WELL-BALANCED SECOND-ORDER APPROXIMATION OF THE SHALLOW WATER EQUATION WITH CONTINUOUS FINITE ELEMENTS*

PASCAL AZERAD†, JEAN-LUC GUERMOND‡ AND BOJAN POPOV‡

Abstract. This paper investigates a first-order and a second-order approximation technique for the shallow water equation with topography using continuous finite elements. Both methods are explicit in time and are shown to be well-balanced. The first-order method is invariant domain preserving and satisfy local entropy inequalities when the bottom is flat. Both methods are positivity preserving. Both techniques are parameter free, work well in the presence of dry states and can be made high-order in time by using strong-stability preserving time stepping algorithms.

Key words. Shallow water, well-balanced approximation, invariant domain, second-order method, finite element method, positivity preserving.

AMS subject classifications. 65M08, 65M60, 65M12, 35L50, 35L65, 76M10

1. Introduction.

The objective of this paper is to develop an invariant domain preserving well-balanced approximation of the shallow water equation with bathymetry using continuous finite elements. There are many finite volume and Discontinuous Galerkin (DG) techniques available in the literature that can solve this problem efficiently up to second and higher-order in space. Examples of schemes that are well balanced at rest and robust in the presence of dry states can be found, for example, in Audusse et al. [2], Audusse and Bristeau [1], Bollermann et al. [6], Gallardo et al. [14], Kurganov and Petrova [23], Perthame and Simeoni [27], Ricchiuto and Bollermann [28]. We refer the reader to the book of Bouchut [7] for a review on this topic, to the paper of Xing and Shu [32] for a survey on finite volume and DG methods, and to the paper [23] for a survey of central-upwind schemes. However, to the best of our knowledge, this type of approximations are not developed in the context of continuous finite elements. Or we should say that no robust continuous finite element technique is yet available in the literature that guarantees second-order accuracy, works properly in every regime (subcritical, transcritical, transcritical with hydraulic jumps, wet and dry regions) and is well-balanced at rest. We propose such a method in the present paper. Two variants of the method are discussed: one variant is first-order accurate in space, positivity preserving and preserves every convex invariant domain of the system in the absence of bathymetry; the other variant is second-order accurate in space and positivity preserving. Both variants are explicit in time and use continuous finite elements on unstructured meshes.

The first building block of the method consists of using the methodology introduced in Guermond and Popov [16]. The second building block consists of making the schemes well-balanced with respect to rest states by using the so-called hydrostatic reconstruction from [2, §2.1] and variations thereof. The technique from [16] is a loose

*This material is based upon work supported in part by the National Science Foundation grants DMS-1619892 and DMS-1620058, by the Air Force Office of Scientific Research, USAF, under grant/contract number FA9550-15-1-0257, and by the Army Research Office under grant/contract number W911NF-15-1-0517.

Draft version, October 22, 2018

†Institut Montpellierain Alexander Grothendieck, UMR 5149, Université de Montpellier, 34095 Montpellier, France

‡Department of Mathematics, Texas A&M University 3368 TAMU, College Station, TX 77843, USA.

This manuscript is for review purposes only.

extension of Lax's scheme [24, p.163] to continuous finite elements; it solves general hyperbolic systems in any space dimension using forward Euler time stepping and continuous finite elements on non-uniform grids. The artificial dissipation is defined so that any convex invariant sets containing the initial data is an invariant domain for the method. The solution thus constructed satisfies a discrete entropy inequality for every admissible entropy of the system. The accuracy in space is formally first-order and the accuracy in time can be made high-order by using Strong Stability Preserving Runge-Kutta time stepping. Some ideas of the method are rooted in the work of Hoff [20, 21], and Frid [13]. The method is made second-order and positivity preserving by using techniques introduced in Guermond and Popov [17].

The paper is organized as follows. The model problem and the finite element setting are introduced in §2. The first-order variant of the method is described in §3. The main results of this section are Propositions 3.9 and 3.11. The second-order variant of the method is described in §4. The key results of this section are Proposition 4.2 and 4.4. The performances of the algorithms introduced in the paper are numerically illustrated in §5 on standard benchmark problems.

2. Preliminaries.

In this section we introduce the model problem, the finite element setting and we define (recall) the concept of well-balancing at rest.

2.1. The model problem.

Let D be a polygonal domain in $\mathbb{R}^d$, with $d \in \{1, 2\}$, occupied by a body of water evolving in time under the action of gravity. Assuming that the deformations of the free surface are small compared to the water elevation and the bottom topography z varies slowly, the problem can be well represented by Saint-Venant's shallow water model. This model describes the time and space evolution of the water height h and flow rate, or discharge, $\boldsymbol{q}$ in the direction parallel to the bottom. Using $\boldsymbol{u} = (h, \boldsymbol{q})^\mathsf{T}$ as dependent variable the model is as follows:

$$\partial_t \boldsymbol{u} + \nabla\cdot \boldsymbol{f}(\boldsymbol{u}) + \boldsymbol{b}(\boldsymbol{u}, \nabla z) = 0, \quad \boldsymbol{x} \in D, t \in \mathbb{R}_+ \tag{2.1}$$

$$\boldsymbol{f}(\boldsymbol{u}) := \begin{pmatrix} \boldsymbol{q}^\mathsf{T} \\ \frac{1}{h}\boldsymbol{q}\otimes\boldsymbol{q} + \frac{1}{2}gh^2\mathbb{I}_d \end{pmatrix} \in \mathbb{R}^{(1+d)\times d}, \quad \boldsymbol{b}(\boldsymbol{u}, \nabla z) := \begin{pmatrix} 0 \\ gh\nabla z \end{pmatrix}. \tag{2.2}$$

The quantity $\boldsymbol{q}$ is related to the horizontal component of the water velocity $\boldsymbol{v}$ by $\boldsymbol{q} = \boldsymbol{v}h$. The function $z : D \ni \boldsymbol{x} \mapsto z(\boldsymbol{x}) \in \mathbb{R}$ is the given topography.

We assume that either the boundary conditions are periodic or the initial data $\boldsymbol{u}_0$ and the bottom topography are constant outside a compact set in D and the solution to (2.1) is constant outside this compact set over some time interval $[0, T]$.

2.2. The finite element space.

We approximate the solution of (2.2) with continuous finite elements. Let $(\mathcal{T}_h)_{h>0}$ be a shape-regular family of matching meshes. (Here we slightly abuse of notation by denoting the meshsize by h. For instance we are going to denote by h_h the finite element approximation of the water height.) The elements in $\mathcal{T}_h$ are assumed to be generated from a finite number of reference elements denoted $\{\widehat{K}_r\}_{1\le r\le\varpi}$. For example, the mesh $\mathcal{T}_h$ could be composed of a combination of triangles and quadrangles ($\varpi = 2$ in this case). Given a set of reference finite elements in the sense of Ciarlet $\{(\widehat{K}_r, \widehat{P}_r, \widehat{\Sigma}_r)\}_{1\le r\le\varpi}$ (the index $r \in \{1{:}\varpi\}$ is omitted in the rest of the paper to alleviate the notation) we introduce the finite element space

$$P(\mathcal{T}_h) := \{v \in C^0(D; \mathbb{R}) \mid v_{|K}\circ T_K \in \widehat{P},\ \forall K \in \mathcal{T}_h\} \tag{2.3}$$

where for any $K \in \mathcal{T}_h$, $T_K : \widehat{K} \to K$ is the geometric bijective transformation that maps the reference element $\widehat{K}$ to the current element K. We do not assume that

T_K is affine. The exact nature of the degrees of freedom in $\widehat{\Sigma}_r$ is not essential, but the reader who is not familiar with finite elements can think of Lagrange elements or Bernstein elements. The reference space $\widehat{P}$ is assumed to be composed of scalar-valued functions (these are polynomials usually). The reference shape functions are denoted $\{\widehat{\theta}_i\}_{i\in\{1:n_{\rm sh}\}}$; recall that they form a basis of $\widehat{P}$. We assume that the basis $\{\widehat{\theta}_i\}_{i\in\{1:n_{\rm sh}\}}$ has the partition of unity property: $\sum_{i\in\{1:n_{\rm sh}\}}\widehat{\theta}_i(\widehat{\boldsymbol{x}})=1$, for all $\widehat{\boldsymbol{x}}\in\widehat{K}$. The approximation in space of $\boldsymbol{u}$ in (2.2) will be done in $\boldsymbol{P}(\mathcal{T}_h):=[P(\mathcal{T}_h)]^{1+d}$. The approximation of the bathymetry map will be done in $P(\mathcal{T}_h)$. The global shape functions in $P(\mathcal{T}_h)$ are denoted by $\{\varphi_i\}_{i\in\{1:I\}}$; the set $\{\varphi_i\}_{i\in\{1:I\}}$ is a basis of $P(\mathcal{T}_h)$. The partition of unity property on the reference shape functions implies that

$$\sum_{i\in\{1:I\}}\varphi_i(\boldsymbol{x})=1,\qquad\forall\boldsymbol{x}\in D. \tag{2.4}$$

Let D_i be the support of φ_i and $|D_i|$ be the measure of D_i, $i\in\{1:I\}$. For any union of cells $E\subset\mathcal{T}_h$, we define $\mathcal{I}(E):=\{j\in\{1:I\}\mid |D_j\cap E|\neq 0\}$ to be the set that contains the indices of all the shape functions whose support on E is of nonzero measure. We are going to regularly invoke $\mathcal{I}(K)$ and $\mathcal{I}(D_i)$ and the partition of unity property: $\sum_{i\in\mathcal{I}(K)}\varphi_i(\boldsymbol{x})=1$ for all $\boldsymbol{x}\in K$.

Let $\mathcal{M}$ be the consistent mass matrix with entries $m_{ij}:=\int_D\varphi_i(\boldsymbol{x})\varphi_j(\boldsymbol{x})\,\mathrm{d}x$, and let $\mathcal{M}^L$ be the diagonal lumped mass matrix with entries $m_i:=\int_D\varphi_i(\boldsymbol{x})\,\mathrm{d}x$. The partition of unity property implies that $m_i=\sum_{j\in\mathcal{I}(D_i)}m_{ij}$. One key assumption that we use in the rest of the chapter is that

$$m_i>0,\quad\forall i\in\{1:I\}. \tag{2.5}$$

The identities (2.4) is satisfied by all the standard finite elements and (2.5) is satisfied by many Lagrange elements and by the Bernstein-Bezier elements of any degree.

Upon denoting by $\|\cdot\|_{\ell^2}$ the Euclidean norm in $\mathbb{R}^d$, we introduce the following two quantities which will play an important in the rest of paper:

$$\boldsymbol{c}_{ij}:=\int_D\varphi_i\nabla\varphi_j\,\mathrm{d}x,\qquad \boldsymbol{n}_{ij}:=\frac{\boldsymbol{c}_{ij}}{\|\boldsymbol{c}_{ij}\|_{\ell^2}}\qquad i,j\in\{1:I\}. \tag{2.6}$$

Note that (2.4) implies $\sum_{j\in\{1:I\}}\boldsymbol{c}_{ij}=\boldsymbol{0}$. Furthermore, if either φ_i or φ_j is zero on ∂D, then $\boldsymbol{c}_{ij}=-\boldsymbol{c}_{ji}$. In particular we have $\sum_{i\in\{1:I\}}\boldsymbol{c}_{ij}=\boldsymbol{0}$ if φ_j is zero on ∂D. This property will be used to establish conservation.

LEMMA 2.1. *Let $\boldsymbol{k}\in C^1(\mathbb{R}^{1+d};\mathbb{R}^{(1+d)\times d})$. Let $\boldsymbol{u}_h=\sum_{j\in\{1:I\}}\mathsf{U}_j\varphi_j\in\boldsymbol{P}(\mathcal{T}_h)$. Then $\sum_{j\in\mathcal{I}(D_i)}\boldsymbol{k}(\mathsf{U}_j)\cdot\boldsymbol{c}_{ij}$, is a second-order approximation of $\int_D\nabla\cdot(\boldsymbol{k}(\boldsymbol{u}_h))\varphi_i\,\mathrm{d}x$.*

Proof. Since we have $\int_{D_i}\nabla\cdot(\boldsymbol{k}(\boldsymbol{u}_h))\varphi_i\,\mathrm{d}x=\sum_{j\in\{1:I\}}\boldsymbol{k}(\mathsf{U}_j)\int_{D_i}\varphi_i\nabla\varphi_j\,\mathrm{d}x$ when $\boldsymbol{k}$ is linear, the quantity $\sum_{j\in\mathcal{I}(D_i)}\boldsymbol{k}(\mathsf{U}_j)\cdot\boldsymbol{c}_{ij}$ is a second-order approximation in space of $\int_D\nabla\cdot(\boldsymbol{k}(\boldsymbol{u}_h))\varphi_i\,\mathrm{d}x$, i.e., the error scales like $\mathcal{O}(h^2)\|\boldsymbol{c}_{ij}\|_{\ell^2}$. ∎

DEFINITION 2.2 (Centro-symmetry). *The mesh $\mathcal{T}_h$ is said to be centro-symmetric if the following conditions hold true:* (i) *For all $i\in\{1:I\}$, there is a permutation $\sigma_i:\mathcal{I}(D_i)\to\mathcal{I}(D_i)$ such $\boldsymbol{c}_{ij}=-\boldsymbol{c}_{i\sigma_i(j)}$,* (ii) *If the function $D_i\ni\boldsymbol{x}\to\sum_{j\in\mathcal{I}(D_i)}\alpha_j\varphi_j(\boldsymbol{x})\in\mathbb{R}$ is linear over D_i then $\alpha_i=\frac{1}{2}(\alpha_j+\alpha_{\sigma_i(j)})$ for all $j\in\mathcal{I}(D_i)$.*

For instance, in the context of Lagrange elements, the centro-symmetric assumption holds if for any $i\in\{1:I\}$ the set of the Lagrange nodes with indices in $\mathcal{I}(D_i)$

can be partitioned into pairs that are symmetric with respect to the Lagrange node of index i. Although at some point in the paper we will invoke centro-symmetry of the mesh to establish formal consistency of some terms, we do not assume that the mesh is centro-symmetric in the rest of the paper.

2.3. Well-balancing properties. The concept of well-balancing originates in the seminal work of Bermudez and Vazquez [4] and Greenberg and Leroux [15]. The idea is that the scheme should at the very least preserve steady states at rest. Of course, it could be desirable to preserve *general* steady solutions, i.e., not necessarily at rest, but this is beyond the scope of the present paper. We refer the reader to Noelle et al. [26] where this question is addressed. Since at rest $\boldsymbol{q} = \mathbf{0}$ the balance of momentum reduces to $\mathbf{0} = g\nabla(\frac{1}{2}h^2) + gh\nabla z = gh\nabla(h+z)$, one should have either $h+z$ is constant (so-called wet state) or h is zero (so-called dry state). Hence a well-balanced scheme in the context of the shallow water equation is one such that, at rest, dry states remain dry and $h+z$ remains constant for wet states. This property is not easy to satisfy for approximation techniques that are second-order and higher in space. We refer the reader to Bouchut [7] for a concise account and further references on well-balanced schemes. In this paper we are going to adapt to continuous finite elements a methodology proposed in Audusse et al. [2], Audusse and Bristeau [1] known as the "hydrostatic reconstruction" technique.

Let $z_h = \sum_{i=1}^{I} \mathsf{Z}_i\varphi_i \in P(\mathcal{T}_h)$ be the approximation of the bathymetry map. Let $h_h = \sum_{i=1}^{I} \mathsf{H}_i\varphi_i \in P(\mathcal{T}_h)$ be the approximation of the water height. Let $\boldsymbol{q}_h = \sum_{i=1}^{I} \mathbf{Q}_i\varphi_i$ be the approximation of the flow rate. Let us now define the rest state. Curiously, defining a rest state is not as trivial as it sounds. We are going to use two definitions. One of them makes use of the following quantity which is known in the literature as the hydrostatic reconstruction of the water height:

$$\mathsf{H}_i^{*,j} := \max(0, \mathsf{H}_i + \mathsf{Z}_i - \max(\mathsf{Z}_i, \mathsf{Z}_j)), \quad \forall i \in \{1{:}I\},\ j \in \mathcal{I}(D_i), \tag{2.7}$$

To better understand this definition, assume that the water is at rest and consider for instance a dry node j in the neighborhood of a wet node i, i.e., $j \in \mathcal{I}(D_i)$, see left panel of Fig 1. In this case $\mathsf{H}_j = 0$ and $\mathsf{Z}_j \geq \mathsf{H}_i + \mathsf{Z}_i$, which then implies $\mathsf{H}_i^{*,j} = \mathsf{H}_j^{*,i}$. Similarly if both i and j are dry states we have $\mathsf{H}_i^{*,j} = \mathsf{H}_j^{*,i}$, and if both i and j are wet states and are such that $\mathsf{H}_j + \mathsf{Z}_j = \mathsf{H}_i + \mathsf{Z}_i$ we also have $\mathsf{H}_i^{*,j} = \mathsf{H}_j^{*,i}$. These observations motivate the following definition.

Definition 2.3 (Rest at large). *A numerical state $(h_h, \boldsymbol{q}_h, z_h)$ is said to be at rest at large if the approximate momentum $\boldsymbol{q}_h$ is zero, and if the approximate water height h_h and the approximate bathymetry map z_h satisfy the following property for all $i \in \{1{:}I\}$: $\mathsf{H}_i^{*,j} = \mathsf{H}_j^{*,i}$ for all $j \in \mathcal{I}(D_i)$.*

Definition 2.4 (Exact rest). *A numerical state $(h_h, \boldsymbol{q}_h, z_h)$ is said to be at exact rest (or exactly at rest) if $\boldsymbol{q}_h$ is zero, and if the approximate water height h_h and the approximate bathymetry map z_h satisfy the following alternative for all $i \in \{1{:}I\}$: for all $j \in \mathcal{I}(D_i)$, either $\mathsf{H}_j = \mathsf{H}_i = 0$ or $\mathsf{H}_j + \mathsf{Z}_j = \mathsf{H}_i + \mathsf{Z}_i$.*

The existence of an exact rest state is a compatibility condition between the mesh and the initial data. This compatibility condition is not satisfied by the configuration depicted in the left panel of Figure 1 whereas it is satisfied by the configuration in the center panel. Exact rest implies rest at large. Note in passing that the zone where $h+z$ is constant may not be connected; that is to say, it is possible to have different free surface heights in disconnected wet zones as shown in the right panel of Figure 1.

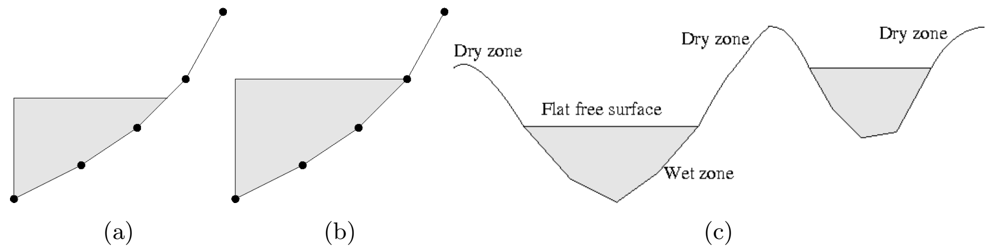

Fig. 1: Configuration (a) is not an exact rest state according to Definition 2.4 whereas configuration (b) is. Both states are at rest at large. Panel (c) shows a typical steady state at rest with wet and dry areas.

DEFINITION 2.5 (Well-balancing at large). (i) *A function* $\mathsf{K} : \boldsymbol{P}(\mathcal{T}_h) \to \mathbb{R}^I \times (\mathbb{R}^I)^d$ *is said to be a well-balanced flux approximation at large if* $\mathsf{K}(\boldsymbol{u}_h) = 0$ *when* $\boldsymbol{u}_h$ *is a rest state at large according to Definition 2.3.* (ii) *A mapping* $\boldsymbol{S} : \boldsymbol{P}(\mathcal{T}_h) \to \boldsymbol{P}(\mathcal{T}_h)$ *is a well-balanced scheme at large if* $\boldsymbol{S}(\boldsymbol{u}_h) = \boldsymbol{u}_h$ *when* $\boldsymbol{u}_h$ *is a rest state at large.*

DEFINITION 2.6 (Exact well-balancing). (i) *A function* $\mathsf{K} : \boldsymbol{P}(\mathcal{T}_h) \to \mathbb{R}^I \times (\mathbb{R}^I)^d$ *is said to be an exactly well-balanced flux approximation if* $\mathsf{K}(\boldsymbol{u}_h) = 0$ *when* $\boldsymbol{u}_h$ *is an exact rest state according to Definition 2.4.* (ii) *A mapping* $\boldsymbol{S} : \boldsymbol{P}(\mathcal{T}_h) \to \boldsymbol{P}(\mathcal{T}_h)$ *is an exactly well-balanced scheme if* $\boldsymbol{S}(\boldsymbol{u}_h^n) = \boldsymbol{u}_h^n$ *when* $\boldsymbol{u}_h^n$ *is an exact rest state.*

DEFINITION 2.7 (Conservation). *We say that* $\boldsymbol{u}_h^n \to \boldsymbol{u}_h^{n+1}$ *is a conservative finite element approximation of* (2.1) *if* $\sum_{i\in\{1:I\}} m_i \mathsf{H}_i^n = \sum_{i\in\{1:I\}} m_i \mathsf{H}_i^{n+1}$ *and if* $\sum_{i\in\{1:I\}} m_i \mathbf{Q}_i^n = \sum_{i\in\{1:I\}} m_i \mathbf{Q}_i^{n+1}$ *when the topography map is constant.*

3. First-order scheme. We describe in this section a time and space approximation of (2.2). The scheme is well-balanced at large but approximates the flux to first-order in space only. This scheme satisfies local invariant domain properties and local discrete entropy inequalities when the bottom is flat. It is an adaptation of the method presented in Audusse et al. [2] to the continuous finite element setting developed in Guermond and Popov [16]. To the best of our knowledge, this is the first result of this type for continuous finite elements.

3.1. Flux approximation. Just like in [2, (2.13)], the key is to consider the hydrostatic reconstruction (2.7) and to observe that $\sum_{j\in\mathcal{I}(D_i)} \frac{1}{2}((\mathsf{H}_j^{*,i})^2 - (\mathsf{H}_i^{*,j})^2)\boldsymbol{c}_{ij}$ is a well-balanced first-order approximation of the flux $\int_{D_i}(\nabla(\frac{1}{2}h^2) + h\nabla z)\varphi_i \,\mathrm{d}x$.

LEMMA 3.1 (Consistency/Well-balancing). (i) *Assume that* $\{\widehat{\theta}_n\}_{n\in\{1:n_{sh}\}}$ *consists of Lagrange or Bernstein functions. Then* $\sum_{j\in\mathcal{I}(D_i)} \frac{1}{2}((\mathsf{H}_j^{*,i})^2 - (\mathsf{H}_i^{*,j})^2)\boldsymbol{c}_{ij}$ *is a first-order approximation of the flux* $\int_{D_i}(\nabla(\frac{1}{2}h^2) + h\nabla z)\varphi_i \,\mathrm{d}x$. (ii) *The mapping* $\boldsymbol{u}_h \to (0, \sum_{j\in\mathcal{I}(D_i)} \frac{1}{2}((\mathsf{H}_j^{*,i})^2 - (\mathsf{H}_i^{*,j})^2)\boldsymbol{c}_{ij})_{i\in\{1:I\}}$ *is well-balanced at large.*

Proof. (i) Let us fix $i \in \{1:I\}$. We slightly abuse the notation by using h to denote the meshsize. For the consistency analysis we assume that the water height and the bathymetry map are smooth and the water height is non-negative. More precisely, we assume that there is C_z such that for all $i \in \{1:I\}$, $|\mathsf{Z}_i - \mathsf{Z}_j| \le C_z h$, for all $j \in \mathcal{I}(D_i)$

Assume first that $\mathsf{Z}_j \ge \mathsf{Z}_i$. We immediately get $\mathsf{H}_j^{*,i} = \mathsf{H}_j$. If in addition $\mathsf{H}_i \ge$

$C_z h$, then $\mathsf{H}_i^{*,j} = \max(0, \mathsf{H}_i + (\mathsf{Z}_i - \mathsf{Z}_j)) = \mathsf{H}_i + (\mathsf{Z}_i - \mathsf{Z}_j)$, and we have $\frac{1}{2}\big((\mathsf{H}_j^{*,i})^2 - (\mathsf{H}_i^{*,j})^2\big) = \frac{1}{2}\mathsf{H}_j^2 - \frac{1}{2}(\mathsf{H}_i + (\mathsf{Z}_i - \mathsf{Z}_j))^2 = \frac{1}{2}\mathsf{H}_j^2 - \frac{1}{2}\mathsf{H}_i^2 + \mathsf{H}_i(\mathsf{Z}_j - \mathsf{Z}_i) + \mathcal{O}(h^2)$. Similarly, if $\mathsf{H}_i \le C_z h$, then $\mathsf{H}_i^{*,j} = \mathcal{O}(h)$ and we again have $\frac{1}{2}\big((\mathsf{H}_j^{*,i})^2 - (\mathsf{H}_i^{*,j})^2\big) = \frac{1}{2}\mathsf{H}_j^2 - \frac{1}{2}\mathsf{H}_i^2 + \mathsf{H}_i(\mathsf{Z}_j - \mathsf{Z}_i) + \mathcal{O}(h^2)$. On the other hand, if $\mathsf{Z}_i \le \mathsf{Z}_j$, we obtain $\frac{1}{2}\big((\mathsf{H}_j^{*,i})^2 - (\mathsf{H}_i^{*,j})^2\big) = \frac{1}{2}\mathsf{H}_j^2 - \frac{1}{2}\mathsf{H}_i^2 + \mathsf{H}_j(\mathsf{Z}_j - \mathsf{Z}_i) + \mathcal{O}(h^2)$. But since $\mathsf{H}_j = \mathsf{H}_i + \mathcal{O}(h)$, (we are using continuous finite elements and the water height is assumed to be smooth) we also have $\frac{1}{2}\big((\mathsf{H}_j^{*,i})^2 - (\mathsf{H}_i^{*,j})^2\big) = \frac{1}{2}\mathsf{H}_j^2 - \frac{1}{2}\mathsf{H}_i^2 + \mathsf{H}_i(\mathsf{Z}_j - \mathsf{Z}_i) + \mathcal{O}(h^2)$ in this case.

Using Lemma 2.1 we infer that $\sum_{j\in\mathcal{I}(D_i)} \big(\frac{1}{2}\mathsf{H}_j^2 - \frac{1}{2}\mathsf{H}_i^2\big)\boldsymbol{c}_{ij}$ is a second-order approximation of $\int_D (\nabla(\frac{1}{2}h^2))\varphi_i \,\mathrm{d}x$. Similarly, $\sum_{j\in\mathcal{I}(D_i)} \big(\mathsf{H}_i(\mathsf{Z}_j - \mathsf{Z}_i)\big)\boldsymbol{c}_{ij}$ is a second-order approximation of $\mathsf{H}_i \int_D (\nabla z)\varphi_i \,\mathrm{d}x$. If z is linear over $\mathcal{D}_i$ (which is a sufficient assumption for the consistency analysis), then $\mathsf{H}_i \int_D (\nabla z)\varphi_i \,\mathrm{d}x = \nabla z_{|D_i} \mathsf{H}_i \int_D \varphi_i \,\mathrm{d}x$. Since $\mathsf{H}_i \int_D \varphi_i \,\mathrm{d}x$ can be shown to be a second-order approximation of $\int_{D_i} h\varphi_i \,\mathrm{d}x$ (at least for Lagrange and Bernstein basis functions), we conclude that $\sum_{j\in\mathcal{I}(D_i)} \big(\mathsf{H}_i(\mathsf{Z}_j - \mathsf{Z}_i)\big)\boldsymbol{c}_{ij}$ is a second-order approximation of $\int_D (h\nabla z)\varphi_i \,\mathrm{d}x$. Combining these observations with the above argument and upon observing that $\|\boldsymbol{c}_{ij}\|_{\ell^2}\mathcal{O}(h^2) = m_i\mathcal{O}(h)$, we conclude that $\sum_{j\in\mathcal{I}(D_i)} \frac{1}{2}((\mathsf{H}_j^{*,i})^2 - (\mathsf{H}_i^{*,j})^2\big)\boldsymbol{c}_{ij}$ is a first-order approximation of $\int_D (\nabla(\frac{1}{2}h^2) + h\nabla z)\varphi_i \,\mathrm{d}x$.

(ii) Let us prove the well-balancing at large. Assume that $\boldsymbol{u}_h$ is a rest state at large, according to Definition 2.3 we have $\mathsf{H}_j^{*,i} = \mathsf{H}_i^{*,j}$, hence $(\mathsf{H}_j^{*,i})^2 - (\mathsf{H}_i^{*,j})^2 = 0$. The conclusion follows immediately. ∎

Let us introduce the gas dynamics flux $\boldsymbol{g}(\boldsymbol{u}) := (\boldsymbol{q}, \frac{1}{h}\boldsymbol{q}\otimes\boldsymbol{q})^\mathsf{T}$. We now need to approximate $\int_{D_i} \boldsymbol{g}(\boldsymbol{u})\varphi_i \,\mathrm{d}x$. Since we have seen above that using H^* is a good idea to guarantee well-balancing at large, one could imagine working with the pair $(\mathsf{H}_i^{*,j}, \mathbf{Q}_i)^\mathsf{T}$. The problem with this choice is that if it happens that $\mathsf{H}_i^{*,j}$ is zero (because $\mathsf{H}_i + \mathsf{Z}_i \le \max(\mathsf{Z}_i, \mathsf{Z}_j)$), there is no reason for the approximate flow rate $\mathbf{Q}_i$ to be zero; hence the quantity $\mathbf{Q}_i/\mathsf{H}_i^{*,j}$ which approximate the velocity could be unbounded. To avoid this problem, we proceed as in [2] by working with the quantities

$$\mathbf{Q}_i^{*,j} := \mathbf{Q}_i \frac{\mathsf{H}_i^{*,j}}{\mathsf{H}_i}, \qquad \mathbf{U}_i^{*,j} := (\mathsf{H}_i^{*,j}, \mathbf{Q}_i^{*,j})^\mathsf{T}, \tag{3.1}$$

with the convention that $\mathbf{Q}_i^{*,j} := 0$ if $\mathsf{H}_i = 0$. Note that we have $\|\mathbf{Q}_i^{*,j}\|_{\ell^2} \le \|\mathbf{Q}_i\|_{\ell^2}$ since $0 \le \mathsf{H}_i^{*,j} \le \mathsf{H}_i$ by definition. We now face the question of constructing a consistent approximation of $\int_{D_i} \boldsymbol{g}(\boldsymbol{u})\varphi_i \,\mathrm{d}x$ using the state variable $\mathbf{U}_i^{*,j}$. To simplify the notation let us introduce the approximate velocity $\boldsymbol{v}_h = \sum_{i\in\{1:I\}} \mathbf{V}_i\varphi_i$ with

$$\mathbf{V}_i := \frac{\mathbf{Q}_i}{\mathsf{H}_i}, \qquad i \in \{1:I\}. \tag{3.2}$$

Definition 3.2 (Shoreline). *We say that a degree of freedom i is away from the shoreline if either $\mathsf{H}_j = 0$ for all $j \in \mathcal{I}(D_i)$ or $\min(\mathsf{H}_j, \mathsf{H}_i) > |\mathsf{Z}_i - \mathsf{Z}_j|$ for all $j \in \mathcal{I}(D_i)$.*

Note that if the bottom topography is smooth, i.e., there is C_z such that for all $i \in \{1:I\}$, $|\mathsf{Z}_i - \mathsf{Z}_j| \le C_z h$, then any degree of freedom i such that $\mathsf{H}_j \ge C_z h$, for all $j \in \mathcal{I}(D_i)$, is away from the shoreline according to the above definition. Roughly speaking, a degree of freedom i is said to be away from the shoreline if either all the degrees of freedom around i are dry or the water depth around i is at least $C_z h$ if the bottom topography is smooth (h being the meshsize).

LEMMA 3.3. *The quantity* $\sum_{j\in\mathcal{I}(D_i)}(\boldsymbol{g}(\mathsf{U}_j^{*,i})+\boldsymbol{g}(\mathsf{U}_i^{*,j}))\cdot\boldsymbol{c}_{ij}$ *is a first-order approximation of* $\int_{D_i}\nabla\cdot\boldsymbol{g}(\boldsymbol{u})\varphi_i\,\mathrm{d}x$ *away from the shoreline if the mesh is centro-symmetric.*

Proof. Let $i\in\{1{:}I\}$ be a degree of freedom away from the shoreline. The approximation of the flux is $\sum_{j\in\mathcal{I}(D_i)}(\mathbf{V}_j\mathsf{H}_j^{*,i}+\mathbf{V}_i\mathsf{H}_i^{*,j}))\cdot\boldsymbol{c}_{ij}$ for the mass conservation equation and $\sum_{j\in\mathcal{I}(D_i)}((\mathbf{V}_j\otimes\mathbf{V}_j)\mathsf{H}_j^{*,i}+(\mathbf{V}_i\otimes\mathbf{V}_i)\mathsf{H}_i^{*,j}))\cdot\boldsymbol{c}_{ij}$ for the flow rate conservation. Let us start with the mass conservation equation. We proceed as in the proof of Lemma 3.1 and again assume that the water height and the bathymetry map are smooth and the water height is non-negative. Since the mesh is centro-symmetric by hypothesis, we can assume without loss of generality that $\mathsf{Z}_j\ge\mathsf{Z}_i\ge\mathsf{Z}_{\sigma_i(j)}$. Then $\mathsf{H}_j^{*,i}=\mathsf{H}_j$ and since i is away from the shoreline we have either $\mathsf{H}_i^{*,j}=\mathsf{H}_i+\mathsf{Z}_i-\mathsf{Z}_j$ if $\mathsf{H}_i\ne0$, or $\mathsf{H}_i^{*,j}=0$ if $\mathsf{H}_i=0$. Similarly, $\mathsf{H}_i^{*,\sigma_i(j)}=\mathsf{H}_i$ and since i is away from the shoreline we have either $\mathsf{H}_{\sigma_i(j)}^{*,i}=\mathsf{H}_{\sigma_i(j)}+\mathsf{Z}_{\sigma_i(j)}-\mathsf{Z}_i$ if $\mathsf{H}_{\sigma_i(j)}\ne0$, or $\mathsf{H}_{\sigma_i(j)}^{*,i}=0$ if $\mathsf{H}_{\sigma_i(j)}=0$. Hence, if i is a wet state (and all the states in $\mathcal{I}(D_i)$ are wet since i is away from the shoreline), we have

$$
\begin{aligned}
&(\mathbf{V}_j\mathsf{H}_j^{*,i}+\mathbf{V}_i\mathsf{H}_i^{*,j})\cdot\boldsymbol{c}_{ij}+(\mathbf{V}_{\sigma_i(j)}\mathsf{H}_{\sigma_i(j)}^{*,i}+\mathbf{V}_i\mathsf{H}_i^{*,\sigma_i(j)})\cdot\boldsymbol{c}_{i\sigma_i(j)}\\
&\qquad=\big(\mathbf{V}_j\mathsf{H}_j+\mathbf{V}_i(\mathsf{H}_i+\mathsf{Z}_i-\mathsf{Z}_j)-(\boldsymbol{V}_{\sigma_i(j)}(\mathsf{H}_{\sigma_i(j)}+\mathsf{Z}_{\sigma_i(j)}-\mathsf{Z}_i)+\boldsymbol{V}_i\mathsf{H}_i)\big)\cdot\boldsymbol{c}_{ij}\\
&\qquad=(\mathbf{V}_j\mathsf{H}_j-\boldsymbol{V}_i\mathsf{H}_i)\cdot\boldsymbol{c}_{ij}+(\boldsymbol{V}_{\sigma_i(j)}\mathsf{H}_{\sigma_i(j)}-\mathbf{V}_i\mathsf{H}_i)\cdot\boldsymbol{c}_{i\sigma_i(j)}\\
&\qquad\quad+\mathbf{V}_i(\mathsf{Z}_i-\mathsf{Z}_j)\cdot\boldsymbol{c}_{ij}+\boldsymbol{V}_{\sigma_i(j)}(\mathsf{Z}_{\sigma_i(j)}-\mathsf{Z}_i)\cdot\boldsymbol{c}_{i\sigma_i(j)},
\end{aligned}
$$

where we have used the centro-symmetry property: $\boldsymbol{c}_{ij}=-\boldsymbol{c}_{i\sigma_i(j)}$. If i is a dry state (recall that j and $\sigma_i(j)$ are also dry states since i is away from the shoreline) then

$$
\begin{aligned}
&(\mathbf{V}_j\mathsf{H}_j^{*,i}+\mathbf{V}_i\mathsf{H}_i^{*,j})\cdot\boldsymbol{c}_{ij}+(\mathbf{V}_{\sigma_i(j)}\mathsf{H}_{\sigma_i(j)}^{*,i}+\mathbf{V}_i\mathsf{H}_i^{*,\sigma_i(j)})\cdot\boldsymbol{c}_{i\sigma_i(j)}\\
&\qquad=(\mathbf{V}_j\mathsf{H}_j-\boldsymbol{V}_i\mathsf{H}_i)\cdot\boldsymbol{c}_{ij}+(\boldsymbol{V}_{\sigma_i(j)}\mathsf{H}_{\sigma_i(j)}-\mathbf{V}_i\mathsf{H}_i)\cdot\boldsymbol{c}_{i\sigma_i(j)}.
\end{aligned}
$$

Since according to Lemma 2.1, $\sum_{j\in\mathcal{I}(D_i)}(\mathbf{V}_j\mathsf{H}_j-\boldsymbol{V}_i\mathsf{H}_i)\cdot\boldsymbol{c}_{ij}=\sum_{j\in\mathcal{I}(D_i)}\mathbf{V}_j\mathsf{H}_j\cdot\boldsymbol{c}_{ij}$ is a second-order approximation of $\int_D\nabla\cdot(\boldsymbol{v}_hh_h)\varphi_i\,\mathrm{d}x$, we have to show that the contribution of the extra term $\mathbf{V}_i(\mathsf{Z}_i-\mathsf{Z}_j)\cdot\boldsymbol{c}_{ij}-\boldsymbol{V}_{\sigma_i(j)}(\mathsf{Z}_{\sigma_i(j)}-\mathsf{Z}_i)\cdot\boldsymbol{c}_{ij}$ that arises when i is a wet state is small. Assuming that the velocity is smooth, we have $\boldsymbol{V}_{\sigma_i(j)}=\boldsymbol{V}_i+\mathcal{O}(h)$, which shows that $\mathbf{V}_i(\mathsf{Z}_i-\mathsf{Z}_j)\cdot\boldsymbol{c}_{ij}-\boldsymbol{V}_{\sigma_i(j)}(\mathsf{Z}_{\sigma_i(j)}-\mathsf{Z}_i)\cdot\boldsymbol{c}_{ij}=\mathbf{V}_i(2\mathsf{Z}_i-\mathsf{Z}_j-\mathsf{Z}_{\sigma_i(j)})\cdot\boldsymbol{c}_{ij}+\|\boldsymbol{c}_{ij}\|_{\ell^2}\mathcal{O}(h^2)$. The centro-symmetry assumption implies that $2\mathsf{Z}_i-\mathsf{Z}_j-\mathsf{Z}_{\sigma_i(j)}=\mathcal{O}(h^2)$ if the bathymetry map is smooth. In conclusion $\sum_{j\in\mathcal{I}(D_i)}(\mathbf{V}_j\mathsf{H}_j^{*,i}+\mathbf{V}_i\mathsf{H}_i^{*,j}))\cdot\boldsymbol{c}_{ij}=\sum_{j\in\mathcal{I}(D_i)}\mathbf{V}_j\mathsf{H}_j\cdot\boldsymbol{c}_{ij}+m_i\mathcal{O}(h)$ away from the shoreline. Using the same argument one proves that $\sum_{j\in\mathcal{I}(D_i)}((\mathbf{V}_j\otimes\mathbf{V}_j)\mathsf{H}_j^{*,i}+(\mathbf{V}_i\otimes\mathbf{V}_i)\mathsf{H}_i^{*,j}))\cdot\boldsymbol{c}_{ij}=\sum_{j\in\mathcal{I}(D_i)}(\mathbf{V}_j\otimes\mathbf{V}_j)\mathsf{H}_j+m_i\mathcal{O}(h)$. This concludes the proof. ∎

Remark 3.4 (hydrostatic reconstruction). The lack of consistency of the hydrostatic reconstruction at the shoreline or in presence of large gradients in the topography map has been identified in Delestre et al. [10, Prop. 2.1]. Various alternatives to the hydrostatic reconstruction have since been proposed like in Berthon and Foucher [5], Bryson et al. [9], Duran et al. [12] where the authors propose to work with the free surface elevation instead of the water height. □

3.2. Full time and space approximation. Let $\boldsymbol{u}_h^0=\sum_{i=1}^I\mathbf{U}_i^0\varphi_i\in\boldsymbol{P}(\mathcal{T}_h)$ be a reasonable approximation of $\boldsymbol{u}_0$. Let $n\in\mathbb{N}$, τ be the time step, t_n be the current time, and let us set $t_{n+1}=t_n+\tau$. Let $\boldsymbol{u}_h^n=\sum_{i=1}^I\mathbf{U}_i^n\varphi_i\in\boldsymbol{P}(\mathcal{T}_h)$ be the space approximation of $\boldsymbol{u}$ at time t_n. Upon denoting $\mathsf{H}_i^{*,j,n} := \max(0, \mathsf{H}_i^n + \mathsf{Z}_i - \max(\mathsf{Z}_i, \mathsf{Z}_j))$, we propose to estimate $\mathbf{U}_i^{n+1}$ as follows:

$$
\begin{aligned}
(3.3)\quad m_i \frac{\mathbf{U}_i^{n+1} - \mathbf{U}_i^n}{\tau} &+ \sum_{j\in\mathcal{I}(D_i)} (\boldsymbol{g}(\mathbf{U}_j^{*,i,n}) + \boldsymbol{g}(\mathbf{U}_i^{*,j,n}))\cdot\boldsymbol{c}_{ij} \\
&+ \begin{pmatrix} 0 \\ \frac12 g\big((\mathsf{H}_j^{*,i,n})^2 - (\mathsf{H}_i^{*,j,n})^2\big)\boldsymbol{c}_{ij} \end{pmatrix} - \sum_{i\neq j\in\mathcal{I}(D_i)} d_{ij}^n (\mathbf{U}_j^{*,i,n} - \mathbf{U}_i^{*,j,n}) = 0,
\end{aligned}
$$

where the artificial viscosity coefficient d_{ij}^n is defined by

$$
(3.4)\qquad d_{ij}^n := \max(d_{ij}^{\boldsymbol{f},n}, d_{ji}^{\boldsymbol{f},n}),
$$

$$
(3.5)\qquad d_{ij}^{\boldsymbol{f},n} := \max\Big(\lambda_{\max}^{\boldsymbol{f}}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^{*,i,n}), \lambda_{\max}^{\boldsymbol{f}}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_i^{*,j,n})\Big)\, \|\boldsymbol{c}_{ij}\|_{\ell^2},
$$

and $\lambda_{\max}^{\boldsymbol{f}}(\boldsymbol{n}, \mathbf{U}_L, \mathbf{U}_R)$ is the maximum wave speed in the Riemann problem:

$$
(3.6)\qquad \partial_t \boldsymbol{u} + \partial_x(\boldsymbol{f}(\boldsymbol{u})\cdot\boldsymbol{n}) = 0, \qquad \boldsymbol{u}(x,0) = (1-H(x))\mathbf{U}_L + H(x)\mathbf{U}_R,
$$

where $H(x)$ is the Heaviside function. Note that $d_{ij}^n \geq 0$ and $d_{ij}^n = d_{ji}^n$ for all $j \neq i$ in $\mathcal{I}(D_i)$. For convenience we denote $d_{ii}^n := -\sum_{i\neq j\in\mathcal{I}(D_i)} d_{ij}^n$. Therefore we have $\sum_{j\in\mathcal{I}(D_i)} d_{ij}^n = \sum_{j\in\mathcal{I}(D_i)} d_{ji}^n = 0$; this property will be used in the rest of the paper.

3.3. Reduction to the 1D Riemann problem. For completeness, we show how the estimation of $\lambda_{\max}^{\boldsymbol{f}}(\boldsymbol{n}, \mathbf{U}_L, \mathbf{U}_R)$ can be reduced to estimating the maximum wave speed in a one-dimensional Riemann problem independent of $\boldsymbol{n}$. Similarly to [16], we make a change of basis and introduce $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_{d-1} \in \mathbb{R}^d$ so that $\{\boldsymbol{n}, \boldsymbol{t}_1, \ldots, \boldsymbol{t}_{d-1}\}$ is an orthonormal basis of $\mathbb{R}^d$. With respect to this basis we have that $\boldsymbol{q} = (q, \boldsymbol{q}^\perp)$ where $q := \boldsymbol{q}\cdot\boldsymbol{n}$, and $\boldsymbol{q}^\perp := (\boldsymbol{q}\cdot\boldsymbol{t}_1, \ldots, \boldsymbol{q}\cdot\boldsymbol{t}_{d-1})^\mathsf{T}$. Then, with the notation $v = q/h$, the Riemann problem (3.6) can be rewritten in the new orthonormal basis as follows:

$$
(3.7)\qquad \partial_t \boldsymbol{u} + \partial_x(\boldsymbol{n}\cdot\boldsymbol{f}(\boldsymbol{u})) = \boldsymbol{0}, \quad \boldsymbol{u} = \begin{pmatrix} h \\ q \\ \boldsymbol{q}^\perp \end{pmatrix}, \qquad \boldsymbol{f}(\boldsymbol{u})\cdot\boldsymbol{n} = \begin{pmatrix} q \\ vq + \frac{g}{2}h^2 \\ v\boldsymbol{q}^\perp \end{pmatrix},
$$

with data $\mathbf{U}_L = (h_L, q_L, \boldsymbol{q}_L^\perp)^\mathsf{T}$, $\mathbf{U}_R = (h_R, q_R, \boldsymbol{q}_R^\perp)^\mathsf{T}$. The solution to (3.7) is henceforth denoted $\boldsymbol{u}(\boldsymbol{n}, \mathbf{U}_L, \mathbf{U}_R)(x,t)$. Following [16], we introduce the following definition.

Definition 3.5 (Invariant set). *A convex set $A \subset \mathcal{A}$ is said to be invariant for the flat bottom system, i.e., (2.1) with $\boldsymbol{b} = 0$, if for any admissible pair $(\mathbf{U}_L, \mathbf{U}_R) \in A \times A$ and any unit vector $\boldsymbol{n} \in \mathbb{R}^d$, we have $\boldsymbol{u}(\boldsymbol{n}, \mathbf{U}_L, \mathbf{U}_R)(x,t) \in A$ for a.e. $x \in \mathbb{R}$, $t > 0$.*

Let us $\overline{\boldsymbol{u}}(t, \boldsymbol{n}, \mathbf{U}_L, \mathbf{U}_R) := \int_{-\frac12}^{\frac12} \boldsymbol{u}(\boldsymbol{n}, \mathbf{U}_L, \mathbf{U}_R)(x,t)\,dx$. Then, the following result is a consequence of $\lambda_{\max}^{\boldsymbol{f}}(\boldsymbol{n}, \mathbf{U}_L, \mathbf{U}_R)$ being finite, see [16, Lem. 2.1].

Lemma 3.6 (Invariant set and average). (i) *Let $A \subset \mathcal{A}$ be an invariant set for the flat bottom system. If $(\mathbf{U}_L, \mathbf{U}_R) \in A$, then $\overline{\boldsymbol{u}}(t, \boldsymbol{n}, \mathbf{U}_L, \mathbf{U}_R) \in A$.* (ii) *Assume that $2t\,\lambda_{\max}(\boldsymbol{n}, \mathbf{U}_L, \mathbf{U}_R) \leq 1$, then $\overline{\boldsymbol{u}}(t, \boldsymbol{n}, \mathbf{U}_L, \mathbf{U}_R) = \frac12(\mathbf{U}_L + \mathbf{U}_R) - t(\boldsymbol{f}(\mathbf{U}_R) - \boldsymbol{f}(\mathbf{U}_L))\cdot\boldsymbol{n}$.*

This lemma is the key motivation for the definition of the viscosity coefficients $d_{ij}^{\boldsymbol{f},n}$ in (3.5) (see [16, §3.3] for more details).

The maximum wave speed in the Riemann problem (3.7) is determined by the one-dimensional shallow water system for the component $(h,q)^\mathsf{T}$ because the last component is just passively transported and does not influence the first two equations of the system. That is to say (3.7) reduces to solving the Riemann problem

$$\partial_t(h,q)^\mathsf{T} + \partial_x(\boldsymbol{f}_{1\text{D}}(h,q)) = 0, \tag{3.8}$$

with data $\boldsymbol{u}_L := (h_L, q_L)$, $\boldsymbol{u}_R := (h_R, q_R)$ and flux $\boldsymbol{f}_{1\text{D}}(h,q) := (q, vq + \frac{g}{2}h^2)^\mathsf{T}$. This establishes the following result which will be useful to estimate $d_{ij}^{\boldsymbol{f},n}$ in (3.5). When using a SSP RK method, This is done at the end of every substep of the SSP RK method

Proposition 3.7 (Maximum wave speed). *Let $\lambda_{\max}^{\boldsymbol{f}}(\boldsymbol{n}, \mathbf{U}_L, \mathbf{U}_R)$, $\lambda_{\max}^{\boldsymbol{f}_{1\text{D}}}(\boldsymbol{u}_L, \boldsymbol{u}_R)$ be the maximum wave speed in the Riemann problems (3.7) and (3.8), respectively. Then $\lambda_{\max}^{\boldsymbol{f}}(\boldsymbol{n}, \mathbf{U}_L, \mathbf{U}_R) = \lambda_{\max}^{\boldsymbol{f}_{1\text{D}}}(\boldsymbol{u}_L, \boldsymbol{u}_R)$.*

In order to estimate $\lambda_{\max}^{\boldsymbol{f}_{1\text{D}}}(\boldsymbol{u}_L, \boldsymbol{u}_R)$ from above, we introduce

$$\lambda_1^-(h_*) := v_L - \sqrt{gh_L}\left(1 + \left(\frac{h_* - h_L}{2h_L}\right)_+\right)^{\frac12}\left(1 + \left(\frac{h_* - h_L}{h_L}\right)_+\right)^{\frac12}, \tag{3.9}$$

$$\lambda_2^+(h_*) := v_R + \sqrt{gh_R}\left(1 + \left(\frac{h_* - h_R}{2h_R}\right)_+\right)^{\frac12}\left(1 + \left(\frac{h_* - h_R}{h_R}\right)_+\right)^{\frac12}. \tag{3.10}$$

The following result is proved in Guermond and Popov [18]:

Lemma 3.8. *Let $h_{\min} = \min(h_L, h_R)$, $h_{\max} = \max(h_L, h_R)$, $x_0 = (2\sqrt{2}-1)^2$, and*

$$\overline{h}_* := \begin{cases} \frac{(v_L - v_R + 2\sqrt{gh_L} + 2\sqrt{gh_R})_+^2}{16g}, & \textit{if case 1,} \\ \left(-\sqrt{2h_{\min}} + \sqrt{3h_{\min} + 2\sqrt{2h_{\min}h_{\max}} + \sqrt{\tfrac{2}{g}}(v_L - v_R)\sqrt{h_{\min}}}\right)^2 & \textit{if case 2,} \\ \sqrt{h_{\min}h_{\max}}\left(1 + \frac{\sqrt{2}(v_L - v_R)}{\sqrt{gh_{\min}} + \sqrt{gh_{\max}}}.\right) & \textit{if case 3,} \end{cases}$$

where case 1 is $0 \le f(x_0 h_{\min})$, case 2 is $f(x_0 h_{\min}) < 0 \le f(x_0 h_{\max})$ and case 3 is $f(x_0 h_{\max}) < 0$. Then $\lambda_{\max}^{\boldsymbol{f}}(\boldsymbol{n}, \mathbf{U}_L, \mathbf{U}_R) = \lambda_{\max}^{\boldsymbol{f}_{1\text{D}}}(\boldsymbol{u}_L, \boldsymbol{u}_R) \le \max(|\lambda_1^-(\overline{h}_*)|, |\lambda_2^+(\overline{h}_*)|)$.

3.4. Stability properties. We collect in his section some remarkable stability properties of the scheme defined by (3.3)–(3.5).

Proposition 3.9 (Well-balancing/conservation). *The scheme defined in (3.3) is well-balanced at large, and it is conservative in the sense of Definition 2.7.*

Proof. Let $\boldsymbol{u}_h^n$ be a rest sate at large, then $\mathsf{H}_j^{*,i,n} = \mathsf{H}_i^{*,j,n}$ for all $i \in \{1{:}I\}$ and all $j \in \mathcal{I}(D_i)$; this identity implies well-balancing at large. Let us now establish conservation. Since $\boldsymbol{c}_{ij} = -\boldsymbol{c}_{ji}$ and $d_{ij}^n = d_{ji}^n$ we have

$$\sum_{i\in\{1:I\}} \sum_{j\in\mathcal{I}(D_i)} \boldsymbol{c}_{ji}\alpha_{ij} = 0, \qquad \sum_{i\in\{1:I\}} \sum_{j\in\mathcal{I}(D_i)} d_{ji}^n \beta_{ij} = 0,$$

for any symmetric field $\alpha_{ij} = \alpha_{ji}$ and any skew-symmetric field $\beta_{ij} = -\beta_{ij}$. Hence, we only have to deal with the nonconservative flux in (3.3) $\frac12 g\big((\mathsf{H}_j^{*,i,n})^2 - (\mathsf{H}_i^{*,j,n})^2\big)\boldsymbol{c}_{ij}$. This quantity is zero when the topography map is constant. This concludes the proof. ∎

Since the shallow water system makes sense only for nonnegative water heights, and the water discharge should be zero in dry states, we are lead to consider the following definition for the admissibility of shallow water states.

Definition 3.10 (Admissible water states). *A shallow water state* $\mathbf{U} = (\mathsf{H}, \mathbf{Q})^\mathsf{T}$ *is admissible if* $\mathsf{H} \ge 0$ *and* $\mathbf{Q} = \mathbf{0}$ *if* $\mathsf{H} = 0$. *The set of admissible states is denoted* $\mathcal{A}$.

Note that a convex combination of admissible states is always an admissible state.

Proposition 3.11 (Invariant domain). *Let* $\boldsymbol{u}_h^{n+1}$ *be given by* (3.3)–(3.5), $n \ge 0$. *Let* $\in \{1{:}I\}$. *Assume that* $1 + 4\frac{\tau}{m_i} d_{ii}^n \ge 0$. *Let* A_i^n *be an invariant set of the shallow water equation that contains* $\{\mathbf{U}_j^n\}_{j\in\mathcal{I}(D_i)}$. *Then the following properties hold true:*
(i) *If the bathymetry map is constant then* $\mathbf{U}_i^{n+1} \in A_i^n$*;*
(ii) *If the bathymetry is not constant, let* $\Delta \boldsymbol{Z}_i^n := \frac{\tau}{m_i}\sum_{i\ne j\in\mathcal{I}(D_i)} g((\mathsf{H}_i^n)^2 - (\mathsf{H}_i^{*,j,n})^2)\boldsymbol{c}_{ij}$ *and* $\Delta\mathbf{U}_i^{*,n} := \frac{2\tau}{m_i}\sum_{i\ne j\in\mathcal{I}(D_i)} d_{ij}^n \Big(1 - \frac{\mathsf{H}_i^{*,j,n}}{\mathsf{H}_i^n}\Big)\mathbf{U}_i^n$*, then* $\mathbf{U}_i^{n+1} \in \operatorname{conv}(A_i^n, \mathbf{0}) + (0, \Delta\boldsymbol{Z}_i^n)^\mathsf{T} + \Delta\mathbf{U}_i^{*,n}$*; in particular the scheme preserves the non-negativity of the water height;*
(iii) *If the states* $\{\mathbf{U}_i^n\}$ *are admissible then the state* $\{\mathbf{U}_i^{n+1}\}$ *are also admissible.*

Proof. Recalling that $\boldsymbol{f}(\boldsymbol{u}) = \boldsymbol{g}(\boldsymbol{u}) + (0, \frac{1}{2}gh^2\mathbb{I}_d)^\mathsf{T}$, then (3.3) can also be rewritten

$$
\begin{aligned}
\frac{m_i}{\tau}(\mathbf{U}_i^{n+1} - \mathbf{U}_i^n) &+ \sum_{j\in\mathcal{I}(D_i)} \boldsymbol{f}(\mathbf{U}_j^{*,i,n})\cdot\boldsymbol{c}_{ij} - d_{ij}^n\mathbf{U}_j^{*,i,n} + \boldsymbol{f}(\mathbf{U}_i^{*,j,n})\cdot\boldsymbol{c}_{ij} - d_{ij}^n\mathbf{U}_i^{*,j,n} \\
&\qquad + \sum_{j\in\mathcal{I}(D_i)} \big(0, -g(\mathsf{H}_i^{*,j,n})^2\boldsymbol{c}_{ij}\big)^\mathsf{T} + (d_{ij}^n + d_{ij}^n)\mathbf{U}_i^{*,j,n} = \mathbf{0}.
\end{aligned}
$$

Using conservation, i.e., $\boldsymbol{c}_{ii} = -\sum_{i\ne j\in\mathcal{I}(D_i)}\boldsymbol{c}_{ij}$, this equation can be recast into

$$
\begin{aligned}
\frac{m_i}{\tau}(\mathbf{U}_i^{n+1} - \mathbf{U}_i^n) = &\sum_{i\ne j\in\mathcal{I}(D_i)} -(\boldsymbol{f}(\mathbf{U}_j^{*,i,n}) - \boldsymbol{f}(\mathbf{U}_i^n))\cdot\boldsymbol{c}_{ij} + d_{ij}^n(\mathbf{U}_j^{*,i,n} + \mathbf{U}_i^n) \\
&+ \sum_{i\ne j\in\mathcal{I}(D_i)} -(\boldsymbol{f}(\mathbf{U}_i^{*,j,n}) - \boldsymbol{f}(\mathbf{U}_i^n))\cdot\boldsymbol{c}_{ij} + d_{ij}^n(\mathbf{U}_i^{*,j,n} + \mathbf{U}_i^n) \\
&+ \sum_{i\ne j\in\mathcal{I}(D_i)} \big(0, g((\mathsf{H}_i^n)^2 - (\mathsf{H}_i^{*,j,n})^2)\boldsymbol{c}_{ij}\big)^\mathsf{T} - (d_{ij}^n + d_{ij}^n)(\mathbf{U}_i^{*,j,n} + \mathbf{U}_i^n).
\end{aligned}
$$

Upon introducing the vectors $\overline{\mathbf{U}_{ij}^n} \in \mathbb{R}^{1+d}$, $\overline{\mathbf{W}_{ij}^n} \in \mathbb{R}^{1+d}$ and $\Delta\boldsymbol{Z}_i^n \in \mathbb{R}^d$ defined by

$$
\begin{aligned}
\overline{\mathbf{U}_{ij}^n} &:= -\frac{\|\boldsymbol{c}_{ij}\|_{\ell^2}}{2d_{ij}^n}(\boldsymbol{f}(\mathbf{U}_j^{*,i,n}) - \boldsymbol{f}(\mathbf{U}_i^n))\cdot\boldsymbol{n}_{ij} + \frac{1}{2}(\mathbf{U}_j^{*,i,n} + \mathbf{U}_i^n) \\
\overline{\mathbf{W}_{ij}^n} &:= -\frac{\|\boldsymbol{c}_{ij}\|_{\ell^2}}{2d_{ij}^n}(\boldsymbol{f}(\mathbf{U}_i^{*,j,n}) - \boldsymbol{f}(\mathbf{U}_i^n))\cdot\boldsymbol{n}_{ij} + \frac{1}{2}(\mathbf{U}_i^{*,j,n} + \mathbf{U}_i^n) \\
\Delta\boldsymbol{Z}_i^n &:= \sum_{i\ne j\in\mathcal{I}(D_i)} g((\mathsf{H}_i^n)^2 - (\mathsf{H}_i^{*,j,n})^2)\boldsymbol{c}_{ij},
\end{aligned}
$$

we finally obtain

$$
\begin{aligned}
\mathbf{U}_i^{n+1} = &\Big(1 - \sum_{i\ne j\in\mathcal{I}(D_i)}\frac{4\tau}{m_i}d_{ij}^n\Big)\mathbf{U}_i^n + \sum_{i\ne j\in\mathcal{I}(D_i)}\frac{2\tau}{m_i}d_{ij}^n(\overline{\mathbf{U}_{ij}^n} + \overline{\mathbf{W}_{ij}^n}) \\
&+ \frac{\tau}{m_i}\big(0, \Delta\boldsymbol{Z}_i^n\big)^\mathsf{T} + \frac{2\tau}{m_i}\sum_{i\ne j\in\mathcal{I}(D_i)} d_{ij}^n\Big(1 - \frac{\mathsf{H}_i^{*,j,n}}{\mathsf{H}_i^n}\Big)\mathbf{U}_i^n.
\end{aligned}
$$

Upon introducing the fake time $t = \frac{\|\boldsymbol{c}_{ij}\|_{\ell^2}}{2d_{ij}^n}$ and observing that the definition of d_{ij}^n implies that $2t\lambda_{\max}^{\boldsymbol{f}}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^{*,i,n}) \le 1$ and $2t\lambda_{\max}^{\boldsymbol{f}}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_i^{*,j,n}) \le 1$, we infer from

Lemma 3.6 that $\overline{\mathsf{U}^n_{ij}} \in \operatorname{conv}_{j\in\mathcal{I}(D_i)}(\mathsf{U}^{*,i,n}_j)$ and $\overline{\mathsf{W}^n_{ij}} \in \operatorname{conv}_{j\in\mathcal{I}(D_i)}(\mathsf{U}^{*,j,n}_i)$; hence, $\frac{\overline{\mathsf{U}^n_{ij}}+\overline{\mathsf{W}^n_{ij}}}{2} \in \operatorname{conv}_{j\in\mathcal{I}(D_i)}(\mathsf{U}^{*,i,n}_j, \mathsf{U}^{*,j,n}_i)$. In conclusion, under the CFL condition $1 + 4\frac{\tau}{m_i}d^n_{ii} \geq 0$, the state $\widetilde{\mathsf{U}}^{n+1}_i := \big(1 + \frac{4\tau}{m_i}d^n_{ii}\big)\mathsf{U}^n_i + \sum_{i\neq j\in\mathcal{I}(D_i)}\frac{2\tau}{m_i}d^n_{ij}(\overline{\mathsf{U}^n_{ij}} + \overline{\mathsf{W}^n_{ij}})$ belongs to $\operatorname{conv}_{j\in\mathcal{I}(D_i)}(\mathsf{U}^{*,i,n}_j, \mathsf{U}^{*,j,n}_i)$. If the bathymetry map is flat then $\mathsf{H}^n_i = \mathsf{H}^{*,j,n}_i$ and we obtain $\mathsf{U}^{n+1}_i = \widetilde{\mathsf{U}}^{n+1}_i \in \operatorname{conv}_{j\in\mathcal{I}(D_i)}(\mathsf{U}^n_j) \subset A^n_i$ and this proves (i). If the bathymetry is not flat, then $\mathsf{U}^{*,i,n}_j$ is in the convex hull of U^n_j and $\mathbf{0}$ for all $j \in \mathcal{I}(D_i)$ and $\mathsf{U}^{*,j,n}_i$ is in the convex hull of U^n_i and $\mathbf{0}$ for all $j \in \mathcal{I}(D_i)$; this proves that $\widetilde{\mathsf{U}}^{n+1}_i \in \operatorname{conv}(A^n_i, \mathbf{0})$. Hence, if the bathymetry is not flat we get $\mathsf{U}^{n+1}_i \in \operatorname{conv}(A^n_i, \mathbf{0}) + (0, \Delta \boldsymbol{Z}^n_i)^\mathsf{T} + \Delta\mathsf{U}^{*,n}_i$ as announced. The water height in $\Delta\mathsf{U}^{*,n}_i$ is $\frac{2\tau}{m_i}\sum_{i\neq j\in\mathcal{I}(D_i)} d^n_{ij}\Big(\mathsf{H}^n_i - \mathsf{H}^{*,j,n}_i\Big) \geq 0$. Since all the states in A^n_i have non-negative water height, we conclude that $\mathsf{H}^{n+1}_i \geq 0$ and this proves (ii). Finally, fix $n \geq 0$ and assume that all states $\{\mathsf{U}^n_j\}$ are admissible in the sense of Definition 3.10. If $\mathsf{H}^n_i > 0$ then we have that

$$\mathsf{H}^{n+1}_i \geq \Big(1 - \sum_{i\neq j\in\mathcal{I}(D_i)} \frac{4\tau}{m_i} d^n_{ij}\Big)\mathsf{H}^n_i > 0,$$

and this proves that U^{n+1}_i is admissible. In the remaining case $\mathsf{H}^n_i = 0$, we have that $\mathsf{H}^{*,j,n}_i = 0$ for all $j \in \mathcal{I}(D_i)$ and $\Delta\boldsymbol{Z}^n_i = 0$. Hence $\mathsf{U}^{n+1}_j = \widetilde{\mathsf{U}}^{n+1}_i$ and using that $\widetilde{\mathsf{U}}^{n+1}_i$ is a convex combination of admissible states we conclude that the state U^{n+1}_i is admissible and this proves (iii). ∎

We finish with a discrete inequality which reduces to a standard discrete entropy inequality when the bottom topography is flat. The proof is omitted for brevity.

Proposition 3.12. *Let $\boldsymbol{u}^{n+1}_h$ be given by (3.3)–(3.5). Assume the CFL condition $1 + 4\frac{\tau}{m_i}d^n_{ii} \geq 0$. Then for any flat bed shallow water entropy pair $(\eta, \boldsymbol{G})$, we have the following discrete entropy inequality*

$$\begin{aligned}(3.11)\quad \frac{m_i}{\tau}(\eta(\mathsf{U}^{n+1}_i) - \eta(\mathsf{U}^n_i)) &+ \sum_{i\neq j\in\mathcal{I}(D_i)} (\boldsymbol{G}(\mathsf{U}^{*,i,n}_j) + \boldsymbol{G}(\mathsf{U}^{*,j,n}_i))\cdot\boldsymbol{c}_{ij} \\ &\leq \sum_{i\neq j\in\mathcal{I}(D_i)} d^n_{ij}\Big(\eta(\mathsf{U}^{*,i,n}_j) + \eta(\mathsf{U}^{*,j,n}_i) - 2\eta(\mathsf{U}^n_i)\Big) \\ &\quad + \Big((0, \Delta\boldsymbol{Z}^n_i)^\mathsf{T} + \sum_{i\neq j\in\mathcal{I}(D_i)} 2d^n_{ij}\Big(1 - \frac{\mathsf{H}^{*,j,n}_i}{\mathsf{H}^n_i}\Big)\mathsf{U}^n_i\Big)\cdot\nabla\eta(\mathsf{U}^{n+1}_i).\end{aligned}$$

Remark 3.13 (Literature). We refer the reader to Bouchut and Frid [8, §2] for an alternative point of view to derive the invariant domain property and entropy inequality obtained above. □

4. Second-order extension.

In this section we propose a scheme that is second-order accurate in space, is exactly well-balanced, and is positivity preserving.

4.1. Flux approximation.

We start by constructing a well-balanced second-order approximation of the quantity $\int_{D_i}(\nabla(\frac{1}{2}h^2) + h\nabla z)\varphi_i \,\mathrm{d}x$.

Lemma 4.1 (Consistency/Well-balancing). (i) *Assume that $\{\widehat{\theta}_n\}_{n\in\{1:n_{sh}\}}$ consists of Lagrange or Bernstein basis functions. The expression* $\sum_{j\in\mathcal{I}(D_i)} \mathsf{H}_i(\mathsf{H}_j + \mathsf{Z}_j)\boldsymbol{c}_{ij}$

is a second-order approximation of $\int_D(\nabla(\frac{1}{2}h^2)+h\nabla z)\varphi_i\,\mathrm{d}x$. (ii) *The mapping* $\boldsymbol{u}_h \to (0,\sum_{j\in\mathcal{I}(D_i)}\mathsf{H}_i(\mathsf{H}_j+\mathsf{Z}_j)\boldsymbol{c}_{ij})_{i\in\{1:I\}}$ *is an exactly well-balanced flux.*

Proof. (i) If $h+z$ is linear over $K\in\mathcal{T}_h$, then $\int_K h\nabla(h+z)\varphi_i\,\mathrm{d}x = \nabla(h+z)|_K\int_K h\varphi_i\,\mathrm{d}x$ and the approximation $\int_K h\varphi_i\,\mathrm{d}x \approx \mathsf{H}_i\frac{1}{d}|K|$ is second-order accurate, at least for Lagrange and Bernstein basis functions. Hence, upon noticing that $\sum_{K\subset D_i}\nabla(h+z)_{|K}\frac{1}{d}|K| = \int_{D_i}\nabla(h+z)\varphi_i\,\mathrm{d}x = \sum_{j\in\mathcal{I}(D_i)}(\mathsf{H}_j+\mathsf{Z}_j)\boldsymbol{c}_{ij}$, the expression $\int_D h\nabla(h+z)\varphi_i\,\mathrm{d}x \approx \sum_{j\in\mathcal{I}(D_i)}\mathsf{H}_i(\mathsf{H}_j+\mathsf{Z}_j)\boldsymbol{c}_{ij}$ is formally second-order accurate.

(ii) Let us now prove well-balancing. Let us assume exact rest. Let us fix $i\in\{1:I\}$. Notice that owing to the partition of unity property we have $\sum_{j\in\mathcal{I}(D_i)}\boldsymbol{c}_{ij}=0$; hence $\sum_{j\in\mathcal{I}(D_i)}\mathsf{H}_i(\mathsf{H}_j+\mathsf{Z}_j)\boldsymbol{c}_{ij} = \sum_{j\in\mathcal{I}(D_i)}\mathsf{H}_i(\mathsf{H}_j+\mathsf{Z}_j-\mathsf{H}_i-\mathsf{Z}_i)\boldsymbol{c}_{ij}$. Consider $j\in\mathcal{I}(D_i)$. According to our definition of the exact rest state (see Definition 2.4), either $\mathsf{H}_i=0$ and $\mathsf{H}_j=0$, or $\mathsf{H}_j+\mathsf{Z}_j-\mathsf{H}_i-\mathsf{Z}_i=0$; whence the conclusion. ☐

Let us introduce the gas dynamics flux $\boldsymbol{g}(\boldsymbol{u}) := (\boldsymbol{q}, \frac{1}{h}\boldsymbol{q}\otimes\boldsymbol{q})^\mathsf{T}$, then upon invoking Lemma 2.1, $\sum_{j\in\mathcal{I}(D_i)}\boldsymbol{g}(\mathbf{U}_j)\cdot\boldsymbol{c}_{ij}$ is a second-order approximation of $\int_{D_i}\nabla\cdot(\boldsymbol{g}(\boldsymbol{u}))\varphi_i\,\mathrm{d}x$.

4.2. Full time and space approximation. Let $\boldsymbol{u}_h^0 = \sum_{i=1}^I \mathbf{U}_i^0\varphi_i \in \boldsymbol{P}(\mathcal{T}_h)$ be a reasonable approximation of $\boldsymbol{u}_0$. Let $n\in\mathbb{N}$, τ be the time step, t_n be the current time, and $t_{n+1}:=t_n+\tau$. Let $\boldsymbol{u}_h^n = \sum_{i=1}^I\mathbf{U}_i^n\varphi_i\in\boldsymbol{P}(\mathcal{T}_h)$ be the space approximation of $\boldsymbol{u}$ at time t_n and let $\boldsymbol{u}_h^{n+1} := \sum_{i=1}^I\mathbf{U}_i^{n+1}\varphi_i$. We estimate $\mathbf{U}_i^{n+1}$ as follows:

$$
\begin{aligned}
\frac{m_i}{\tau}(\mathbf{U}_i^{n+1}-\mathbf{U}_i^n) = &\sum_{j\in\mathcal{I}(D_i)} -\boldsymbol{g}(\mathbf{U}_j^n)\cdot\boldsymbol{c}_{ij} - \big(0, g\mathsf{H}_i^n(\mathsf{H}_j^n+Z_j)\boldsymbol{c}_{ij})\big)^\mathsf{T} \\
&+\sum_{i\neq j\in\mathcal{I}(D_i)} d_{ij}^n(\mathbf{U}_j^{*,i,n}-\mathbf{U}_i^{*,j,n}) + \mu_{ij}^n\big(\mathbf{U}_j^n-\mathbf{U}_j^{*,i,n}-(\mathbf{U}_i^n-\mathbf{U}_i^{*,j,n})\big)
\end{aligned}
\tag{4.1}
$$

$$
\mu_{ij}^n := \max((\mathbf{V}_i\cdot\boldsymbol{n}_{ij})_-, (\mathbf{V}_j\cdot\boldsymbol{n}_{ij})_+)\|\boldsymbol{c}_{ij}\|_{\ell^2}, \qquad d_{ij}^n\ge\mu_{ij}^n, \qquad i\neq j. \tag{4.2}
$$

Here we use the notation $a_+ := \max(a,0)$ and $a_- = -\min(a,0)$. In the above scheme $d_{ij}^n = d_{ji}^n$ can be any non-negative number larger than μ_{ij}^n when $i\neq j$. One could just take $d_{ij}^n=\mu_{ij}^n$, but a more robust choice consists of using $d_{ij}^n = \max(d_{ij}^{\boldsymbol{f},n}, d_{ji}^{\boldsymbol{f},n})$; note that in this case the local maximum wave speed formulae (3.9) and (3.10) used with $\boldsymbol{u}_L := (\mathsf{H}_i^n, \mathbf{Q}_i^n\cdot\boldsymbol{n}_{ij})$ and $\boldsymbol{u}_R = (\mathsf{H}_j^n, \mathbf{Q}_i^n\cdot\boldsymbol{n}_{ij})$ imply that $d_{ij}^n\ge\mu_{ij}^n$. Notice that $\mu_{ij}^n=\mu_{ji}^n$ because $\boldsymbol{n}_{ij}=-\boldsymbol{n}_{ji}$ owing to the assumed boundary condition. We adopt again the convention $d_{ii}^n := -\sum_{i\neq j\in\mathcal{I}(D_i)}d_{ij}^n$.

PROPOSITION 4.2. *The scheme* (4.1)-(4.2) *is exactly well-balanced and conservative. It is positivity preserving provided* $1+2d_{ii}^n\frac{\tau}{m_i}\ge 0$ *for all* $i\in\{1:I\}$.

Proof. The artificial viscosity term on the right-hand side of (4.1) at exact rest is $\sum_{i\neq j\in\mathcal{I}(D_i)} -\mu_{ij}^n\big(-\mathsf{H}_j^n+\mathsf{H}_i^n, 0\big)^\mathsf{T}=0$, since $\mu_{ij}^n=0$ at rest state (at large). The remainder of the proof is a consequence Lemma 4.1, which establishes exact well-balancing. Since $\sum_{j\in\mathcal{I}(D_i)} -\boldsymbol{g}(\mathbf{U}_j^n)\cdot\boldsymbol{c}_{ij} = \sum_{j\in\mathcal{I}(D_i)}(\boldsymbol{g}(\mathbf{U}_i^n)-\boldsymbol{g}(\mathbf{U}_j^n))\cdot\boldsymbol{c}_{ij}$, the conservation can be shown like in the proof of Proposition 3.9. Finally, to prove positivity, let us fix i and assume that $\mathsf{H}_j^n\ge0$, for all $j\in\mathcal{I}(D_i)$. The water height update is

$$
\begin{aligned}
\mathsf{H}_i^{n+1} = \mathsf{H}_i^n &- \frac{\tau}{m_i}\sum_{i\neq j}\big(\mu_{ij}^n\mathsf{H}_i^n + (d_{ij}^n-\mu_{ij}^n)\mathsf{H}_i^{*,j,n}\big) \\
&+ \frac{\tau}{m_i}\sum_{i\neq j}\big((\mu_{ij}^n-\boldsymbol{c}_{ij}\cdot\boldsymbol{V}_j^n)\mathsf{H}_j^n + (d_{ij}^n-\mu_{ij}^n)\mathsf{H}_j^{*,i,n}\big).
\end{aligned}
$$

Using that $d_{ij}^n - \mu_{ij}^n \geq 0$, $\mu_{ij}^n \geq 0$, $\mathsf{H}_i^n \geq \mathsf{H}_i^{*,j,n} \geq 0$ and $\mathsf{H}_j^{*,i,n} \geq 0$ we obtain

$$\mathsf{H}_i^{n+1} \geq \mathsf{H}_i^n\Big(1 - \frac{\tau}{m_i}\sum_{i\neq j} d_{ij}^n\Big) + \frac{\tau}{m_i}\sum_{i\neq j}(\mu_{ij}^n - \boldsymbol{c}_{ij}\cdot\boldsymbol{V}_j^n)\mathsf{H}_j^n.$$

The conclusion follows from the assumption on the CFL number and the definition of μ_{ij}^n which implies that $\mu_{ij}^n - \boldsymbol{c}_{ij}\cdot\boldsymbol{V}_j^n \geq ((\boldsymbol{V}_j^n\cdot\boldsymbol{n}_{ij})_+ - \boldsymbol{V}_j^n\cdot\boldsymbol{n}_{ij})\|\boldsymbol{c}_{ij}\|_{\ell^2} \geq 0$. ∎

Remark 4.3. Note that the approximation of the flux in the scheme (4.1) is formally second-order accurate in space and contrary to (3.3) does not suffer from the small inconsistency of the hydrostatic reconstruction, since the hydrostatic reconstruction is used only in the artificial viscosity. In particular (4.1) is formally second-order accurate in space when the artificial viscosity is set to zero. □

4.3. Second-order positivity preserving viscosity. In order to make the proposed method fully second-order accurate in space, we now propose a new definition of the viscosity along the line of Guermond and Popov [17]. Namely, we choose the viscous terms d_{ij}^n and μ_{ij}^n in the scheme (4.1) to be $d_{ij}^n := \alpha_{ij}^n d_{ij}^{v,n}$ and $\mu_{ij}^n := \alpha_{ij}^n \mu_{ij}^{v,n}$ where $d_{ij}^{v,n} := \max(d_{ij}^{\boldsymbol{f},n}, d_{ji}^{\boldsymbol{f},n})$ is the first-order viscosity based on the maximum wave speed, $\mu_{ij}^{v,n} := \max((\mathbf{V}_i\cdot\boldsymbol{n}_{ij})_-, (\mathbf{V}_j\cdot\boldsymbol{n}_{ij})_+)\|\boldsymbol{c}_{ij}\|_{\ell^2}$ and $\alpha_{ij}^n \in [0,1]$ is appropriately chosen. More precisely, the proposed second-order scheme is

$$\begin{aligned}\frac{m_i}{\tau}(\mathbf{U}_i^{n+1} - \mathbf{U}_i^n) = &\sum_{j\in\mathcal{I}(D_i)} -\boldsymbol{g}(\mathbf{U}_j^n)\cdot\boldsymbol{c}_{ij} - \big(0, g\mathsf{H}_i^n(\mathsf{H}_j^n + Z_j)\boldsymbol{c}_{ij}\big)^{\mathsf{T}} \\ &+ \sum_{i\neq j\in\mathcal{I}(D_i)} d_{ij}^n(\mathbf{U}_j^{*,i,n} - \mathbf{U}_i^{*,j,n}) + \mu_{ij}^n\big(\mathbf{U}_j^n - \mathbf{U}_j^{*,i,n} - (\mathbf{U}_i^n - \mathbf{U}_i^{*,j,n})\big),\end{aligned} \tag{4.3}$$

$$\mu_{ij}^n := \max(\psi_i^n, \psi_j^n)\mu_{ij}^{v,n}, \quad i \neq j, \tag{4.4}$$

$$d_{ij}^n := \max(\psi_i^n, \psi_j^n)d_{ij}^{v,n}, \quad i \neq j, \tag{4.5}$$

with $\psi_i^n \in [0,1]$ yet to be determined. One possible choice for the second-order coefficient ψ_i^n consists of setting $\psi_i^n = \psi(\alpha_i^n)$ where we define

$$\alpha_i^n := \frac{|\sum_{j\in\mathcal{I}(D_i)} \mathsf{H}_j^n - \mathsf{H}_i^n|}{\sum_{j\in\mathcal{I}(D_i)} |\mathsf{H}_j^n - \mathsf{H}_i^n|}. \tag{4.6}$$

It is shown in Guermond and Popov [19] that any function ψ in $C^{0,1}([0,1];[0,1])$ with $\psi(1) = 1$ gives an algorithm that is positivity preserving up to a CFL condition, (see also [17] for the scalar version of the method and other possible choices for ψ_i^n). We take $\psi(\alpha) = \alpha^2$ in all the numerical simulations reported at the end of the paper.

PROPOSITION 4.4. *Let k_ψ be the Lipschitz constant of ψ. The scheme* (4.3)-(4.4)-(4.5) *is positivity preserving provided that* $\frac{\tau}{m_i}(-d_{ii}^n + \sum_{j\in\mathcal{I}(D_i)}(\boldsymbol{c}_{ij}\cdot\boldsymbol{V}_j^n)_-) \leq \frac{1}{2}$ *and* $\frac{\tau}{m_i}\max_{i\neq j\in\mathcal{I}(D_i)}(\boldsymbol{c}_{ij}\cdot V_j)_- \leq \frac{1}{4k_\psi c^\sharp}$ *where* $c_\sharp = \max_{i\in\{1:I\}} \mathrm{card}(\mathcal{I}(D_i))$.

Proof. By proceeding as in the proof of Proposition 4.2, we obtain

$$\begin{aligned}\mathsf{H}_i^{n+1} = \mathsf{H}_i^n &- \frac{\tau}{m_i}\sum_{i\neq j}\big(\mu_{ij}^n\mathsf{H}_i^n + (d_{ij}^n - \mu_{ij}^n)\mathsf{H}_i^{*,j,n}\big) \\ &+ \frac{\tau}{m_i}\sum_{i\neq j}\big((\mu_{ij}^n - \boldsymbol{c}_{ij}\cdot\boldsymbol{V}_j^n)\mathsf{H}_j^n + (d_{ij}^n - \mu_{ij}^n)\mathsf{H}_j^{*,i,n}\big).\end{aligned}$$

Using that $d^n_{ij} \geq \mu^n_{ij}$ and $\mathsf{H}^{*,i,n}_j \geq 0$, $\mathsf{H}^n_i \geq \mathsf{H}^{*,j,n}_i$ for all j, we obtain

$$\mathsf{H}^{n+1}_i \geq \mathsf{H}^n_i(1 - \frac{\tau}{m_i}\sum_{i\neq j} d^n_{ij}) + \frac{\tau}{m_i}\sum_{i\neq j}(\mu^n_{ij} - \boldsymbol{c}_{ij}\cdot\boldsymbol{V}^n_j)\mathsf{H}^n_j.$$

To finish the proof, it remains to show that the right-hand side is nonnegative under the appropriate CFL condition. The reader is referred to [19] for the proof of this result and for other choices for α^n_{ij} that also make the scheme (4.3) positivity preserving. ◻

Remark 4.5 (Linearity-preserving). It is possible to modify the definition of α^n_i in (4.6) to make the method linearity-preserving (the reader is referred to Berger et al. [3] for a review on linearity-preserving limiters in the finite volume literature). More precisely, when the shape functions are Lagrange-based, one can set $\alpha^n_i := \left|\sum_{j\in\mathcal{I}(D_i)} \beta_{ij}(\mathsf{H}^n_j - \mathsf{H}^n_i)\right| / \sum_{j\in\mathcal{I}(D_i)} \beta_{ij}|\mathsf{H}^n_j - \mathsf{H}^n_i|$ where the coefficients β_{ij} are generalized barycentric coordinates; see Guermond and Popov [17] for details. We take $\beta_{ij} = 1$ in all the numerical simulations reported at the end of the paper. □

5. Numerical illustrations. In this section we illustrate the performance of the various algorithms introduced in the paper. Most of the test cases are taken from the so-called SWASHES suite from Delestre et al. [11].

5.1. Technical details. All the numerical simulations are done in two space dimensions even when the problem under consideration has a one-dimensional solution. In order to avoid extraneous super-convergence effects we use unstructured, non-nested, Delaunay meshes composed of triangles. The computations are done with continuous Lagrange $\mathbb{P}_1$ finite elements. The time stepping is done with the SSP RK(3,3) method (three stages, third-order), see Shu and Osher [30, Eq. (2.18)] and Kraaijevanger [22, Thm. 9.4]. All the computations reported in this section have been done with the upper bound on $\lambda^{\boldsymbol{f}_{1\mathrm{D}}}_{\max}(\boldsymbol{v}_L, \boldsymbol{v}_R)$ given by Lemma 3.8.

To avoid division by zero in the presence of dry states we introduce $h_\epsilon := \epsilon \max_{\boldsymbol{x}\in D} h_0(\boldsymbol{x})$ with $\epsilon = 10^{-16}$, where h_0 is the initial water height. That is to say, we approximate the 0 water height by 10^{-16} times the maximum water height at the initial time. Then we regularize the gas dynamics flux $\boldsymbol{g}$ as follows: $\boldsymbol{g}_\epsilon(\boldsymbol{u}) := (\boldsymbol{q}, \frac{2h}{h^2+\max(h,h_\epsilon)^2}\boldsymbol{q}\otimes\boldsymbol{q})^\mathsf{T}$. That is to say the speed $\boldsymbol{v} := \boldsymbol{g}/h$ is regularized by setting $\boldsymbol{v}_\epsilon := \frac{2h}{h^2+\max(h,h_\epsilon)^2}\boldsymbol{q}$. Note that we obtain $\boldsymbol{g}(\boldsymbol{u}) = \boldsymbol{g}_\epsilon(\boldsymbol{u})$ and $\boldsymbol{v}_\epsilon = \boldsymbol{v}$ when $h \geq h_\epsilon$; that is, the regularization is active only when $h \leq h_\epsilon$.

All the schemes proposed in this paper are positivity preserving on the water height provided they are programmed correctly. Hence provided the initial water is non-negative, the water height should never become negative up to roundoff errors. We have observed that is is possible to avoid the effects of roundoff errors in the presence of dry regions by programming the update of the water height as follows:

$$\begin{aligned}(5.1)\quad \mathsf{H}^{n+1}_i = \mathsf{H}^n_i\Bigg(1 - \frac{\tau}{m_i}\Big(\boldsymbol{c}_{ii}\cdot\mathbf{V}^n_i + \sum_{i\neq j}\mu^n_{ij} + (d^n_{ij} - \mu^n_{ij})\frac{\mathsf{H}^{*,j,n}_i}{\mathsf{H}^n_i}\Big)\Bigg)\\ + \frac{\tau}{m_i}\sum_{i\neq j}\big(-\boldsymbol{c}_{ij}\cdot\mathbf{Q}^n_j + \mu^n_{ij}\mathsf{H}^n_j + (d^n_{ij} - \mu^n_{ij})\mathsf{H}^{*,i,n}_j\big),\end{aligned}$$

instead of setting $H^{n+1}_i = H^n_i + \frac{\tau}{m_i}\Delta R^n_i$ with

$$\Delta R^n_i := \sum_{j\in\mathcal{I}(D_i)} -\boldsymbol{c}_{ij}\cdot\mathbf{Q}^n_j + \mu^n_{ij}(\mathsf{H}^n_j - \mathsf{H}^n_i) + (d^n_{ij} - \mu^n_{ij})\big(\mathsf{H}^{*,i,n}_j - \mathsf{H}^{*,j,n}_i\big).$$

When doing convergence tests over meshes of different meshsize, the convergence rates are estimated as follows: given two errors e_1, e_2 obtained on two meshes $\mathcal{T}_{h1}$, $\mathcal{T}_{h2}$, and denoting $I_1 := \dim P(\mathcal{T}_{h1})$ $I_2 := \dim P(\mathcal{T}_{h2})$, the convergence rate is defined to be the ratio $d \log(e_1/e_2)/\log(I_2/I_1)$ since the quantity $I^{-\frac{1}{d}}$ scales like the meshsize. In all the test cases we take $g = 9.81\,\mathrm{m\,s^{-1}}$ and $d = 2$.

5.2. Well-balancing. We have verified on various tests, not reported here for brevity, that the proposed methods are well-balanced. More precisely, the first-order algorithm (3.3)–(3.5) is well-balanced irrespective of the structure of the mesh, i.e., the discharge stays close to the roundoff error indefinitely. The well-balancing of the second-order algorithm depends whether exact rest is possible or not as defined in Definition 2.4. If the mesh is such that exact rest is possible, then the algorithm is well-balanced up to machine accuracy indefinitely. If exact rest is not supported by the mesh, approximate well-balancing is achieved up to truncation error indefinitely.

5.3. Flows over a bump. We consider in this section several classical test cases detailed in [11, §3.1]. The domain is a one-dimensional channel $[0, L]$ with length $L = 25\,\mathrm{m}$. The bathymetry profile proposed in [11, §3.1] is flat with a parabolic bump, but to increase the smoothness of the solution in order to estimate the convergence rate properly, we modify a little bit the profile as follows:

$$z(x) = \begin{cases} \frac{0.2}{64}(x-8)^3(12-x)^3 & \text{if } 8 \le x \le 12 \\ 0, & \text{otherwise.} \end{cases} \tag{5.2}$$

Steady solutions satisfy mass conservation $q(x) = q(0)$ and the Bernoulli relation

$$\frac{q^2}{2gh^2} + h(x) + z(x) = C_{\mathrm{Ber}}. \tag{5.3}$$

where the Bernoulli constant C_{Ber} depends on the data. All the computations in §5.3 are done in two dimensions in the channel $D = [0, L]\times[0, 1]$.

5.3.1. Subcritical flow. We now consider a steady state solution with the inflow discharge $-\boldsymbol{q}\cdot\boldsymbol{n} = q_{\mathrm{in}} = 4.42\,\mathrm{m^2\,s^{-1}}$ imposed at $\{x = 0\}$ and $\boldsymbol{q}\cdot\boldsymbol{n} = 0$ on the sides of the channel $\{y = 0\} \cup \{y = 1\}$. The water height is enforced to be equal to $h_L = 2\,\mathrm{m}$ at $\{x = L\}$; hence $C_{\mathrm{Ber}} := \frac{q_{\mathrm{in}}^2}{2gh_L^2} + h_L$. The initial condition is $\boldsymbol{q}_0(\boldsymbol{x}) = 0$ and $h_0(\boldsymbol{x}) = h_L - z(x)$. We look for the solution at $t = 80\,\mathrm{s}$ which should be close to steady state. From Bernoulli's relation (5.3), $z(x) + h(x) + \frac{q_{\mathrm{in}}^2}{2gh^2(x)} = C_{\mathrm{Ber}}$ one gets that the exact steady state solution $h(x)$ solves the algebraic equation

$$h^3(x) + (z(x) - C_{\mathrm{Ber}})\, h^2(x) + \frac{q_{\mathrm{in}}^2}{2g} = 0, \quad \forall x \in [0, L]. \tag{5.4}$$

Let $b(x) := z(x) - C_{\mathrm{Ber}}$ and $d := \frac{q_{\mathrm{in}}^2}{2g}$. With the considered data, the cubic equation $h^3 + bh^2 + d = 0$ has three real zeros. The one that corresponds to the steady state solution is the largest root. Upon defining

$$Q(x) := -\frac{b^2(x)}{9}, \quad R(x) := -\frac{27d + 2b^3(x)}{54}, \quad \cos(\theta(x)) = (-Q(x))^{-\frac{3}{2}} R(x), \tag{5.5}$$

the water height is given by the trigonometric form of Cardano's formula:

$$h(x) = 2\sqrt{-Q(x)}\cos(\frac{\theta(x)}{3}) - \frac{b(x)}{3}. \tag{5.6}$$

Two types of computations are done with the scheme (4.3)–(4.5) using either the second-order viscosity $\psi(\alpha) = \alpha^2$ or the first-order viscosity $\psi(\alpha) = 1$. We use CFL = 1.25. In order to speedup the convergence to steady state we additionally impose the exact water height at $x = 0$. This artifact is used only to observe the theoretical convergence rate in space at $t = 80$. We show in Table 1 the error on the water height measured in the L^1-norm and in the L^2-norm. All the errors are relative to the corresponding norm of the exact solution. We observe that the convergence rates exceeds 2 both in the L^1-norm and in the L^2-norm for the viscosity $\psi(\alpha) = \alpha^2$. This is a super-convergence effect that we do not really understand at the moment. Let us recall that the meshes that are used here are non-nested, unstructured and the initial condition is rest. As expected the asymptotic convergence rate of the solution obtained with the first-order viscosity $\psi(\alpha) = 1$ is 1 irrespective of the norm.

Table 1: Subcritical flow over a bump with h given by (5.6). Computation done at $t = 80\,$s with initial data at rest; CFL=1.25. L^1-norm (rows 2–6), L^2-norm (rows 7–11). Viscosities are: $\psi(\alpha) = \alpha^2$ (columns 3–4); first-order viscosity (columns 5–6).

Norm	I	$\psi(\alpha) = \alpha^2$		$\psi(\alpha) = 1$	
L^1	248	1.46E-03	Rate	4.99E-03	Rate
	885	2.57E-04	2.73	3.39E-03	0.61
	3069	3.44E-05	3.08	1.95E-03	0.84
	12189	1.21E-06	3.09	1.03E-03	0.98
	48053	7.47E-07	2.66	5.19E-04	1.00
L^2	248	2.91E-3	Rate	9.57E-03	Rate
	885	6.48E-04	2.35	6.36E-03	0.64
	3069	1.25E-04	2.52	3.62E-03	0.86
	12189	2.31E-05	2.59	1.90E-03	0.99
	48053	4.04E-06	2.55	9.57E-04	1.00

5.3.2. Transcritical flow. We run again the above test in the transcritical regime. Given q_{in}, we set the Bernoulli constant C_{Ber} so that the Bernoulli relation (5.4) has two identical positive roots at the top of the bump, meaning that the discriminant of the equation (5.4), $Q^3 + R^2$, is zero, where Q and R are defined in (5.5). This fixes the Bernoulli constant C_{Ber} to be equal to $z_M + \frac{3}{2}(\frac{q_{\text{in}}^2}{g})^{\frac{1}{3}}$, where z_M is the height of the bump. The flow is fluvial (subsonic) upstream and becomes torrential (supersonic) at the top of the bump. The exact water height is the largest root of (5.4) when $x \le x_M$ and is the other positive root of (5.4) in the other case:

$$h(x) = \begin{cases} 2\sqrt{-Q(x)}\cos(\frac{\theta(x)}{3}) - \frac{b(x)}{3}, & \text{if } x \le x_M \\ 2\sqrt{-Q(x)}\cos(\frac{4\pi+\theta(x)}{3}) - \frac{b(x)}{3}, & \text{otherwise,} \end{cases} \tag{5.7}$$

where $\theta(x)$ is defined in (5.5) and x_M is such that $z(x_M)$ is the maximum of $z(x)$.

We take $q_{\text{in}} = 1.53\,\text{m}^2\,\text{s}^{-1}$. With the bottom topography defined in (5.2), we have $x_M = 10\,\text{m}$ and $z_M = 0.2\,\text{m}$. The flow rate is enforced at $\{x = 0\}$ and the exact water height (given by (5.7)) is enforced at the outflow $\{x = L\}$. We start with the initial condition $q(x) = 0\,\text{m}^2\,\text{s}^{-1}$ and $h(x) + z(x) = 0.66\,\text{m}$. The errors are measured at $t = 80\,$s. All the errors are relative to the corresponding norm of the exact solution. The computational domain is again $D = [0, 25]\times[0, 1]$. Two types of computations are done with the scheme (4.3)–(4.5) using either the second-order viscosity $\psi(\alpha) = \alpha^2$ or the first-order viscosity $\psi(\alpha) = 1$. We use CFL = 0.95. We show in Table 2 the error on the water height measured in the L^1-norm and in the L^2-norm.

Table 2: Transcritical flow over a bump with h given by (5.7). Computation done at $t = 80\,\mathrm{s}$ with initial data at rest; CFL=0.95. L^1-norm (rows 2–6), L^2-norm (rows 7–11). Viscosities are: $\psi(\alpha) = \alpha^2$ (columns 3–4); first-order viscosity (columns 5–6).

Norm	I	$\psi(\alpha)=\alpha^2$		$\psi(\alpha)=1$	
L^1	248	2.03E-02	Rate	1.63E-01	Rate
	885	3.49E-03	2.77	9.09E-02	0.92
	3069	4.71E-04	3.08	4.67E-02	1.02
	12189	9.86E-05	2.40	2.35E-02	1.05
	48053	1.95E-05	2.38	1.17E-02	1.02
L^2	248	2.28E-02	Rate	1.57E-01	Rate
	885	4.41E-03	2.58	8.73E-02	0.93
	3069	6.40E-04	2.96	4.49E-02	1.02
	12189	1.30E-04	2.44	2.27E-02	1.05
	48053	2.49E-05	2.42	1.13E-02	1.02

5.3.3. Transcritical flow over a bump with shock. We run again the above test in the transcritical regime with a hydraulic jump (i.e., a shock). To get a shock the flow must at some point become sonic and the water height at the outflow boundary must be larger than the water height at the sonic point. At the sonic point the discriminant of the Bernoulli relation (5.4) is zero. Just like in the test in §5.3.2 we position the sonic point at the top of the bump, i.e., the Bernoulli constant C_{Ber} is equal to $z_M + \frac{3}{2}(\frac{q_{\mathrm{in}}^2}{g})^{\frac{1}{3}}$, where z_M is the height of the bump. The flow is fluvial (subsonic) upstream and becomes torrential (supersonic) at the top of the bump and stays supersonic up to the hydraulic jump. Now we fix the location of the shock $x_S \in (x_M, 12)$. The water height before the hydraulic jump is the second largest root of (5.4): $h(x_S^-) = 2\sqrt{-Q(x_S^-)}\cos(\frac{4\pi+\theta(x_S^-)}{3}) - \frac{b(x_S^-)}{3}$. The water height after the jump is determined by the Rankine-Hugoniot relation: $h(x_S^+) = 0.5(-h(x_S^-) + \sqrt{\Delta})$, where $\Delta = (h(x_S^-))^2 + \frac{8q_{\mathrm{in}}^2}{gh(x_S^-)}$. In conclusion the exact solution for the water height is

$$h(x) = \begin{cases} 2\sqrt{-Q(x)}\cos(\frac{\theta(x)}{3}) - \frac{b(x)}{3}, & \text{if } x \le x_M \\ 2\sqrt{-Q(x)}\cos(\frac{4\pi+\theta(x)}{3}) - \frac{b(x)}{3}, & \text{if } x_M \le x < x_S \\ h(x_S^+) + z(x_S) - z(x), & x_S < x. \end{cases} \tag{5.8}$$

The bottom topography defined in (5.2) gives $x_M = 10\,\mathrm{m}$, $z_M = 0.2\,\mathrm{m}$. In our computations we take $q_{\mathrm{in}} = 0.18\,\mathrm{m}^2\,\mathrm{s}^{-1}$ to be consistent with the literature, Delestre et al. [11], Noelle et al. [26], but we could take any value for q_{in}. We use $x_S = 11.7\,\mathrm{m}$ and compute the water height at the outflow boundary $h_L := h(x_S^+) + z(x_S) - z(L)$ (using $g = 9.81\,\mathrm{m}\,\mathrm{s}^{-2}$, this gives $h_L = 0.282\,052\,798\,138\,021\,81\,\mathrm{m}$). Note that in [11, 26] the topography is different ($z(x) = \max(0, 0.2 - 0.05(x-10)^2)$), the gravity constant is also different ($g = 9.812\,\mathrm{m}\,\mathrm{s}^{-2}$), and the shock location is also different ($x_S = 11.665\,504\,281\,554\,291\,\mathrm{m}$). We insist on using our smooth bottom topography (5.2) instead of the parabolic profile, since it allows us to estimate properly the convergence rate of the method. With the non-smooth topography used in the literature ($z(x) = \max(0, 0.2 - 0.05(x-10)^2)$), the distance between the shock and the kink in the bottom topography is 0.3 m, which represent 1.2% of the length of the domain. To start observing a meaningful convergence rate with this topography using a quasi-uniform mesh would require to have at least 10 grid points between the two singularities, which would require to have at least 833 grid point in the x-direction and 33 points in the y-direction (since $D = [0, 25]\times[0, 1]$). The asymptotic convergence range is reached with far less grid points with our smooth topography.

The flow rate is enforced at $\{x = 0\}$ and the exact water height h_L is enforced at the outflow $\{x = L\}$. The initial condition is $q(x) = q_{\text{in}}$ and $h(x) + z(x) = h_L$. The errors are measured at $t = 80\,\text{s}$. Two types of computations are done with the scheme (4.3)–(4.5) using either the second-order viscosity $\psi(\alpha) = \alpha^2$ or the first-order viscosity $\psi(\alpha) = 1$. We use CFL $= 0.95$. We show in Table 3 the relative error on the water height measured in the L^1-norm and in the L^2-norm. Once again the superiority of the second-order viscosity $\psi(\alpha) = \alpha^2$ is evident.

Table 3: Transcritical flow with a shock, (5.8). Computation done at $t = 80\,\text{s}$ with initial data at rest; CFL=0.95. L^1-norm (rows 2–6), L^2-norm (rows 7–11) Viscosities are: $\psi(\alpha) = \alpha^2$ (columns 3–4); first-order viscosity $\psi(\alpha) = 1$ (columns 5–6).

Norm	I	$\psi(\alpha) = \alpha^2$		$\psi(\alpha) = 1$	
L^1	248	2.79E-02	Rate	7.40E-02	Rate
	885	7,97E-03	1.97	4.43E-02	0.81
	3069	4.03E-03	1.05	2.71E-02	0.75
	12189	2.69E-03	0.62	1.74E-02	0.68
	48053	1.54E-03	0.82	1.15E-02	0.61
L^2	248	6.70E-02	Rate	1.12E-01	Rate
	885	4.81E-02	0.52	8.60E-02	0.42
	3069	3.75E-02	0.38	7.71E-02	0.17
	12189	3.37E-02	0.17	7.19E-02	0.11
	48053	2.55E-02	0.41	6.54E-02	0.14

5.4. Unsteady flows. In the preceding sections, we went through steady-state solutions of increasing difficulties. These solutions are useful to check well-balancing and accuracy in space, but they do not give information about the transient behavior. Thus, in this section, we test transient solutions with wet/dry transitions.

5.4.1. Dam break on a dry bottom. We start with an ideal dam break called Ritter's solution, see [29]. This is a Riemann problem with the initial condition:

$$h(x) = \begin{cases} h_l & \text{if} \quad 0 \le x < x_0 \\ 0 & \text{if} \quad x_0 \le x < L, \end{cases} \tag{5.9}$$

where $h_l > 0$ and $v(x) = 0$ m/s. The analytical solution is

$$h(x,t) = \begin{cases} h_l & \text{if} \quad 0 \le x \le x_A(t) \\ \frac{4}{9g}\left(\sqrt{gh_l} - \frac{x-x_0}{2t}\right)^2 & \text{if} \quad x_A(t) \le x \le x_B(t) \\ 0 & \text{if} \quad x_B(t) \le x \le L, \end{cases} \tag{5.10}$$

$$v(x,t) = \begin{cases} 0 & \text{if} \quad 0 \le x \le x_A(t) \\ \frac{2}{3}\left(\frac{x-x_0}{t} + \sqrt{gh_l}\right) & \text{if} \quad x_A(t) \le x \le x_B(t) \\ 0 & \text{if} \quad x_B(t) \le x \le L, \end{cases} \tag{5.11}$$

where $x_A(t) = x_0 - t\sqrt{gh_l}$, $x_B(t) = x_0 + 2t\sqrt{gh_l}$. This test is used to check if the scheme preserves positivity of the water height and is able to locate and treat correctly the wet/dry transition. As in SWASHES [11], we consider $h_l = 0.005\,\text{m}$, $x_0 = 5\,\text{m}$, $L = 10\,\text{m}$ and $t = 6\,\text{s}$. The computational domain in $D = [0, L]\times[0, 1]$.

We show in Table 4 convergence results on the water height for the solution to the above problem at $t = 6\,\text{s}$ with two different initializations. The results in columns 3–6

Table 4: Problem (5.9) at $t = 6$ with data (5.10)-(5.11) at $t = 1$ (columns 3-6) and $t = 0$ (columns 7-10); CFL=0.5. L^1-norm (rows 2–6), L^2-norm (rows 7–11). Viscosities: $\psi(\alpha) = \alpha^2$ (columns 3–4; 7–8); first-order viscosity (columns 5–6; 9–10).

		Initialization time $t = 1$				Initialization time $t = 0$			
Norm	I	$\psi(\alpha) = \alpha^2$		$\psi(\alpha) = 1$		$\psi(\alpha) = \alpha^2$		$\psi(\alpha) = 1$	
L^1	248	1.52E-02	Rate	3.64E-02	Rate	3.33E-02	Rate	4.82E-02	Rate
	816	7.41E-03	1.20	2.17E-02	0.81	1.82E-02	1.01	3.38E-02	0.56
	3069	3.03E-03	1.35	1.22E-02	0.88	1.08E-02	0.79	2.39E-02	0.53
	12189	1.21E-03	1.34	6.70E-03	0.92	4.81E-03	1.16	1.52E-02	0.69
	48053	4.73E-04	1.37	3.54E-03	0.93	2.65E-03	0.87	9.61E-03	0.67
L^2	248	2.00E-01	–	4.65E-02	–	4.31E-01	–	6.14E-02	–
	816	1.10E-02	1.01	2.97E-02	0.70	2.45E-02	0.95	4.36E-02	0.54
	3069	5.42E-03	1.06	1.82E-02	0.76	1.40E-02	0.84	3.11E-02	0.52
	12189	2.65E-03	1.04	1.11E-02	0.76	7.13E-03	0.98	2.06E-02	0.63
	48053	1.28E-03	1.06	6.64E-03	0.75	3.83E-03	0.91	1.34E-02	0.63

have been obtained with the initial data given by (5.10)-(5.11) with the initial time $t = 1\,$s. This test is meant to estimate the accuracy on the method with a solution whose partial derivatives are in $\mathrm{BV}(D)$. We observe the rates $\frac{4}{3}$ in the L^1-norm and 1 in the L^2-norm with the viscosity $\psi(\alpha) = \alpha^2$. The rates are 1 and $\frac{3}{4}$ for the first-order viscosity, $\psi(\alpha) = 1$. The results on the discharge (not shown) give exactly the same convergence rates. The results in columns 7–10 have been obtained by using the Riemann data (5.9) at $t = 0\,$s. There is a loss of accuracy since the initial data is now only in $\mathrm{BV}(D)$. We observe the convergence rate 1 in the L^1-norm and the L^2-norm for the viscosity $\psi(\alpha) = \alpha^2$ and $\frac{2}{3}$ in the L^1-norm and the L^2-norm with the viscosity first-order viscosity $\psi(\alpha) = 1$. The results on the discharge (not shown) give exactly the same convergence rates. Note that with both initializations the $\psi(\alpha) = \alpha^2$ viscosity performs better than the first-order viscosity $\psi(\alpha) = 1$. We have also performed the above tests with the first-oder scheme (3.3)–(3.5) and the results (not shown) are are almost undistinguishable from those given by the scheme (4.3)–(4.5) with the first-order viscosity $\psi(\alpha) = 1$.

5.5. Planar surface in a paraboloid. We now consider a two-dimensional solution with moving shoreline developed by Thacker, see [31]. It is periodic in time with moving wet/dry transitions. It provides a perfect test for shallow water codes as it deals with bed slope and wetting/drying with two-dimensional effects. Moreover, as the gradient of the solution has BV regularity, it is appropriate to verify the accuracy of a numerical method up to second-order in $L^1(D)$. The topography is a paraboloid of revolution defined by

$$z(\boldsymbol{x}) = -h_0\left(1 - \left(\frac{r(\boldsymbol{x})}{a}\right)^2\right),$$

with $r(\boldsymbol{x}) = \sqrt{(x - L/2)^2 + (y - L/2)^2}$ for each $\boldsymbol{x} := (x, y) \in [0, L]\times[0, L]$. When the water is at rest, h_0 is the water height at the central point of the domain and a is the radius of the circular free surface. An analytical solution with a moving shoreline and a free surface that remains planar in time is given by

$$\text{(5.12)} \quad \begin{cases} h(\boldsymbol{x}, t) = \max(\frac{\eta h_0}{a^2}\left(2(x - \frac{L}{2})\cos(\omega t) + 2(y - \frac{L}{2})\sin(\omega t)\right) - z(x, y), 0), \\ v_x(\boldsymbol{x}, t) = -\eta\omega\sin(\omega t), \\ v_y(\boldsymbol{x}, t) = \eta\omega\cos(\omega t), \end{cases}$$

where the frequency is defined by $\omega = \sqrt{2gh_0}/a$ and η is a free parameter. To visualize this case, one can think of a glass with some liquid in rotation inside.

Table 5: Planar free surface in a paraboloid vessel with exact solution (5.12). Computations done at $t = 3 \times 2\pi/\omega$ with initial data (5.12) at $t = 0$; CFL=0.3. L^1-norm (rows 2–6); Second-order method with $\psi(\alpha) = \alpha^2$ (columns 3–4); Second-order method with $\psi(\alpha) = 1$ (columns 5–6); First-order method (columns 7–8).

Norm	I	Mthd. 2, $\psi(\alpha) = \alpha^2$		Mthd. 2, $\psi(\alpha) = 1$		Mthd. 1	
L^1	508	2.71E-01	Rate	6.25E-01	Rate	7.85E-01	Rate
	1926	6.51E-02	2.13	4.27E-01	0.57	7.44E-01	0.08
	7553	1.58E-02	2.08	2.54E-01	0.76	5.46E-01	0.45
	29870	4.46E-03	1.83	1.49E-01	0.88	3.33E-01	0.72
	118851	1.50E-03	1.58	7.26E-02	0.94	1.82R-01	0.87

The initial condition is the analytic solution at $t = 0$. Boundary conditions are natural, i.e., nothing is enforced. Typical values of parameters are the same as in SWASH [11] $a = 1$ m, $h_0 = 0.1$ m, $L = 4$ m, $\eta = 0.5$. The solution is computed up to time $t = 3 \times 2\pi/\omega$. The computational domain is $D = [0, L]\times[0, L]$.

5.6. Tidal wave over an island. We finish with a simulation of an experiment reported in Liu et al. [25], which consists of a water tank $D = [0, 30]\times[0, 25]$ with a conical island. The topography is

$$z(\boldsymbol{x}) := \min(h_{\text{top}}, (h_{\text{cone}} - r(\boldsymbol{x})/s_{\text{cone}})_+), \quad r(\boldsymbol{x}) := \sqrt{(x-15)^2 + (y-13)^2}, \tag{5.13}$$

where $h_{\text{top}} = 0.625\,\text{m}$, $h_{\text{cone}} = 0.9\,\text{m}$, $s_{\text{cone}} = 4\,\text{m}$. All the dimensions are in meters. We do not use the experimental set up for the initial conditions since there is no real consensus in the literature on the setup of the initial data. Instead, we set the initial condition to be a (solitary) wave big enough to overtop the island to demonstrate that the method is robust with respect to the presence of dry states. Moreover, we impose transparent boundary conditions to show that they are easy to enforce in the finite element setting. Essentially, imposing transparent boundary conditions consists of not doing anything (these are the so-called natural boundary conditions). The initial condition is given by $h(\boldsymbol{x}, 0) = h_{\text{init}}(\boldsymbol{x})$, $\boldsymbol{q}(\boldsymbol{x}, 0) = (u_{\text{init}}(\boldsymbol{x})h_{\text{init}}(\boldsymbol{x}), 0)$ where

$$h_{\text{init}}(\boldsymbol{x}) := \left(h_0 + \frac{A}{\cosh^2\left(\sqrt{\frac{3A}{4h_0^3}}(x - x_{\text{s}})\right)} - z(\boldsymbol{x})\right)_+, \tag{5.14}$$

$$u_{\text{init}}(\boldsymbol{x}) := \frac{A}{\cosh^2\left(\sqrt{\frac{3A}{4h_0^3}}(x - x_{\text{s}})\right)}\sqrt{\frac{g}{h_0}}, \tag{5.15}$$

with $h_0 = 0.32\,\text{m}$, $A = h_0$ and $x_{\text{s}} = 2.04\,\text{m}$. The computations are done on an unstructured Delaunay mesh composed of 174432 triangles and 87767 grid points. The average meshsize is 0.1 m. We report in Figure 2 the water elevation at 6 different times 4.08 s, 4.92 s, 5.88 s, 6.96 s, 9.72 s, 14.52 s showing the various stages of the overtopping of the island. To visualize properly the dry areas, the water height is set to zero in the images (not in the computations) when $h \leq 10^{-3}h_0$. For rendering purposes, the elevation map and the water height in the images are scaled by 3.

References.

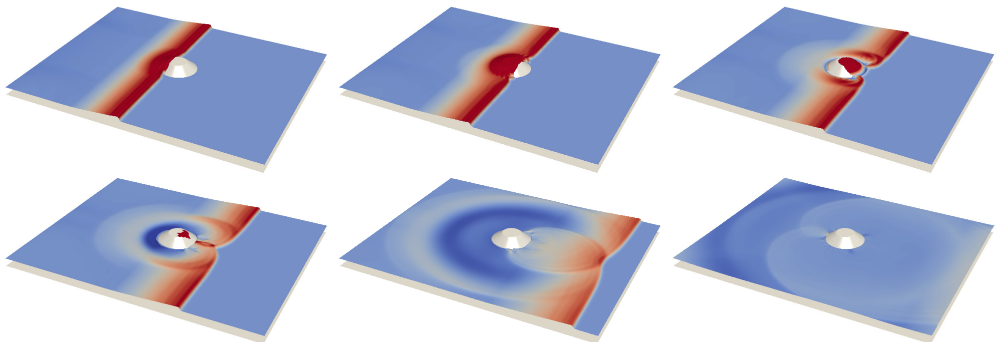

Fig. 2: Tidal wave overtopping a conical island.

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

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

[3] M. Berger, M. J. Aftosmis, and S. M. Murman. Analysis of slope limiters on irregular grids. AIAA Paper 2005-0490, American Institute for Aeronautics and Astronautics, Reno, NV, USA, May 2005. Also NASA TM NAS-05-007.

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

[5] C. Berthon and F. Foucher. Efficient well-balanced hydrostatic upwind schemes for shallow-water equations. *J. of Comput. Phys.*, 231(15):4993 – 5015, 2012.

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

[7] F. Bouchut. *Nonlinear stability of finite volume methods for hyperbolic conservation laws and well-balanced schemes for sources.* Frontiers in Mathematics. Birkhäuser Verlag, Basel, 2004.

[8] F. Bouchut and H. Frid. Finite difference schemes with cross derivatives correctors for multidimensional parabolic systems. *J. Hyperbolic Differ. Equ.*, 3(1): 27–52, 2006.

[9] S. Bryson, Y. Epshteyn, A. Kurganov, and G. Petrova. Well-balanced positivity preserving central-upwind scheme on triangular grids for the Saint-Venant system. *ESAIM Math. Model. Numer. Anal.*, 45(3):423–446, 2011.

[10] O. Delestre, S. Cordier, F. Darboux, and F. James. A limitation of the hydrostatic reconstruction technique for shallow water equations. *Comptes Rendus Mathematique*, 350(1314):677 – 681, 2012.

[11] O. Delestre, C. Lucas, P.-A. Ksinant, S. Cordier, F. Darboux, C. Laguerre, T.-N.-T. Vo, and F. James. SWASHES: a compilation of shallow water analytic solutions for hydraulic and environmental studies. *Int. J. Numer. Meth. Fluids*, 72:269 – 300, 2013.

[12] A. Duran, Q. Liang, and F. Marche. On the well-balanced numerical discretization of shallow water equations on unstructured meshes. *J. Comput. Phys.*, 235: 565–586, 2013.

[13] H. Frid. Maps of convex sets and invariant regions for finite-difference systems of conservation laws. *Arch. Ration. Mech. Anal.*, 160(3):245–269, 2001.
[14] J. M. Gallardo, C. Parés, and M. Castro. On a well-balanced high-order finite volume scheme for shallow water equations with topography and dry areas. *J. of Comput. Phys.*, 227(1):574 – 601, 2007.
[15] J. M. Greenberg and A. Y. Leroux. A well-balanced scheme for the numerical processing of source terms in hyperbolic equations. *SIAM J. Numer. Anal.*, 33 (1):1–16, 1996.
[16] J.-L. Guermond and B. Popov. Invariant domains and first-order continuous finite element approximation for hyperbolic systems. *SIAM J. on Numer. Anal.*, 54(4):2466–2489, 2016.
[17] J.-L. Guermond and B. Popov. Invariant domains and second-order continuous finite element approximation for scalar conservation equations. *SIAM J. on Numer. Anal.*, 2016, submitted.
[18] J.-L. Guermond and B. Popov. Estimation from above of the maximum wave speed in the Riemann problem for the Euler equations and related problems. 2017, preprint.
[19] J.-L. Guermond and B. Popov. Second-order positivity preserving continuous finite element approximation of hyperbolic systems of conservation laws. 2017, preprint.
[20] D. Hoff. A finite difference scheme for a system of two conservation laws with artificial viscosity. *Math. Comp.*, 33(148):1171–1193, 1979.
[21] D. Hoff. Invariant regions for systems of conservation laws. *Trans. Amer. Math. Soc.*, 289(2):591–610, 1985.
[22] J. F. B. M. Kraaijevanger. Contractivity of Runge-Kutta methods. *BIT*, 31(3): 482–528, 1991.
[23] A. Kurganov and G. Petrova. A second-order well-balanced positivity preserving central-upwind scheme for the Saint-Venant system. *Commun. Math. Sci.*, 5(1): 133–160, 2007.
[24] P. D. Lax. Weak solutions of nonlinear hyperbolic equations and their numerical computation. *Comm. Pure Appl. Math.*, 7:159–193, 1954.
[25] P. L. F. Liu, Y.-S. Cho, M. J. Briggs, U. Kanoglu, and C. E. Synolakis. Runup of solitary waves on a circular island. *J. Fluid Mech.*, 302:259–285, 11 1995.
[26] S. Noelle, Y. Xing, and C.-W. Shu. High-order well-balanced finite volume WENO schemes for shallow water equation with moving water. *J. Comput. Phys.*, 226(1):29–58, 2007.
[27] B. Perthame and C. Simeoni. A kinetic scheme for the Saint-Venant system with a source term. *Calcolo*, 38(4):201–231, 2001.
[28] M. Ricchiuto and A. Bollermann. Stabilized residual distribution for shallow water simulations. *J. Comput. Phys.*, 228(4):1071–1115, 2009.
[29] A. Ritter. Die fortpflanzung der wasserwellen. *Zeitschrift des Vereines Deutscher Ingenieure*, 36:947–954, 1892.
[30] C.-W. Shu and S. Osher. Efficient implementation of essentially non-oscillatory shock-capturing schemes. *J. Comput. Phys.*, 77(2):439 – 471, 1988.
[31] W. C. Thacker. Some exact solutions to the nonlinear shallow-water wave equations. *J. Fluid Mech.*, 107:499–508, 1981.
[32] Y. Xing and C.-W. Shu. A survey of high order schemes for the shallow water equations. *J. Math. Study*, 47(3):221–249, 2014.