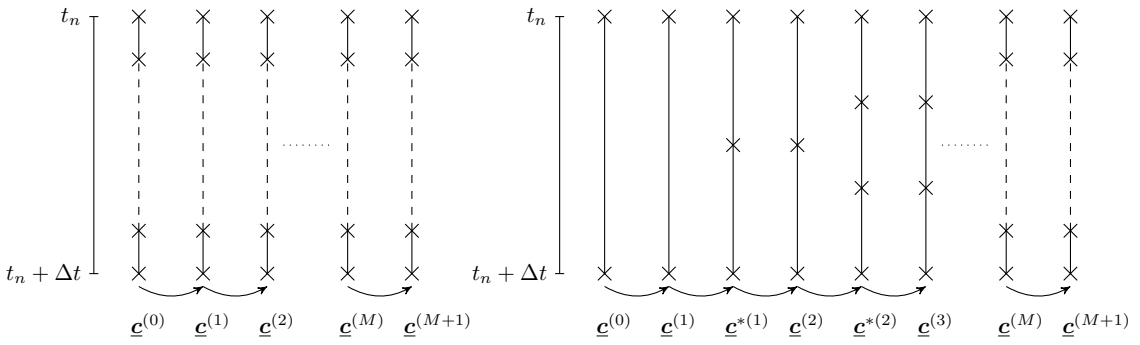


Figure 1: Sketch of the original DeC method on the left and of the modified one on the right

6 Numerical results

In this section, we will numerically investigate the elements previously introduced. Three different polynomial bases will be considered: the Bernstein polynomials, Bn $n = 1, 2, 3, 4$; the Lagrange polynomials associated to equispaced nodes, Pn $n = 1, 2, 3$; the Lagrange polynomials associated to GL nodes, PGLn $n = 1, 2, 3, 4$. More precisely, P1 and B1 coincide and P4 is not present because unstable. Moreover, for each basis PGLn, we adopt the associated GL quadrature formula in order to achieve a natural high order mass lumping.

In all the tests, the domain is $\Omega := (0, 25)$ and we assume no friction ($n_M = 0$), unless differently specified. For the convergence analyses, we will consider the following C^∞ bathymetry

$$B(x) := \begin{cases} 0.2 \exp\left(1 - \frac{1}{1-\left(\frac{x-10}{5}\right)^2}\right), & \text{if } 5 < x < 15, \\ 0, & \text{otherwise,} \end{cases} \tag{59}$$

while, in all the other cases, we will consider the C^0 bathymetry

$$B(x) := \begin{cases} 0.2 - 0.05(x-10)^2, & \text{if } 8 < x < 12, \\ 0, & \text{otherwise.} \end{cases} \tag{60}$$

Concerning the parameters $\alpha_{f,r} := \delta_r \bar{\rho}_f h_f^2$ in (39) and (40) and $\alpha_f := \delta \bar{\rho}_f h_f^2$ in (41), (43) and (45), we set $\bar{\rho}_f$ to be the spectral radius of the Jacobian $\boldsymbol{J} := \frac{\partial \boldsymbol{F}}{\partial \boldsymbol{u}}$ of the flux at the interface f and

$$h_f := \left(\frac{1}{2} \sum_{x_i \in \left(\cup_{K \in K_f} K\right)} \left| \left[\!\left[\frac{\partial}{\partial x} \varphi_i \right]\!\right] \right| \right)^{-1}, \tag{61}$$

where K_f denotes the set of the two elements containing the interface f. For the parameters δ_r and δ, we adopt the values reported in Table 1. Let us notice that, in the context of jc (39) and jt (40), we consider here the stabilization on the jump of the first and of the second derivatives.

In all the tests, we set CFL:=0.1, except for the ones involving B4 and PGL4, in which we adopt CFL:=0.05. In the context of unsteady tests, we will report results obtained with the basis

	P1=B1,PGL1	B2,P2,PGL2	B3,P3,PGL3	B4,PGL4
$\delta_1 = \delta$	0.05	0.3	0.15	0.5
δ_2	0.5	0.2	0.2	0.01

Table 1: Coefficients adopted for δ_r and δ in the CIP stabilizations for the different basis functions

functions PGLn only, since, as underlined in [7, 1, 36, 37, 34], the DeC methods for CG involving the low order mass lumping require more iterations than what expected from theory for discretizations from order 4 on, for unsteady simulations, in order to attain the formal order of accuracy.

Remark 6.1 (On the choice of the parameters)**.** *One must be very careful in tuning the coefficients δ_r and δ. A wrong choice can lead to unstable schemes, characterized by lower orders of convergence, with respect to the ones expected from theory, or even blow-ups. For further details, the reader is referred to the study of the linear stability presented in [36], where a collection of optimal settings in terms of CFL and stabilization parameters is reported. Nevertheless, the analysis behind the definition of such optimal settings does not directly apply to the context of this work for several reasons: the model problem was the linear advection equation, while, here we deal with the nonlinear SW equations; the DeC time integration method considered was the original formulation of the DeC presented in [7] and not the novel modification introduced in [34]; only the jump of the first derivative was taken into account, instead, here we consider the stabilization on the jump of the second derivative in the context of two stabilizations, i.e., jc* (39) *and jt* (40)*.*

Remark 6.2 (On the coupling of jr and jg with a WB space discretization)**.** *As already pointed out, the main focus of the stabilizations jr* (43) *and jg* (45) *is a consistent approximation of the same quantity: $\frac{\partial}{\partial x}\boldsymbol{F} - \boldsymbol{S}$. Generally speaking, there is no reason to use the stabilization based on the jump of the residual when one has the global flux at hand, e.g., in the context of the space discretization WB-GF. Similarly, when the global flux is not available, as in the context of the space discretization WB-HS, it is not reasonable to compute it and use it just for the stabilization. To sum up, as a general rule, we will couple the stabilization jr with the space discretization WB-HS and the stabilization jg with the space discretization WB-GF.*

Due to the huge amount of possible combinations bewteen the considered elements (in the context of every test, apart from the reference non-WB approach, we have two WB space discretizations, four WB CIP stabilizations, three types of basis functions and, for each of them, different degrees), we will not systematically present all the results, but rather some representatives, to allow a meaningful comprehension.

The numerical results are organized as follows. We will start by testing, in Section 6.1, the exact well-balancing with respect to the lake at rest steady state. We will collect the results, for the basis functions of highest degree, B4, PGL4 and P3, in tables. We will continue, in Section 6.2, with some convergence analyses to check the arbitary high order accuracy of all the elements introduced on smooth steady states: a supercritical flow, a subcritical flow and a transcritical flow. In particular, in order to provide results for all the different types of polynomial bases, the results of the convergence analysis on the three steady states will be shown respectively for B4, PGL4 and P3. Nevertheless, several extra comparisons will be reported, concerning the settings with the best performances. In Section 6.3, we will report the results of simulations involving the evolution of small perturbations of the lake at rest steady state. Finally, in Section 6.4, we will focus on

		L^1 error H	L^1 error q
Reference non-WB		1.577E-002	1.169E-003
WB-HS	jc (non-WB)	9.568E-004	2.060E-003
	jt	3.786E-015	1.084E-013
	je	9.028E-015	8.644E-014
	jr	2.515E-015	9.935E-014
WB-GF	jc (non-WB)	9.608E-004	2.080E-003
	jt	5.215E-015	1.441E-014
	je	4.510E-015	9.251E-015
	jg	4.393E-015	1.303E-014

Table 2: Lake at rest, B4

		L^1 error H	L^1 error q
Reference non-WB		1.028E-002	1.879E-003
WB-HS	jc (non-WB)	3.007E-004	5.963E-004
	jt	9.403E-013	4.418E-012
	je	9.396E-013	4.415E-012
	jr	9.409E-013	4.415E-012
WB-GF	jc (non-WB)	3.016E-004	6.168E-004
	jt	6.431E-013	2.659E-012
	je	6.423E-013	2.659E-012
	jg	6.431E-013	2.651E-012

Table 3: Lake at rest, PGL4

the evolution of small perturbations of general steady states, whose analytical expression is not available in closed-form, with and without friction.

6.1 Exact well-balancing for lake at rest

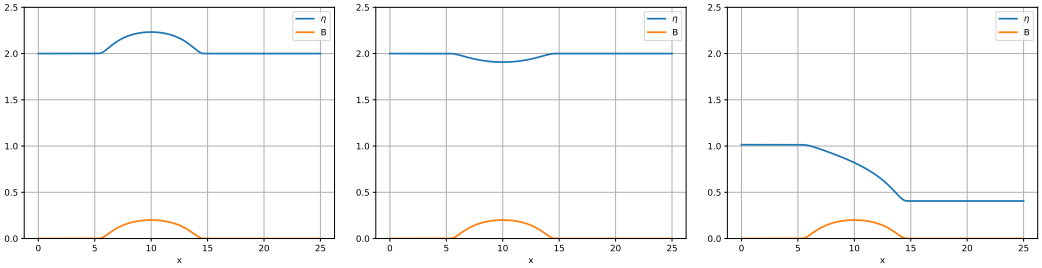
The simulations in this section are meant to test the WB feature with respect to the lake at rest steady state. We assume, in this context, the C^0 bathymetry (60), as all the WB elements that we introduced do not require any smoothness assumption. Let us consider the lake at rest steady state given by

$$\eta = H + B \equiv \overline{\eta}, \quad v \equiv 0, \tag{62}$$

with $\overline{\eta} := 0.5$ and a final time $T_f := 10$ with strong boundary conditions. The results got for $B4$, $PGL4$ and $P3$ and 100 elements are respectively reported in Tables 2, 3 and 4. As expected from theory, the reference non-WB approach gives an error which is far from machine precision. The same holds for schemes obtained by coupling a WB space discretization with the orginal non-WB stabilization jc. In all the remaining cases, we have combinations of WB elements and, in fact, the related errors are around machine precision.

		L^1 error H	L^1 error q
Reference non-WB		2.551E-002	2.931E-003
WB-HS	jc (non-WB)	4.185E-004	7.355E-004
	jt	7.342E-015	1.170E-014
	je	8.673E-015	1.451E-014
	jr	8.004E-015	1.382E-014
WB-GF	jc (non-WB)	4.206E-004	7.420E-004
	jt	6.213E-014	2.522E-013
	je	6.153E-014	2.502E-013
	jg	6.104E-014	2.502E-013

Table 4: Lake at rest, P3



(a) Supercritical (b) Subcritical (c) Transcritical

Figure 2: Smooth steady states

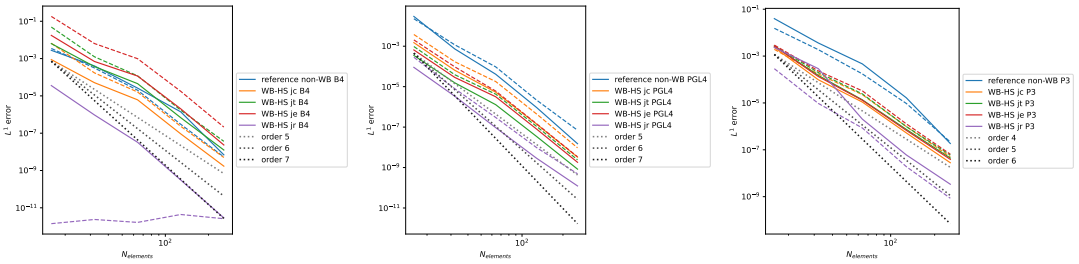
6.2 Arbitrary high order accuracy

In this section, we aim at numerically confirming the arbitrary high order accuracy of the considered space discretizations and of the novel jump stabilizations on smooth solutions. Therefore, for the tests presented here, we assume the C^∞ bathymetry (59). Let us consider the three frictionless isoenergetic smooth steady states [19] satisfying (8) with the following boundary conditions

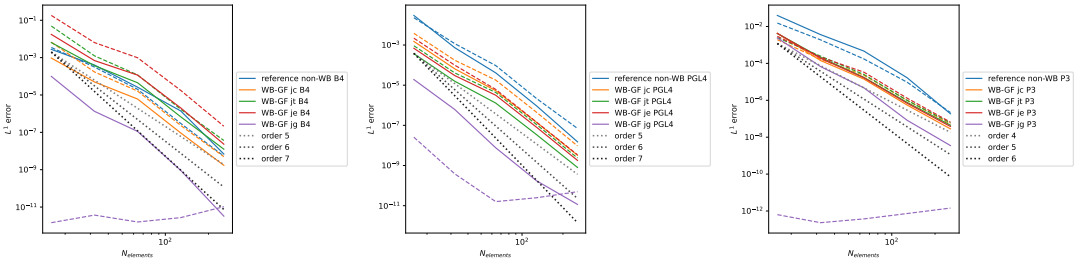
- **Supercritical**
$$q_L := 24, \quad H_L := 2, \tag{63}$$
- **Subcritical**
$$q_L := 4.42, \quad H_R := 2, \tag{64}$$
- **Transcritical**
$$q_L := 1.53, \tag{65}$$

where, due to the fact that the momentum of the flow must be constant, the value at the boundary prescribes also the value in the interior of the domain: $q \equiv q_L$. Since B is given, the total water height in each point, for the three steady states, can be (exactly) computed by solving (8) with respect to H and is depicted in Figure 2. We consider a final time $T_f := 100$.

The results of the convergence analysis for the three steady states, respectively with B4, PGL4 and P3, are reported in Figure 3. We can see how the formal order of accuracy is always recovered,



(a) Supercritical, WB-HS, B4 (b) Subcritical, WB-HS, PGL4 (c) Transcritical, WB-HS, P3



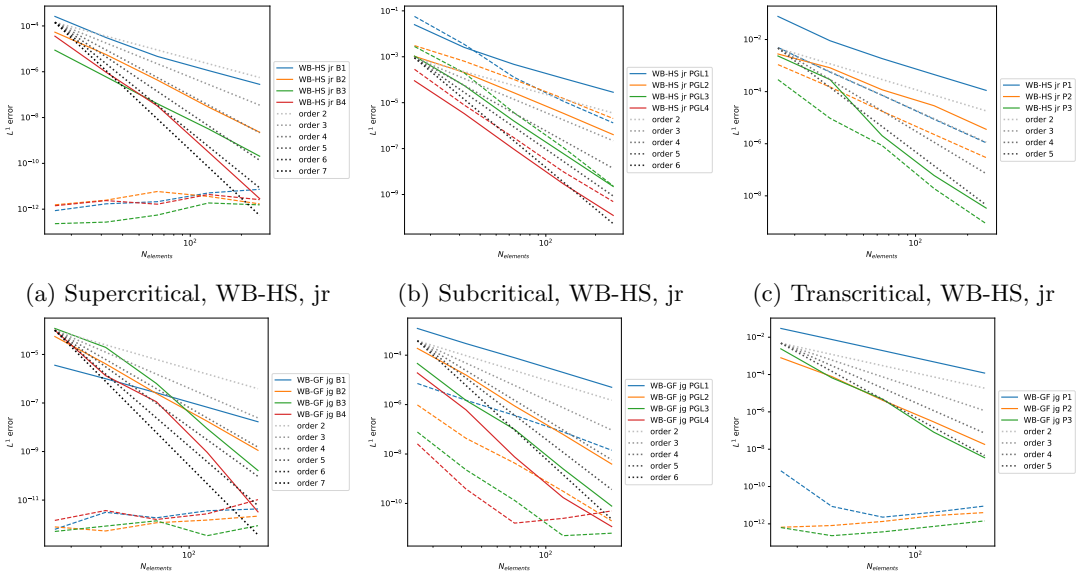
(d) Supercritical, WB-GF, B4 (e) Subcritical, WB-GF, PGL4 (f) Transcritical, WB-GF, P3

Figure 3: Convergence analysis: supercritical with B4, subcritical with PGL4 and transcritical with P3. L^1 error on H in continuous line, on q in dashed line

with very evident superconvergences for the stabilizations involving the jump of the residual, jr (43), and of the derivative of the global flux, jg (45). The errors obtained with such stabilizations are always much smaller than the ones obtained with the other schemes: roughly speaking, the difference is at least one order of magnitude, even more in the supercritical case. Further, the two stabilizations are characterized by steeper convergence slopes, with respect to the ones expected from theory, and a strong propensity to capture the constant momentum up to machine precision, see for example the supercritical tests.

We will focus now on the two best performing jump stabilizations, neglecting the other ones for the sake of compactness. In Figure 4, we display the results on the same tests for basis functions of different degrees. We can see that the superconvergences are not strictly related to the basis functions of highest degree. Apart from B1 (equivalent to P1) and PGL1, whose results confirm the expected second order accuracy, in almost all the other cases we experience convergence slopes steeper than the ones expected from theory:

- in the supercritical case, B2 converges with order 4 rather than 3, B4 with 7 rather than 5; further, only for jg, B3 converges with order 6 rather than 4;
- in the subcritical case, PGL3 converges with order 5 rather than 4; further, only for jg, PGL2 converges with order 4 rather than 3 and PGL4 with order 6 rather than 5;
- in the transcritical case, P3 converges with order 5 rather than 4 and, only for jg, P2 converges with order 4 rather than 3.



(a) Supercritical, WB-HS, jr (b) Subcritical, WB-HS, jr (c) Transcritical, WB-HS, jr

(d) Supercritical, WB-GF, jg (e) Subcritical, WB-GF, jg (f) Transcritical, WB-GF, jg

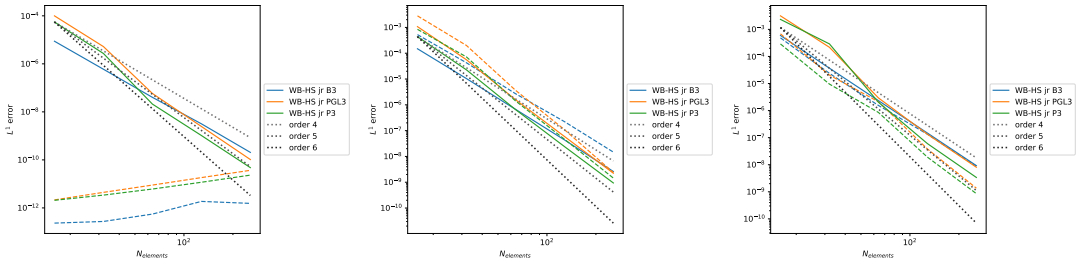
Figure 4: Convergence analysis: best performing settings for basis functions with different degrees. Supercritical with Bn, subcritical with PGLn and transcritical with Pn. L^1 error on H in continuous line, on q in dashed line

Moreover, also in this case, the ability of capturing exactly the constant momentum is very remarkable, see the supercritical case or the transcritical case with jg.

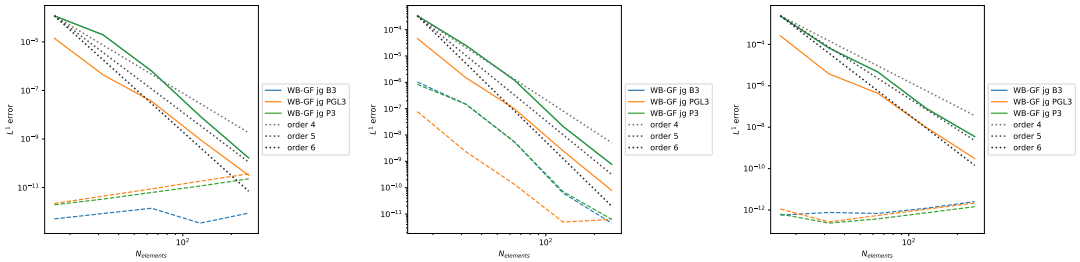
A comparison between B3, PGL3 and P3 on the three steady states is reported in Figure 5. We can see that in the context of the setting WB-HS-jr, P3 performs better than B3 and PGL3; this does not hold for WB-GF-jg, for which PGL3 is the best performing basis among the ones considered and the results of B3 and P3 are very similar.

Finally, we present a comparison between the best performing settings for the basis functions of highest degree B4, PGL4 and P3 in Figure 6. We already underlined, in Remark 6.2, that WB-HS-jr and WB-GF-jg represent the most natural couplings. Nevertheless, for the sake of curiosity, we will consider also the other two possible combinations. For the supercritical flow, WB-HS-jg is the best performing setting followed by WB-HS-jr; for the subcritical and the transcritical flows, we can see how WB-GF-jg is by far the best combination in terms of capturing of the constant momentum. In the context of the subcritical flow, such setting is also characterized by smaller errors on the water height, while, in the context of the transcritical flow the performance of all the settings under this point of view is not significatively different.

As already remarked, reporting the results for all the possible combinations of basis functions, space discretizations and jump stabilizations for any test would have been rather chaotic. For this reason, only the most significative ones have been selected. Nevertheless, we have tried, through several comparisons, to provide a wide variety of results for all the settings, focusing more on the best performing ones. Summarizing, the results seen in this subsection confirm the advantages in adopting the stabilizations jr (43) and jg (45) in the context of smooth steady states.

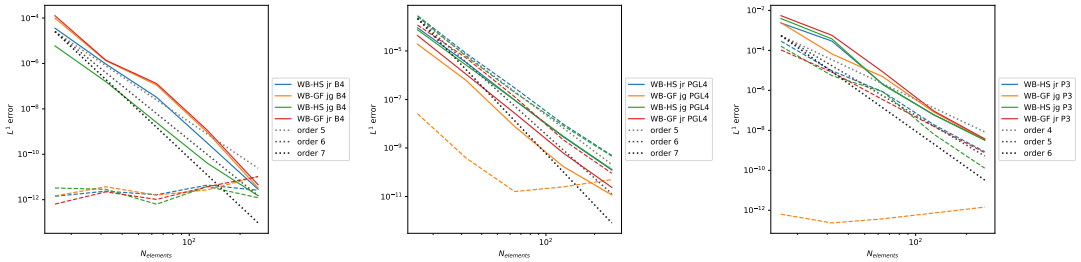


(a) Supercritical, WB-HS, jr (b) Subcritical, WB-HS, jr (c) Transcritical, WB-HS, jr



(d) Supercritical, WB-GF, jg (e) Subcritical, WB-GF, jg (f) Transcritical, WB-GF, jg

Figure 5: Convergence analysis: comparison between B3, PGL3 and P3 with the best performing settings. L^1 error on H in continuous line, on q in dashed line



(a) Supercritical (b) Subcritical (c) Transcritical

Figure 6: Convergence analysis: comparison between the best performing settings. L^1 error on H in continuous line, on q in dashed line

6.3 Evolution of small perturbations of lake at rest

In this section, we test the ability of the WB space discretizations and stabilizations to capture the evolution of small perturbations of the lake at rest steady state.

We consider again the reference test in Section 6.1 but we introduce the following small perturbation

$$\eta(x) := \begin{cases} \eta_s + A \exp\left(1 - \frac{1}{1-\left(\frac{x-6}{0.5}\right)^2}\right), & \text{if } 5.5 < x < 6.5, \\ \eta_s, & \text{otherwise,} \end{cases} \tag{66}$$

with $A := 5 \cdot 10^{-5}$, where $\eta_s \equiv \overline{\eta}$ represents the total water height of the steady state.

The initial condition and the evolution of the pertubation at the time $T_f := 1.5$, obtained with PGL4 and adopting non-WB settings, are depicted in Figure 7. For each setting, two results are plotted, one obtained with a coarse mesh with 30 elements, the other one obtained with a refined mesh with 128 elements. One can see that there is indeed an advantage in adopting a WB space discretization: in the context of the reference non-WB framework, the discretization error completely overwhelms the perturbation, while, in the other two cases, one gets spurious oscillations which are much smaller. Nevertheless, the presence of such oscillations testifies the non-WB character of the schemes obtained by coupling a WB space discretization, WB-HS or WB-GF, with a non-WB stabilization, jc.

In Figure 8, instead, one can see how the adoption of a fully WB scheme, i.e., for which also the stabilization is WB, is able to completely remove the spurious oscillations.

Analogous results have been got for PGLn with $n = 1, 2, 3$. For what concerns Bn and Pn, instead, the results are similar only up to order 3. For $n \geq 3$, the pathology of the time-stepping method for low order mass lumpings, already mentioned, prevents from recovering the formal order of accuracy without increasing the number of iterations with respect to what theoretically predicted. For the sake of compactness such results have been omitted.

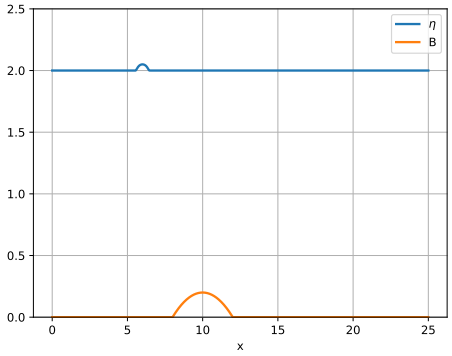
6.4 Evolution of small perturbations of moving equilibria

In this section, we test the WB properties of the introduced elements with respect to general steady states not known in closed-form. We remark that the two WB space discretizations here adopted, as well as all the novel CIP stabilizations, have been designed ad hoc to exactly preserve the lake at rest steady state. However, the last two stabilizations, jr (43) and jg (45), also address the problem of the preservation of general steady states, being based on discretizations of $\frac{\partial}{\partial x}\boldsymbol{F} - \boldsymbol{S}$. In fact, as already shown in the convergence analyses, such stabilizations are characterized by strong superconvergences towards steady states. This section is divided in two parts: in the first one we assume no friction, instead, in the second one we assume a Manning friction coefficient $n_M := 0.03$.

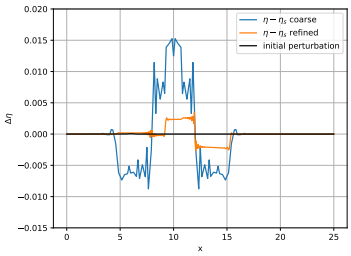
6.4.1 Tests without friction

We consider here the three non-smooth steady states characterized by the boundary conditions (63), (64) and (65) but with the C^0 bathymetry (60). We will analyze them separately in the following. Again, the water height can be retrieved via the (exact) solution of (8).

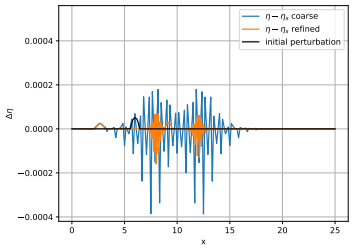
- **Supercritical flow**
  We consider in this case the same small perturbation (66) adopted for the lake at rest steady state but a different final time $T_f := 1$. Indeed, in this case η_s is not constant. The initial



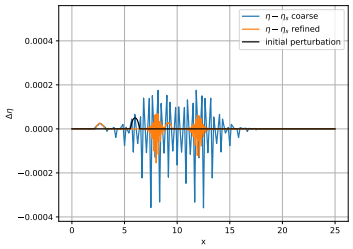
(a) Initial total height and bathymetry. The perturbation is amplified by a factor 10^3 in order to make it visible



(b) Reference non-WB setting

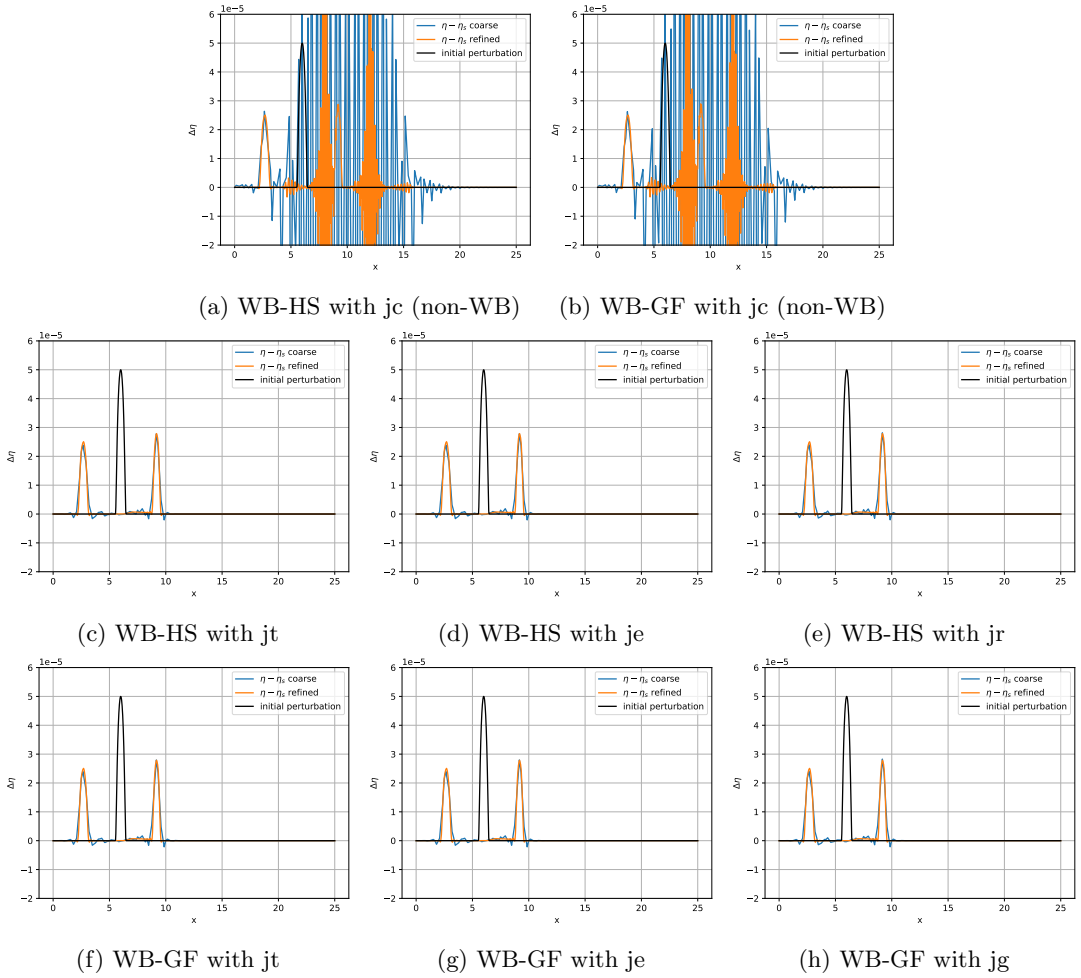


(c) WB-HS with jc (non-WB)



(d) WB-GF with jc (non-WB)

Figure 7: Perturbation of lake at rest: initial condition and results obtained with non-WB settings. Results referred to PGL4 with 30 elements for the coarse mesh and 128 elements for the refined mesh. A different scale has been used for the reference non-WB setting



(a) WB-HS with jc (non-WB)

(b) WB-GF with jc (non-WB)

(c) WB-HS with jt

(d) WB-HS with je

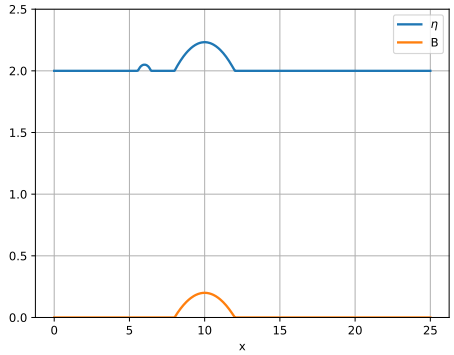
(e) WB-HS with jr

(f) WB-GF with jt

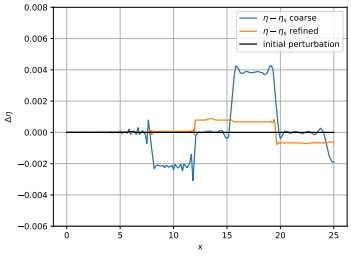
(g) WB-GF with je

(h) WB-GF with jg

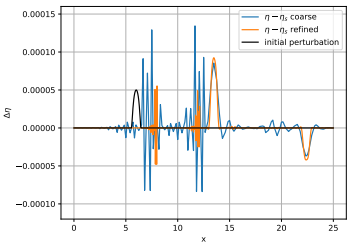
Figure 8: Perturbation of lake at rest: comparison between WB and non-WB settings. Results referred to PGL4 with 30 elements for the coarse mesh and 128 elements for the refined mesh



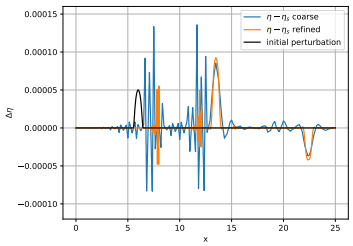
(a) Initial total height and bathymetry. The perturbation is amplified by a factor 10^3 in order to make it visible



(b) Reference non-WB setting



(c) WB-HS with jc

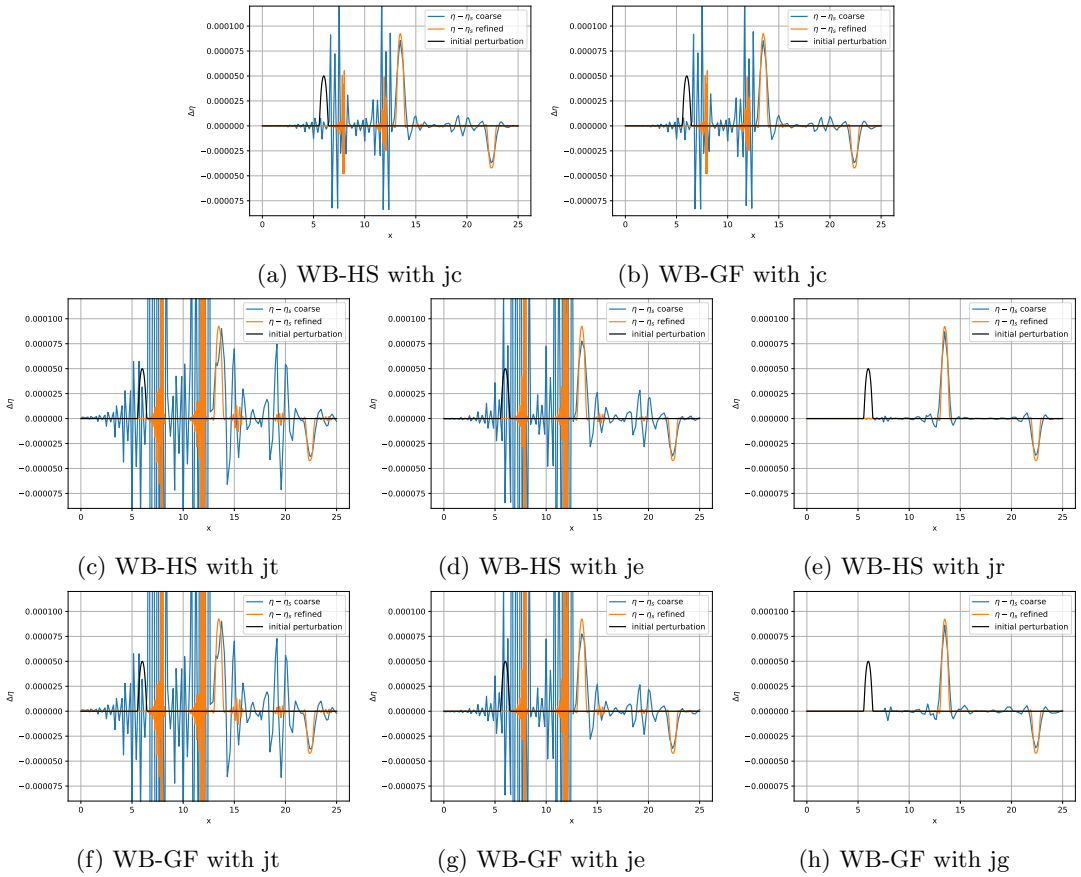


(d) WB-GF with jc

Figure 9: Perturbation of non-smooth frictionless supercritical steady state: initial condition and results obtained with non-WB settings. Results referred to PGL4 with 30 elements for the coarse mesh and 128 elements for the refined mesh. A different scale has been used for the reference non-WB setting

condition and the results got with the non-WB (with respect to lake at rest) settings are reported in Figure 9. Coherently with the previous case, the results are referred to PGL4 with 30 and 128 elements respectively for the coarse and the refined meshes. Again, there is a certain advantage in adopting the WB space discretizations, which seem to be more capable to handle a non-smooth bathymetry even for steady states different from the lake at rest. However, still they are characterized by spurious oscillations and this is not surprising as the elements have not been designed to achieve well-balancing with respect to a general steady state.

In Figure 10, we see the effect of the different stabilizations. It is immediately noticeable the ability of jr and jg to capture in a polite way the evolution of the perturbation without spurious oscillations. This feature can be somehow expected since, as already remarked, the two stabilizations are designed to stabilize the quantity $\frac{\partial}{\partial x}\boldsymbol{F} - \boldsymbol{S}$. Nevertheless, let us notice that no particular discretization has been adopted to make sure that they are exactly zero with respect to the investigated steady state. One can observe that very little spurious oscillations are present also for jr and jg in the results obtained with the coarse mesh, but this is normal,



(a) WB-HS with jc (b) WB-GF with jc

(c) WB-HS with jt (d) WB-HS with je (e) WB-HS with jr

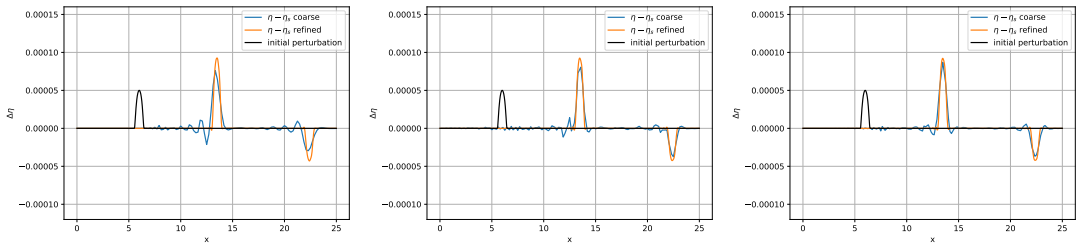
(f) WB-GF with jt (g) WB-GF with je (h) WB-GF with jg

Figure 10: Perturbation of non-smooth frictionless supercritical steady state: comparison between the different stabilizations. Results referred to PGL4 with 30 elements for the coarse mesh and 128 elements for the refined mesh

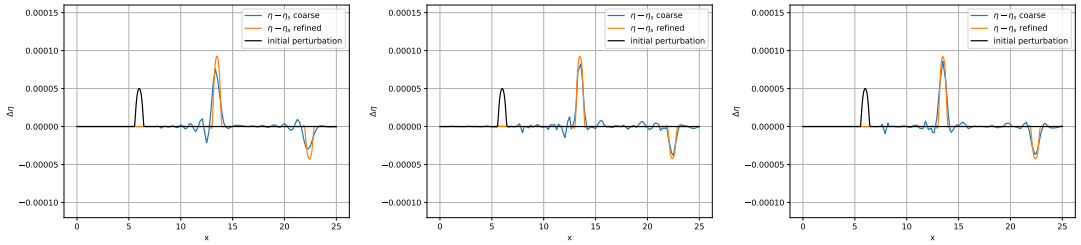
due to the lack of a limiting strategy. Such oscillations completely disappear in the mesh refinement.

An interesting "fair" comparison between basis functions with different degrees is displayed in Figure 11. The number of elements in the coarse meshes has been chosen in such a way that the total number of DoFs is constant. One can clearly see the effect of increasing the order of accuracy in the diminishing of the spurious oscillations both in number and magnitude, as well as in the better capturing of the peaks.

The ability of nicely capturing the evolution of the perturbation even with order 3 on a coarse mesh should not be taken for granted, see for example Figure 12, where the results of the same test with PGL2 and 60 elements are reported for the other jump stabilizations. Let us notice that the scale used for the plot completely differs from the one adopted in Figure 11 in order to entirely capture the spurious oscillations.

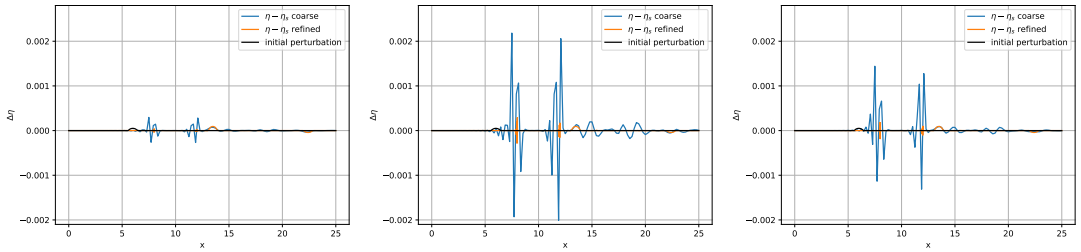


(a) WB-HS with jr, PGL2 (b) WB-HS with jr, PGL3 (c) WB-HS with jr, PGL4

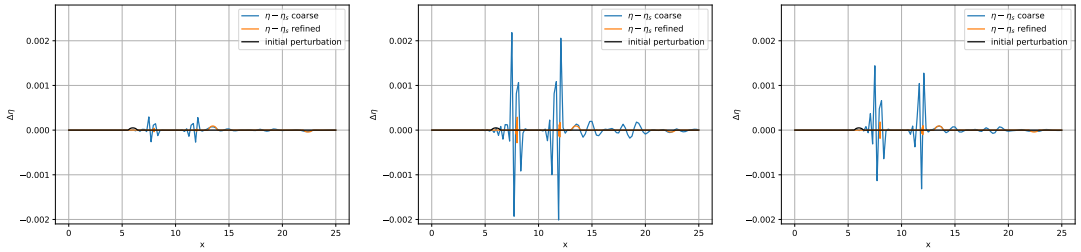


(d) WB-GF with jg, PGL2 (e) WB-GF with jg, PGL3 (f) WB-GF with jg, PGL4

Figure 11: Perturbation of non-smooth frictionless supercritical steady state: fair comparison between basis functions of different degree with jr and jg. Respectively 30, 40 and 60 elements for PGL4, PGL3 and PGL2 for the coarse meshes and 128, 256 and 512 elements for the refined ones

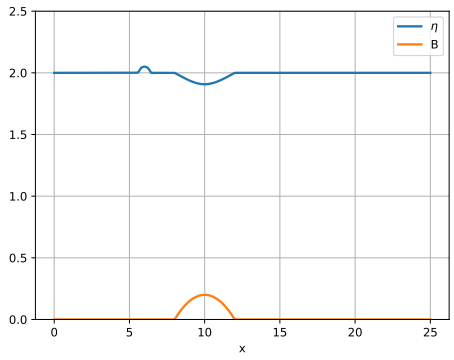


(a) WB-HS with jc (b) WB-HS with jt (c) WB-HS with je

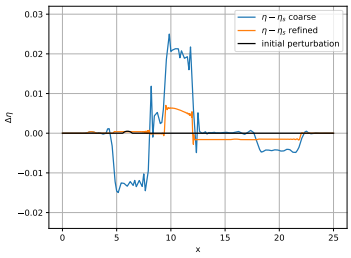


(d) WB-GF with jc (e) WB-GF with jt (f) WB-GF with je

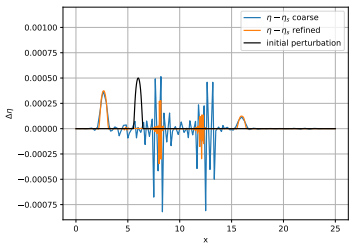
Figure 12: Perturbation of non-smooth frictionless supercritical steady state: results with all the stabilizations but jr and jg for PGL2 with 60 elements for the coarse mesh and 512 elements for the refined one



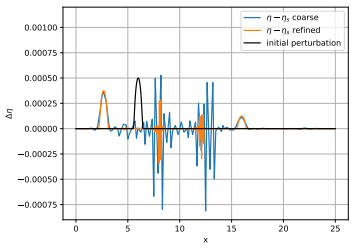
(a) Initial total height and bathymetry. The perturbation is amplified by a factor 10^2 in order to make it visible



(b) Reference non-WB setting



(c) WB-HS with jc



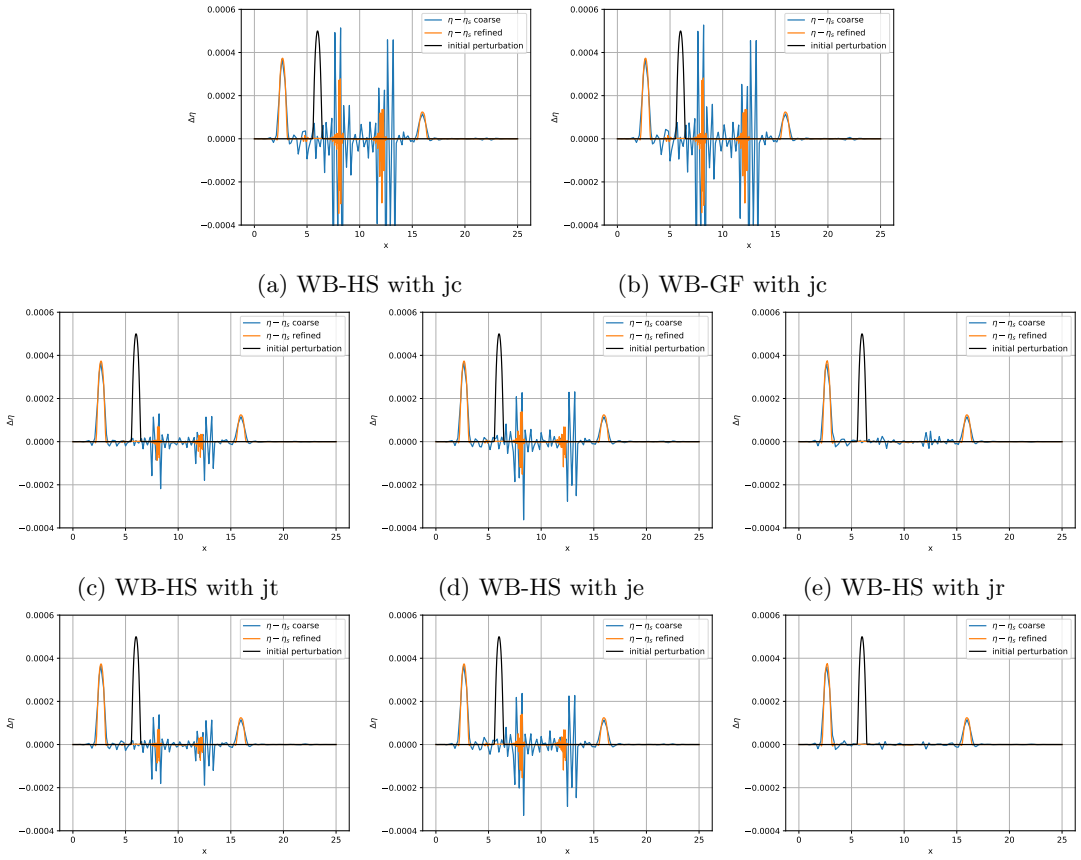
(d) WB-GF with jc

Figure 13: Perturbation of non-smooth frictionless subcritical steady state: initial condition and results obtained with non-WB settings. Results referred to PGL4 with 30 elements for the coarse mesh and 128 elements for the refined mesh. A different scale has been used for the reference non-WB setting

Again, we omit the results obtained with B2 and P2 and the ones got with all the basis functions of degree one. The former ones are analogous to the results obtained with PGL2, the latter ones are also qualitatively similar for a suitable amplitude of the perturbation.

- **Subcritical flow**
  The results got in this context are qualitatively similar to the ones obtained in the previous test, up to the fact that in this case we choose the same perturbation (66) but with $A := 5{\cdot}10^{-4}$ and a final time $T_f := 1.5$.

  Again, we start by showing, in Figure 13, the unsatisfactory results that one gets in the context of the reference non-WB framework and with the two WB space discretizations coupled with the original stabilization jc. However, also in this case we underline how the WB space discretizations are definitely more suitable when one has to deal with a non-smooth bathymetry. Another common feature shared with the supercritical case is the presence of spurious oscillations due to the lack of any particular attention to well-balancing toward a general steady state.



(a) WB-HS with jc (b) WB-GF with jc

(c) WB-HS with jt (d) WB-HS with je (e) WB-HS with jr

(f) WB-GF with jt (g) WB-GF with je (h) WB-GF with jg

Figure 14: Perturbation of non-smooth frictionless subcritical steady state: comparison between the different stabilizations. Results referred to PGL4 with 30 elements for the coarse mesh and 128 elements for the refined mesh

The advantages of adopting an approach oriented towards the preservation of a general steady state can bee seen in Figure 14. Again, the WB stabilizations jr and jg manage to remove almost completely the non-physical oscillations, which totally disappear in the mesh refinement. Also in this case, we remark that no limiting strategy has been adopted and this is the reason for the little fluctuations that one can see in the results associated to the coarse meshes.

For the sake of compactness, we omit here other results but analogous considerations to the ones reported at the end of the tests concerning the perturbation of the supercritical case hold.

- **Transcritical flow**
  The perturbation assumed in this context is identical to the one assumed in the subcritical

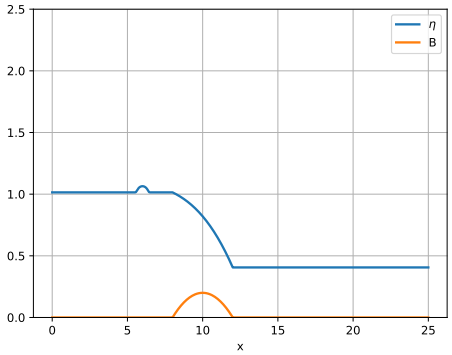


Figure 15: Perturbation of non-smooth frictionless transcritical steady state: initial total height and bathymetry. The perturbation is amplified by a factor 10^2 in order to make it visible

case, i.e., (66) with $A := 5 \cdot 10^{-4}$. We consider the same final time $T_f := 1.5$. The initial condition is displayed in Figure 15.

The results got for this steady state are analogous to the ones obtained in the previous cases and, therefore, for the sake of compactness, we directly focus on the comparison between the different stabilizations, which is reported in Figure 16. The advantages of basing the stabilization on $\frac{\partial}{\partial x}\boldsymbol{F} - \boldsymbol{S}$ are pretty evident. The spurious oscillations obtained with jr and jg are much smaller and controlled in terms of number and magnitude. The little ones still present in such cases, due to a lack of limiting, fade away in the mesh refinement.

For the sake of compactness, we omit other results but we remark that there are no significative differences with respect to the other non-smooth frictionless steady states.

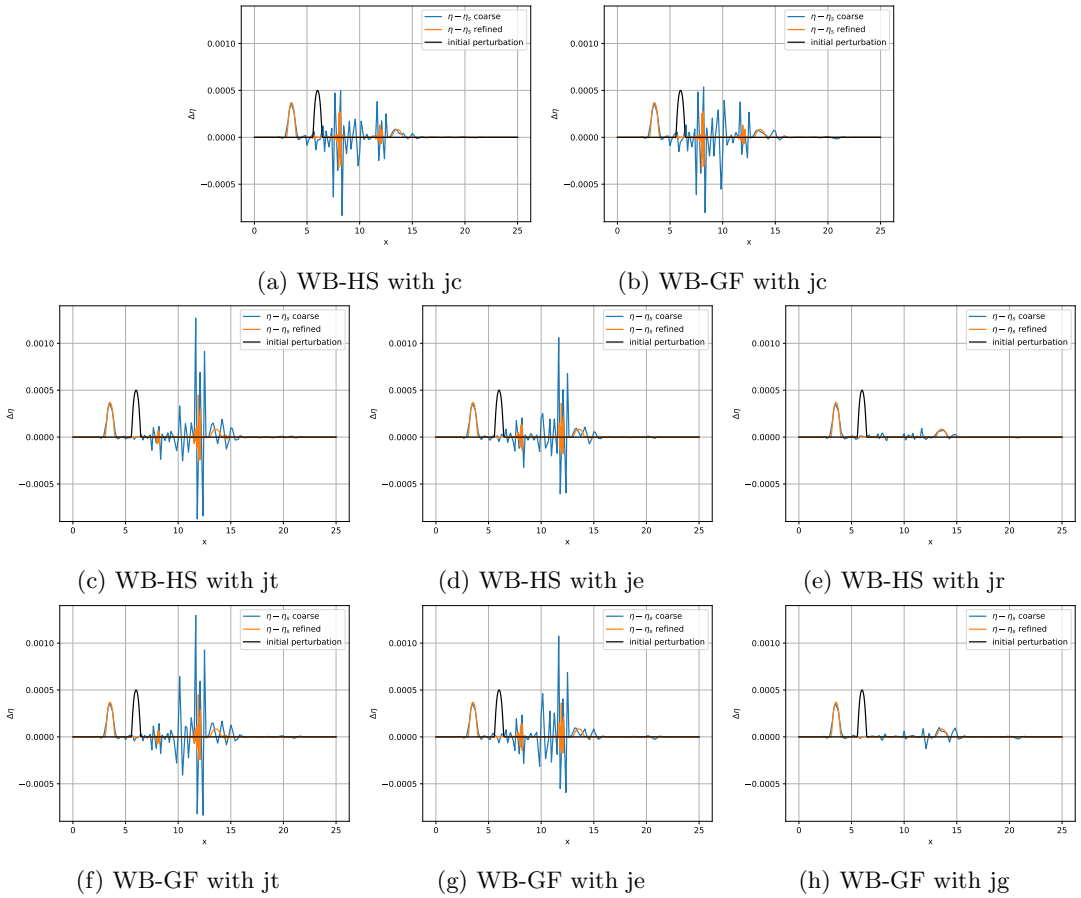
6.4.2 Tests with friction

In this last section, we will only focus on the supercritical and on the subcritical flows respectively characterized by the boundary conditions (63) and (64). In particular, we assume the usual C^0 bathymetry (60) and $n_M := 0.03$.

Just like in the frictionless tests of the previous section, the steady states are not available in closed-form. Moreover, in this context, (8) does not hold and there is no way to exactly compute the water height. Therefore, the steady states have been obtained by running simulations with very refined meshes, with 2048 elements and P1 basis functions, for time long enough and, finally, transferred to the meshes used for the tests through interpolation of η and q. In this context, as initial conditions, we adopted the frictionless steady states of the previous section. Further, for coherence, for each simulation involving the evolution of the perturbation of a steady state with a specific setting, the steady state obtained through the same setting is adopted as reference.

The perturbations and the final times assumed here are the same as the ones assumed in the frictionless case in the analogous tests of the previous section.

- **Supercritical case**



(a) WB-HS with jc (b) WB-GF with jc

(c) WB-HS with jt (d) WB-HS with je (e) WB-HS with jr

(f) WB-GF with jt (g) WB-GF with je (h) WB-GF with jg

Figure 16: Perturbation of non-smooth frictionless transcritical steady state: comparison between the different stabilizations. Results referred to PGL4 with 30 elements for the coarse mesh and 128 elements for the refined mesh

We start by showing, in Figure 17, the numerical steady states obtained with different settings. We can see how all the results provided are consistent: the friction causes a speed decrease in the direction of the flow, which, due to the constant momentum, is responsible for the general increase of the total height, from left to right, not present in the frictionless case. Concerning η, we cannot appreciate any macroscopical difference between the approximations provided by the different schemes. For what concerns q, instead, the reader is invited to notice the different scales used for the different settings: jr and jg are the only stabilizations able to capture the constant momentum up to machine precision. In all the other cases, the oscillations in correspondence of the discontinuities of the first derivative of the bathymetry are of the order of $10^{-3} \sim 10^{-4}$.

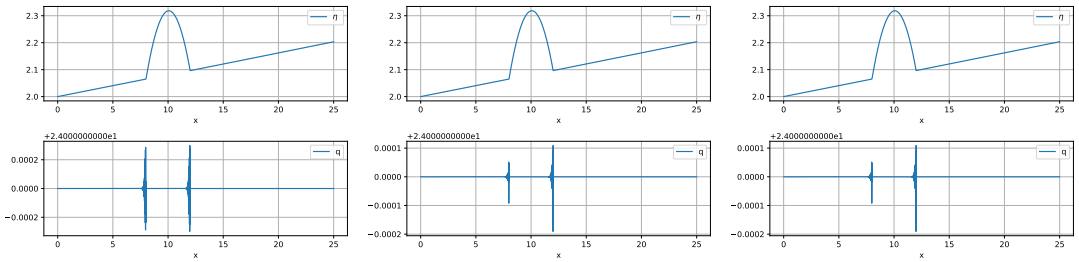
We continue now with the perturbation analysis. In Figures 18 and 19, one can see the evolution of the perturbation of the steady state obtained with several settings. The results are analogous to the ones obtained in the frictionless case: we can see an advantage in adopting the WB space discretizations in combination with jc with respect to the reference non-WB setting. As usual, the best results are the ones related to jr and jg. Analogous considerations hold with respect to the frictionless case. Let us notice, that in Figure 19, we had to choose a different scale for the results obtained with jt and je in order to capture the spurious oscillations in their entirety.

- **Subcritical case**

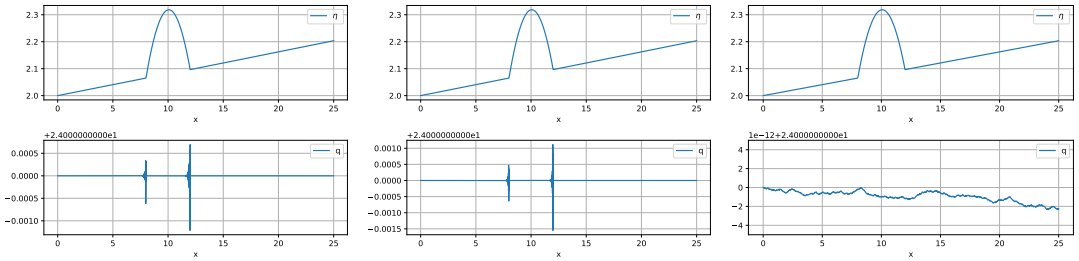
Also in this case, we focus first on the steady state obtained with different settings, reported in Figure 20. As confirmed by the numerical results, in this context we have a general increase in the velocity of the flow from left to right with consequent decrease of the water height. The numerical results confirm the consistency of all the settings but, again, there is remarkable difference between jr and jg and the other stabilizations in capturing the constant momentum. The amplitude of the spurious oscillations due to the non-smooth bathymetry, around $10^{-3} \sim 10^{-4}$ for all the other settings, jump down to 10^{-7} with jr and to machine precision with jg.

The comparison between the reference non-WB approach and the WB space discretizations coupled with the original jump jc gives results analogous to the ones obtained for all the previous tests, confirming the preferability of the latter settings to the original one, and it is, therefore, omitted. We directly focus on the comparison between the different CIP stabilizations, shown in Figure 21, from which we can further confirm the advantages in the adoption of jr and jg.

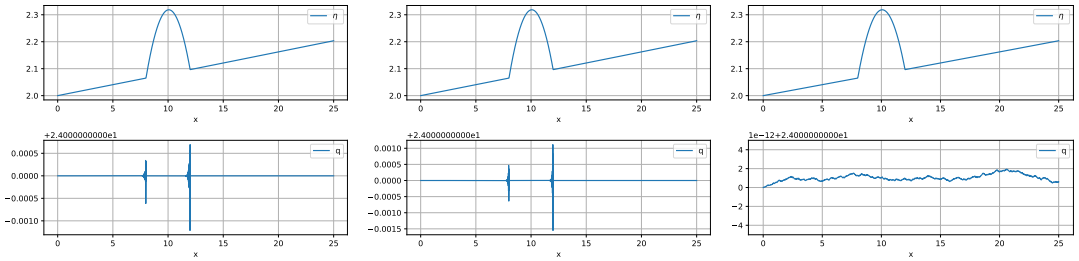
We close this section with a comparison, in Figure 22, between basis functions of different degrees for the best performing settings. Like in the context of the frictionless supercritical case, the number of the elements in the different coarse meshes is selected in such a way to have a constant number of DoFs. The results are analogous: the quality of the results improves, as the degree increases, in terms of ability to capture the peaks. Further, amplitude and number of the spurious oscillations decrease. Under this point of view, we remark that the remaining spurious oscillations, whose amplitude is however very small, are due to the fact that here we do not adopt any limiting technique and, moreover, they disappear in the mesh refinement. Indeed, an "unfair" comparison, with a constant number of elements, would give even better results.



(a) Reference non-WB (b) WB-HS with jc (c) WB-GF with jc

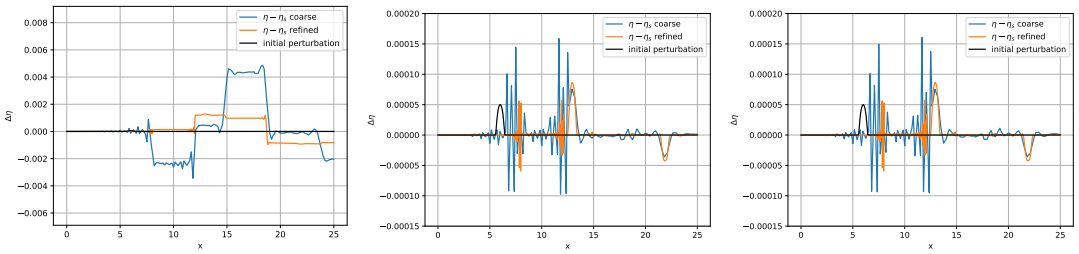


(d) WB-HS with jt (e) WB-HS with je (f) WB-HS with jr



(g) WB-GF with jt (h) WB-GF with je (i) WB-GF with jg

Figure 17: Non-smooth supercritical steady state with friction: numerical steady state obtained with different settings. Results referred to P1 with 2048 elements. Different scales have been used for q



(a) Reference non-WB setting (b) WB-HS with jc (c) WB-GF with jc

Figure 18: Perturbation of non-smooth supercritical steady state with friction: results obtained with non-WB settings. Results referred to PGL4 with 30 elements for the coarse mesh and 128 elements for the refined mesh. A different scale has been used for the reference non-WB setting

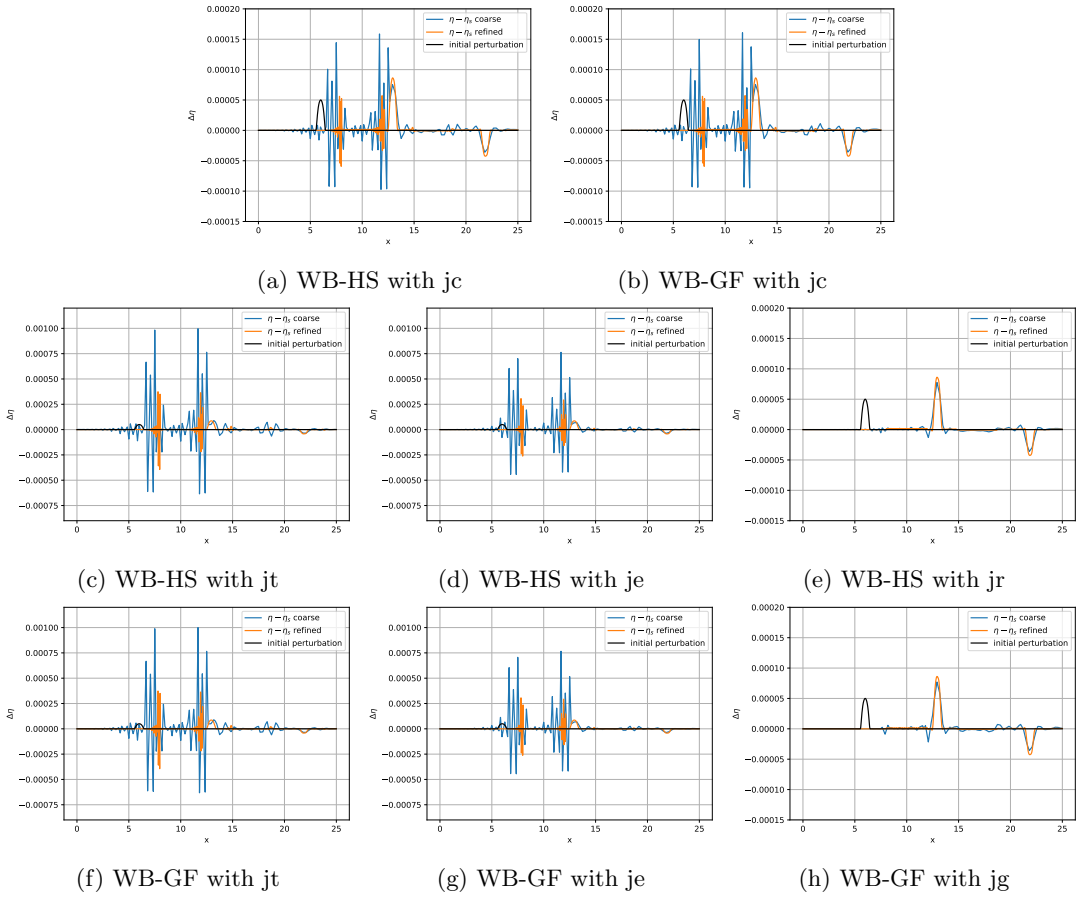
7 Conclusions and further developments

In this work, we have analyzed the performance of two WB space discretizations and four novel WB CIP stabilizations. All the elements are specifically designed to exactly preserve the lake at rest steady state. In particular, two stabilizations, jr and jg, address the problem of well-balancing toward a general steady state, being they based on a discretization of the steady equilibrium. The numerical results confirm the exact well-balancing with respect to the lake at rest and the arbitrary high order accuracy. Further, jr and jg have been shown able to better handle other general steady states in terms of superconvergences on smooth tests, with a strong propension in retrieving the constant momentum up to machine precision, and ability to capture the evolution of small perturbations of such steady states.

Some of the elements can be generalized: the stabilization based on the entropy variables je, the one based on the jump of the space residual jr, the notion of global flux and, hence, the related space discretization and stabilization jg can be extended to the Euler equations in a natural way; further, all the elements, apart from the ones relying on the notion of global flux, can be extended to a multidimensional unstructured setting in a straightforward way.

Acknowledgements

L. Micalizzi and R. Abgrall have been funded by the SNF grant 200020_204917. M. Ricchiuto is a member of the CARDAMOM team at INRIA University of Bordeaux.

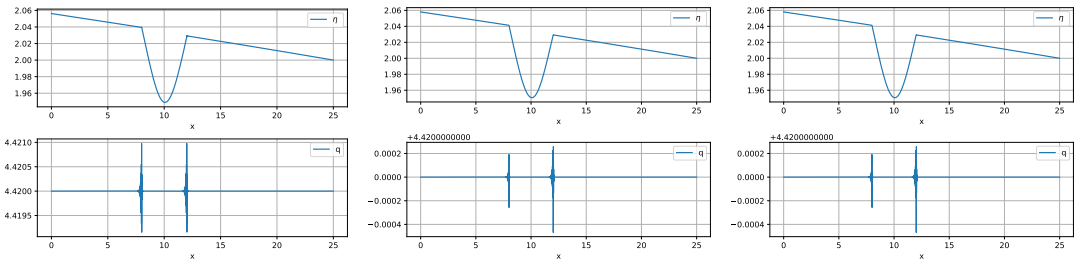


(a) WB-HS with jc (b) WB-GF with jc

(c) WB-HS with jt (d) WB-HS with je (e) WB-HS with jr

(f) WB-GF with jt (g) WB-GF with je (h) WB-GF with jg

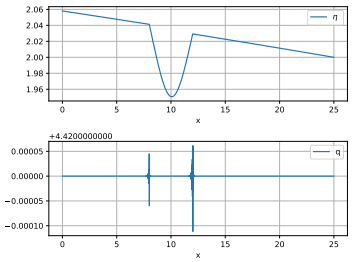
Figure 19: Perturbation of non-smooth supercritical steady state with friction: comparison between the different stabilizations. Results referred to PGL4 with 30 elements for the coarse mesh and 128 elements for the refined mesh. A different scale has been used for the results involving jt and je in order to capture the spurious oscillations in their entirety



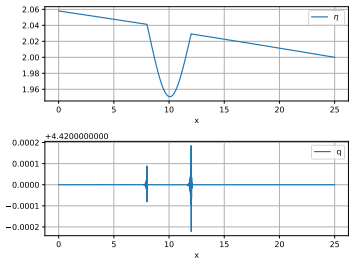
(a) Reference non-WB

(b) WB-HS with jc

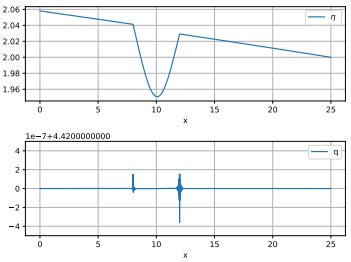
(c) WB-GF with jc



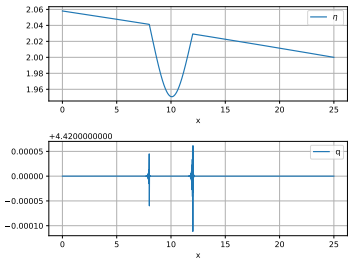
(d) WB-HS with jt



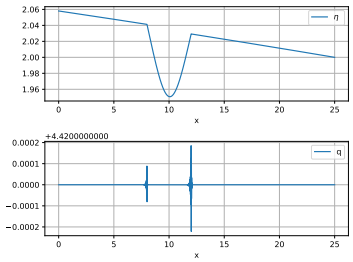
(e) WB-HS with je



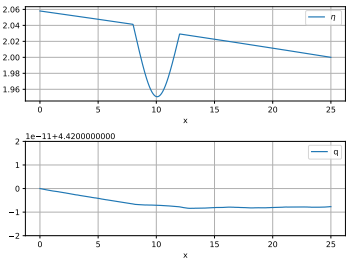
(f) WB-HS with jr



(g) WB-GF with jt

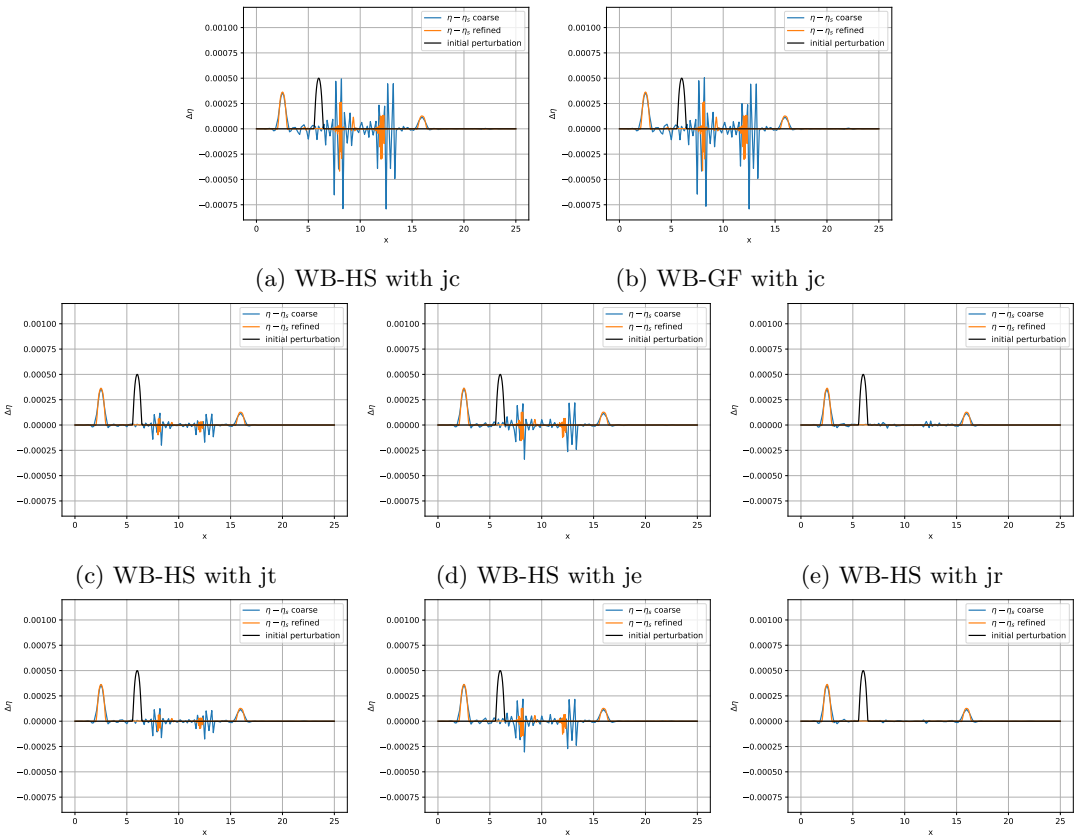


(h) WB-GF with je



(i) WB-GF with jg

Figure 20: Non-smooth subcritical steady state with friction: numerical steady state obtained with different settings. Results referred to P1 with 2048 elements. Different scales have been used for q

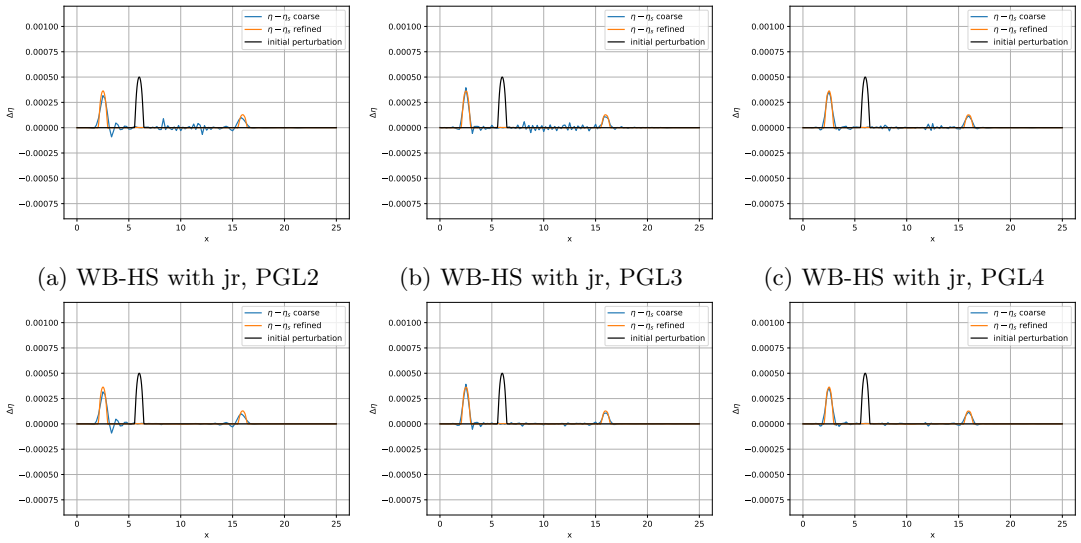


(a) WB-HS with jc (b) WB-GF with jc

(c) WB-HS with jt (d) WB-HS with je (e) WB-HS with jr

(f) WB-GF with jt (g) WB-GF with je (h) WB-GF with jg

Figure 21: Perturbation of non-smooth subcritical steady state with friction: comparison between the different stabilizations. Results referred to PGL4 with 30 elements for the coarse mesh and 128 elements for the refined mesh



(a) WB-HS with jr, PGL2 (b) WB-HS with jr, PGL3 (c) WB-HS with jr, PGL4

(d) WB-GF with jg, PGL2 (e) WB-GF with jg, PGL3 (f) WB-GF with jg, PGL4

Figure 22: Perturbation of non-smooth subcritical steady state with friction: fair comparison between basis functions of different degree with jr and jg. Respectively 30, 40 and 60 elements for PGL4, PGL3 and PGL2 for the coarse meshes and 128, 256 and 512 elements for the refined ones

A Proof of Proposition 4.1

Proof. The first point is a straightforward consequence of the assumptions made on the basis functions. In particular, we can write

$$\begin{aligned}\sum_{\boldsymbol{x}_i \in K} \mathbf{ST}_i^K(\boldsymbol{u}_h) &= \sum_{\boldsymbol{x}_i \in K} \sum_{\substack{f \subset \partial K \\ f \in \mathcal{F}_h}} \sum_{r=1}^{R} \alpha_{f,r} \int_f \nabla^r_{\boldsymbol{\nu}_f} \varphi_i|_K \left[\!\left[\nabla^r_{\boldsymbol{\nu}_f} \boldsymbol{u}_h \right]\!\right] d\sigma(\boldsymbol{x}) \\ &= \sum_{\substack{f \subset \partial K \\ f \in \mathcal{F}_h}} \sum_{r=1}^{R} \alpha_{f,r} \int_f \nabla^r_{\boldsymbol{\nu}_f} \left(\sum_{\boldsymbol{x}_i \in K} \varphi_i|_K \right) \left[\!\left[\nabla^r_{\boldsymbol{\nu}_f} \boldsymbol{u}_h \right]\!\right] d\sigma(\boldsymbol{x}),\end{aligned} \tag{67}$$

which is indeed equal to zero because of (10).

Now, let us deal with the second point. In order to light the notation, without loss of generality, we will focus on the r-th derivative only, dropping the sum over the orders, and we will neglect the factor $\alpha_{f,r}$. Then, what we want to show is the equivalence

$$\sum_{f \in \mathcal{F}_h} \int_f \left[\!\left[\nabla^r_{\boldsymbol{\nu}_f} \varphi_i \right]\!\right] \left[\!\left[\nabla^r_{\boldsymbol{\nu}_f} \boldsymbol{u}_h \right]\!\right] d\sigma(\boldsymbol{x}) = \sum_{K \in K_i} \sum_{\substack{f \subset \partial K \\ f \in \mathcal{F}_h}} \int_f \nabla^r_{\boldsymbol{\nu}_f} \varphi_i|_K \left[\!\left[\nabla^r_{\boldsymbol{\nu}_f} \boldsymbol{u}_h \right]\!\right] d\sigma(\boldsymbol{x}). \tag{68}$$

We start by observing that, in a conformal tessellation, all the faces $f \in \mathcal{F}_h$ shared by the elements are given by the intersection between two neighboring elements K and K'. Thanks to

this, the left-hand side of (68) can be written as

$$\sum_{f\in\mathcal{F}_h}\int_f \left[\!\left[\nabla^r_{\boldsymbol{\nu}_f}\varphi_i\right]\!\right]\left[\!\left[\nabla^r_{\boldsymbol{\nu}_f}\boldsymbol{u}_h\right]\!\right]d\sigma(\boldsymbol{x}) = \sum_{K\in\mathcal{T}_h}\sum_{\substack{K'\in\mathcal{T}_h\\K'\neq K}}\frac{1}{2}\int_{K\cap K'}\left[\!\left[\nabla^r_{\boldsymbol{\nu}_f}\varphi_i\right]\!\right]\left[\!\left[\nabla^r_{\boldsymbol{\nu}_f}\boldsymbol{u}_h\right]\!\right]d\sigma(\boldsymbol{x}), \tag{69}$$

where, with abuse of notation, we stick to $\boldsymbol{\nu}_f$ to indicate the normal to the face shared by K and K' at the right-hand side. We remark that the orientation of $\boldsymbol{\nu}_f$ is not relevant in this context.

Remark A.1. *In the previous equation* (69)*, the sums over K and K' are meant over all the elements of the tessellation: whenever two different elements K^a and K^b do not share any face, then $K^a\cap K^b=\emptyset$ and their contribution is zero. Instead, when they share a face f, the contribution of that face is counted twice: once when $K=K^a$ and $K'=K^b$, once when $K=K^b$ and $K'=K^a$. This is why we have to put $\frac{1}{2}$. We remark that, due to the assumption of conformal tessellation, any face f not belonging to the boundary is shared exactly by two elements.*

Concerning the direction of evaluation of the jump, not relevant in (19), we assume here $[\![z]\!] := z|_K - z|_{K'}$. Thus, from (69), one gets

$$\sum_{K\in\mathcal{T}_h}\sum_{\substack{K'\in\mathcal{T}_h\\K'\neq K}}\frac{1}{2}\int_{K\cap K'}\left[\!\left[\nabla^r_{\boldsymbol{\nu}_f}\varphi_i\right]\!\right]\left[\!\left[\nabla^r_{\boldsymbol{\nu}_f}\boldsymbol{u}_h\right]\!\right]d\sigma(\boldsymbol{x}) \tag{70a}$$

$$= \sum_{K\in\mathcal{T}_h}\sum_{\substack{K'\in\mathcal{T}_h\\K'\neq K}}\frac{1}{2}\int_{K\cap K'}\left(\nabla^r_{\boldsymbol{\nu}_f}\varphi_i|_K - \nabla^r_{\boldsymbol{\nu}_f}\varphi_i|_{K'}\right)\left[\!\left[\nabla^r_{\boldsymbol{\nu}_f}\boldsymbol{u}_h\right]\!\right]d\sigma(\boldsymbol{x}) \tag{70b}$$

$$= \sum_{K\in\mathcal{T}_h}\sum_{\substack{K'\in\mathcal{T}_h\\K'\neq K}}\frac{1}{2}\int_{K\cap K'}\nabla^r_{\boldsymbol{\nu}_f}\varphi_i|_K\left[\!\left[\nabla^r_{\boldsymbol{\nu}_f}\boldsymbol{u}_h\right]\!\right]d\sigma(\boldsymbol{x}) \tag{70c}$$

$$- \sum_{K\in\mathcal{T}_h}\sum_{\substack{K'\in\mathcal{T}_h\\K'\neq K}}\frac{1}{2}\int_{K\cap K'}\nabla^r_{\boldsymbol{\nu}_f}\varphi_i|_{K'}\left[\!\left[\nabla^r_{\boldsymbol{\nu}_f}\boldsymbol{u}_h\right]\!\right]d\sigma(\boldsymbol{x}) \tag{70d}$$

$$= \sum_{K\in\mathcal{T}_h}\sum_{\substack{K'\in\mathcal{T}_h\\K'\neq K}}\frac{1}{2}\int_{K\cap K'}\nabla^r_{\boldsymbol{\nu}_f}\varphi_i|_K\left(\nabla^r_{\boldsymbol{\nu}_f}\boldsymbol{u}_h|_K - \nabla^r_{\boldsymbol{\nu}_f}\boldsymbol{u}_h|_{K'}\right)d\sigma(\boldsymbol{x}) \tag{70e}$$

$$- \sum_{K\in\mathcal{T}_h}\sum_{\substack{K'\in\mathcal{T}_h\\K'\neq K}}\frac{1}{2}\int_{K\cap K'}\nabla^r_{\boldsymbol{\nu}_f}\varphi_i|_{K'}\left(\nabla^r_{\boldsymbol{\nu}_f}\boldsymbol{u}_h|_K - \nabla^r_{\boldsymbol{\nu}_f}\boldsymbol{u}_h|_{K'}\right)d\sigma(\boldsymbol{x}). \tag{70f}$$

Let us focus on the term at (70f). By a simple renaming of K and K' in such a way to switch the indices of the sums, by entering the sign $-$ inside the integral and from the fact that

$K \cap K' = K' \cap K$, we get

$$
\begin{aligned}
&- \sum_{K\in\mathcal{T}_h} \sum_{\substack{K'\in\mathcal{T}_h \\ K'\neq K}} \frac{1}{2} \int_{K\cap K'} \nabla^r_{\boldsymbol{\nu}_f} \varphi_i|_{K'} \left(\nabla^r_{\boldsymbol{\nu}_f} \boldsymbol{u}_h|_K - \nabla^r_{\boldsymbol{\nu}_f} \boldsymbol{u}_h|_{K'} \right) d\sigma(\boldsymbol{x}) && \text{(71a)} \\
&= - \sum_{K'\in\mathcal{T}_h} \sum_{\substack{K\in\mathcal{T}_h \\ K\neq K'}} \frac{1}{2} \int_{K'\cap K} \nabla^r_{\boldsymbol{\nu}_f} \varphi_i|_K \left(\nabla^r_{\boldsymbol{\nu}_f} \boldsymbol{u}_h|_{K'} - \nabla^r_{\boldsymbol{\nu}_f} \boldsymbol{u}_h|_K \right) d\sigma(\boldsymbol{x}) && \text{(71b)} \\
&= \sum_{K'\in\mathcal{T}_h} \sum_{\substack{K\in\mathcal{T}_h \\ K\neq K'}} \frac{1}{2} \int_{K'\cap K} \nabla^r_{\boldsymbol{\nu}_f} \varphi_i|_K \left(\nabla^r_{\boldsymbol{\nu}_f} \boldsymbol{u}_h|_K - \nabla^r_{\boldsymbol{\nu}_f} \boldsymbol{u}_h|_{K'} \right) d\sigma(\boldsymbol{x}) && \text{(71c)} \\
&= \sum_{K'\in\mathcal{T}_h} \sum_{\substack{K\in\mathcal{T}_h \\ K\neq K'}} \frac{1}{2} \int_{K\cap K'} \nabla^r_{\boldsymbol{\nu}_f} \varphi_i|_K \left(\nabla^r_{\boldsymbol{\nu}_f} \boldsymbol{u}_h|_K - \nabla^r_{\boldsymbol{\nu}_f} \boldsymbol{u}_h|_{K'} \right) d\sigma(\boldsymbol{x}) && \text{(71d)} \\
&= \sum_{K\in\mathcal{T}_h} \sum_{\substack{K'\in\mathcal{T}_h \\ K'\neq K}} \frac{1}{2} \int_{K\cap K'} \nabla^r_{\boldsymbol{\nu}_f} \varphi_i|_K \left(\nabla^r_{\boldsymbol{\nu}_f} \boldsymbol{u}_h|_K - \nabla^r_{\boldsymbol{\nu}_f} \boldsymbol{u}_h|_{K'} \right) d\sigma(\boldsymbol{x}), && \text{(71e)}
\end{aligned}
$$

where the last equality comes from the fact that in this case $\sum_{K\in\mathcal{T}_h} \sum_{\substack{K'\in\mathcal{T}_h \\ K'\neq K}}$ is equivalent to $\sum_{K'\in\mathcal{T}_h} \sum_{\substack{K\in\mathcal{T}_h \\ K\neq K'}}$. By replacing then (70f) with (71e), we get

$$
\begin{aligned}
&\sum_{K\in\mathcal{T}_h} \sum_{\substack{K'\in\mathcal{T}_h \\ K'\neq K}} \frac{1}{2} \int_{K\cap K'} \nabla^r_{\boldsymbol{\nu}_f} \varphi_i|_K \left(\nabla^r_{\boldsymbol{\nu}_f} \boldsymbol{u}_h|_K - \nabla^r_{\boldsymbol{\nu}_f} \boldsymbol{u}_h|_{K'} \right) d\sigma(\boldsymbol{x}) \\
&\quad - \sum_{K\in\mathcal{T}_h} \sum_{\substack{K'\in\mathcal{T}_h \\ K'\neq K}} \frac{1}{2} \int_{K\cap K'} \nabla^r_{\boldsymbol{\nu}_f} \varphi_i|_{K'} \left(\nabla^r_{\boldsymbol{\nu}_f} \boldsymbol{u}_h|_K - \nabla^r_{\boldsymbol{\nu}_f} \boldsymbol{u}_h|_{K'} \right) d\sigma(\boldsymbol{x}) \\
&= \sum_{K\in\mathcal{T}_h} \sum_{\substack{K'\in\mathcal{T}_h \\ K'\neq K}} \frac{1}{2} \int_{K\cap K'} \nabla^r_{\boldsymbol{\nu}_f} \varphi_i|_K \left(\nabla^r_{\boldsymbol{\nu}_f} \boldsymbol{u}_h|_K - \nabla^r_{\boldsymbol{\nu}_f} \boldsymbol{u}_h|_{K'} \right) d\sigma(\boldsymbol{x}) \\
&\quad + \sum_{K\in\mathcal{T}_h} \sum_{\substack{K'\in\mathcal{T}_h \\ K'\neq K}} \frac{1}{2} \int_{K\cap K'} \nabla^r_{\boldsymbol{\nu}_f} \varphi_i|_K \left(\nabla^r_{\boldsymbol{\nu}_f} \boldsymbol{u}_h|_K - \nabla^r_{\boldsymbol{\nu}_f} \boldsymbol{u}_h|_{K'} \right) d\sigma(\boldsymbol{x}) \qquad \text{(71f)} \\
&= \sum_{K\in\mathcal{T}_h} \sum_{\substack{K'\in\mathcal{T}_h \\ K'\neq K}} \int_{K\cap K'} \nabla^r_{\boldsymbol{\nu}_f} \varphi_i|_K \left(\nabla^r_{\boldsymbol{\nu}_f} \boldsymbol{u}_h|_K - \nabla^r_{\boldsymbol{\nu}_f} \boldsymbol{u}_h|_{K'} \right) d\sigma(\boldsymbol{x}) \\
&= \sum_{K\in\mathcal{T}_h} \sum_{\substack{K'\in\mathcal{T}_h \\ K'\neq K}} \int_{K\cap K'} \nabla^r_{\boldsymbol{\nu}_f} \varphi_i|_K \left[\!\left[\nabla^r_{\boldsymbol{\nu}_f} \boldsymbol{u}_h \right]\!\right] d\sigma(\boldsymbol{x}) \\
&= \sum_{K\in\mathcal{T}_h} \sum_{\substack{f\subset\partial K \\ f\in\mathcal{F}_h}} \int_{f} \nabla^r_{\boldsymbol{\nu}_f} \varphi_i|_K \left[\!\left[\nabla^r_{\boldsymbol{\nu}_f} \boldsymbol{u}_h \right]\!\right] d\sigma(\boldsymbol{x}).
\end{aligned}
$$

Now, since φ_i has support in the union of elements containing the node $\boldsymbol{x}_i$ to which it is associated, i.e., it is not identically zero just in the elements $K \in K_i$, we can write

$$\sum_{K\in\mathcal{T}_h}\sum_{\substack{f\subset\partial K\\ f\in\mathcal{F}_h}}\int_f \nabla^r_{\boldsymbol{\nu}_f}\varphi_i|_K \left[\!\left[\nabla^r_{\boldsymbol{\nu}_f}\boldsymbol{u}_h\right]\!\right] d\sigma(\boldsymbol{x}) = \sum_{K\in K_i}\sum_{\substack{f\subset\partial K\\ f\in\mathcal{F}_h}}\int_f \nabla^r_{\boldsymbol{\nu}_f}\varphi_i|_K \left[\!\left[\nabla^r_{\boldsymbol{\nu}_f}\boldsymbol{u}_h\right]\!\right] d\sigma(\boldsymbol{x}). \tag{71g}$$

With this, we have completed the proof of the equivalence (68). □

B Proof of Proposition 4.2

Proof. In the context of a lake at rest steady state, the velocity part of the flux and of the source are zero and the considered global flux values (35)-(36) reduce to

$$\boldsymbol{G}_h(x_i) = \boldsymbol{F}_h(x_i) + \boldsymbol{R}_h(x_i), \tag{72a}$$

$$\boldsymbol{F}_h(x_i) = \sum_{j=1}^{I} \boldsymbol{F}_j^{HS}\varphi_j(x_i) = \begin{pmatrix} 0 \\ \left[\frac{gH^2}{2}\right]_h(x_i) \end{pmatrix}, \quad \boldsymbol{R}_h(x_i) = -\int_{x_L}^{x_i} \left[-\begin{pmatrix} 0 \\ \left[gH\frac{\partial}{\partial x}B\right]_h(s) \end{pmatrix}\right] ds. \tag{72b}$$

We want to prove that, in such a case, $\boldsymbol{G}_h(x_i) \equiv const\ \forall i$. Actually, we can see that the first component is identically zero; therefore, let us consider the second component only

$$G_{h,2}(x_i) = \left[\frac{gH^2}{2}\right]_h(x_i) + \int_{x_L}^{x_i}\left[gH\frac{\partial}{\partial x}B\right]_h(s)ds \tag{72c}$$

$$= \left[\frac{gH^2}{2}\right]_h(x_i) + \int_{x_L}^{x_i}\left(\left[g(H_h+B_h)\frac{\partial}{\partial x}B_h\right]_h(s) - \frac{\partial}{\partial x}\left[\frac{gB^2}{2}\right]_h(s)\right) ds. \tag{72d}$$

Let us focus on K_1, the leftmost element of the tessellation, and let us consider $x_i \in K_1$. Through basic analysis, thanks to the linearity of the interpolation and to the fact that the total height is constant ($\eta \equiv \overline{\eta}$) for lake at rest, the integral in (72d) can be rewritten as

$$\int_{x_L}^{x_i}\left(\left[g(H_h+B_h)\frac{\partial}{\partial x}B_h\right]_h(s) - \frac{\partial}{\partial x}\left[\frac{gB^2}{2}\right]_h(s)\right) ds \tag{72e}$$

$$= \int_{x_L}^{x_i}\left(\left[g(H_h+B_h)\frac{\partial}{\partial x}B_h\right]_h(s)\right) ds - \left[\frac{gB^2}{2}\right]_h(x_i) + \left[\frac{gB^2}{2}\right]_h(x_L) \tag{72f}$$

$$= \int_{x_L}^{x_i}\left(\left[g\overline{\eta}\frac{\partial}{\partial x}B_h\right]_h(s)\right) ds - \left[\frac{gB^2}{2}\right]_h(x_i) + \left[\frac{gB^2}{2}\right]_h(x_L) \tag{72g}$$

$$= g\overline{\eta}\int_{x_L}^{x_i}\left(\left[\frac{\partial}{\partial x}B_h\right]_h(s)\right) ds - \left[\frac{gB^2}{2}\right]_h(x_i) + \left[\frac{gB^2}{2}\right]_h(x_L). \tag{72h}$$

Now, we have a crucial passage: since the interpolation B_h, restricted to K_1, lives in the space of the polynomials of degree M, its derivative lives in the space of the polynomials of degree $M-1$ and, hence, it can be interpolated exactly through the basis functions φ_i of degree M with support

in K_1 and so $\left[\frac{\partial}{\partial x}B_h\right]_h = \frac{\partial}{\partial x}B_h$. This allows to recast (72h) as

$$\begin{aligned} & g\overline{\eta}\int_{x_L}^{x_i}\left(\left[\frac{\partial}{\partial x}B_h\right]_h(s)\right)ds - \left[\frac{gB^2}{2}\right]_h(x_i) + \left[\frac{gB^2}{2}\right]_h(x_L) \\ &= g\overline{\eta}\int_{x_L}^{x_i}\left(\frac{\partial}{\partial x}B_h(s)\right)ds - \left[\frac{gB^2}{2}\right]_h(x_i) + \left[\frac{gB^2}{2}\right]_h(x_L) \\ &= g\overline{\eta}\left(B_h(x_i) - B_h(x_L)\right) - \left[\frac{gB^2}{2}\right]_h(x_i) + \left[\frac{gB^2}{2}\right]_h(x_L). \end{aligned} \tag{72i}$$

Coming back to (72d), thanks to the fact that $H = \overline{\eta} - B$, we have

$$\begin{aligned} G_{h,2}(x_i) &= \left[\frac{gH^2}{2}\right]_h(x_i) + \int_{x_L}^{x_i}\left(\left[g(H_h+B_h)\frac{\partial}{\partial x}B_h\right]_h(s) - \frac{\partial}{\partial x}\left[\frac{gB^2}{2}\right]_h(s)\right)ds \\ &= \left[\frac{gH^2}{2}\right]_h(x_i) + g\overline{\eta}\left(B_h(x_i) - B_h(x_L)\right) - \left[\frac{gB^2}{2}\right]_h(x_i) + \left[\frac{gB^2}{2}\right]_h(x_L) \\ &= \left[\frac{g(\overline{\eta}-B)^2}{2}\right]_h(x_i) + g\overline{\eta}\left(B_h(x_i) - B_h(x_L)\right) - \left[\frac{gB^2}{2}\right]_h(x_i) + \left[\frac{gB^2}{2}\right]_h(x_L) \\ &= \left[\frac{g(\overline{\eta}^2+B^2-2\overline{\eta}B)}{2}\right]_h(x_i) + g\overline{\eta}\left(B_h(x_i) - B_h(x_L)\right) - \left[\frac{gB^2}{2}\right]_h(x_i) + \left[\frac{gB^2}{2}\right]_h(x_L) \\ &= \frac{g\overline{\eta}^2}{2} + \left[\frac{gB^2}{2}\right]_h(x_i) - g\overline{\eta}B_h(x_i) + g\overline{\eta}\left(B_h(x_i) - B_h(x_L)\right) - \left[\frac{gB^2}{2}\right]_h(x_i) + \left[\frac{gB^2}{2}\right]_h(x_L) \\ &= \frac{g\overline{\eta}^2}{2} - g\overline{\eta}B_h(x_L) + \left[\frac{gB^2}{2}\right]_h(x_L) = \left[\frac{g\overline{\eta}^2+gB^2-2g\overline{\eta}B}{2}\right]_h(x_L) = const, \quad \forall x_i \in K^1. \end{aligned} \tag{72j}$$

We proved that, for the DoFs x_i in the first element, the global flux is equal to a constant independent of the specific DoF. Actually, exactly through the same computations, one can show that this holds more in general for any DoF

$$G_{h,2}(x_i) = \left[\frac{g\overline{\eta}^2+gB^2-2g\overline{\eta}B}{2}\right]_h(x_L) = const, \quad \forall i = 1,\dots,I. \tag{72k}$$

The key point is that, despite $\frac{\partial}{\partial x}B_h$ being discontinuous across the interfaces of the elements, B_h is continuous, leading to a cancellation effect in the integration over subsequent elements. This completes the proof. □

References

[1] Rémi Abgrall, Paola Bacigaluppi, and Svetlana Tokareva. High-order residual distribution scheme for the time-dependent euler equations of fluid dynamics. *Computers & Mathematics with Applications*, 78(2):274–297, 2019.

[2] Rémi Abgrall, Konstantin Lipnikov, Nathaniel Morgan, and Svetlana Tokareva. Multidimensional staggered grid residual distribution scheme for lagrangian hydrodynamics. *SIAM Journal on Scientific Computing*, 42(1):A343–A370, 2020.

[3] Rémi Abgrall, Élise Le Mélédo, Philipp Öffner, and Davide Torlo. Relaxation Deferred Correction Methods and their Applications to Residual Distribution Schemes. *The SMAI Journal of computational mathematics*, 8:125–160, 2022.

[4] Remi Abgrall and Mario Ricchiuto. High order methods for cfd, 2017.

[5] Remi Abgrall and Mario Ricchiuto. Hyperbolic balance laws: residual distribution, local and global fluxes. *Numerical Fluid Dynamics: Methods and Computations*, pages 177–222, 2022.

[6] Rémi Abgrall and Davide Torlo. High order asymptotic preserving deferred correction implicit-explicit schemes for kinetic models. *SIAM Journal on Scientific Computing*, 42(3):B816–B845, 2020.

[7] Rémi Abgrall. High order schemes for hyperbolic problems using globally continuous approximation and avoiding mass matrices. *J. Sci. Comput.*, 73(2-3):461–494, 2017.

[8] Jonas P Berberich, Praveen Chandrashekar, and Christian Klingenberg. High order well-balanced finite volume methods for multi-dimensional systems of hyperbolic balance laws. *Computers & Fluids*, 219:104858, 2021.

[9] Alexander N Brooks and Thomas JR Hughes. Streamline upwind/petrov-galerkin formulations for convection dominated flows with particular emphasis on the incompressible navier-stokes equations. *Computer methods in applied mechanics and engineering*, 32(1-3):199–259, 1982.

[10] Erik Burman and Peter Hansbo. Edge stabilization for galerkin approximations of convection–diffusion–reaction problems. *Computer Methods in Applied Mechanics and Engineering*, 193(15):1437–1453, 2004. Recent Advances in Stabilized and Multiscale Finite Element Methods.

[11] Yangyang Cao, Alexander Kurganov, Yongle Liu, and Ruixiao Xin. Flux globalization based well-balanced path-conservative central-upwind schemes for shallow water models. *Journal of Scientific Computing*, 92(2):69, 2022.

[12] Yangyang Cao, Alexander Kurganov, Yongle Liu, and Vladimir Zeitlin. Flux globalization based well-balanced path-conservative central-upwind scheme for two-layer thermal rotating shallow water equations. *Journal of Computational Physics*, 474:111790, 2023.

[13] Manuel J. Castro and Carlos Parés. Well-balanced high-order finite volume methods for systems of balance laws. *J. Sci. Comput.*, 82(2), 2020.

[14] Alina Chertock, Shumo Cui, Alexander Kurganov, Şeyma Nur Özcan, and Eitan Tadmor. Well-balanced schemes for the Euler equations with gravitation: conservative formulation using global fluxes. *J. Comput. Phys.*, 358:36–52, 2018.

[15] Alina Chertock, Alexander Kurganov, Xin Liu, Yongle Liu, and Tong Wu. Well-balancing via flux globalization: Applications to shallow water equations with wet/dry fronts. *Journal of Scientific Computing*, 90:1–21, 2022.

[16] Mirco Ciallella, Lorenzo Micalizzi, Philipp Öffner, and Davide Torlo. An arbitrary high order and positivity preserving method for the shallow water equations. *Computers & Fluids*, 247:105630, 2022.

[17] Mirco Ciallella, Davide Torlo, and Mario Ricchiuto. Arbitrary high order weno finite volume scheme with flux globalization for moving equilibria preservation. *Journal of Scientific Computing*, 96(2):53, 2023.

[18] Ramon Codina and Jordi Blasco. A finite element formulation for the stokes problem allowing equal velocity-pressure interpolation. *Computer Methods in Applied Mechanics and Engineering*, 143(3-4):373–391, 1997.

[19] Olivier Delestre, Carine Lucas, Pierre-Antoine Ksinant, Frédéric Darboux, Christian Laguerre, T-N-Tuoi Vo, Francois James, and Stéphane Cordier. Swashes: a compilation of shallow water analytic solutions for hydraulic and environmental studies. *International Journal for Numerical Methods in Fluids*, 72(3):269–300, 2013.

[20] Vivien Desveaux, Markus Zenk, Christophe Berthon, and Christian Klingenberg. A well-balanced scheme to capture non-explicit steady states in the Euler equations with gravity. *Internat. J. Numer. Methods Fluids*, 81(2):104–127, 2016.

[21] Jim Douglas and Todd Dupont. Interior penalty procedures for elliptic and parabolic galerkin methods. In *Computing methods in applied sciences*, pages 207–216. Springer, 1976.

[22] Alok Dutt, Leslie Greengard, and Vladimir Rokhlin. Spectral deferred correction methods for ordinary differential equations. *BIT*, 40(2):241–266, 2000.

[23] Manuel J. Castro Elena Gaburro and Michael Dumbser. Well-balanced arbitrary-lagrangian-eulerian finite volume schemes on moving nonconforming meshes for the euler equations of gas dynamics with gravity. *Monthly Notices of the Royal Astronomical Society*, 477(2):2251–2275, 2018.

[24] Leslie Fox and ET Goodwin. Some new methods for the numerical integration of ordinary differential equations. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 45, pages 373–388. Cambridge University Press, 1949.

[25] Ll Gascón and JM Corberán. Construction of second-order tvd schemes for nonhomogeneous hyperbolic conservation laws. *Journal of computational physics*, 172(1):261–297, 2001.

[26] Gregor J Gassner, Andrew R Winters, and David A Kopriva. A well balanced and entropy conservative discontinuous galerkin spectral element method for the shallow water equations. *Applied Mathematics and Computation*, 272:291–308, 2016.

[27] Maria Han Veiga, Philipp Öffner, and Davide Torlo. Dec and ader: similarities, differences and a unified framework. *Journal of Scientific Computing*, 87(1):1–35, 2021.

[28] Maria Han Veiga, David A. Velasco-Romero, Rémi Abgrall, and Romain Teyssier. Capturing near-equilibrium solutions: a comparison between high-order discontinuous Galerkin methods and well-balanced schemes. *Commun. Comput. Phys.*, 26(1):1–34, 2019.

[29] Alexander Kurganov, Yongle Liu, and Ruixiao Xin. Well-balanced path-conservative central-upwind schemes based on flux globalization. *Journal of Computational Physics*, 474:111773, 2023.

[30] Mats G Larson and Sara Zahedi. Stabilization of high order cut finite element methods on surfaces. *IMA Journal of Numerical Analysis*, 40(3):1702–1745, 2020.

[31] I MacDonald, MJ Baines, NK Nichols, and PG Samuels. Analytic benchmark solutions for open-channel flows. *Journal of Hydraulic Engineering*, 123(11):1041–1045, 1997.

[32] Ian MacDonald. *Analysis and computation of steady open channel flow*. PhD thesis, Citeseer, 1996.

[33] Yogiraj Mantri, Philipp Öffner, and Mario Ricchiuto. Fully well balanced entropy controlled dgsem for shallow water flows: global flux quadrature and cell entropy correction. *arXiv preprint arXiv:2212.11931*, 2022.

[34] Lorenzo Micalizzi and Davide Torlo. A new efficient explicit deferred correction framework: analysis and applications to hyperbolic pdes and adaptivity. *arXiv preprint arXiv:2210.02976*, 2022.

[35] Lorenzo Micalizzi, Davide Torlo, and Walter Boscheri. Efficient iterative arbitrary high order methods: an adaptive bridge between low and high order. *arXiv preprint arXiv:2212.07783*, 2022.

[36] Sixtine Michel, Davide Torlo, Mario Ricchiuto, and Rémi Abgrall. Spectral analysis of continuous FEM for hyperbolic PDEs: influence of approximation, stabilization, and time-stepping. *Journal of Scientific Computing*, 89(2):1–41, 2021.

[37] Sixtine Michel, Davide Torlo, Mario Ricchiuto, and Rémi Abgrall. Spectral analysis of high order continuous FEM for hyperbolic PDEs on triangular meshes: influence of approximation, stabilization, and time-stepping. *arXiv preprint arXiv:2206.06150*, 2022.

[38] Michael L Minion. Semi-implicit spectral deferred correction methods for ordinary differential equations. *Communications in Mathematical Sciences*, 1(3):471–500, 2003.

[39] Philipp Öffner and Davide Torlo. Arbitrary high-order, conservative and positivity preserving Patankar-type deferred correction schemes. *Appl. Numer. Math.*, 153:15–34, 2020.

[40] Hendrik Ranocha. Shallow water equations: split-form, entropy stable, well-balanced, and positivity preserving numerical methods. *GEM-International Journal on Geomathematics*, 8(1):85–133, 2017.

[41] Mario Ricchiuto. On the C-property and generalized C-property of residual distribution for the shallow water equations. *J. Sci. Comput.*, 48(1-3):304–318, 2011.

[42] Mario Ricchiuto and Andreas Bollermann. Stabilized residual distribution for shallow water simulations. *Journal of Computational Physics*, 228(4):1071–1115, 2009.

[43] Deepak Varma and Praveen Chandrashekar. A second-order, discretely well-balanced finite volume scheme for Euler equations with gravity. *Comput. & Fluids*, 181:292–313, 2019.

[44] Maria Han Veiga, Lorenzo Micalizzi, and Davide Torlo. On improving the efficiency of ader methods. *arXiv preprint arXiv:2305.13065*, 2023.

[45] Yulong Xing and Chi-Wang Shu. High order finite difference weno schemes with the exact conservation property for the shallow water equations. *Journal of Computational Physics*, 208(1):206–227, 2005.