SIAM J. SCI. COMPUT.
Vol. 40, No. 6, pp. A3873–A3901

WELL-BALANCED SECOND-ORDER FINITE ELEMENT APPROXIMATION OF THE SHALLOW WATER EQUATIONS WITH FRICTION*

JEAN-LUC GUERMOND†, MANUEL QUEZADA DE LUNA‡, BOJAN POPOV†, CHRISTOPHER E. KEES‡, AND MATTHEW W. FARTHING‡

Abstract. This paper investigates the approximation of the shallow water equations with topography and friction, using continuous finite elements. A new, second-order, parameter-free, well-balanced and positivity preserving explicit approximation technique is introduced. The novelties of the method are the explicit treatment of the friction term, the robust approximation of dry states, a commutator-based, high-order, entropy viscosity, and a local limiting procedure. The computational method is illustrated on various benchmark tests.

Key words. shallow water, well-balanced approximation, invariant domain, friction term, second-order accuracy, finite element method, positivity preserving

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

DOI. 10.1137/17M1156162

1. Introduction.

The present paper is concerned with the approximation of the time-dependent shallow water equations with friction, using explicit time stepping and continuous finite elements. The objective is to construct a method that is at least second order in space and third or higher order in time, well-balanced with respect to rest states (see Bermúdez and Vázquez [6], Greenberg and Le Roux [22]), well-balanced with respect to time-independent sliding solutions on inclined planes (see Chertock et al. [13]), and robust with respect to dry states. Such solution methods are available in the literature for finite volumes techniques either on structured meshes, as in [13], or unstructured meshes as in Duran et al. [17]. We also refer the reader to Audusse et al. [1], Bollermann, Noelle, and Lukáčová-Medvidová [8], Gallardo, Parés, and Castro [21], Kurganov and Petrova [38], Perthame and Simeoni [43], and Ricchiuto and Bollermann [44] for examples of schemes that are well-balanced at rest and robust with respect to dry states. A good review of well-balancing can be found in the book of Bouchut [10]. We also refer the reader to the paper of Xing and Shu [53] for a survey on well-balanced schemes in the context of finite volume and discontinuous Galerkin methods, and to [38] for a survey of central upwind schemes. However, to the best of our knowledge, approximations techniques with the above properties are not well developed in the context of continuous finite elements.

The starting point of the present paper is a method introduced in Azerad, Guermond, and Popov [3]. It is a finite element technique that is second-order accurate

*Submitted to the journal's Methods and Algorithms for Scientific Computing section November 9, 2017; accepted for publication (in revised form) August 30, 2018; published electronically November 27, 2018.

http://www.siam.org/journals/sisc/40-6/M115616.html

Funding: 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 FA9550-15-1-0257, and by the Army Research Office under grant W911NF-15-1-0517. Permission was granted by the Chief of Engineers to publish this information.

†Department of Mathematics, Texas A&M University, 3368 TAMU, College Station, TX 77843 (guermond@math.tamu.edu, popov@math.tamu.edu).

‡U.S. Army Engineer Research and Development Center, Coastal and Hydraulics Laboratory (ERDC-CHL), Vicksburg, MS 39180 (manuel.quezada.dl@gmail.com, Christopher.E.Kees@erdc.dren.mil, matthew.w.farthing@usace.army.mil).

in space, positivity preserving, and well-balanced with respect to rest states. The way the numerical viscosity is constructed in [3], though, makes the accuracy of the method limited to second order at best with a loss of accuracy at extrema in the water height. The goal of the present work is to go beyond the method described in [3]. More specifically, we propose extensions in the following four directions: (i) We introduce a singular Manning friction term and show that after proper regularization this term can be treated explicitly under the usual CFL condition. To the best of our knowledge, the proposed technique seems to be one of the first explicit methods available in the literature to handle this type of singular source term; (ii) we make the proposed method high order in space by adapting the entropy viscosity methodology introduced in Guermond, Pasquetti, and Popov [28]. One particular innovation consists in bypassing one mass matrix inversion by estimating a commutator; (iii) the high-order method is made positivity preserving via a new local limiting process. We adapt the technique introduced in Guermond et al. [31] to the shallow water setting and extract local lower bounds on the water heights and, more generally, exact local bounds that are used to limit the high-order solution. The resulting method is parameter free and is numerically shown to be accurate and robust in all problems we have solved.

The paper is organized as follows. The description of the problem and the finite element setting are introduced in section 2. The extension to the shallow water equations with friction of the method introduced in Azerad, Guermond, and Popov [3] is done in section 3. The explicit treatment and the regularization of the friction term are explained in section 3.2. The guaranteed maximum wave speed method that is used to extract exact bounds for limiting the high-order solution is exposed in section 4. The smoothness-based viscosity introduced in [3] is recalled in section 5 for completeness. The higher-order entropy viscosity extension of the method and the novel commutator technique that is used to estimate the entropy residual are introduced in section 6. The novel limiting technique announced above is exposed in section 6.4. The method is numerically illustrated on various benchmark tests in section 7; this section finishes with the simulation of the Malpasset dam break.

2. Preliminaries. We introduce the model problem and the finite element setting in this section.

2.1. The model problem. Let us consider a body of water evolving under the action of gravity and friction effects. Under the assumption that the deformations of the free surface are small compared to the water height and the bottom topography z varies slowly with respect to horizontal displacements, the problem can be well represented by the Saint-Venant shallow water model

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

where $D\subset\mathbb{R}^d$ is henceforth called the computational domain and d is the space dimension, which is either 1 or 2. The dependent variable is $\boldsymbol{u}=(\mathsf{h},\boldsymbol{q})^{\mathsf{T}}$, where h is the water height, and the vector field $\boldsymbol{q}$ is the flow rate or discharge. The ratio $\boldsymbol{v}:=\frac{1}{\mathsf{h}}\boldsymbol{q}$ is the depth averaged velocity. We refer to $\boldsymbol{v}$ as the velocity. The flux $\mathbb{f}(\boldsymbol{u})$ and $\boldsymbol{b}(\boldsymbol{u},\nabla z)$ are given by

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

where $\mathbb{I}_d$ is the $d\times d$ identity matrix. The mapping $z: D\ni\boldsymbol{x}\mapsto z(\boldsymbol{x})\in\mathbb{R}$ is the bottom topography and is assumed to be known. Finally to account for loss of

discharge through friction effects, we adopt Manning's friction law

$$\boldsymbol{S}(\boldsymbol{u}) := (0, -g\mathrm{n}^2\mathsf{h}^{-\gamma}\boldsymbol{q}\|\boldsymbol{v}\|_{\ell^2})^\mathsf{T}. \tag{2.3}$$

In the applications reported in the paper we take $\gamma = \frac{4}{3}$. The parameter n is Manning's roughness coefficient; it has units $\mathrm{m}^{\frac{\gamma-2}{2}}\mathrm{s}$.

2.2. Finite element setting. We are going to approximate the solution of (2.2) with continuous finite elements. For this purpose we introduce a shape-regular family of matching meshes $(\mathcal{T}_h)_{h>0}$. We are going to abuse the notation slightly by denoting the water height by h and the mesh size by h when the context is unambiguous. For instance, the finite element approximation of the water height is denoted h_h. The elements in $\mathcal{T}_h$ are assumed to be generated from a reference element denoted $\widehat{K}$. For any $K \in \mathcal{T}_h$, we denote by $T_K : \widehat{K} \to K$ 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. Let $(\widehat{K}, \widehat{P}, \widehat{\Sigma})$ be a reference finite element. The reference space $\widehat{P}$ is assumed to be composed of scalar-valued functions; we denote by k the largest natural number such that $\mathbb{P}_k \subset \widehat{P}$, where $\mathbb{P}_k$ is the space of the d-variate polynomials of degree at most k. In the applications reported at the end of the paper we are going to take $k = 1$, but most of what is said in the paper holds true for higher polynomial degree. The reference shape functions are denoted $\{\widehat{\theta}_i\}_{i\in\mathcal{N}}$; recall that $\widehat{P} = \mathrm{span}\{\widehat{\theta}_i\}_{i\in\mathcal{N}}$. We assume that the basis $\{\widehat{\theta}_i\}_{i\in\mathcal{N}}$ has the partition of unity property, $\sum_{i\in\mathcal{N}} \widehat{\theta}_i(\widehat{\boldsymbol{x}}) = 1$, for all $\widehat{\boldsymbol{x}} \in \widehat{K}$. We then introduce the finite element space

$$P(\mathcal{T}_h) := \{v \in C^0(D;\mathbb{R}) \mid v_{|K} \circ T_K \in \widehat{P}\ \forall K \in \mathcal{T}_h\}. \tag{2.4}$$

We are going to use $\boldsymbol{P}(\mathcal{T}_h) := [P(\mathcal{T}_h)]^{1+d}$ to approximate the conservative variable $\boldsymbol{u}$ in space. The approximation of the bottom topography will be done in $P(\mathcal{T}_h)$. The global shape functions in $P(\mathcal{T}_h)$ are denoted by $\{\varphi_i\}_{i\in\mathcal{V}}$, i.e., $\dim(P(\mathcal{T}_h)) = \mathrm{card}(\mathcal{V})$. Recall that $\varphi_{\mathtt{j\_dof}(K,i)|K} = \widehat{\theta}_i((T_K)^{-1})$ for all $i \in \mathcal{N}$ and all $K \in \mathcal{T}_h$, where $\mathtt{j\_dof}$: $\mathcal{T}_h \times \mathcal{N} \longrightarrow \mathcal{V}$ is the connectivity mapping. This identity together with the partition of unity property implies that $\sum_{i\in\mathcal{V}} \varphi_i(\boldsymbol{x}) = 1$ for all $\boldsymbol{x} \in D$.

For any $i \in \mathcal{V}$, we set $\mathcal{I}(i) := \{j \in \mathcal{V} \mid \varphi_i\varphi_j \not\equiv 0\}$ and refer to $\mathcal{I}(i)$ as the stencil of the shape function φ_i.

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 $m_i = \sum_{j\in\mathcal{I}(i)} m_{ij}$. We henceforth assume that

$$m_i > 0 \quad \forall i \in \mathcal{V}. \tag{2.5}$$

This assumption is satisfied by many finite element families: $\mathbb{P}_1$ elements on simplicies, $\mathbb{Q}_1$ elements on quadrangles and hexahedrons, prismatic elements, Bernstein finite elements, etc.

Finally, we define the following two quantities which will play an important role in the rest of the paper:

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

Note that the partition of unity property implies $\sum_{j\in\mathcal{V}} \boldsymbol{c}_{ij} = \boldsymbol{0}$. Furthermore, if either φ_i or φ_j is zero on ∂D, then $\boldsymbol{c}_{ij} = -\boldsymbol{c}_{ji}$. In particular we have $\sum_{i\in\mathcal{V}} \boldsymbol{c}_{ij} = \boldsymbol{0}$ if φ_j is zero on ∂D. This property will be used to establish conservation.

DEFINITION 2.1 (centrosymmetry). *The mesh $\mathcal{T}_h$ is said to be centrosymmetric if the following condition holds true: For all $i \in \mathcal{V}$, there is a permutation $\sigma_i : \mathcal{I}(i) \to \mathcal{I}(i)$ such that for every function $D_i \ni \boldsymbol{x} \to \sum_{j\in\mathcal{I}(i)} \mathsf{A}_j\varphi_j(\boldsymbol{x}) \in \mathbb{R}$ that is linear over D_i we have $\mathsf{A}_i = \frac{1}{2}(\mathsf{A}_j + \mathsf{A}_{\sigma_i(j)})$ for all $j \in \mathcal{I}(i)$.*

Although at some point in the paper we will invoke centrosymmetry of the mesh to establish formal consistency of some terms, we do not assume that the mesh is centrosymmetric in the rest of the paper.

3. Generic algorithm. We describe in this section a generic algorithm that is well-balanced. The positivity on the water height and the accuracy in space will depend on the definition of the artificial viscosity. Three possible definitions of the artificial viscosity are presented in sections 4, 5, and 6. The bottom topography is henceforth approximated by $z_h = \sum_{i\in\mathcal{V}} \mathsf{Z}_i\varphi_i$. We also use the following notation for the approximation of the water height and discharge: $\mathsf{h}_h = \sum_{i\in\mathcal{V}} \mathsf{H}_i\varphi_i$, $\boldsymbol{q}_h = \sum_{i\in\mathcal{V}} \mathbf{Q}_i\varphi_i$, respectively. The approximate conservative variable $\boldsymbol{u}_h := (\mathsf{h}_h, \boldsymbol{q}_h)^{\mathsf{T}}$ is represented as follows: $\boldsymbol{u}_h = \sum_{i\in\mathcal{V}} \mathbf{U}_i\varphi_i$, i.e., $\mathbf{U}_i := (\mathsf{H}_i, \mathbf{Q}_i)^{\mathsf{T}}$. The key novelty in this section with respect to Azerad, Guermond, and Popov [3] is the handling of the Manning friction.

3.1. Velocity regularization. The velocity, $\boldsymbol{v}_h := \frac{1}{\mathsf{h}_h}\boldsymbol{q}_h$, which is invoked to compute the flux and other related quantities, is approximated as follows: $\boldsymbol{v}_h = \sum_{i\in\mathcal{V}} \mathbf{V}_i\varphi_i$, where $\mathbf{V}_i$ is related to the water height and the discharge by using a formula that avoids division by zero in dry regions. There are many ways to remove the dry state singularity; in this paper we are going to use a formula similar in spirit to that in Kurganov and Petrova [38, Eq. (2.17)], Chertock et al. [13, Eq. (3.10)], [3, section 5.1]. Recalling that h_0 is the initial water height, we set $\mathsf{H}_{0,\max} = \operatorname{ess\,sup}_{\boldsymbol{x}\in D} \mathsf{h}_0(\boldsymbol{x})$, and we define the following regularization length scale and regularized velocity:

$$\mathsf{H}^\epsilon := \epsilon\, \mathsf{H}_{0,\max}, \qquad \mathbf{V}_i := \frac{2\mathsf{H}_i}{\mathsf{H}_i^2 + \max(\mathsf{H}_i, \mathsf{H}^\epsilon)^2}\mathbf{Q}_i, \tag{3.1}$$

where ϵ is a tiny parameter that takes care of roundoff errors. In the applications reported at the end of the paper we take $\epsilon = 10^{-13}$ to account for double precision arithmetic. Note that with the above definition we have $\mathbf{V}_i = \frac{1}{\mathsf{H}_i}\mathbf{Q}_i$ if $\mathsf{H}_i \geq \mathsf{H}^\epsilon$; hence, the proposed regularization is active only in situations where genuine dry states occur. Whether this type of regularization can be physically justified is not clear. Further studies in this direction should be done in the future.

3.2. Full time and space approximation. The notion of well-balancing is rooted in the seminal work of Bermúdez and Vázquez [6, Def. 1] and Greenberg and Le Roux [22]. The idea is that well-balanced schemes should at the very least preserve steady states at rest. Following Audusse et al. [1], we are going to make use of the so-called hydrostatic reconstruction of the water height to make the method well-balanced with respect to rest states:

$$\mathsf{H}_i^{*,j} := \max(0, \mathsf{H}_i + \mathsf{Z}_i - \max(\mathsf{Z}_i, \mathsf{Z}_j)) \quad \forall i \in \mathcal{V},\ j \in \mathcal{I}(i). \tag{3.2}$$

This leads us to introduce the auxiliary state $\mathbf{U}_i^{*,j} := \frac{\mathsf{H}_i^{*,j}}{\mathsf{H}_i}\mathbf{U}_i$.

Let $\boldsymbol{u}_h^0 = \sum_{i\in\mathcal{V}} \mathbf{U}_i^0\varphi_i \in \boldsymbol{P}(\mathcal{T}_h)$ be a reasonable approximation of $\boldsymbol{u}_0$, where we recall that $\boldsymbol{u}_0$ is the initial datum. Let τ be the time step at the current time t_n, $n \in \mathbb{N}$, and let us set $t_{n+1} := t_n + \tau$. Let $\boldsymbol{u}_h^n := \sum_{i\in\mathcal{V}} \mathbf{U}_i^n\varphi_i \in \boldsymbol{P}(\mathcal{T}_h)$ be the space approximation of $\boldsymbol{u}$ at time t_n and let us set $\boldsymbol{u}_h^{n+1} := \sum_{i\in\mathcal{V}} \mathbf{U}_i^{n+1}\varphi_i$.

We construct the update $\boldsymbol{u}_h^{n+1}$ by following the same strategy as in [3]; the key difference is in the treatment of the Manning friction. The idea is to use the Galerkin method augmented with some artificial viscosity. The time stepping is done with a strong stability preserving (SSP) Runge–Kutta method with the mass matrix lumped. Regarding the friction term, we approximate $\int_D g\mathrm{n}^2\mathsf{h}^{-\gamma}\boldsymbol{q}\|\boldsymbol{v}\|_{\ell^2}\varphi_i\,\mathrm{d}x$ as follows:

$$\int_D g\mathrm{n}^2\mathsf{h}^{-\gamma}\boldsymbol{q}\|\boldsymbol{v}\|_{\ell^2}\varphi_i\,\mathrm{d}x \approx \frac{2g\mathrm{n}^2\mathbf{Q}_i^n\|\mathbf{V}_i\|_{\ell^2}m_i}{(\mathsf{H}_i^n)^\gamma + \max\left((\mathsf{H}_i^n)^\gamma, 2\tau g\mathrm{n}^2\|\mathbf{V}_i\|_{\ell^2}\right)}. \tag{3.3}$$

Remark 3.1 (friction regularization). Observe that the above definition of the friction reduces to $g\mathrm{n}^2(\mathsf{H}_i^n)^{-\gamma}\mathbf{Q}_i^n\|\mathbf{V}_i\|_{\ell^2}m_i$ when $\mathsf{H}_i^n \geq (2\tau g\mathrm{n}^2\|\mathbf{V}_i\|_{\ell^2})^{\frac{1}{\gamma}}$; that is to say, the regularization is inactive away from dry states when the time step τ is small enough. This regularization allows us to use explicit time stepping with a standard CFL restriction and makes the scheme well-balanced with respect to rest states and sliding steady states. A precise statement about well-balancing is made in Proposition 3.7. Notice that in the literature the Manning friction is usually treated semi-implicitly. Since the proposed regularization allows for explicit treatment, the present scheme is easier to implement and possibly faster. ∎

We approximate the term $\int_D(\nabla\cdot(\mathbb{f}(\boldsymbol{u})) + (0, gh\nabla z)^\mathsf{T})\varphi_i\,\mathrm{d}x$ as follows:

$$\int_D \left(\nabla\cdot(\mathbb{f}(\boldsymbol{u})) + \begin{pmatrix}0\\ gh\nabla z\end{pmatrix}\right)\varphi_i\,\mathrm{d}x \approx \sum_{j\in\mathcal{I}(i)} \begin{pmatrix}\mathsf{H}_j^n\mathbf{V}_j^n\cdot\boldsymbol{c}_{ij}\\ \mathbf{V}_j^n\mathbf{Q}_j^n\cdot\boldsymbol{c}_{ij} + g\mathsf{H}_i^n(\mathsf{H}_j^n + Z_j)\boldsymbol{c}_{ij}\end{pmatrix}. \tag{3.4}$$

Since all the SSP Runge–Kutta methods are composed of convex combinations of the forward Euler method, we restrict the presentation of the scheme to the forward Euler method. Recalling that the mass matrix is lumped in [3], we update $\mathbf{U}_i^{n+1}$, $i\in\mathcal{V}$, as follows:

$$\begin{aligned} m_i\frac{\mathbf{U}_i^{n+1}-\mathbf{U}_i^n}{\tau} = {} & \sum_{j\in\mathcal{I}(i)} -\left(\mathsf{H}_j^n\mathbf{V}_j^n\cdot\boldsymbol{c}_{ij}, \mathbf{V}_j^n\mathbf{Q}_j^n\cdot\boldsymbol{c}_{ij} + g\mathsf{H}_i^n(\mathsf{H}_j^n+Z_j)\boldsymbol{c}_{ij}\right)^\mathsf{T} \\ & -\left(0, \frac{2g\mathrm{n}^2\mathbf{Q}_i^n\|\mathbf{V}_i\|_{\ell^2}m_i}{(\mathsf{H}_i^n)^\gamma + \max((\mathsf{H}_i^n)^\gamma, 2g\mathrm{n}^2\tau\|\mathbf{V}_i\|_{\ell^2})}\right)^\mathsf{T} \\ & + \sum_{j\in\mathcal{I}(i)\setminus\{i\}} (d_{ij}^n-\mu_{ij}^n)\left(\mathbf{U}_j^{*,i,n}-\mathbf{U}_i^{*,j,n}\right) + \mu_{ij}^n(\mathbf{U}_j^n-\mathbf{U}_i^n), \end{aligned} \tag{3.5}$$

$$d_{ij}^n = d_{ji}^n, \quad \mu_{ij}^n = \mu_{ji}^n, \quad d_{ij}^n \geq \mu_{ij}^n \geq 0, \quad i\neq j. \tag{3.6}$$

Here d_{ij}^n, μ_{ij}^n are the artificial viscosity coefficients. Setting $d_{ij}^n = \mu_{ij}^n$ gives the Galerkin method with lumped mass matrix; the scheme is then high-order accurate in space but positivity is a priori lost. The purpose of μ_{ij}^n and d_{ij}^n is to introduce numerical dissipation in a controlled way in order to preserve well-balancing and positivity of the water height. Precise definitions of d_{ij}^n and μ_{ij}^n are given in sections 4, 5, and 6. Well-balancing is established in section 3.3 and positivity is established in sections 4, 5, and 6.

Remark 3.2 (Dirichlet boundary conditions). The question of boundary conditions for hyperbolic systems is highly nontrivial, but whenever Dirichlet boundary conditions have to be enforced either on the water height or on the normal component of the discharge, we enforce them a posteriori on $\mathbf{U}_i^{n+1}$. When using an SSP

Runge–Kutta method, the enforcement of Dirichlet conditions is done at the end of every Runge–Kutta substep. ∎

3.3. Conservation and well-balancing. In this section we investigate some properties of the scheme (3.5)–(3.6) that are independent of the definition of the numerical viscosity coefficients μ_{ij}^n and d_{ij}^n.

DEFINITION 3.3 (conservation). *We say that a mapping* $\boldsymbol{R} : \boldsymbol{P}(\mathcal{T}_h) \to \boldsymbol{P}(\mathcal{T}_h)$ *is a conservative approximation of* (2.1) *if* $\sum_{i\in\mathcal{V}} m_i \boldsymbol{R}(\mathbf{U}) = \sum_{i\in\mathcal{V}} m_i \mathbf{U}$ *when the topography map is constant and there is no friction.*

PROPOSITION 3.4. *The mapping* $\mathbf{U}^n \to \mathbf{U}^{n+1}$ *defined by the scheme* (3.5)–(3.6) *is conservative.*

Proof. For the proof, see Azerad, Guermond, and Popov [3]; see also Remark 3.9. ∎

The key idea advocated by Bermúdez and Vázquez [6] and Greenberg and Le Roux [22] is that good numerical schemes should at the very least preserve steady states at rest and possibly preserve some other time-independent solutions. We refer the reader to Noelle, Xing, and Shu [42] where this question is addressed in the context of high-order methods.

DEFINITION 3.5 (exact well-balancing at rest). (i) *A numerical state* $(\mathsf{h}_h, \boldsymbol{q}_h, z_h)$ *is said to be at exact rest (or exactly at rest) if the approximate discharge* $\boldsymbol{q}_h$ *is zero, and if the approximate water height* h_h *and the approximate bottom topography map* z_h *satisfy the following alternative for all* $i \in \mathcal{V}$*: either* $\mathsf{H}_j = \mathsf{H}_i = 0$ *or* $\mathsf{H}_j + \mathsf{Z}_j = \mathsf{H}_i + \mathsf{Z}_i$ *for all* $j \in \mathcal{I}(i)$. (ii) *A mapping* $\boldsymbol{R} : \boldsymbol{P}(\mathcal{T}_h) \to \boldsymbol{P}(\mathcal{T}_h)$ *is said to be an exactly well-balanced approximation of* (2.1) *if* $\boldsymbol{R}(\boldsymbol{u}_h) = \boldsymbol{u}_h$ *when* $\boldsymbol{u}_h$ *is an exact rest state.*

PROPOSITION 3.6. *The scheme* (3.5)–(3.6) *is exactly well-balanced if* $\mu_{ij}^n = 0$ *when* $\mathbf{V}_i^n = \mathbf{V}_j^n = \mathbf{0}$.

Proof. The statement has been proved in [3] when there is no friction. But at rest, the friction term is zero, so its contribution to the discharge equation is zero; hence the argument from [3] still holds in this case as well. ∎

Now following Chertock et al. [13], we observe that when the bottom is an infinite inclined plane, the system (2.1) admits a steady state solution that solves $\boldsymbol{b}(\boldsymbol{u}, \nabla z) = \boldsymbol{S}(\boldsymbol{u})$. More precisely, assuming that the plane has two tangent orthonormal vectors $\boldsymbol{t}_1, \boldsymbol{t}_2$ with $\boldsymbol{t}_2$ being horizontal and $\boldsymbol{t}_1$ pointing downward, we have $\nabla z = -b\boldsymbol{t}_1$ with $b > 0$. The steady state solution to (2.1) is given by

$$\boldsymbol{q}(\boldsymbol{x},t)\cdot\boldsymbol{t}_2 = 0, \quad \boldsymbol{q}(\boldsymbol{x},t)\cdot\boldsymbol{t}_1 = q_0, \quad \mathsf{h}(\boldsymbol{x},t) = \mathsf{h}_0 := \left(\frac{\mathrm{n}^2 q_0^2}{b}\right)^{\frac{1}{2+\gamma}}. \tag{3.7}$$

PROPOSITION 3.7 (well-balanced steady state). *Let* q_0, h_0 *be defined in* (3.7). *Assume that* $\tau \le \frac{q_0}{2gb\mathsf{h}_0}$. *Suppose also that the following alternative holds:* (i) *The mesh is centrosymmetric and fine enough, and the artificial viscosity is defined so that* d_{ij}^n *and* μ_{ij}^n *are constant when* $\mathbf{U}_j^n = \mathbf{U}_i^n$ *for all* $j \in \mathcal{I}(i)$*; or* (ii) *the mesh is nonuniform but the artificial viscosity is defined so that* $d_{ij}^n = 0$ *and* $\mu_{ij}^n = 0$ *when* $\mathbf{U}_i^n = \mathbf{U}_j^n$ *for all* $j \in \mathcal{I}(i)$ *and all* $i \in \mathcal{V}$. *Then the field* $\mathbf{U}^{\mathrm{st}} := (\mathsf{H}^{\mathrm{st}}, \mathbf{Q}^{\mathrm{st}})^{\mathsf{T}}$, *where* $\mathbf{Q}_i^{\mathrm{st}}\cdot\boldsymbol{t}_2 = 0$, $\mathbf{Q}_i^{\mathrm{st}}\cdot\boldsymbol{t}_1 = q_0$, $\mathsf{H}_i^{\mathrm{st}} = \mathsf{h}_0 = \left(b^{-1}\mathrm{n}^2 q_0^2\right)^{\frac{1}{2+\gamma}}$, *is a steady state solution to* (3.5)–(3.6).

Proof. Suppose that $\mathsf{H}_i^n = \mathsf{h}_0$ and $\mathbf{Q}_i^n = q_0 \boldsymbol{t}_1$ for all $i \in \mathcal{V}$. Let us prove that $\frac{1}{\tau}(\mathbf{U}_i^{n+1} - \mathbf{U}_i^n) = 0$. Since the bottom topography is linear, we have

$$\sum_{j \in \mathcal{I}(i)} Z_j \boldsymbol{c}_{ij} = \int_D \nabla z \varphi_i \, \mathrm{d}x = m_i \nabla z = -b m_i \boldsymbol{t}_1. \tag{3.8}$$

Since the discrete field $\mathbf{U}^n$ is constant is space, using the property $\sum_{j \in \mathcal{I}(i)} \boldsymbol{c}_{ij} = 0$ we infer that

$$\sum_{j \in \mathcal{I}(i)} -\left(\mathsf{H}_j^n \mathbf{V}_j^n \cdot \boldsymbol{c}_{ij}, \mathbf{V}_j^n \mathbf{Q}_j^n \cdot \boldsymbol{c}_{ij} + g \mathsf{H}_i^n (\mathsf{H}_j^n + Z_j) \boldsymbol{c}_{ij}\right)^{\mathsf{T}} = (0, g \mathsf{h}_0 m_i b \boldsymbol{t}_1)^{\mathsf{T}}. \tag{3.9}$$

Moreover, using that $2g\tau \le \frac{q_0}{b\mathsf{h}_0}$, we infer that $\mathsf{h}_0^\gamma = \frac{\mathrm{n}^2 q_0}{\mathsf{h}_0} \frac{q_0}{b\mathsf{h}_0} \ge 2g\mathrm{n}^2 \tau \frac{q_0}{\mathsf{h}_0}$. Hence the regularization in the approximation of the Manning friction (3.3) is inactive, and the friction term reduces to $-g\mathrm{n}^2 q_0^2 \mathsf{h}_0^{-1-\gamma} m_i \boldsymbol{t}_1$. As a result, denoting by r.h.s. the right-hand side of (3.5), we have

$$\begin{aligned} \text{r.h.s.} &= \left(0, -g\mathrm{n}^2 q_0^2 \mathsf{h}_0^{-1-\gamma} m_i \boldsymbol{t}_1 + g \mathsf{h}_0 m_i b \boldsymbol{t}_1\right)^{\mathsf{T}} \\ &\quad + \sum_{j \in \mathcal{I}(i) \setminus \{i\}} (d_{ij}^n - \mu_{ij}^n) \left(\mathbf{U}_j^{*,i,n} - \mathbf{U}_i^{*,j,n}\right) + \mu_{ij}^n (\mathbf{U}_j^n - \mathbf{U}_i^n) \\ &= \sum_{j \in \mathcal{I}(i) \setminus \{i\}} (d_{ij}^n - \mu_{ij}^n) \left(\mathbf{U}_j^{*,i,n} - \mathbf{U}_i^{*,j,n}\right). \end{aligned} \tag{3.10}$$

In alternative (i), we have $\min(\mathsf{h}_0 + \mathsf{Z}_i - \max(\mathsf{Z}_i, \mathsf{Z}_j), \mathsf{h}_0 + \mathsf{Z}_j - \max(\mathsf{Z}_i, \mathsf{Z}_j)) > 0$ since the mesh is supposed to be fine enough; hence $\mathsf{H}_j^{*,i,n} - \mathsf{H}_i^{*,j,n} = \mathsf{Z}_j - \mathsf{Z}_i$. This means that $\mathbf{U}_j^{*,i,n} - \mathbf{U}_i^{*,j,n} = (\mathsf{Z}_j - \mathsf{Z}_i)(1, \mathsf{h}_0^{-1} q_0 \boldsymbol{t}_1)^{\mathsf{T}}$. Since by assumption, the coefficients $d := d_{ij}^n$ and $\mu := \mu_{ij}^n$ do not depend on $j \in \mathcal{I}(i)$, we have

$$\text{r.h.s.} = \alpha(d - \mu) \begin{pmatrix} 1 \\ \mathsf{h}_0^{-1} q_0 \boldsymbol{t}_1 \end{pmatrix} \sum_{j \in \mathcal{I}(i) \setminus \{i\}} (\mathsf{Z}_j - \mathsf{Z}_i). \tag{3.11}$$

Let $\sigma_i : \mathcal{I}(i) \to \mathcal{I}(i)$ be the permutation introduced in the centrosymmetry definition (see Definition 2.1); we have $\sum_{j \in \mathcal{I}(i) \setminus \{i\}} \mathsf{Z}_j = \frac{1}{2} \sum_{j \in \mathcal{I}(i) \setminus \{i\}} \left(\mathsf{Z}_j + \mathsf{Z}_{\sigma_i(j)}\right)$. The centrosymmetry assumption implies that

$$\text{r.h.s.} = \alpha(d - \mu) \begin{pmatrix} 1 \\ \mathsf{h}_0^{-1} q_0 \boldsymbol{t}_1 \end{pmatrix} \frac{1}{2} \sum_{j \in \mathcal{I}(i) \setminus \{i\}} \left(\mathsf{Z}_j + \mathsf{Z}_{\sigma_i(j)} - 2\mathsf{Z}_i\right) = 0, \tag{3.12}$$

since the function $\boldsymbol{x} \to \sum_{j \in \mathcal{I}(i)} \mathsf{Z}_j \varphi_j(\boldsymbol{x})$ is linear over D_i for all $i \in \mathcal{V}$. Hence $\frac{1}{\tau}(\mathbf{U}_i^{n+1} - \mathbf{U}_i^n) = 0$ as desired. In the other alternative (ii), we suppose that $d_{ij}^n = \mu_{ij}^n = 0$ when $\mathbf{U}_i^n = \mathbf{U}_j^n$ for all $j \in \mathcal{I}(i)$; this means that the right-hand side of (3.5) is zero, whence $\frac{1}{\tau}(\mathbf{U}_i^{n+1} - \mathbf{U}_i^n) = 0$ as desired. This completes the proof. ∎

Remark 3.8. The first-order artificial viscosity proposed in section 4 satisfies the assumption of the alternative (i). The higher-order viscosities proposed in sections 5 and 6 satisfy the assumption of alternative (ii). ∎

Remark 3.9 (local conservation). The notion of conservation introduced in Definition 3.3 may look inappropriate to the reader who is more familiar with the finite volume or/and discontinuous Galerkin literature since our definition of conservation is global. Actually, assuming that the topography map is constant and there is no friction, the update (3.5) can be rewritten $m_i\mathbf{U}_i^{n+1} = m_i\mathbf{U}_i^n + \tau\sum_{j\in\mathcal{I}(i)}\mathbb{F}_{ij}^n$, where $\mathbb{F}_{ij}^n := -\big((\mathsf{H}_j^n\mathbf{V}_j^n + \mathsf{H}_i^n\mathbf{V}_i^n)\cdot\boldsymbol{c}_{ij}, (\mathbf{V}_j^n\mathbf{Q}_j^n + \mathbf{V}_i^n\mathbf{Q}_i^n)\cdot\boldsymbol{c}_{ij} + g\mathsf{H}_i^n\mathsf{H}_j^n\boldsymbol{c}_{ij}\big)^{\mathsf{T}} + \mu_{ij}^n(\mathbf{U}_j^n - \mathbf{U}_i^n) + (d_{ij}^n - \mu_{ij}^n)(\mathbf{U}_j^{*,i,n} - \mathbf{U}_i^{*,j,n})$ since $\sum_{j\in\mathcal{I}(i)}\mathbf{V}_i^n\mathbf{Q}_i^n\cdot\boldsymbol{c}_{ij} = \mathbf{V}_i^n\mathbf{Q}_i^n\cdot(\sum_{j\in\mathcal{I}(i)}\boldsymbol{c}_{ij}) = \mathbf{0}$ and $\sum_{j\in\mathcal{I}(i)}\mathsf{H}_i^n\mathbf{V}_i^n\cdot\boldsymbol{c}_{ij} = \mathsf{H}_i^n\mathbf{V}_i^n\cdot(\sum_{j\in\mathcal{I}(i)}\boldsymbol{c}_{ij}) = 0$. The key property we are using here is that definition (2.6) implies that $\sum_{j\in\mathcal{I}(i)}\boldsymbol{c}_{ij} = \mathbf{0}$, which is a consequence of the partition of unity property. Using definition (2.6) again we observe that $\boldsymbol{c}_{ij} = -\boldsymbol{c}_{ji}$ if either φ_i or φ_j is zero on ∂D. As a result, away from the boundary, we have $\mathbb{F}_{ij}^n = -\mathbb{F}_{ji}^n$; that is to say, the mass flux from the degree of freedom j to the degree of freedom i is opposite to the mass flux from the degree of freedom i to the degree of freedom j. This is the property that is usually understood as "local mass conservation" in the finite volume or/and discontinuous Galerkin literature. In conclusion, the update (3.5), i.e., $m_i\frac{\mathbf{U}_i^{n+1}-\mathbf{U}_i^n}{\tau} = \sum_{j\in\mathcal{I}(i)}\mathbb{F}_{ij}^n$, is "locally conservative" in the sense that $\mathbb{F}_{ij}^n = -\mathbb{F}_{ji}^n$ for any $j\in\mathcal{I}(i)$ and any $i\in\mathcal{V}$ (and for for any $i\in\mathcal{I}(j)$ and any $j\in\mathcal{V}$) away from the boundary and if the topography map is constant and there is no friction. ∎

Remark 3.10 (hydrostatic reconstruction). It has been observed in Delestre et al. [15] that the hydrostatic reconstruction may not be appropriate to approximate steady state solutions that are not at rest "for certain combinations of water height, slope, and mesh size." This issue has been investigated in Audusse et al. [2], Chen and Noelle [12]. We are not going to address this problem in the present paper. ∎

4. Guaranteed maximum speed (GMS) viscosity.

In this section we give a definition of the artificial viscosity coefficients μ_{ij}^n, d_{ij}^n that makes the method positive and entropy satisfying (when the bottom is flat), but also reduces the accuracy in space to first order (see [3]). More accurate definitions of the viscosities μ_{ij}^n and d_{ij}^n, producing higher-order methods, are given in sections 5 and 6.

4.1. Definition of the GMS viscosity.

In order to distinguish the low-order viscosity from other higher-order variants, we now use the superscript $^{\mathrm{V}}$ and define μ_{ij}^{V} and $d_{ij}^{\mathrm{V},n}$ as follows:

$$\mu_{ij}^{\mathrm{V},n} := \max((\mathbf{V}_i\cdot\boldsymbol{n}_{ij})_-, (\mathbf{V}_j\cdot\boldsymbol{n}_{ij})_+)\|\boldsymbol{c}_{ij}\|_{\ell^2}, \qquad i\neq j, \tag{4.1}$$

$$d_{ij}^{\mathrm{V},n} := \max\left(\overline{\lambda}_{\max}^{\mathbb{f}_{1\mathrm{D}}}(\boldsymbol{n}_{ij},\mathbf{U}_i^n,\mathbf{U}_j^n)\|\boldsymbol{c}_{ij}\|_{\ell^2}, \overline{\lambda}_{\max}^{\mathbb{f}_{1\mathrm{D}}}(\boldsymbol{n}_{ji},\mathbf{U}_j^n,\mathbf{U}_i^n)\|\boldsymbol{c}_{ji}\|_{\ell^2}\right), \qquad i\neq j, \tag{4.2}$$

with the notation $a_+ := \max(a,0)$ and $a_- = -\min(a,0)$. Here $\overline{\lambda}_{\max}^{\mathbb{f}_{1\mathrm{D}}}(\boldsymbol{n},\mathbf{U}_L,\mathbf{U}_R)$ is an upper bound for the maximum wave speed in the following one-dimensional Riemann problem $\partial_t\boldsymbol{w} + \partial_x(\mathbb{f}_{1\mathrm{D}}(\boldsymbol{w})) = 0$ with Riemann data $\boldsymbol{w}_L := (h_L, \boldsymbol{n}\cdot\mathbf{Q}_L)$ and $\boldsymbol{w}_R := (h_R, \boldsymbol{n}\cdot\mathbf{Q}_R)$ and the restricted flux $\mathbb{f}_{1\mathrm{D}}(\boldsymbol{w}) := (q, \frac{1}{h}q^2 + \frac{g}{2}h^2)^{\mathsf{T}}$, where $\boldsymbol{w} := (\mathsf{h}, q)^{\mathsf{T}}$. It is shown in [3, Props. 3.7 and 3.8] that if the bottom is flat and if there if no friction, then the scheme (3.5)–(3.6) with the above viscosities preserves all the convex invariant domains of the PDE and satisfies discrete entropy inequalities for every admissible entropy pair.

4.2. Maximum wave speed. For completeness, we now give an explicit upper bound on the maximum wave speed. Let $\lambda_1^-(\mathsf{h}_*)$ and $\lambda_2^+(\mathsf{h}_*)$ be defined as follows:

$$\lambda_1^-(\mathsf{h}_*) = u_L - \sqrt{g\mathsf{h}_L}\sqrt{\left(1+\left(\frac{\mathsf{h}_*-\mathsf{h}_L}{2\mathsf{h}_L}\right)_+\right)\left(1+\left(\frac{\mathsf{h}_*-\mathsf{h}_L}{\mathsf{h}_L}\right)_+\right)}, \tag{4.3}$$

$$\lambda_2^+(\mathsf{h}_*) = u_R + \sqrt{g\mathsf{h}_R}\sqrt{\left(1+\left(\frac{\mathsf{h}_*-\mathsf{h}_R}{2\mathsf{h}_R}\right)_+\right)\left(1+\left(\frac{\mathsf{h}_*-\mathsf{h}_R}{\mathsf{h}_R}\right)_+\right)}. \tag{4.4}$$

Let $f(\mathsf{h}) := f_L(\mathsf{h},\mathsf{h}_L) + f_R(\mathsf{h},\mathsf{h}_R) + u_R - u_L$, where for $Z \in \{L,R\}$ we have defined

$$f_Z(\mathsf{h},\mathsf{h}_Z) = \begin{cases} 2(\sqrt{g\mathsf{h}} - \sqrt{g\mathsf{h}_Z}) & \text{if} \quad \mathsf{h} \le \mathsf{h}_Z, \\ (\mathsf{h}-\mathsf{h}_Z)\sqrt{\frac{g(\mathsf{h}+\mathsf{h}_Z)}{2\mathsf{h}\mathsf{h}_Z}} & \text{if} \quad \mathsf{h} > \mathsf{h}_Z. \end{cases} \tag{4.5}$$

Let $\mathsf{h}_{\min} = \min(\mathsf{h}_L,\mathsf{h}_R)$, $\mathsf{h}_{\max} = \max(\mathsf{h}_L,\mathsf{h}_R)$, $x_0 = (2\sqrt{2}-1)^2$, and $\overline{\mathsf{h}}_*$ be defined by

$$\overline{\mathsf{h}}_* = \begin{cases} \frac{1}{16g}\big(\max(0, u_L - u_R + 2\sqrt{g\mathsf{h}_L} + 2\sqrt{g\mathsf{h}_R})\big)^2 & \text{if } 0 \le f(x_0\mathsf{h}_{\min}), \\ \sqrt{\mathsf{h}_{\min}\mathsf{h}_{\max}}\left(1 + \frac{\sqrt{2}(u_L-u_R)}{\sqrt{g\mathsf{h}_{\min}}+\sqrt{g\mathsf{h}_{\max}}}\right) & \text{if } f(x_0\mathsf{h}_{\max}) < 0, \\ \mathsf{h}_{\min}\left(-\sqrt{2}+\sqrt{3+2\sqrt{2\frac{\mathsf{h}_{\max}}{\mathsf{h}_{\min}}}+\sqrt{\frac{2}{g\mathsf{h}_{\min}}}(u_L-u_R)}\right)^2 & \text{otherwise.} \end{cases} \tag{4.6}$$

Note that to avoid division by zero in the third case, $f(x_0\mathsf{h}_{\min}) < 0 \le f(x_0\mathsf{h}_{\max})$, it is better to use the expression

$$\overline{\mathsf{h}}_* = \left(-\sqrt{2\mathsf{h}_{\min}} + \sqrt{3\mathsf{h}_{\min} + 2\sqrt{2\mathsf{h}_{\min}\mathsf{h}_{\max}} + \sqrt{2g^{-1}}(u_L-u_R)\sqrt{\mathsf{h}_{\min}}}\right)^2. \tag{4.7}$$

The following result, which gives an upper bound on $\lambda_{\max}^{\mathbb{f}_{1D}}(\boldsymbol{n},\mathbf{U}_L,\mathbf{U}_R)$, is proved in Guermond and Popov [27].

LEMMA 4.1. *Let $\lambda_{\max}^{\mathbb{f}_{1D}}(\boldsymbol{n},\mathbf{U}_L,\mathbf{U}_R)$ be the maximum wave speed in the following one-dimensional Riemann problem $\partial_t \boldsymbol{v} + \partial_x(\mathbb{f}_{1D}(\boldsymbol{v})) = 0$, then $\lambda_{\max}^{\mathbb{f}_{1D}}(\boldsymbol{n},\mathbf{U}_L,\mathbf{U}_R) \le \overline{\lambda}_{\max}^{\mathbb{f}_{1D}}(\boldsymbol{n},\mathbf{U}_L,\mathbf{U}_R) := \max(\lambda_1^-(\overline{\mathsf{h}}_*),\lambda_2^+(\overline{\mathsf{h}}_*))$.*

Remark 4.2 (GMS). Using an artificial viscosity based on a guaranteed maximum wave speed is important for two reasons. First, it makes the method invariant domain preserving. Without a guaranteed maximum wave speed the low-order solution could produce unphysical states (see, e.g., Einfeldt et al. [18]). Second, using a guaranteed maximum wave is important to guarantee that the scheme satisfies discrete entropy inequalities for every admissible entropy pair. We refer the reader to [3, Props. 3.7 and 3.8] and to [25, Thm. 4.7] for precise statements. We stress that being invariant domain preserving is not enough to guarantee convergence to the proper weak solution. A counterexample is given in Guermond and Popov [26, section 3.2] for the Burgers equation. It is shown therein that the so-called "LED (local extremum diminishing) algebraic upwinding" technique, which is a first-order viscosity method that guarantees the maximum principle locally and maintains the total variation of the initial data, converges to a weak solution that is not entropic. The origin of the problem is that the artificial viscosity is not based on a guaranteed maximum wave speed. This problem is often referred to in the literature as "entropy glitch". For

instance the wave speed produced by Roe's average does not give a guaranteed maximum wave speed. Let us finally add that the GMS-based scheme will be used in section 6.4 to extract local bounds which will be enforced on the high-order solution to reduce oscillations and correct unphysical states. ∎

5. Smoothness-based positivity preserving viscosity. In order to make the method (3.5) second-order accurate in space, we now adopt a strategy based on a smoothness indicator first suggested in Jameson, Schmidt, and Turkel [35, Eq. (12)], Jameson [34, p. 1490] (see also Burman [11, Thm. 4.1]). We define the viscosity coefficients

$$\mu_{ij}^{\mathrm{S},n} = \alpha_{ij}^n \mu_{ij}^{\mathrm{V},n}, \qquad d_{ij}^{\mathrm{S},n} = \alpha_{ij}^n d_{ij}^{\mathrm{V},n}, \tag{5.1}$$

where α_{ij}^n is the smoothness indicator. More precisely, let β_{ij} be some positive coefficients yet to be defined. For any $i \in \mathcal{V}$ we introduce the smoothness indicator based on the water height:

$$\alpha_i^n := \frac{|\sum_{j\in\mathcal{I}(i)} \beta_{ij}(\mathsf{H}_j^n - \mathsf{H}_i^n)|}{\sum_{j\in\mathcal{I}(S_i)} \beta_{ij}|\mathsf{H}_j^n - \mathsf{H}_i^n|}. \tag{5.2}$$

The purpose of the parameters β_{ij} is to make the method linearity preserving; that is, $\alpha_i^n = 0$ when the water height is linear on the support of the shape function φ_i. The linearity-preserving property implies that the numerator of (5.2) behaves like $h^2\|D^2\mathsf{h}(\boldsymbol{\xi},t^n)\|_{\ell^2(\mathbb{R}^{d\times d})}$ at some point $\boldsymbol{\xi}$, whereas the denominator behaves like $h\|\nabla\mathsf{h}(\boldsymbol{\zeta},t^n)\|_{\ell^2(\mathbb{R}^d)}$ at some point $\boldsymbol{\zeta}$. In these conditions, assuming that h is not maximum or minimum in the neighborhood of $\boldsymbol{\zeta}$, the smoothness indicator α_i^n behaves like $h\|D^2\mathsf{h}(\boldsymbol{\xi},t^n)\|_{\ell^2(\mathbb{R}^{d\times d})}/\|\nabla\mathsf{h}(\boldsymbol{\zeta},t^n)\|_{\ell^2(\mathbb{R}^d)}$, that is to say, α_i^n is of the order of the nondimensional ratio $h/\operatorname{diam}(D)$ in smooth regions and away from the extrema of the water height. We now define

$$\alpha_{ij}^n = \max(\psi(\alpha_i^n), \psi(\alpha_j^n)), \tag{5.3}$$

where $\psi : [0,1] \to [0,1]$ can be any Lipschitz function such that $\psi(1) = 1$. In the numerical simulations reported at the end of the paper we use

$$\psi(\alpha) = \left(\frac{(\alpha-\alpha_0)_+}{1-\alpha_0}\right)^p, \quad \alpha_0 \in [0,1), \tag{5.4}$$

with $p = 2$ and $\alpha_0 = 0$. Choosing $\alpha_0 \in (0,1)$ makes the viscosity zero in regions where the water height is very smooth and is not at an extremum. The linearity-preserving property implies that, away from extrema, $\mu_{ij}^n = \frac{\mathcal{O}(h)}{\operatorname{diam}(D)}\mu_{ij}^{\mathrm{V},n}$, $d_{ij}^n = \frac{\mathcal{O}(h)}{\operatorname{diam}(D)}d_{ij}^{\mathrm{V},n}$, which makes the method formally second-order accurate in space. The algorithm (3.5) with (5.1)–(5.2)–(5.3)–(5.4) is referred to as the α^2-method in the remainder of the paper.

THEOREM 5.1. *Let $\psi : [0,1] \to [0,1]$ be any Lipschitz function such that $\psi(1) = 1$. Up to the appropriate CFL condition depending on the Lipschitz constant of ψ (see Proposition* 3.7 *and* [3, Prop. 4.4]*), the algorithm* (3.5) *with* (5.1)–(5.2)–(5.3)–(5.4) *is positivity preserving.*

Proof. Since the mass conservation equation is the equation that controls the positivity property of the water height, and the friction term modifies the momentum equations but does not affect the mass conservation equation, the proof is exactly the same as in Guermond and Popov [26, Thm. 4.4]. ∎

Remark 5.2 (linearity preserving). There are many ways to define the weights β_{ij}, $j \in \mathcal{I}(i)$. One option consists of using generalized barycentric coordinates as discussed in [26, section 4.3] and [3, Rem. 4.5]; one can use, for instance, the so-called mean value coordinates; see, e.g., Floater [20, Eq. (5.1)]. A second option consists of reconstructing gradients as proposed in Kuzmin, Basting, and Shadid [40, Eq. (33)]; the method achieves linearity preservation with a $\mathbb{P}_1$ approximation. A third option, prefigured in Burman [11, Thm. 4.1], using the jump and the average of the gradient across the mesh interfaces is proposed in Badia and Bonilla [4, Eq. (7)]. Finally, a fourth option is given in Barrenechea, Burman, and Karakatsani [5, Eq. (2.16)] where local optimization problems are solved to estimate β_{ij}, $j \in \mathcal{I}(i)$. ∎

Remark 5.3 (other viscosities). Other possibilities to define α_{ij}^n that are not based on smoothness but also guarantee positivity of the water height are described in [26, sections 4.4 and 4.5]. ∎

6. Entropy viscosity. We present in this section a variation of the method that is higher-order accurate in space when combined with quadratic and higher-order finite elements. We define μ_{ij}^n and d_{ij}^n via the entropy viscosity technique introduced by Guermond, Pasquetti, and Popov [28] and reduce the dispersive errors induced by the lumped mass matrix by making use of the consistent mass matrix. These two modifications introduce violations in the positivity. We correct those violations via the flux corrected transport method first introduced by Boris and Book [9] and then generalized by Zalesak [54].

6.1. Commutator-based entropy viscosity. The idea behind the entropy viscosity (EV) method is to introduce a nonlinear artificial viscosity based on an estimate of the entropy production. Given a hyperbolic system of the form $\partial_t \boldsymbol{u} + \nabla\cdot(\mathbb{f}(\boldsymbol{u})) = 0$, $\boldsymbol{u} \in \mathbb{R}^m$, with an entropy pair $(\eta, \boldsymbol{F})$, it is proposed in Guermond and Popov [26] to estimate the entropy production at time t_n as follows: (i) First compute the (inviscid) Galerkin solution u_h^{Gal} obtained by solving

$$\int_D \left(\frac{\boldsymbol{u}_h^{\mathrm{Gal}} - \boldsymbol{u}_h^n}{\tau} + \nabla\cdot(\mathbb{f}(\boldsymbol{u}_h^n)) \right) \varphi_i \,\mathrm{d}x = 0 \qquad \forall i \in \mathcal{V} \tag{6.1}$$

(note that a linear system involving the consistent mass matrix has to be solved here). (ii) Then, upon setting $\eta_i^{\max,n} := \max_{j\in\mathcal{I}(S_i)} \eta(\mathbf{U}_j^n)$, $\eta_i^{\min,n} := \min_{j\in\mathcal{I}(S_i)} \eta(\mathbf{U}_j^n)$, $\epsilon_i = \epsilon \max_{j\in\mathcal{I}(S_i)} |\eta(\mathbf{U}_j^n)|$, and $\Delta\eta_i^n = \max(\frac{1}{2}(\eta_i^{\max,n} - \eta_i^{\min,n}), \epsilon_i)$, the entropy residual is defined by

$$R_i^n := \frac{1}{\Delta\eta_i^n} \int_D \nabla\eta(\boldsymbol{u}_h^n)\cdot \left(\frac{\boldsymbol{u}_h^{\mathrm{Gal}} - \boldsymbol{u}_h^n}{\tau} + \nabla\mathbb{f}(\boldsymbol{u}_h^n){:}\nabla\boldsymbol{u}_h^n \right) \phi_i \,\mathrm{d}x \tag{6.2}$$

with the convention $\nabla\mathbb{f}(\boldsymbol{v}){:}\nabla\boldsymbol{w} = \sum_{j=1}^m \sum_{k=1}^d \partial_{v_j} f_{ik}(\boldsymbol{v}) \partial_{x_k} w_j(\boldsymbol{x})$ for all $i \in \{1{:}m\}$. (iii) Finally, the viscosity is defined by setting $\mu_{ij}^{\mathrm{EV},n} := \min\left(\mu_{ij}^{\mathrm{V},n}, \max(|R_i^n|, |R_j^n|)\right)$, $d_{ij}^{\mathrm{EV},n} := \min\left(d_{ij}^{\mathrm{V},n}, \max(|R_i^n|, |R_j^n|)\right)$. This methodology is shown in [26] to work very well, but its main inconvenience is that it requires us to compute the function $\boldsymbol{u}_h^{\mathrm{Gal}}$ by solving the mass matrix problem (6.1).

We now propose a route somewhat equivalent to the one above that circumvents the mass matrix problem. We first observe that instead of solving (6.1), one can use

$$m_i \frac{\mathbf{U}_i^{\mathrm{Gal}} - \mathbf{U}_i^n}{\tau} + \sum_{j\in\mathcal{I}(i)} \mathbb{f}(\mathbf{U}_j^n)\cdot\boldsymbol{c}_{ij} = 0. \tag{6.3}$$

Then upon observing that $\nabla\eta(\mathbf{U}_i^n)\cdot(m_i\frac{\mathbf{U}_i^{\mathrm{Gal}}-\mathbf{U}_i^n}{\tau}) = -\sum_{j\in\mathcal{I}(i)}\nabla\eta(\mathbf{U}_i^n)\cdot(\mathbb{f}(\mathbf{U}_j^n)\cdot\boldsymbol{c}_{ij})$ and $\nabla\eta(\boldsymbol{u}^n)\cdot(\nabla\mathbb{f}(\boldsymbol{u}^n):\nabla\boldsymbol{u}^n) = \nabla\cdot\boldsymbol{F}(\boldsymbol{u}^n)$, we can estimate the entropy residual by setting

$$R_i^n := \frac{1}{\Delta\eta_i^n}\sum_{j\in\mathcal{I}(i)}\big(\boldsymbol{F}(\mathbf{U}_j^n) - \nabla\eta(\mathbf{U}_i^n)\cdot\mathbb{f}(\mathbf{U}_j^n)\big)\cdot\boldsymbol{c}_{ij}. \tag{6.4}$$

Note that $R_i^n = 0$ if η is a linear form. Note also that, contrary to other variants of the EV method presented in previous works [28, 29], the above definition is invariant by any scaling and translation on the entropy; that is, replacing $\eta(\boldsymbol{v})$ by $\lambda_i\eta(\boldsymbol{v})+\mu_i$ for any $\lambda_i\in\mathbb{R}\backslash\{0\}$, $\mu_i\in\mathbb{R}$, does not change R_i^n. Moreover, using the property $\sum_{j\in\mathcal{I}(i)}\boldsymbol{c}_{ij}=0$, and using the Frechet differential notation, we have

$$\begin{aligned}
R_i^n &= \frac{1}{\Delta\eta_i^n}\sum_{j\in\mathcal{I}(i)}\big(\boldsymbol{F}(\mathbf{U}_j^n) - \boldsymbol{F}(\mathbf{U}_i^n) - \nabla\eta(\mathbf{U}_i^n)\cdot(\mathbb{f}(\mathbf{U}_j^n)-\mathbb{f}(\mathbf{U}_i^n))\big)\cdot\boldsymbol{c}_{ij}\\
&= \frac{1}{\Delta\eta_i^n}\sum_{j\in\mathcal{I}(i)}\Big(D\boldsymbol{F}(\mathbf{U}_i^n)(\mathbf{U}_j^n-\mathbf{U}_i^n) + \tfrac{1}{2}D^2\boldsymbol{F}(\widetilde{\mathbf{U}}_{ij})(\mathbf{U}_j^n-\mathbf{U}_i^n)^2\\
&\qquad - \nabla\eta(\mathbf{U}_i^n)\cdot(D\mathbb{f}(\mathbf{U}_i^n)(\mathbf{U}_j^n-\mathbf{U}_i^n)) - \tfrac{1}{2}\nabla\eta(\mathbf{U}_i^n)\cdot\big(D^2\mathbb{f}(\widehat{\mathbf{U}}_{ij})(\mathbf{U}_j^n-\mathbf{U}_i^n)^2\big)\Big)\cdot\boldsymbol{c}_{ij}
\end{aligned} \tag{6.5}$$

for some $\widetilde{\mathbf{U}}_{ij},\widehat{\mathbf{U}}_{ij}$ in the convex hull of $\{\mathbf{U}_j \mid j\in\mathcal{I}(i)\}$. Using that $D\boldsymbol{F}(\boldsymbol{v}) = \nabla\eta(\boldsymbol{v})\cdot D\mathbb{f}(\boldsymbol{v})$, we obtain

$$|R_i^n| \le \frac{1}{2\Delta\eta_i^n}(\|D^2\eta\|\|D\mathbb{f}\| + 2\|\nabla\eta\|\|D^2\mathbb{f}\|)\sum_{j\in\mathcal{I}(i)}\|\mathbf{U}_j^n-\mathbf{U}_i^n\|_{\ell^2(\mathbb{R}^m)}^2\|\boldsymbol{c}_{ij}\|_{\ell^2(\mathbb{R}^d)}, \tag{6.6}$$

where $\|D^2\mathbb{f}\| := \sup_{\boldsymbol{v}\in\mathcal{C}}\sup_{0\neq\boldsymbol{w}_1,\boldsymbol{w}_2\in\mathbb{R}^m}\frac{\|D^2F(\boldsymbol{v})(\boldsymbol{w}_1,\boldsymbol{w}_2)\|_{\ell^2(\mathbb{R}^{m\times d})}}{\|\boldsymbol{w}_1\|_{\ell^2(\mathbb{R}^m)}\|\boldsymbol{w}_2\|_{\ell^2(\mathbb{R}^m)}}$ and we have adopted a similar definition for $\|\nabla\eta\|$. In conclusion, the entropy residual $|R_i^n|$ behaves like $\lambda\|\boldsymbol{c}_{ij}\|_{\ell^2(\mathbb{R}^d)}\frac{\|D^2\eta\|}{\|D\eta\|}\|\nabla\mathbf{U}^n\|_{\ell^2(\mathbb{R}^{m\times d})}h_i$ for any $j\in\mathcal{I}(i)$, where $\lambda := \|D\mathbb{f}\|$ is a local wave speed and h_i is the diameter of D_i. Hence, if $\boldsymbol{u}_h^n$ is smooth over D_i, the entropy residual is (at least) one order smaller than $d_{ij}^{\mathrm{V},n}$. Actually, when $\widehat{P}$ is composed of polynomials of degree k, $|R_i^n|$ behaves like $\lambda\|\boldsymbol{c}_{ij}\|_{\ell^2(\mathbb{R}^d)}\frac{\|D^2\eta\|}{\|D\eta\|}\|D^k\mathbf{U}^n\|_{\ell^2(\mathbb{R}^{m\times d^k})}h_i^k$.

We now apply the above idea to the system (2.1). The shallow water equations without friction admit an entropy pair

$$\eta(\boldsymbol{u}) = g(\tfrac{1}{2}h^2+hz)+\tfrac{1}{2}h\|\boldsymbol{v}\|_{\ell^2}^2,\quad F(\boldsymbol{u}) = (\tfrac{1}{2}h\|\boldsymbol{v}\|_{\ell^2}^2 + g(h^2+hz))\boldsymbol{v}; \tag{6.7}$$

see, e.g., Audusse et al. [1, Eq. (1.3)], Bouchut [10, Eq. (3.15)], LeVeque and George [41], Toro [51]. Then using the definition (2.2) for the flux $\mathbb{f}$, a first way to estimate the entropy residual consists of setting

$$|R_i^n| := \frac{1}{\Delta\eta_i^n}\sum_{j\in\mathcal{I}(i)}\big(\boldsymbol{F}(\mathbf{U}_j^n) - \nabla\eta(\mathbf{U}_i^n)\cdot\mathbb{f}(\mathbf{U}_j^n)\big)\cdot\boldsymbol{c}_{ij} - g\mathsf{H}_i^n\mathsf{Z}_j\mathbf{V}_j^n\cdot\boldsymbol{c}_{ij}. \tag{6.8}$$

Since we are just estimating commutators, a second way to define an entropy residual consists of using the entropy pair for the shallow water equation without friction and flat bottom:

$$\eta^{\mathrm{flat}}(\boldsymbol{u}) = g\tfrac{1}{2}h^2+\tfrac{1}{2}h\|\boldsymbol{v}\|_{\ell^2}^2,\quad \boldsymbol{F}^{\mathrm{flat}}(\boldsymbol{u}) = (\tfrac{1}{2}h\|\boldsymbol{v}\|_{\ell^2}^2+gh^2)\boldsymbol{v}, \tag{6.9}$$

and setting

$$|R_i^n| := \frac{1}{2\Delta\eta_i^{\mathrm{flat},n}} \sum_{j\in\mathcal{I}(i)} \Big(\boldsymbol{F}^{\mathrm{flat}}(\mathbf{U}_j^n) - \nabla\eta^{\mathrm{flat}}(\mathbf{U}_i^n)\cdot\mathbb{f}(\mathbf{U}_j^n) \Big)\cdot\boldsymbol{c}_{ij}. \tag{6.10}$$

Then the numerical EVs are defined as follows:

$$\mu_{ij}^{\mathrm{EV},n} := \min\Big(\mu_{ij}^{\mathrm{V},n}, \max(|R_i^n|,|R_j^n|)\Big), \tag{6.11}$$

$$d_{ij}^{\mathrm{EV},n} := \min\Big(d_{ij}^{\mathrm{V},n}, \max(|R_i^n|,|R_j^n|)\Big), \tag{6.12}$$

where $\mu_{ij}^{\mathrm{V},n}$ and $d_{ij}^{\mathrm{V},n}$ are defined in (4.2) and (4.2), respectively. We use the second definition (6.10) in the numerical tests reported at the end of the paper.

Remark 6.1 (loss of positivity). The above definition of $\mu_{ij}^{\mathrm{EV},n}$ and $d_{ij}^{\mathrm{EV},n}$ no longer guarantees positivity of the water height when used in the scheme (3.5). We remedy this situation in section 6.3. ∎

6.2. Dispersion correction. Recall that the method (3.5) uses the lumped mass matrix. This is not fully satisfactory, at least for piecewise linear approximation, since it is well known that lumping the mass matrix induces dispersion errors that have adverse effects when solving problems with nonsmooth initial data. We refer the reader to Christon, Martinez, and Voth [14], Gresho and Sani [23], Guermond and Pasquetti [24], Thompson [50] where this question is thoroughly investigated. As shown in the above references, one can remove the dispersion error by using the consistent mass matrix. The beneficial effects of the consistent mass matrix are particularly visible when working with nonsmooth solutions. In the rest of the paper, the algorithm that we refer to as the EV method consists of using the viscosities (6.11)–(6.12) and computing the update $\boldsymbol{u}_h^{n+1}$ by solving the following mass matrix problem

$$\begin{aligned}\frac{(\mathcal{M}(\mathbf{U}^{n+1}-\mathbf{U}^n))_i}{\tau} = \sum_{j\in\mathcal{I}(i)} &-\big(\mathsf{H}_j^n\mathbf{V}_j^n\cdot\boldsymbol{c}_{ij}, \mathbf{V}_j^n\mathbf{Q}_j^n\cdot\boldsymbol{c}_{ij} + g\mathsf{H}_i^n(\mathsf{H}_j^n+Z_j)\boldsymbol{c}_{ij}\big)^{\mathsf{T}} \\ &- \left(0, \frac{2g\mathrm{n}^2\mathbf{Q}_i^n\|\mathbf{V}_i\|_{\ell^2}m_i}{(\mathsf{H}_i^n)^\gamma + \max((\mathsf{H}_i^n)^\gamma, 2g\mathrm{n}^2\tau\|\mathbf{V}_i\|_{\ell^2})}\right)^{\mathsf{T}} \\ + \sum_{j\in\mathcal{I}(i)\setminus\{i\}} &(d_{ij}^{\mathrm{EV},n} - \mu_{ij}^{\mathrm{EV},n})\Big(\mathbf{U}_j^{*,i,n} - \mathbf{U}_i^{*,j,n}\Big) + \mu_{ij}^{\mathrm{EV},n}(\mathbf{U}_j^n - \mathbf{U}_i^n).\end{aligned} \tag{6.13}$$

Remark 6.2 (loss of positivity). The presence of the consistent mass matrix makes the above scheme non-positivity-preserving irrespective of the definition of the viscosities μ_{ij}^n and d_{ij}^n. More precisely, it is proved in Guermond, Popov, and Yang [30] that the continuous finite element method based on artificial viscosity in space and explicit time stepping cannot satisfy the maximum principle when using the consistent mass matrix. Again, we remedy this problem in section 6.3. ∎

Remark 6.3 (alternative to the consistent mass matrix). An alternative strategy to correct the effects of lumping the mass matrix consists of replacing the inverse of $\mathcal{M}$ by a Neumann series. It is shown in Guermond and Pasquetti [24] that one correction exactly corrects the dispersion errors for $\mathbb{P}_1$ elements; that is, one can legitimately replace $\mathcal{M}^{-1}$ by its approximation $(\mathcal{M}^L)^{-1}(\mathcal{I} + (\mathcal{M}^L - \mathcal{M})(\mathcal{M}^L)^{-1})$. ∎

6.3. Flux corrected transport. An easy way to correct the possible loss of positivity on the water height induced by the use of the EVs (6.11)–(6.12) and the use of the consistent mass matrix consists of invoking the flux corrected transport (FCT) methodology of Boris and Book [9] and Zalesak [54]. For completeness, and without claiming originality, we now explain how we deploy the FCT technique in the context of the time stepping method described in (6.13). We essentially paraphrase Kuzmin, Löhner, and Turek [39, section 6.1].

6.3.1. The method. We start by renaming the updates given by (3.5) and (6.13). We set $\mathbf{U}^{\mathrm{L}} := \mathbf{U}^{n+1}$, where $\mathbf{U}^{n+1}$ is the (L)ow-order solution defined in (3.5), and we rename the corresponding viscosities as follows: $\mu_{ij}^{\mathrm{L},n} := \alpha_{ij}^n \mu_{ij}^{\mathrm{V},n}$ and $d_{ij}^{\mathrm{L},n} := \alpha_{ij}^n d_{ij}^{\mathrm{V},n}$. The viscosities could be defined with $\alpha_{ij}^n = 1$ (the low-order solution is first-order accurate in space), or it could be the α^2-viscosity defined in section 5 (the low-order solution is then second-order accurate in space); the property that matters for the time being is that we are guaranteed that the water height given by $\mathbf{U}^{\mathrm{L}}$ is positive. We set $\mathbf{U}^{\mathrm{H}} := \mathbf{U}^{n+1}$, where $\mathbf{U}^{n+1}$ is the (H)igh-order solution defined in (6.13), and we rename the EVs defined in (6.11)–(6.12) by setting $\mu_{ij}^{\mathrm{H},n} = \mu_{ij}^{\mathrm{EV},n}$ and $d_{ij}^{\mathrm{H},n} = d_{ij}^{\mathrm{EV},n}$. Then, subtracting (3.5) from (6.13) and denoting $\delta\mathbf{U} := \mathbf{U}^{\mathrm{H}} - \mathbf{U}^n$, we obtain the following identity:

$$
\begin{aligned}
(6.14)\quad m_i(\mathbf{U}_i^{\mathrm{H}} - \mathbf{U}_i^{\mathrm{L}}) &= \sum_{j\in\mathcal{I}(i)} (\mathcal{M}^L - \mathcal{M})_{ij}\delta\mathbf{U}_j \\
&+ \tau \sum_{j\in\mathcal{I}(i)\setminus\{i\}} (d_{ij}^{\mathrm{H},n} - d_{ij}^{\mathrm{L},n} - \mu_{ij}^{\mathrm{H},n} + \mu_{ij}^{\mathrm{L},n})(\mathbf{U}_j^{*,i,n} - \mathbf{U}_i^{*,j,n}) + (\mu_{ij}^{\mathrm{H},n} - \mu_{ij}^{\mathrm{L},n})(\mathbf{U}_j^n - \mathbf{U}_i^n),
\end{aligned}
$$

where recall that $\mathcal{M}^L$ is the lumped mass matrix. Note that by definition we have $\sum_{j\in\mathcal{I}(i)}(\mathcal{M}^L - \mathcal{M})_{ij} = 0$, which in turn implies that $\sum_{j\in\mathcal{I}(i)}(\mathcal{M}^L - \mathcal{M})_{ij}\delta\mathbf{U}_j = \sum_{j\in\mathcal{I}(i)}(\mathcal{M}^L - \mathcal{M})_{ij}(\delta\mathbf{U}_j - \delta\mathbf{U}_i)$. Therefore, the above identity can be rewritten in the following alternative form:

$$
\begin{aligned}
(6.15)\quad m_i(\mathbf{U}_i^{\mathrm{H}} - \mathbf{U}_i^{\mathrm{L}}) &= \sum_{j\in\mathcal{I}(i)} (\mathcal{M}^L - \mathcal{M})_{ij}(\delta\mathbf{U}_j - \delta\mathbf{U}_i) \\
&+ \tau \sum_{j\in\mathcal{I}(i)\setminus\{i\}} (d_{ij}^{\mathrm{H},n} - d_{ij}^{\mathrm{L},n} - \mu_{ij}^{\mathrm{H},n} + \mu_{ij}^{\mathrm{L},n})(\mathbf{U}_j^{*,i,n} - \mathbf{U}_i^{*,j,n}) + (\mu_{ij}^{\mathrm{H},n} - \mu_{ij}^{\mathrm{L},n})(\mathbf{U}_j^n - \mathbf{U}_i^n).
\end{aligned}
$$

Upon denoting

$$
\begin{aligned}
(6.16)\quad \mathbf{A}_{ij} := (\mathcal{M}^L - \mathcal{M})_{ij}(\delta\mathbf{U}_j - \delta\mathbf{U}_i) &+ \tau(\mu_{ij}^{\mathrm{H},n} - \mu_{ij}^{\mathrm{L},n})(\mathbf{U}_j^n - \mathbf{U}_i^n) \\
&+ \tau(d_{ij}^{\mathrm{H},n} - d_{ij}^{\mathrm{L},n} - \mu_{ij}^{\mathrm{H},n} + \mu_{ij}^{\mathrm{L},n})(\mathbf{U}_j^{*,i,n} - \mathbf{U}_i^{*,j,n}),
\end{aligned}
$$

we finally obtain the following expression that relates $\mathbf{U}_i^{\mathrm{H}}$ and $\mathbf{U}_i^{\mathrm{L}}$:

$$
(6.17)\qquad m_i(\mathbf{U}_i^{\mathrm{H}} - \mathbf{U}_i^{\mathrm{L}}) = \sum_{j\in\mathcal{I}(i)} \mathbf{A}_{ij}.
$$

Note that the coefficients $\mathbf{A}_{ij}$ are skew symmetric, i.e., $\mathbf{A}_{ij} = -\mathbf{A}_{ji}$; this is exactly the structure that is needed to apply the FCT limiting. In particular this property implies mass conservation: $\sum_{i\in\mathcal{V}} m_i\mathbf{U}_i^{\mathrm{H}} = \sum_{i\in\mathcal{V}} m_i\mathbf{U}_i^{\mathrm{L}}$. The EV solution is enforced

to have positive water height by introducing limiters $\ell_{ij} \in [0,1]$, $\ell_{ij} = \ell_{ji}$, as follows:

$$m_i(\mathbf{U}_i^{n+1} - \mathbf{U}_i^{\mathrm{L}}) = \sum_{j\in\mathcal{I}(i)} \ell_{ij}\mathbf{A}_{ij}. \tag{6.18}$$

Note that the symmetry property of the limiters $\ell_{ij} = \ell_{ji}$ again ensures mass conservation: $\sum_{i\in\mathcal{V}} m_i\mathbf{U}_i^{n+1} = \sum_{i\in\mathcal{V}} m_i\mathbf{U}_i^{\mathrm{L}}$.

PROPOSITION 6.4 (conservation). *Assume that $\ell_{ij} = \ell_{ji}$, then the algorithm* (6.18) *is conservative, that is to say, $\sum_{i\in\mathcal{V}} m_i\mathbf{U}_i^{n+1} = \sum_{i\in\mathcal{V}} m_i\mathbf{U}_i^{\mathrm{L}}$.*

Remark 6.5 (limiters). Note that in (6.18) the limiters are applied on all the fields, i.e., they are applied both on the water height and on the discharge. ∎

6.4. Limiting with exact bounds. We proceed as in Guermond et al. [31] to extract exact local lower bounds on the water heights and, more generally, exact limiting bounds on the solution. The key idea consists in separating the hyperbolic part of the PDE and the source term and to limit the hyperbolic part.

Let $i \in \mathcal{V}$ and let us set

$$\mathbf{S}_i^n := \left(0, \frac{-2g\mathrm{n}^2\mathbf{Q}_i^n\|\mathbf{V}_i\|_{\ell^2}m_i}{(\mathsf{H}_i^n)^\gamma + \max((\mathsf{H}_i^n)^\gamma, 2g\mathrm{n}^2\tau\|\mathbf{V}_i\|_{\ell^2})} + \sum_{j\in\mathcal{I}(i)} g\left(-\mathsf{H}_i^n Z_j + \tfrac{1}{2}(\mathsf{H}_j - \mathsf{H}_i)^2\right)\boldsymbol{c}_{ij}\right)^{\mathsf{T}}, \tag{6.19}$$

then using $\sum_{j\in\mathcal{I}(i)\setminus\{i\}} -g\boldsymbol{c}_{ij}\mathsf{H}_j\mathsf{H}_i = \sum_{j\in\mathcal{I}(i)\setminus\{i\}} \frac{1}{2}g\boldsymbol{c}_{ij}(\mathsf{H}_j - \mathsf{H}_i)^2 - \sum_{j\in\mathcal{I}(i)\setminus\{i\}} \frac{1}{2}g\mathsf{H}_j^2$, (recall that $\sum_{j\in\mathcal{I}(i)\setminus\{i\}} \boldsymbol{c}_{ij} = \mathbf{0}$), the scheme (3.5) can be rewritten as follows:

$$\begin{aligned}\frac{m_i}{\tau}(\mathbf{U}_i^{n+1} - \mathbf{U}_i^n) = \mathbf{S}_i^n &+ \sum_{j\in\mathcal{I}(i)\setminus\{i\}} -\boldsymbol{c}_{ij}\cdot(\mathbb{f}(\mathbf{U}_j^n) - \mathbb{f}(\mathbf{U}_i^n)) \\ &+ \sum_{j\in\mathcal{I}(i)\setminus\{i\}} (d_{ij}^n - \mu_{ij}^n)\left(\mathbf{U}_j^{*,i,n} - \mathbf{U}_i^{*,j,n}\right) + \mu_{ij}^n(\mathbf{U}_j^n - \mathbf{U}_i^n).\end{aligned} \tag{6.20}$$

We now introduce the following auxiliary states for all $j \in \mathcal{I}(i)$:

$$\overline{\mathbf{U}_{ij}^n} = -\frac{\boldsymbol{c}_{ij}}{2d_{ij}^{\mathrm{L},n}}\cdot(\mathbb{f}(\mathbf{U}_j^n) - \mathbb{f}(\mathbf{U}_i^n)) + \frac{1}{2}(\mathbf{U}_j^n + \mathbf{U}_i^n), \tag{6.21}$$

$$\widetilde{\mathbf{U}_{ij}^n} = \frac{d_{ij}^{\mathrm{L},n} - \mu_{ij}^{\mathrm{L},n}}{2d_{ij}^{\mathrm{L},n}}(\mathbf{U}_j^{*,i,n} - \mathbf{U}_j^n - (\mathbf{U}_i^{*,j,n} - \mathbf{U}_i^n)). \tag{6.22}$$

Let $\mathbf{U}_i^{\mathrm{L},n+1}$ be the low-order update corresponding to taking $d_{ij}^n = d_{ij}^{\mathrm{L},n}$ and $\mu_{ij}^n = \mu_{ij}^{\mathrm{L},n}$ in the above equation. The following result will enable us to derive exact bounds.

LEMMA 6.6. *Let $\mathbf{W}_i^{\mathrm{L},n+1} := \mathbf{U}_i^{\mathrm{L},n+1} - \frac{\tau}{m_i}\mathbf{S}_i^n$, then, if $1 - \frac{2\tau}{m_i}\sum_{j\in\mathcal{I}(i)\setminus\{i\}} d_{ij}^{\mathrm{L},n} \geq 0$, the following convex combination holds true:*

$$\mathbf{W}_i^{\mathrm{L},n+1} = \mathbf{U}_i^n\Big(1 - \frac{\tau}{m_i}\sum_{j\in\mathcal{I}(i)\setminus\{i\}} 2d_{ij}^{\mathrm{L},n}\Big) + \frac{\tau}{m_i}\sum_{j\in\mathcal{I}(i)\setminus\{i\}} 2d_{ij}^{\mathrm{L},n}\left(\overline{\mathbf{U}_{ij}^n} + \widetilde{\mathbf{U}_{ij}^n}\right). \tag{6.23}$$

Furthermore we have $\overline{\mathsf{H}_{ij}^n} + \widetilde{\mathsf{H}_{ij}^n} \geq 0$ for all $j \in \mathcal{I}(i)$.

Proof. We rewrite (6.20)

$$\begin{aligned}(6.24)\quad \mathbf{W}_i^{\mathrm{L},n+1} &= \mathbf{U}_i^n + \frac{\tau}{m_i}\sum_{j\in\mathcal{I}(i)\backslash\{i\}} -\boldsymbol{c}_{ij}\cdot(\mathbb{f}(\mathbf{U}_j^n)-\mathbb{f}(\mathbf{U}_i^n)) + d_{ij}^{\mathrm{L},n}(\mathbf{U}_j^n-\mathbf{U}_i^n)\\ &+\frac{\tau}{m_i}\sum_{j\in\mathcal{I}(i)\backslash\{i\}}(d_{ij}^{\mathrm{L},n}-\mu_{ij}^{\mathrm{L},n})\left(\mathbf{U}_j^{*,i,n}-\mathbf{U}_i^{*,j,n}\right)+\mu_{ij}^{\mathrm{L},n}(\mathbf{U}_j^n-\mathbf{U}_i^n)-d_{ij}^{\mathrm{L},n}(\mathbf{U}_j^n-\mathbf{U}_i^n)\\ &= \mathbf{U}_i^n + \frac{\tau}{m_i}\sum_{j\in\mathcal{I}(i)\backslash\{i\}} -\boldsymbol{c}_{ij}\cdot(\mathbb{f}(\mathbf{U}_j^n)-\mathbb{f}(\mathbf{U}_i^n)) + d_{ij}^{\mathrm{L},n}(\mathbf{U}_j^n+\mathbf{U}_i^n)\\ &-\frac{\tau}{m_i}\sum_{j\in\mathcal{I}(i)\backslash\{i\}}2d_{ij}^{\mathrm{L},n}\mathbf{U}_i^n+\frac{\tau}{m_i}\sum_{j\in\mathcal{I}(i)\backslash\{i\}}(d_{ij}^{\mathrm{L},n}-\mu_{ij}^{\mathrm{L},n})\left(\mathbf{U}_j^{*,i,n}-\mathbf{U}_j^n-(\mathbf{U}_i^{*,j,n}-\mathbf{U}_i^n)\right).\end{aligned}$$

Then (6.23) follows naturally. The nonnegativity of $\overline{\mathsf{H}_{ij}^n}+\widetilde{\mathsf{H}_{ij}^n}$ is a consequence of the following series of inequalities:

$$\begin{aligned}(6.25)\quad 2d_{ij}^{\mathrm{L},n}(\overline{\mathsf{H}_{ij}^n}+\widetilde{\mathsf{H}_{ij}^n}) &= -\boldsymbol{c}_{ij}\cdot(\mathbf{V}_j\mathsf{H}_j^n-\mathbf{V}_i\mathsf{H}_i^n)+d_{ij}^{\mathrm{L},n}(\mathsf{H}_j^n+\mathsf{H}_i^n)\\ &\quad+(d_{ij}^{\mathrm{L},n}-\mu_{ij}^{\mathrm{L},n})\left(\mathsf{H}_j^{*,i,n}-\mathsf{H}_j^n-(\mathsf{H}_i^{*,j,n}-\mathsf{H}_i^n)\right)\\ &= -\boldsymbol{c}_{ij}\cdot(\mathbf{V}_j\mathsf{H}_j^n-\mathbf{V}_i\mathsf{H}_i^n)+\mu_{ij}^{\mathrm{L},n}(\mathsf{H}_j^n+\mathsf{H}_i^n)+(d_{ij}^{\mathrm{L},n}-\mu_{ij}^{\mathrm{L}})(\mathsf{H}_j^n+\mathsf{H}_i^n)\\ &\quad+(d_{ij}^{\mathrm{L},n}-\mu_{ij}^{\mathrm{L},n})\left(\mathsf{H}_j^{*,i,n}-\mathsf{H}_j^n-(\mathsf{H}_i^{*,j,n}-\mathsf{H}_i^n)\right)\\ &\quad\geq(d_{ij}^{\mathrm{L},n}-\mu_{ij}^{\mathrm{L},n})\left(\mathsf{H}_j^{*,i,n}+2\mathsf{H}_i^n-\mathsf{H}_i^{*,j,n}\right)\geq 0.\end{aligned}$$

This completes the proof. ∎

Realizing that $\overline{\mathsf{H}_{ii}^n}+\widetilde{\mathsf{H}_{ii}^n}=\mathsf{H}_i^n$, an immediate consequence of the above result is

$$(6.26)\qquad 0\leq \mathsf{H}_i^{\min}:=\min_{j\in\mathcal{I}(i)}\left(\overline{\mathsf{H}_{ij}^n}+\widetilde{\mathsf{H}_{ij}^n}\right)\leq \mathsf{H}_i^{\mathrm{L},n+1}\leq\max_{j\in\mathcal{I}(i)}\left(\overline{\mathsf{H}_{ij}^n}+\widetilde{\mathsf{H}_{ij}^n}\right)=:\mathsf{H}_i^{\max},$$

that is to say, $\mathsf{H}_i^{\min}$ and $\mathsf{H}_i^{\max}$, as defined above, are legitimate lower and upper bounds, which we would like the high-order solution to satisfy. More precisely, the bounds $\mathsf{H}_i^{\min}$, $\mathsf{H}_i^{\max}$ are legitimate in the sense that the set of limiters $\{\ell_{ij}\}_{j\in\mathcal{I}(i)\backslash\{i\}}$ such that the update computed with (6.18) has a water height H_i^{n+1} in the interval $[\mathsf{H}_i^{\min},\mathsf{H}_i^{\max}]$ is not empty; actually, the purpose of the inequalities (6.26) is to show that $\ell_{ij}=0$, $j\in\mathcal{I}(i)\backslash\{i\}$, is in this set. Moreover, since it is our experience that limiting should switch the method to lower order at the shoreline and in dry regions, we introduce the following dry state indicator to detect these regions: $\mathsf{H}_i^{\mathrm{dry}}:=\mathsf{H}_i^{\mathrm{L},n}-\frac{1}{2}(\max_{j\in\mathcal{I}(i)}\mathsf{H}_j^n-\min_{j\in\mathcal{I}(i)}\mathsf{H}_j^n)$. Let $\mathbf{A}_{ij}^{\mathsf{h}}$ be the first component of the column vector $\mathbf{A}_{ij}\in\mathbb{R}^{1+d}$, then we propose to limit the water height as follows:

$$(6.27)\qquad Q_i^-:=m_i(\mathsf{H}_i^{\min}-\mathsf{H}_i^{\mathrm{L},n+1}),\qquad Q_i^+:=m_i(\mathsf{H}_i^{\max}-\mathsf{H}_i^{\mathrm{L},n+1}),$$

$$(6.28)\qquad P_i^-:=\sum_{j\in\mathcal{I}(i)\backslash\{i\}}(\mathbf{A}_{ij}^{\mathsf{h}})_-,\qquad P_i^+:=\sum_{j\in\mathcal{I}(i)\backslash\{i\}}(\mathbf{A}_{ij}^{\mathsf{h}})_+,$$

$$(6.29)\qquad R_i^-:=\begin{cases}0 & \text{if } \mathsf{H}_i^{\mathrm{dry}}\leq 0,\\ 1 & \text{if } P_i=0,\ \mathsf{H}_i^{\mathrm{dry}}>0,\\ \frac{Q_i^-}{P_i^-} & \text{if } P_i\neq 0,\ \mathsf{H}_i^{\mathrm{dry}}>0.\end{cases}\qquad R_i^+:=\begin{cases}0 & \text{if } \mathsf{H}_i^{\mathrm{dry}}\leq 0,\\ 1 & \text{if } P_i=0,\ \mathsf{H}_i^{\mathrm{dry}}>0,\\ \frac{Q_i^+}{P_i^+} & \text{if } P_i\neq 0,\ \mathsf{H}_i^{\mathrm{dry}}>0,\end{cases}$$

$$(6.30)\qquad \ell_{ij}:=\min(R_i^+,R_j^-)\text{ if }\mathbf{A}_{ij}^{\mathsf{h}}\geq 0,\qquad \ell_{ij}:=\min(R_i^-,R_j^+)\text{ if }\mathbf{A}_{ij}^{\mathsf{h}}<0.$$

LEMMA 6.7. *Under the CFL condition* $1-\frac{2\tau}{m_i}\sum_{j\in\mathcal{I}(i)\setminus\{i\}} d_{ij}^{\mathrm{L},n}\geq 0$, *the update* $\mathbf{U}_i^{n+1}$ *given by* (6.18) *with the limiting defined by* (6.27)–(6.28)–(6.29)–(6.30) *satisfies the bounds* $0\leq \mathsf{H}_i^{\min}\leq \mathsf{H}_i^{n+1}\leq \mathsf{H}_i^{\max}$.

Proof. This is a direct consequence of the definitions (6.27)–(6.28)–(6.29)–(6.30). ∎

We finish this section by mentioning that, as in [31], we can limit any convex or concave functional of the solution. As an example, we show how to limit the kinetic energy. At this point we assume that limiting on the water height as described in (6.27)–(6.28)–(6.29) has been done, and we assume that $\mathsf{H}_i^{\mathrm{L},n}>0$. If it is not the case, we set $\ell_{ij}=0$ for all $j\in\mathcal{I}(i)$. Let us set $(\mathsf{H}(\mathbf{W}_i),\mathbf{Q}(\mathbf{W}_i))^\mathsf{T}:=\mathbf{W}_i:=\mathbf{U}_i-\frac{\tau}{m_i}\mathbf{S}_i^n$, where $\mathbf{U}_i$ is any state given by (6.18) when the limiters ℓ_{ij} span $[0,1]$. Let us consider the kinetic energy $\psi(\mathbf{W}):=\frac{1}{2}\frac{1}{\mathsf{H}(\mathbf{W})}\|\mathbf{Q}(\mathbf{W})\|_{\ell^2}^2$. It is well known that the functional $\psi(\mathbf{W})$ is convex; more precisely, we have $D^2\psi(\mathbf{W})(a,\boldsymbol{b})(a,\boldsymbol{b})=\frac{1}{2\mathsf{H}}(a\frac{\mathbf{Q}}{\mathsf{H}}-\boldsymbol{b})^2\geq 0$, for all $a\in\mathbb{R}$, $\boldsymbol{b}\in\mathbb{R}^d$. As a result of the convex combination (6.23), the following holds:

$$\psi(\mathbf{W}_i^{\mathrm{L},n+1})\leq \max_{j\in\mathcal{I}(i)}\psi(\overline{\mathbf{U}_{ij}^n}+\widetilde{\mathbf{U}_{ij}^n})=:\mathsf{K}_i^{\max}. \tag{6.31}$$

We now describe a process to estimate the limiters ℓ_{ij} so that $\psi(\mathbf{W}_i^{n+1})=\psi(\mathbf{U}_i^{n+1}-\frac{\tau}{m_i}\mathbf{S}_i^n)\leq\mathsf{K}_i^{\max}$. Let us set $\Psi(\mathbf{W}):=\mathsf{K}_i^{\max}\mathsf{H}-\mathsf{H}\psi(\mathbf{W})$. It is clear that, provided the water height H is nonzero, $\psi(\mathbf{W})\leq\mathsf{K}_i^{\max}$ if and only if $\Psi(\mathbf{W})\geq 0$. (Note in passing that we have already assumed that $\mathsf{H}_i^{\mathrm{L},n}>0$.) We are going to compute the limiters so that $\Psi(\mathbf{W}_i^{n+1})\geq 0$. We observe first that Ψ is the sum of a quadratic form and a linear form: $\Psi(\mathbf{W})=\mathsf{K}_i^{\max}\mathsf{H}-\frac{1}{2}\|\mathbf{Q}\|_{\ell^2}^2$. Then similarly to [31, section 4.5] we proceed as follows. Let $\{\lambda_j\mid j\in\mathcal{I}(i)\setminus\{i\}\}$ be some positive numbers adding up to 1. For instance one can take $\lambda_j=\frac{m_{ij}}{m_i-m_{ii}}$. In the simulations reported at the end of the paper we take $\lambda_j:=\frac{1}{\mathrm{card}(\mathcal{I}(i))-1}$, $j\in\mathcal{I}(i)\setminus\{i\}$. For all $j\in\mathcal{I}(i)$, $j\neq i$, we set

$$\mathsf{H}_i^{\mathrm{W,L}}:=\mathsf{H}(\mathbf{W}_i^{\mathrm{L},n+1}),\qquad \mathbf{Q}_i^{\mathrm{W,L}}:=\mathbf{Q}(\mathbf{W}_i^{\mathrm{L},n+1}), \tag{6.32}$$

$$\mathbf{P}_{ij}:=(\mathbf{P}_{ij}^{\mathsf{h}},\mathbf{P}_{ij}^{\mathbf{q}})^\mathsf{T}:=\frac{1}{m_i\lambda_j}\mathbf{A}_{ij},\qquad a:=-\frac{1}{2}\|\mathbf{P}_{ij}^{\mathbf{q}}\|_{\ell^2}^2, \tag{6.33}$$

$$b:=\mathsf{K}_i^{\max}\mathbf{P}_{ij}^{\mathsf{h}}-\mathbf{Q}_i^{\mathrm{W,L}}\cdot\mathbf{P}_{ij}^{\mathbf{q}},\qquad c:=\mathsf{K}_i^{\max}\mathsf{H}_i^{\mathrm{W,L}}-\frac{1}{2}\|\mathbf{Q}_i^{\mathrm{W,L}}\|_{\ell^2}^2. \tag{6.34}$$

Let r be the largest positive root of the quadratic equation $ax^2+bx+c=0$ with the convention that $r=1$ if the equation has no positive root. Let ℓ_{ij}^{h} be the limiter obtained after applying (6.27)–(6.28)–(6.29). Then we are guaranteed that $\Psi(\mathbf{W}_i^{\mathrm{L},n+1}+t\mathbf{P}_{ij})\geq 0$ for all $t\in[0,\ell_j^{i,\mathsf{K}}]$, and $\Psi(\mathbf{W}_i^{\mathrm{L},n+1}+\ell_{ij}\mathbf{P}_{ij})\geq 0$ by setting

$$\ell_j^{i,\mathsf{K}}:=\min(r,\ell_{ij}^{\mathsf{h}}),\qquad \ell_{ij}=\min(\ell_j^{i,\mathsf{K}},\ell_i^{j,\mathsf{K}}). \tag{6.35}$$

Indeed, if the quadratic equation has no positive root, then $\Psi(\mathbf{W}_i^{\mathrm{L},n+1}+t\mathbf{P}_{ij})\geq 0$ for all $t\in[0,\ell_{ij}^{\mathsf{h}}]=[0,\min(1,\ell_{ij}^{\mathsf{h}})]=[0,\ell_j^{i,\mathsf{K}}]$. If there is at least one positive root, then $\Psi(\mathbf{W}_i^{\mathrm{L},n+1}+t\mathbf{P}_{ij})\geq 0$ for all $t\in[0,\ell_j^{i,\mathsf{K}}]=[0,\min(r,\ell_{ij}^{\mathsf{h}})]\subset[0,r]$. As a result $\Psi(\mathbf{W}_i^{\mathrm{L},n+1}+\ell_{ij}\mathbf{P}_{ij})\geq 0$ because $\ell_{ij}\in[0,\ell_j^{i,\mathsf{K}}]$.

LEMMA 6.8. *Under the CFL condition* $1-\frac{2\tau}{m_i}\sum_{j\in\mathcal{I}(i)\setminus\{i\}} d_{ij}^{\mathrm{L},n}\geq 0$, *the update* $\mathbf{U}_i^{n+1}$ *given by* (6.18) *with the limiting defined by* (6.32)–(6.35) *satisfies the bound* $\psi(\mathbf{U}_i^{n+1}-\frac{\tau}{m_i}\mathbf{S}_i^n)\leq\mathsf{K}_i^{\max}$.

Proof. Let us set $\mathbf{W}_i^{n+1} := \mathbf{U}_i^{n+1} - \frac{\tau}{m_i}\mathbf{S}_i^n$, then (6.18) can be rewritten $\mathbf{W}_i^{n+1} = \mathbf{W}_i^{\mathrm{L},n+1} + \sum_{j\in\mathcal{I}(i)\setminus\{i\}} \ell_{ij} m_i^{-1}\mathbf{A}_{ij} = \sum_{j\in\mathcal{I}(i)\setminus\{i\}} \lambda_j(\mathbf{W}_i^{\mathrm{L},n+1} + \ell_{ij} m_i^{-1}\lambda_j^{-1}\mathbf{A}_{ij})$. Hence, Jansen's inequality together with (6.32)–(6.35) gives

$$\psi(\mathbf{W}_i^{n+1}) \le \sum_{j\in\mathcal{I}(i)\setminus\{i\}} \lambda_j \psi(\mathbf{W}_i^{\mathrm{L},n+1} + \ell_{ij}\mathbf{P}_{ij}) \le \sum_{j\in\mathcal{I}(i)\setminus\{i\}} \lambda_j \mathsf{K}_i^{\max} = \mathsf{K}_i^{\max}. \tag{6.36}$$

This concludes the proof. ∎

Remark 6.9 (convex limiting). Note that by proceeding as in [31, section 4.6], one can limit the entropy $\eta(\boldsymbol{u})$ defined in (6.7) or the flat bottom entropy $\eta(\boldsymbol{u})^{\mathrm{flat}}$ defined in (6.7) since these two functional are convex. In this case one can invoke a Newton-secant technique to estimate the limiters. The Newton-secant technique is guaranteed to converge to a unique solution since these functionals are convex. We leave the details to the reader. ∎

7. Numerical illustrations. In this section we illustrate the performance of the various algorithms introduced in the paper.

7.1. Technical details. All the numerical simulations are done in two space dimensions even when the problem under consideration has a one-dimensional solution; that is, in all the test cases we take $d = 2$. Unstructured, nonnested, Delaunay meshes composed of triangles are used in order to avoid extraneous super-convergence effects. The computations are done with continuous $\mathbb{P}_1$ finite elements. The time stepping is done with the SSP RK(3,3) method (three stages, third order). Notice that one could use any SSP RK(s,p) technique with $p \ge 2$ since the space approximation is second order. The time step is estimated at each time step by using $\tau = \mathrm{CFL}(\max_{i\in\mathcal{V}} m_i^{-1}\sum_{j\in\mathcal{I}(i)\setminus\{i\}} d_{ij}^{\mathrm{V},n})^{-1}$. All the computations reported in this section have been done with the upper bound on $\lambda_{\max}^{\mathrm{f_{1D}}}(\boldsymbol{v}_L, \boldsymbol{v}_R)$ given by Lemma 4.1. This bound is used to define the artificial viscosity $d_{ij}^{\mathrm{V},n}$ in (4.2). All the computations done with EV use the definition of the residual given in (6.10). We have verified in tests not reported here that the definitions (6.8) and (6.10) give results that are quantitatively very similar. The function $\psi(\alpha)$ that we use in (5.1) is defined in (5.4) and, unless specified otherwise, we use $\alpha_0 = 0$ and $p = 2$. In tests not reported here we have verified that using $\alpha_0 = 0.25$ and $p = 4$ gives slightly better errors, in general, but similar convergence rates.

When doing convergence tests over meshes of different mesh size, 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 $d\log(e_1/e_2)/\log(I_2/I_1)$; note that the quantity $I^{-\frac{1}{d}}$ scales like the mesh size. All the errors are relative with respect to the corresponding norm of the exact solution. We use SI units everywhere. In all the test cases we take $g = 9.81\,\mathrm{m\,s^{-1}}$.

In an effort to demonstrate reproducibility, two demonstration codes have been written. One code has been written at ERDC using the Proteus toolkit, (the reader is referred to Kees and Farthing [37] for the details) and the second one has been developed at TAMU using Fortran95/2003. Both codes use continuous $\mathbb{P}_1$ Lagrange elements on triangles. Proteus uses a reduced version of the limiting on the water height described in section 6.4. In particular, only positivity on the water height is enforced, i.e., (6.30) is used with $\mathsf{H}_i^{\min} = 0$, $R_j^+ = R_i^+ = 1$, and the dry state detector is defined by setting $\mathsf{H}_i^{\mathrm{dry}} = \mathsf{H}_i^n - (|D_i|^{\frac{1}{d}}/\operatorname{diam}(D))^{k+1}\mathsf{H}_{0,\max}$, where $\mathsf{H}_{0,\max} = \operatorname{ess\,sup}_{\boldsymbol{x}\in D} \mathsf{h}_0(\boldsymbol{x})$, $\operatorname{diam}(D)$ is the diameter of D, and $|D_i|$ is the d-dimensional measure

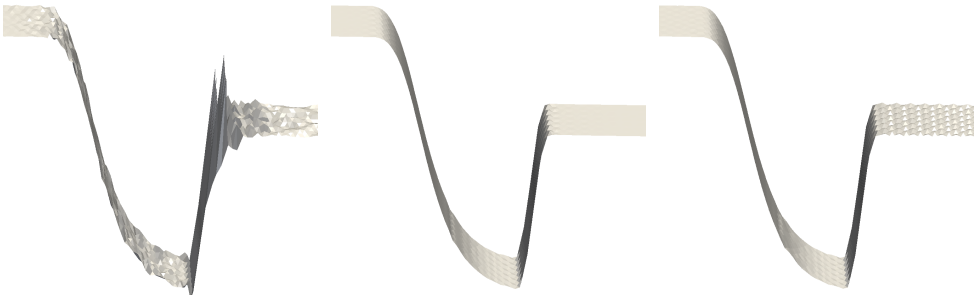

FIG. 1. *Hydraulic jump. Left: Galerkin + full limiting; Center: EV + full limiting; Right: EV without any limiting.*

of D_i. Recall that D_i is the support of the shape function φ_i and that $|D_i|^{\frac{1}{d}}$ is of the order of the local mesh size h; hence $(|D_i|^{\frac{1}{d}}/\operatorname{diam}(D))^{k+1}\mathsf{H}_{0,\max}$ is of the same order as the interpolation error. The TAMU code uses the limiting on the water height described in section 6.4 with the exception of section 7.2, where only limiting on the water height is used. The computations done with the α^2-viscosity in both codes use $\beta_{ij} = 1$. We have verified that using the generalized barycentric coordinates does not make any significant difference.

7.2. EV versus Galerkin + limiting. It is sometimes advocated in the finite element literature that one can use the Galerkin method, or some linearly stabilized version thereof, as the high-order method and rely on some kind of limiting (FCT or otherwise) to obtain the right solution. We have demonstrated in the past that this idea is not robust and sometimes outright wrong (see Ern and Guermond [19], Guermond et al. [29, section 4.2.3], Guermond and Popov [26, Lem. 4.4]), and we want to illustrate this point again. We consider a transcritical flow over a bump with a hydraulic jump. The problem is described in Noelle, Xing, and Shu [42, p. 49], but we run it with the parameters given in Azerad, Guermond, and Popov [3, section 5.3.3]. The flow domain is $D = [0, 25]\times[0, 1]$, the bottom topography is $z(x) = \frac{0.2}{64}(x-8)^3(12-x)^3$ if $8 \le x \le 12$ and $z(x) = 0$ otherwise. The flow rate $q_{\text{in}} = 0.18\,\text{m}^2\,\text{s}^{-1}$ is enforced at $\{x = 0\}$ and the exact water $h_L = 0.282\,052\,798\,138\,021\,81\,\text{m}$ is enforced at the outflow $\{x = L\}$. The initial condition is $q(x) = q_{\text{in}}$ and $h(x) + z(x) = h_L$.

We show in Figure 1 a close-up view of the water height at time $t = 80\,\text{s}$ in the vicinity of the hydraulic jump. The water height is scaled 20 times to amplify the defects. The mesh is nonuniform and composed of 5998 triangles with 3260 $\mathbb{P}_1$ nodes. There are approximately 13 nodes in the y-direction and 250 nodes in the x-direction. The solution shown in the leftmost panel has been obtained with the TAMU code with $d_{ij}^{\text{H},n} = \mu_{ij}^{\text{H},n} = 0$ with limiting on the water height plus limiting on the kinetic energy as described in section 6.4. Spurious oscillations occurring at the hydraulic jump are clearly visible. Note that at each time step, the limiting implies that the solution is always within the bounds prescribed by the low-order solution. Tests done without limiting the kinetic energy (not shown here) give a solution that is barely recognizable. The panel in the center shows the solution obtained with the EV plus the same limiting as in the previous case. The solution is clean. The solution obtained in the rightmost panel has been obtained with the EV solution without any limiting. Here again the solution is clean; one distinguishes tiny (innocuous) oscillations though. This example

illustrates the use of the EV strategy that we have adopted: The EV method gives a reasonable solution without any limiting; limiting is used for robustness only since it guarantees that the solution satisfies physical bounds after limiting. In conclusion the Galerkin method or linearly stabilized versions thereof should not be used as a high-order method when solving nonlinear hyperbolic systems. One should not try to use limiting to fix a bad (possibly nonentropic) approximation technique.

7.3. One-dimensional dam break without friction. We now assess the convergence properties of the method without friction. For this purpose we consider Ritter's solution [45] of the dam break problem without friction and with flat bottom. This is a one-dimensional Riemann problem with the initial condition

$$\boldsymbol{v}_0 = 0, \qquad h_0(x) = \begin{cases} h_l & \text{if } 0 \le x < x_0, \\ 0 & \text{if } x_0 \le x < L, \end{cases} \tag{7.1}$$

where $h_l > 0$. The solution is as follows (see Toro [51], SWASHES in [16]):

$$(h, v) = \begin{cases} (h_l, 0) & \text{if } 0 \le x \le x_A(t), \\ \left(\frac{4}{9g}\left(\sqrt{gh_l} - \frac{x-x_0}{2t}\right)^2, \frac{2}{3}\left(\frac{x-x_0}{t} + \sqrt{gh_l}\right)\right) & \text{if } x_A(t) \le x \le x_B(t), \\ (0, 0) & \text{if } x_B(t) \le x \le L, \end{cases} \tag{7.2}$$

where $x_A(t) = x_0 - t\sqrt{gh_l}$ and $x_B(t) = x_0 + 2t\sqrt{gh_l}$. The computational domain is $D = [0, L]\times[0, 1\,\mathrm{m}]$. We consider $h_l = 0.005\,\mathrm{m}$, $x_0 = 5\,\mathrm{m}$, $L = 10\,\mathrm{m}$. In order to estimate the accuracy of the method with a solution whose partial derivatives are in $\mathrm{BV}(D)$, we select the initial condition to be the solution to the above problem at $t = 1\,\mathrm{s}$ and estimate the relative L^1-norm of the error at $t = 6\,\mathrm{s}$.

Our objective is to compare the α^2-method with the EV method using the α^2-method for the low-order viscosity. In order to demonstrate also the effectiveness of the consistent mass matrix, the EV solution is computed with either the lumped mass matrix (term $\mathcal{M}^L - \mathcal{M}$ not present in (6.16)) or with the consistent mass matrix (term $\mathcal{M}^L - \mathcal{M}$ present in (6.16)). The results are reported in Table 1. This series of tests clearly shows that the EV method is superior to the the α^2-method whether the mass matrix is lumped or not. This test shows also that the EV method is indeed second-order accurate in space; recall that the solution (7.3) is not in $W^{2,1}(D)$, but the gradient is in $\mathrm{BV}(D)$.

TABLE 1
Convergence tests, one-dimensional dam break without friction, flat bathymetry.

		α^2-method		EV–α^2, lumped		EV–α^2, consistent	
	I	L^1-error	Rate	L^1-error	Rate	L^1-error	Rate
Proteus	205	1.72E-02	–	2.00E-02	–	2.14E-02	–
	729	8.50E-03	1.11	7.89E-03	1.47	7.75E-03	1.60
	2737	3.70E-03	1.26	2.98E-03	1.47	2.31E-03	1.83
	10593	1.57E-03	1.27	1.31E-03	1.22	6.88E-04	1.79
	41665	6.37E-04	1.31	5.72E-04	1.21	2.23E-04	1.64
TAMU	248	1.44E-02	–	1.42E-02	–	1.10E-02	–
	816	7.15E-03	1.18	6.89E-03	1.14	3.40E-03	1.85
	3069	2.91E-03	1.35	2.77E-03	1.37	1.04E-03	1.79
	12189	1.17E-03	1.32	1.08E-03	1.36	3.12E-04	1.74
	48053	4.60E-04	1.36	4.21E-04	1.38	8.75E-05	1.85

TABLE 2
Well-balancing tests, TAMU code (EV-α^2, consistent), $\gamma = \frac{4}{3}$.

h_0 (m)	q_0 ($\mathrm{m}^2\mathrm{s}^{-1}$)	n ($\mathrm{m}^{-1/3}\mathrm{s}$)	b	Error
5.7708E-01	2.0E-00	2.0E-02	-1.E-02	4.26E-14
9.5635E-02	1.0E-01	2.0E-02	-1.E-02	1.82E-15
2.5119E-01	1.0E-01	1.0E-01	-1.E-02	9.04E-15
2.4022E-02	2.0E-03	1.0E-01	-1.E-02	1.49E-14
4.4894E-01	2.0E-00	1.0E-01	$-1/\sqrt{3}$	1.86E-14

TABLE 3
Convergence to steady state with friction.

	I	L^1-error	Rate	L^∞-error	Rate
TAMU	258	1.36E-04	–	4.05E-04	–
	885	2.91E-05	2.50	3.78E-05	3.85
	3260	7.15E-06	2.15	9.44E-06	2.13
	12023	1.82E-06	2.10	2.38E-06	2.11
	47043	4.44E-07	2.07	5.82E-07	2.06

7.4. Well-balancing w.r.t. friction. We continue with a series tests suggested in Chertock et al. [13] to check well-balancing. We consider the domain $D = (0, 25\,\mathrm{m}) \times (0, 1\,\mathrm{m})$. The bottom is an inclined plane defined by $z(\boldsymbol{x}) = bx$, where b is the slope. We run the five same experiments as in [13, section 4.1, Ex. 1]. The mesh is a nonuniform Delaunay triangulation composed of 382 triangles; the mesh size is approximately 0.33 m. Dirichlet boundary conditions are enforced at the inflow boundary and no condition is enforced at the outflow. We use $\gamma = \frac{4}{3}$ and let n, b and q_0 vary. The initial data are chosen to be $\boldsymbol{q}_0 := q_0 \boldsymbol{e}_x$ and $\mathsf{h}_0 := \left(\mathrm{n}^2 q_0^2 b^{-1}\right)^{\frac{1}{2+\gamma}}$. We compute $\sup_{0 \le t_n \le T} \|\boldsymbol{q}_h^n - \boldsymbol{q}_0\|_{\boldsymbol{L}^\infty(D)} / \|\boldsymbol{q}_0\|_{\boldsymbol{L}^\infty(D)}$ with $T = 100\,\mathrm{s}$. The results obtained with the TAMU code (EV-α^2, consistent) are reported in Table 2. It is clear that well-balancing is achieved.

7.5. Convergence to steady state with friction. We solve a one-dimensional steady state friction problem with an analytical solution. Assuming that the solution to (2.1) is one-dimensional, time-independent, and the discharge is constant, $\boldsymbol{q} = q_0 \boldsymbol{e}_x$, the bathymetry map and the water height are related through the following identity: $z'(x) = h'(x)(\frac{q_0^2}{g h^3(x)} - 1) - \frac{\mathrm{n}^2 q_0^2}{\mathsf{h}^{\gamma+2}}$. Choosing $h(x) := (2 - \sin^4(\pi x L^{-1}))^{-\frac{1}{\gamma+2}}$, we can solve the above ODE, and, upon setting $z(0) = 0$, the solution turns out to be

$$
\begin{aligned}
(7.3)\quad z(x) = {} & q_0^2 (2g)^{-1} \left(h^{-2}(x) - h^{-2}(0)\right) - h(x) + h(0) \\
& - \mathrm{n}^2 q_0^2 \left(2x - (16\pi)^{-1} \left(-4L \sin^3(\tfrac{\pi x}{L}) \cos(\tfrac{\pi x}{L}) + 3\big(2\pi x - L \sin(\tfrac{2\pi x}{L})\big)\right)\right).
\end{aligned}
$$

We use the same data as in Chertock et al. [13, p. 370]. We consider the rectangular domain $D = (0, 150\,\mathrm{m}) \times (0, 1\,\mathrm{m})$. We set $q_0 = 1\,\mathrm{m}^2\,\mathrm{s}^{-1}$ ($q_0 = 2\,\mathrm{m}^2\,\mathrm{s}^{-1}$), $\mathrm{n} = 0.03\,\mathrm{m}^{\frac{\gamma-2}{2}}\mathrm{s}$ ($\mathrm{n} = 0.03\,\mathrm{m}^{\frac{\gamma-2}{2}}\mathrm{s}$), and $\gamma = \frac{4}{3}$. We enforce $\boldsymbol{q} = q_0 \boldsymbol{e}_x$ at $x = 0$ and the exact value of h at the outflow boundary $x = 25\,\mathrm{m}$. We also enforce $\boldsymbol{q} \cdot \boldsymbol{e}_y = 0$ on the sides of the domain, $y = 0$ and $y = 1\,\mathrm{m}$. The initial conditions are $\boldsymbol{q} = 0$ and $h = 1\,\mathrm{m}$. The simulation is stopped at $T = 1000\,\mathrm{s}$ and we compute the L^1-norm of the error on the water height and the discharge. The results obtained with the TAMU code (EV-α^2, consistent) are shown in Table 3. The optimal $\mathcal{O}(h^2)$ convergence rate is achieved both in the L^1 and L^∞ norms.

TABLE 4
Convergence tests on a one-dimensional paraboloid.

		α^2-method		EV–α^2, lumped		EV–α^2, consistent			
	I	L^1-error	Rate	L^1-error	Rate	L^1-error	Rate	L^∞-error	Rate
Proteus	205	4.58E-03	–	5.48E-03	–	6.93E-03	–	3.58E-02	–
	729	1.72E-03	1.54	2.23E-03	1.42	2.81E-03	1.42	1.65E-02	1.21
	2737	6.60E-04	1.45	8.78E-04	1.41	9.85E-04	1.58	8.66E-3	0.97
	10593	2.56E-04	1.40	3.14E-04	1.52	3.24E-04	1.64	4.91E-03	0.83
	41665	9.78E-05	1.41	1.13E-04	1.50	1.01E-04	1.70	2.94E-03	0.74
TAMU	248	5.38E-03	–	4.02E-03	–	4.06E-03	–	2.41E-02	–
	816	1.59E-03	2.04	1.37E-03	1.81	1.32E-03	1.89	1.21E-02	1.16
	3069	5.38E-04	1.64	5.03E-04	1.51	4.76E-04	1.54	6.69E-03	0.90
	12191	1.72E-04	1.65	1.80E-04	1.47	1.50E-04	1.67	2.98E-03	1.17
	48053	6.66E-05	1.3	7.21E-05	1.33	5.64E-05	1.43	1.29E-03	1.22

7.6. Planar free surface in a one-dimensional paraboloid. We consider a planar surface moving in a one-dimensional paraboloid with the linear friction $\boldsymbol{S}(\boldsymbol{u}) = -k\boldsymbol{q}$ (see Sampson, Easton, and Singh [47, section 4.1]). The bathymetry is $z(\boldsymbol{x}) = \frac{\mathsf{h}_0}{a^2}\left(x - \frac{1}{2}L\right)^2$. The exact solution $(h(\boldsymbol{x},t), q(x,t))^\mathsf{T}$ is given by

$$h(\boldsymbol{x},t) = \max(\tilde{h}(\boldsymbol{x},t), 0), \tag{7.4}$$

$$\begin{aligned}\tilde{h}(\boldsymbol{x},t) = \mathsf{h}_0 - z(\boldsymbol{x}) + (a^2B^2/8g^2\mathsf{h}_0)\mathrm{e}^{-kt}\Big[\left(\tfrac{1}{4}k^2 - s^2\right)\cos(2st) - sk\sin(2st)\Big] \\ - (B^2/4g)\mathrm{e}^{-kt} - (B/g)\mathrm{e}^{-\frac{1}{2}kt}\left[s\cos(st) + \tfrac{1}{2}k\sin(st)\right]\left(x - \tfrac{1}{2}L\right),\end{aligned} \tag{7.5}$$

$$\boldsymbol{q}(\boldsymbol{x},t) = (h(\boldsymbol{x},t)B\mathrm{e}^{-\frac{1}{2}kt}\sin(st), 0)^\mathsf{T}, \tag{7.6}$$

where $p = \sqrt{8g\mathsf{h}_0}/a$ and $s = \sqrt{p^2 - k^2}/2$. We do the computation in the same configuration as in Duran et al. [17].

We consider a rectangular domain $D \in (0, 10000\,\mathrm{m})\times(0, 1000\,\mathrm{m})$ and set $L = 10000\,\mathrm{m}$, $g = 9.81\,\mathrm{ms}^{-2}$, $\mathsf{h}_0 = 10\,\mathrm{m}$, $a = 3000\,\mathrm{m}$, $B = 2\,\mathrm{ms}^{-1}$, $k = 0.001\,\mathrm{s}^{-1}$. Since the paraboloid is large enough so that no water reaches the boundaries, we do not enforce any boundary condition. The initial condition is given by (7.4) at $t = 0$. We show in Table 4 the results of convergence tests on the water height. The errors are computed after one period $t = \frac{4\pi}{p} \approx 1345.71\,\mathrm{s}$. The EV-$\alpha^2$ method delivers the rate $\mathcal{O}(h^{1.5})$ in the L^1-norm and $\mathcal{O}(h)$ in the L^∞-norm. The rate in the L^∞-norm is optimal since the water height in only in $W^{1,\infty}(D\times(0,\infty))$. Using the consistent mass matrix improves the convergence rates.

7.7. Planar free surface in a two-dimensional paraboloid. We now use a variation of Thacker's solution in a paraboloid developed in Sampson, Easton, and Singh [46] (see (20)–(21)–(28) therein). The friction is linear, $\boldsymbol{S}(\boldsymbol{u}) = -k\boldsymbol{q}$, and the topography of the bottom is a paraboloid of revolution defined by $z(\boldsymbol{x}) = r^2\left(\frac{h_0}{a^2}\right)$, where $r^2 = \left(x - \frac{L}{2}\right)^2 + \left(y - \frac{L}{2}\right)^2$. The exact solution $(h(\boldsymbol{x},t), \boldsymbol{q}(\boldsymbol{x},t))^\mathsf{T}$ is given by

$$h(\boldsymbol{x},t) = \max(\tilde{h}(\boldsymbol{x},t), 0), \tag{7.7}$$

$$\begin{aligned}\tilde{h}(\boldsymbol{x},t) = \mathsf{h}_0 - z(\boldsymbol{x}) - (B/g)\exp(-\tfrac{1}{2}kt)\left[\tfrac{1}{2}k\sin(st) + s\cos(st)\right]\left(x - \tfrac{1}{2}L\right) \\ - (B^2/2g)\exp(-kt) - (B/g)\exp(-\tfrac{1}{2}kt)\left[\tfrac{1}{2}k\cos(st) - s\sin(st)\right]\left(y - \tfrac{1}{2}L\right),\end{aligned} \tag{7.8}$$

$$\boldsymbol{q}(\boldsymbol{x},t) = h(\boldsymbol{x},t)B\exp(-\tfrac{1}{2}kt)(\sin(st), \cos(st))^\mathsf{T}. \tag{7.9}$$

TABLE 5
L^1 convergence of planar surface on a two-dimensional paraboloid.

Proteus				
I	L^1-error	Rate	L^∞-error	Rate
441	4.58E-02		8.41E-02	
1681	1.45E-02	1.72	4.07E-02	1.08
6561	6.30E-03	1.22	2.64E-02	0.63
25921	2.24E-03	1.50	1.34E-02	0.99
103041	7.52E-04	1.58	6.46E-03	1.06

TAMU				
I	L^1-error	Rate	L^∞-error	Rate
508	4.70E-02		8.12E-02	
1926	1.95E-02	1.32	4.55E-02	0.87
7553	7.67E-03	1.37	1.98E-02	1.22
29870	2.91E-03	1.41	1.11E-02	0.84
118851	1.09E-03	1.43	6.21E-03	0.85

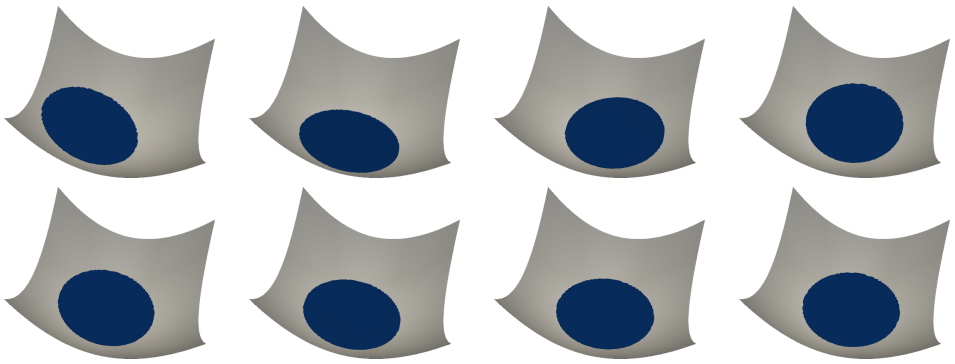

FIG. 2. *Surface plot of water elevation $h(\boldsymbol{x},t)+z(\boldsymbol{x})$ at $t=\{0,1,\dots,7\}$.*

Here $p=\sqrt{8g\mathsf{h}_0}/a$ and $s=\sqrt{p^2-k^2}/2$. We consider $D=[L=10\,\mathrm{m},10\,\mathrm{m}]$ and set $g=9.81\,\mathrm{ms}^{-2}$, $\mathsf{h}_0=1\,\mathrm{m}$, $a=3\,\mathrm{m}$, $B=2\,\mathrm{ms}^{-1}$, $k=0.5\,\mathrm{s}^{-1}$. We do not impose any boundary condition since the water never reaches the boundary of the computational domain. The initial condition is given by (7.8) at $t=0$. We show in Table 5 the errors and the corresponding convergence rates computed at $t=1$. This time we just consider the EV method with the consistent mass matrix. The rates are consistent with those obtained in section 7.6, it is $\mathcal{O}(h)^{1.5}$ in the L^1 norm and $\mathcal{O}(h)$ in the L^∞ norm.

We shown in Figure 2 the results of a computation done with a mesh containing 6561 $\mathbb{P}_1$ nodes. The solution is represented at $t=\{0,1,\dots,7\}$. The free surface keeps its circular shape over the entire duration of the simulation.

7.8. Dam break over three bumps. We consider now a test case proposed in Kawahara and Umetsu [36]. (See also Huang, Zhang, and Pei [33], Song et al. [49] and references therein.) This problem tests the numerical method for complex wetting/drying processes. The problem consists of a channel with the following dimensions $75\,\mathrm{m}\times 30\,\mathrm{m}$. A dam is located at $x=16\,\mathrm{m}$. The initial water height behind the dam is 1.875 m and 0 m beyond the dam. There are three obstacles sitting on the dry bottom; the bathymetry is given by

$$z(\boldsymbol{x})=\max\{0,z_1(\boldsymbol{x}),z_2(\boldsymbol{x}),z_3(\boldsymbol{x})\}, \tag{7.10}$$

$$z_1(\boldsymbol{x})=1-1/8\sqrt{(x-30)^2+(y-6)^2}, \tag{7.11}$$

$$z_2(\boldsymbol{x})=1-1/8\sqrt{(x-30)^2+(y-24)^2)}, \tag{7.12}$$

$$z_3(\boldsymbol{x})=3-3/10\sqrt{(x-47.5)^2+(y-15)^2}. \tag{7.13}$$

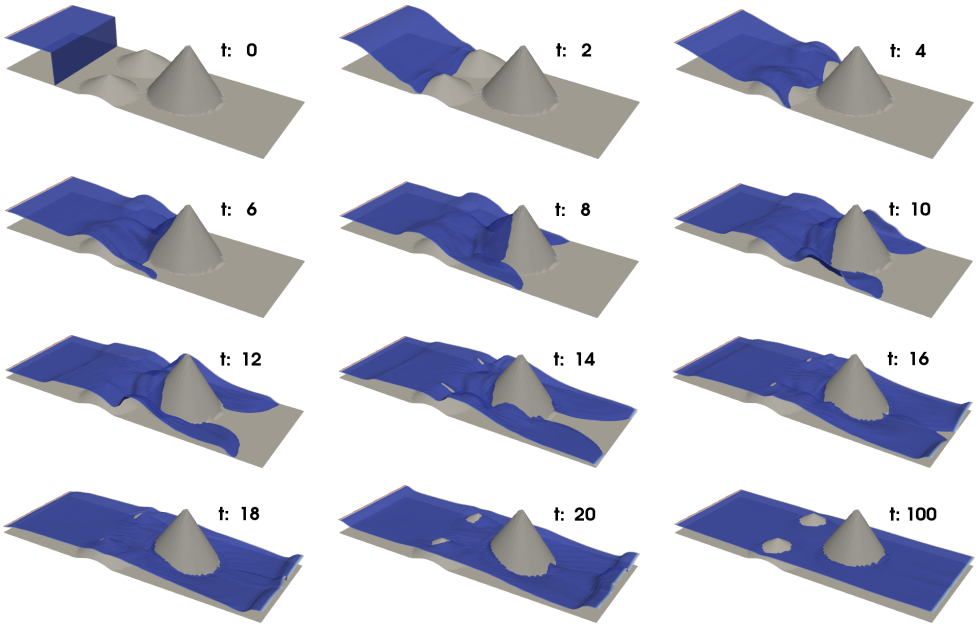

FIG. 3. *Surface plot of the water elevation* $h(\boldsymbol{x}, t) + z(\boldsymbol{x})$ *at different times.*

Gauge	x (m)	y (m)
G6	4947.46	4289.71
G7	5717.30	4407.61
G8	6775.14	3869.23
G9	7128.20	3162.00
G10	8585.30	3443.08
G11	9674.97	3085.89
G12	10939.15	3044.78
G13	11724.37	2810.41
G15	12723.70	2485.08

FIG. 4. *Coordinates (left) and location of the gauges (right).*

The Manning coefficient is $n = 0.02\,\mathrm{m}^{-1/3}\mathrm{s}$. The solution at various times is shown in Figure 3. The computations have been done with a mesh composed of 6561 $\mathbb{P}_1$ nodes.

7.9. Malpasset dam break. We finish with the 1959 Malpasset dam break which caused 423 casualties. In an effort to better understand the origins of this catastrophic event and possibly help prevent such accidents in the future, a scaled model was built in 1964 and calibrated against the available data. We refer the reader to Hervouet and Petitjean [32] for a complete description of this problem and for the data extracted from the the scaled model. Fourteen gauges were placed to measure the maximum water height and arrival time, 9 of which were positioned downstream from the dam. For completeness, we show in the left panel of Figure 4 the coordinates of the nine gauges downstream of the dam; the corresponding geographic locations are show in the right panel of the figure.

TABLE 6
Malpasset dam break problem. Maximum water elevation.

		Proteus			TAMU		
Gauge	Exp.	$n = 0.025$	$n = 0.029$	$n = 0.033$	$n = 0.025$	$n = 0.029$	$n = 0.033$
G6	84.2	86.24	86.17	86.14	86.36	86.24	86.20
G7	49.1	51.49	51.77	52.05	50.66	51.64	51.80
G8	54.0	54.52	54.54	54.53	54.39	54.6	54.7
G9	40.2	48.51	48.59	48.66	48.33	48.39	48.50
G10	34.9	38.11	38.18	38.17	38.21	38.28	38.20
G11	27.4	25.27	25.93	25.91	25.60	25.84	25.97
G12	21.5	18.38	18.31	18.43	18.28	18.27	18.52
G13	16.1	16.65	16.76	16.82	16.44	16.61	16.77
G14	12.9	12.87	12.91	13.06	12.94	12.96	13.13

We follow Hervouet and Petitjean [32] and consider the dam to be a straight line between the points $(4701.18\,\mathrm{m}, 4143.41\,\mathrm{m})$ and $(4655.5\,\mathrm{m}, 4392.1\,\mathrm{m})$. The water level in the reservoir is 100 m. We consider three different Manning coefficients: $0.025\,\mathrm{m}^{-\frac{1}{3}}\mathrm{s}$, $0.029\,\mathrm{m}^{-\frac{1}{3}}\mathrm{s}$, $0.033\,\mathrm{m}^{-\frac{1}{3}}\mathrm{s}$. We use a mesh composed of 56698 triangles and 29391 $\mathbb{P}_1$ nodes. The simulations are run until $t = 3000\,\mathrm{s}$.

The maximum water elevations $\max_{t=[0,3000]} h(g,t) + z(\boldsymbol{x}_g)$ reached during the simulations are shown in Table 6 for all the gauges $g \in \{\mathrm{G6}, \ldots, \mathrm{G14}\}$. Here $z(\boldsymbol{x}_g)$ is the bathymetry at the gauge g. The column denoted "Exp." shows the maximum water elevation estimated from the experiment on the scaled model as reported in Hervouet and Petitjean [32].

In order to put in perspective our results, we show in the left table of Figure 5 the maximum water elevation reported from the following sources: Huang, Zhang, and Pei [33] ("Huang"); Valiani, Caleffi, and Zanni [52] ("Valiani"); Hervouet and Petitjean [32] ("Hervouet"); Biscarini et al. [7] ("Biscarini"); Savant et al. [48] ("Savant"). The symbols connected with solid lines are the results from the TAMU code. With the exception of [7], all the numerical results have been obtained with shallow water codes using one of the Manning coefficients we have considered. Our results agree with those reported in the literature. It seems that the maximum water elevation does not depend very much on the Manning coefficient. We show in the right table of Figure 5 the arrival time at the various gauges for the three Manning coefficients considered.

Finally we show in Figure 6 the water height at various times after the collapse of the dam: $t = 0, 600\,\mathrm{s}, 1200\,\mathrm{s}, 1800\,\mathrm{s}, 2400\,\mathrm{s}, 3000\,\mathrm{s}$. Although the value $0.033\,\mathrm{m}^{-\frac{1}{3}}\mathrm{s}$ best fits the data, we think that it is probably unrealistic to expect that one Manning coefficient can properly model the entire terrain.

8. Conclusions. We have proposed a new numerical method to solve the shallow water equations using continuous finite elements. The properties of the method are based on the introduction of an artificial dissipation that is defined so that the method is positivity preserving, robust with respect to dry states, well-balanced, and can handle explicit treatment of the friction term. The base of the new method is the smoothness-based second-order method introduced in Azerad, Guermond, and Popov [3]. One novelty in this paper is the addition of the friction term to the shallow water system and its explicit treatment and regularization. A new higher-order EV extension of the method is given in section 6. A commutator technique is introduced to increase the efficiency of the EV method. Another novel idea is the derivation of local admissible states from the first-order GMS scheme and the use of these states for local

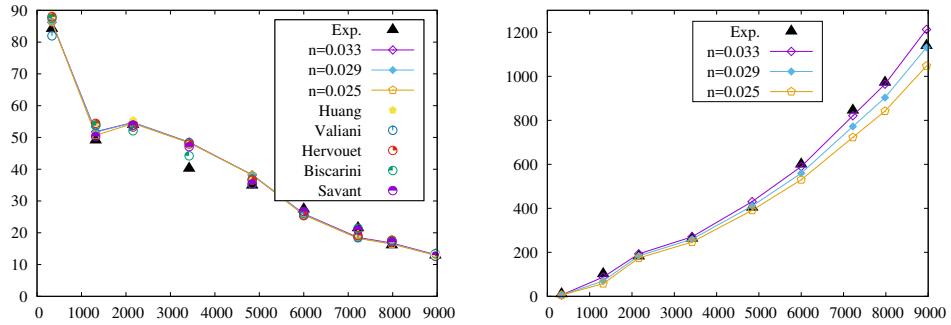

FIG. 5. *Maximum water elevation (left). Arrival time (right).*

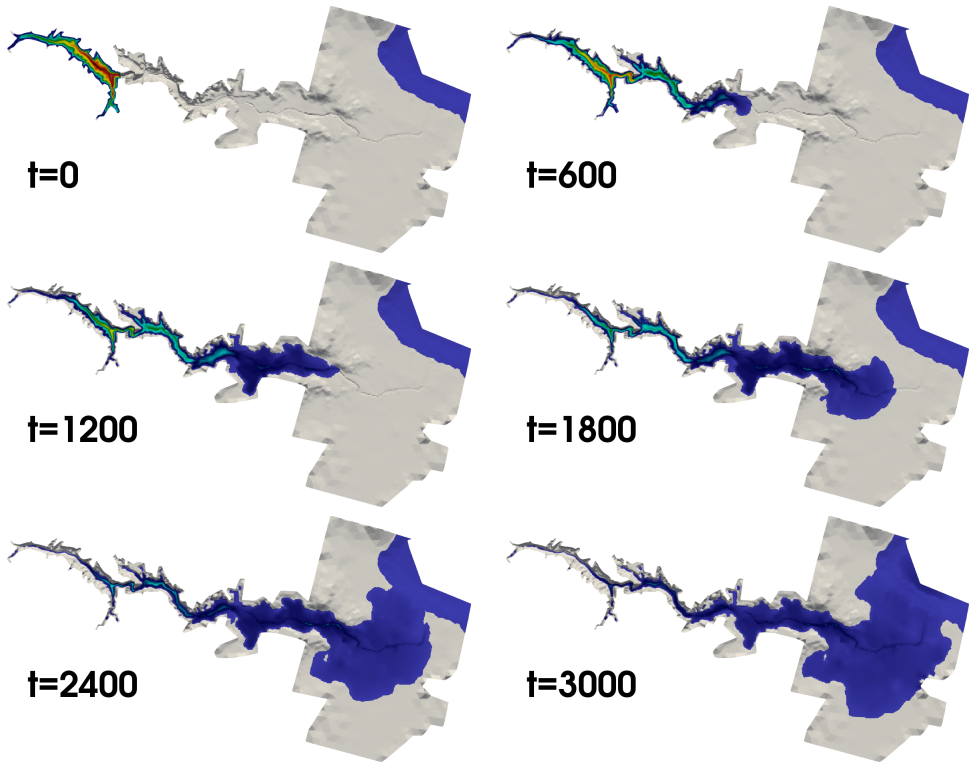

FIG. 6. *Malpasset dam break problem. Water height at different times. The color code (from blue to red) is based on the water height at* $t = 0$.

limiting in order to remove unphysical states in the high-order solution and to control the remaining innocuous oscillations. The new method is numerically illustrated on various benchmark tests including the Malpasset dam break.

REFERENCES

[1] 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 (2004), pp. 2050–2065.

[2] E. Audusse, F. Bouchut, M.-O. Bristeau, and J. Sainte-Marie, *Kinetic entropy inequality and hydrostatic reconstruction scheme for the Saint-Venant system*, Math. Comp., 85 (2016), pp. 2815–2837.
[3] P. Azerad, J.-L. Guermond, and B. Popov, *Well-balanced second-order approximation of the shallow water equation with continuous finite elements*, SIAM J. Numer. Anal., 55 (2017), pp. 3203–3224.
[4] S. Badia and J. Bonilla, *Monotonicity-preserving finite element schemes based on differentiable nonlinear stabilization*, Comput. Methods Appl. Mech. Engrg., 313 (2017), pp. 133–158.
[5] G. R. Barrenechea, E. Burman, and F. Karakatsani, *Edge-based nonlinear diffusion for finite element approximations of convection–diffusion equations and its relation to algebraic flux-correction schemes*, Numer. Math., 135 (2017), 521.
[6] A. Bermúdez and M. E. Vázquez, *Upwind methods for hyperbolic conservation laws with source terms*, Comput. Fluids, 23 (1994), pp. 1049–1071.
[7] C. Biscarini, S. Di Francesco, E. Ridolfi, and P. Manciola, *On the simulation of floods in a narrow bending valley: The Malpasset Dam break case study*, Water, 8 (2016).
[8] 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 (2011), pp. 371–404.
[9] J. P. Boris and D. L. Book, *Flux-corrected transport.* I. *SHASTA, a fluid transport algorithm that works*, J. Comput. Phys., 11 (1973), pp. 38–69.
[10] F. Bouchut, *Nonlinear stability of Finite Volume Methods for Hyperbolic Conservation Laws and well-balanced schemes for sources*, Front. Math., Birkhäuser, Basel, 2004.
[11] E. Burman, *On nonlinear artificial viscosity, discrete maximum principle and hyperbolic conservation laws*, BIT, 47 (2007), pp. 715–733.
[12] G. Chen and S. Noelle, *A new hydrostatic reconstruction scheme based on subcell reconstructions*, SIAM J. Numer. Anal., 55 (2017), pp. 758–784.
[13] A. Chertock, S. Cui, A. Kurganov, and T. Wu, *Well-balanced positivity preserving central-upwind scheme for the shallow water system with friction terms*, Internat. J. Numer. Methods Fluids, 78 (2015), pp. 355–383.
[14] M. A. Christon, M. J. Martinez, and T. E. Voth, *Generalized Fourier analyses of the advection-diffusion equation-part* I*: One-dimensional domains*, Internat. J. Numer. Methods Fluids, 45 (2004), pp. 839–887.
[15] O. Delestre, S. Cordier, F. Darboux, and F. James, *A limitation of the hydrostatic reconstruction technique for shallow water equations*, C. R. Math. Acad. Sci. Paris, 350 (2012), pp. 677–681.
[16] O. Delestre, C. Lucas, P.-A. Ksinant, F. Darboux, C. Laguerre, T.-N. Vo, F. James, S. Cordier, *Swashes: A compilation of shallow water analytic solutions for hydraulic and environmental studies*, Internat. J. Numer. Methods Fluids, 72 (2013), pp. 269–300.
[17] A. Duran, F. Marche, R. Turpault, and C. Berthon, *Asymptotic preserving scheme for the shallow water equations with source terms on unstructured meshes*, J. Comput. Phys., 287 (2015), pp. 184–206.
[18] B. Einfeldt, C.-D. Munz, P. L. Roe, and B. Sjögreen, *On Godunov-type methods near low densities*, J. Comput. Phys., 92 (1991), pp. 273–295.
[19] A. Ern and J.-L. Guermond, *Weighting the edge stabilization*, SIAM J. Numer. Anal., 51 (2013), pp. 1655–1677.
[20] M. S. Floater, *Generalized barycentric coordinates and applications*, Acta Numer., 24 (2015), pp. 161–214 .
[21] 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. Comput. Phys., 227 (2007), pp. 574 – 601.
[22] J. M. Greenberg and A. Y. Le Roux, *A well-balanced scheme for the numerical processing of source terms in hyperbolic equations*, SIAM J. Numer. Anal., 33 (1996), pp. 1–16.
[23] P. M. Gresho and R. L. Sani, *Incompressible Flow and the Finite Element Method. Volume* 1: *Advection-Diffusion and Isothermal Laminar Flow*, Wiley, New York, 1998.
[24] J.-L. Guermond and R. Pasquetti, *A correction technique for the dispersive effects of mass lumping for transport problems*, Comput. Methods Appl. Mech. Engrg., 253 (2013), pp. 186–198.
[25] J.-L. Guermond and B. Popov, *Invariant domains and first-order continuous finite element approximation for hyperbolic systems*, SIAM J. Numer. Anal., 54 (2016), pp. 2466–2489.
[26] J.-L. Guermond and B. Popov, *Invariant domains and second-order continuous finite element approximation for scalar conservation equations*, SIAM J. Numer. Anal., 55 (2017), pp. 3120–3146.

[27] 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*, manuscript.
[28] J.-L. Guermond, R. Pasquetti, and B. Popov, *Entropy viscosity method for nonlinear conservation laws*, J. Comput. Phys., 230 (2011), pp. 4248–4267.
[29] J.-L. Guermond, M. Nazarov, B. Popov, and Y. Yang, *A second-order maximum principle preserving Lagrange finite element technique for nonlinear scalar conservation equations*, SIAM J. Numer. Anal., 52 (2014), pp. 2163–2182.
[30] J.-L. Guermond, B. Popov, and Y. Yang, *The effect of the consistent mass matrix on the maximum-principle for scalar conservation equations*, J. Sci. Comput., 70 (2017), pp. 1358–1366.
[31] J.-L. Guermond, M. Nazarov, B. Popov, and I. Tomas, *Second-order invariant domain preserving approximation of the Euler equations using convex limiting*, SIAM J. Sci. Comput., 40 (2018), pp. A3211–A3239.
[32] J.-M. Hervouet and A. Petitjean, *Malpasset dam-break revisited with two-dimensional computations*, J. Hydraul. Res., 37 (1999), pp. 777–788.
[33] Y. Huang, N. Zhang, and Y. Pei, *Well-balanced finite volume scheme for shallow water flooding and drying over arbitrary topography*, Eng. Appl. Comput. Fluid Mech., 7 (2013), pp. 40–54.
[34] A. Jameson, *Origins and further development of the Jameson-Schmidt-Turkel scheme*, AIAA Journal, 55 (2017), pp. 1487–1510.
[35] A. Jameson, W. Schmidt, and E. Turkel, *Numerical solution of the Euler equations by finite volume. Methods using Runge-Kutta time-stepping schemes*, in 14th AIAA Fluid and Plasma Dynamics Conference, AIAA, New York, 1981, pp. 1981–1259.
[36] M. Kawahara and T. Umetsu, *Finite element method for moving boundary problems in river flow*, Internat. J. Numer. Methods Fluids, 6 (1986), pp. 365–386.
[37] C. E. Kees and M. W. Farthing, *Parallel computational methods and simulation for coastal and hydraulic applications using the Proteus Toolkit*, in Supercomputing11: Proceedings of the PyHPC11 Workshop, Seattle, WA, 2011, https://www.dlr.de/sc/en/Portaldata/15/Resources/dokumente/pyhpc2011/submissions/pyhpc2011_submission_11.pdf.
[38] A. Kurganov and G. Petrova, *A second-order well-balanced positivity preserving central-upwind scheme for the Saint-Venant system*, Commun. Math. Sci., 5 (2007), pp. 133–160.
[39] D. Kuzmin, R. Löhner, and S. Turek, *Flux–Corrected Transport*, Sci. Comput, Springer, Berlin, 2005.
[40] D. Kuzmin, S. Basting, and J. N. Shadid, *Linearity-preserving monotone local projection stabilization schemes for continuous finite elements*, Comput. Methods Appl. Mech. Engrg., 322 (2017), pp. 23–41.
[41] R. J. LeVeque and D. L. George, *High-resolution finite volume methods for the shallow water equations with bathymetry and dry states*, in Advanced Numerical Models for Simulating Tsunami Waves and Runup, Adv. Const. Ocean Eng. 10, 2008, World Scientific, Hackensack, NJ, pp. 43–73.
[42] 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 (2007), pp. 29–58.
[43] B. Perthame and C. Simeoni, *A kinetic scheme for the Saint-Venant system with a source term*, Calcolo, 38 (2001), pp. 201–231.
[44] M. Ricchiuto and A. Bollermann, *Stabilized residual distribution for shallow water simulations*, J. Comput. Phys., 228 (2009), pp. 1071–1115.
[45] A. Ritter, *Die fortpflanzung der wasserwellen*, Z. Vereines Deutscher Ingen., 36 (1892), pp. 947–954.
[46] J. Sampson, A. Easton, M. Singh, *Moving boundary shallow water flow in circular paraboloidal basins*, in Proceedings of the Sixth Engineering Mathematics and Applications Conference, 5th International Congress on Industrial and Applied Mathematics, at the University of Technology, Sydney, Australia, R. L. May and W.F. Blyth, eds., Engineering Mathematical Group, ANZIAM, Sydney, Australia, 2003, pp. 223–227.
[47] J. Sampson, A. Easton, and M. Singh, *Moving boundary shallow water flow above parabolic bottom topography*, ANZIAM J., 47 (2005), pp. C373–C387.
[48] G. Savant, C. Berger, T. O. McAlpin, and J. N. Tate, *Efficient implicit finite-element hydrodynamic model for dam and levee breach*, J. Hydraul. Eng., 137 (2010), pp. 1005–1018.
[49] L. Song, J. Zhou, Q. Li, X. Yang, and Y. Zhang, *An unstructured finite volume model for dam-break floods with wet/dry fronts over complex topography*, Internat. J. Numer. Methods Fluids, 67 (2011), pp. 960–980.

[50] T. Thompson, *A discrete commutator theory for the consistency and phase error analysis of semi-discrete C^0 finite element approximations to the linear transport equation*, J. Comput. Appl. Math., 303 (2016), pp. 229–248.
[51] E. F. Toro, *Shock-Capturing Methods for Free-Surface Shallow Flows*, Wiley, Chichester, England, 2001.
[52] A. Valiani, V. Caleffi, and A. Zanni, *Case study: Malpasset dam-break simulation using a two-dimensional finite volume method*, J. Hydraul. Eng., 128 (2002), pp. 460–472.
[53] Y. Xing and C.-W. Shu, *A survey of high order schemes for the shallow water equations*, J. Math. Study, 47 (2014), pp. 221–249.
[54] S. T. Zalesak, *Fully multidimensional flux-corrected transport algorithms for fluids*, J. Comput. Phys., 31 (1979), pp. 335–362.