Contents lists available at ScienceDirect

Journal of Computational Physics

journal homepage: www.elsevier.com/locate/jcp

Stabilized residual distribution for shallow water simulations

Mario Ricchiuto [a,*], Andreas Bollermann [b]

[a] INRIA Bordeaux – Sud-Ouest, 351 Cours de la Liberation, 33405 Talence Cedex, France
[b] IGPM, RWTH Aachen, Templergraben 55, 52062 Aachen, Germany

ARTICLE INFO

Article history:
Received 10 June 2008
Received in revised form 7 October 2008
Accepted 11 October 2008
Available online 25 October 2008

Keywords:
Residual distribution
Shallow water equations
Well balancedness
Positivity
Unstructured grids
High order

ABSTRACT

We propose a stabilized Residual Distribution ($\mathcal{RD}$) scheme for the simulation of shallow water flows. The final discretization is obtained combining the stabilized $\mathcal{RD}$ approach proposed in (Abgrall, *J. Comp. Phys.* 214, 2006) and (Ricchiuto and Abgrall, *ICCFD4*, Springer-Verlag 2006), with the conservative formulation already used in (Ricchiuto *et al.*, *J. Comp. Phys.* 222, 2007) to simulate shallow water flows. The scheme proposed is a nonlinear variant of a Lax–Friedrichs type discretization. It is well balanced, it actually yields second-order of accuracy in smooth areas, and it preserves the positivity of the height of the water in presence of dry areas. This is made possible by the residual character of the discretization, by properly adapting the stabilization operators proposed in (Abgrall, *J. Comp. Phys.* 214, 2006) and (Ricchiuto and Abgrall, *ICCFD4*, Springer-Verlag, 2006), and thanks to the positivity preserving character of the underlying Lax–Friedrichs scheme. We demonstrate the properties of the discretization proposed on a wide variety of tests.

1. Introduction

We consider the solution of the two dimensional shallow water equations (SWE) by means of conservative Residual Distribution ($\mathcal{RD}$) schemes on unstructured triangular meshes. The SWE model the dynamics of shallow free surface flows under the action of gravity. The model used here does not include the effects of friction or other source terms beside the ground elevation (bathymetry). It constitutes a non-homogeneous system of conservation laws for the water height and its discharge. We will also consider the case of dry bed, where important properties of the system are lost, and one runs into several numerical problems.

When solving the SWE, the discretization should respect a certain number of criteria. The schemes should keep the lake at rest solution, *i.e.* there should be no spurious numerical waves in areas with zero velocity and constant total water height. Schemes which keep the lake at rest solution are called *well balanced*. In presence of a dry/wet interface, the preservation of the positivity of the water level becomes important. Similarly, spurious oscillations near discontinuities are an unwanted effect. Hence, we need schemes that enjoy some kind of positivity preservation property, and that have a non-oscillatory character.

To solve the system numerically, we combine the stabilized formulation of nonlinear limited *Residual Distribution* ($\mathcal{RD}$) schemes proposed in [2,40], with the conservative approach of [18,42], which has been already used in [41] to solve the SWE. In the last reference, however, some issues are left open, or not addressed in detail.

The most important is the lack of iterative convergence encountered when using most nonlinear high order $\mathcal{RD}$ schemes. This hampers grid convergence, leading to sub-optimal accuracy. This issue is analyzed in [2,40], where a cure is proposed. The idea is to add, in smooth regions of the solution, a high order streamline dissipation term. Having a residual character,

[*] Corresponding author. Tel.: +33 524574117; fax: +33 524574038.
E-mail address: Mario.Ricchiuto@inria.fr (M. Ricchiuto).

0021-9991/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcp.2008.10.020

this term does not spoil the accuracy of the (already formally second order) underlying nonlinear scheme. It actually improves it by improving the properties of the algebraic nonlinear equations. Ultimately, this guarantees the existence of a unique solution, and restores iterative and grid convergence. For the steady SWE, preliminary results employing this technique have been already shown in [41]. However, the schemes used in [41] have an upwind character. While somewhat improving the convergence properties, upwinding requires heavily the use of the flux Jacobians. The resulting schemes are quite costly, and one can run into trouble in vicinity of dry/wet interfaces.

In this paper we improve on the work of [41], by using nonlinear discretizations based on a simpler approach. This is achieved by:

- using the stabilized formulation of [2,40]. This allows to build a more flexible discretization, and ultimately allows to achieve the expected grid convergence whenever the solution is smooth;
- using a nonlinear discretization built upon a multidimensional Lax–Friedrichs scheme. The numerical results show that the stabilized nonlinear Lax–Friedrichs scheme yields results as accurate as the ones obtained with the nonlinear variants of the multidimensional upwind N scheme proposed in [41].
- adapting the nonlinear Lax–Friedrichs scheme to the computations of dry/wet interfaces. In this respect, we benefit from the positivity properties of the underlying first order Lax–Friedrichs scheme. However, an *ad hoc* treatment of cells at the wet/dry front is still needed to be able to fully profit of this property, and to guarantee the preservation of the steady lake at rest state.

One of our major objectives is to show how to adapt the existing RD technology to obtain schemes tailored to shallow water simulations, and yielding, on unstructured triangulations, results comparable to the ones given by state of the art finite volume discretizations.

The exposition is organized as follows: In Section 2, we briefly recall the SWE, their properties, and some exact solutions. In Section 3, we discuss the basics of conservative $\mathcal{RD}$ schemes, with a description of how we implemented the stabilization procedure of [2,40] in the steady as well as in the time dependent case. The application to the SWE is then discussed in Section 4. We recall and generalize some results presented in [41] concerning the well balancedness of our approach, and then we analyze the preservation of the positivity of the water height for the Lax–Friedrichs based schemes. The steps undertaken to handle the wetting/drying process are described in the same section. We devote Section 5 to the details of the implementation and choice of the parameters used in the simulations. The effectiveness of the approach proposed is demonstrated by the very extensive numerical validation presented in Section 6. We end the paper with a summary and an outlook on the issues still left open.

2. The shallow water system

2.1. Conservation law form

The shallow water equations (SWE) model the behavior of shallow free surface flows under the action of gravity. In conservation law form they can be written as:

$$\frac{\partial \mathbf{u}}{\partial t} + \nabla \cdot \boldsymbol{\mathcal{F}}(\mathbf{u}) - \boldsymbol{\mathcal{S}}(\mathbf{u}, x, y) = 0 \quad \text{on} \quad \Omega_T = \Omega \times [0, t_f] \subset \mathbb{R}^2 \times \mathbb{R}^+, \tag{1}$$

with conserved variables, flux, and source term given by

$$\mathbf{u} = \begin{bmatrix} H \\ Hu \\ Hv \end{bmatrix} \quad \boldsymbol{\mathcal{F}} = [\boldsymbol{\mathcal{F}}_1, \boldsymbol{\mathcal{F}}_2] = \begin{bmatrix} Hu & Hv \\ Hu^2 + g\frac{H^2}{2}, & Huv \\ Huv & Hv^2 + g\frac{H^2}{2} \end{bmatrix} \quad \boldsymbol{\mathcal{S}} = -gH \begin{bmatrix} 0 \\ \frac{\partial B(x,y)}{\partial x} \\ \frac{\partial B(x,y)}{\partial y} \end{bmatrix}, \tag{2}$$

where H denotes the relative water height, $\vec{u} = (u, v)$ the flow speed, g the (constant) gravity acceleration, and $B(x,y)$ the local *bathymetry* or *bed* height. The source term models the effects on the flow of variations of the bed slope. We also introduce the free surface level, or total water height H_{tot} (see Fig. 1),

$$H_{tot}(x, y, t) = H(x, y, t) + B(x, y). \tag{3}$$

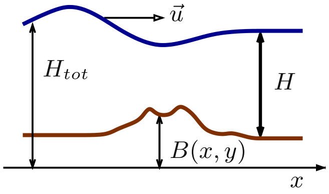


Fig. 1. Shallow water equations: main parameters.

2.2. Symmetric quasi-linear form and total energy equation

Weak solutions of the shallow water system are characterized by the entropy inequality [28,50]

$$\frac{\partial E}{\partial t} + \nabla \cdot (\vec{u}E) + \nabla \cdot \left(\vec{u}\frac{gH^2}{2}\right) \leqslant 0, \tag{4}$$

where E is the total energy given by

$$E(\mathbf{u}) = H\left(\frac{1}{2}gH + gB + \frac{\vec{u}\cdot\vec{u}}{2}\right). \tag{5}$$

The $\leqslant$ sign in (4) becomes a strict inequality across discontinuities, and an equality on smooth classical solutions. The energy E is convex in $\mathbf{u}$, and acts for the system as a mathematical entropy, in the sense of Harten [27]. In particular, introducing the vector of *symmetrizing variables* $\mathbf{v}$ given by [28]

$$\mathbf{v}^t = \frac{\partial E(\mathbf{u})}{\partial \mathbf{u}} = [p \;\; u \;\; v] \quad p = gH - \frac{\vec{u}\cdot\vec{u}}{2}, \tag{6}$$

the system can be written in the symmetric quasi-linear form [28]

$$A_0\frac{\partial \mathbf{v}}{\partial t} + A_1\frac{\partial \mathbf{v}}{\partial x} + A_2\frac{\partial \mathbf{v}}{\partial y} - \boldsymbol{\mathcal{S}}(\mathbf{v},x,y) = 0, \tag{7}$$

with the notation $\boldsymbol{\mathcal{S}}(\mathbf{v},x,y) = \boldsymbol{\mathcal{S}}(\mathbf{u}(\mathbf{v}),x,y)$ and with symmetric Jacobians $\{A_k\}_{k=0}^2$

$$A_0 = \frac{\partial \mathbf{u}}{\partial \mathbf{v}}, \quad A_1 = \frac{\partial \boldsymbol{\mathcal{F}}_1}{\partial \mathbf{v}}, \quad A_2 = \frac{\partial \boldsymbol{\mathcal{F}}_2}{\partial \mathbf{v}}. \tag{8}$$

The total energy equation is recovered multiplying on the left (7) by $\mathbf{v}^t$ [28]

$$\mathbf{v}^tA_0\frac{\partial \mathbf{v}}{\partial t} + \mathbf{v}^tA_1\frac{\partial \mathbf{v}}{\partial x} + \mathbf{v}^tA_2\frac{\partial \mathbf{v}}{\partial y} - \mathbf{v}^t\boldsymbol{\mathcal{S}}(\mathbf{v},x,y) = \mathbf{v}^t\left(\frac{\partial \mathbf{u}(\mathbf{v})}{\partial t} + \nabla\cdot\boldsymbol{\mathcal{F}}(\mathbf{v}) - \boldsymbol{\mathcal{S}}(\mathbf{v},x,y)\right) \leqslant 0. \tag{9}$$

Being symmetric, the matrix

$$K_\xi = A_1\xi_1 + A_2\xi_2, \tag{10}$$

has real eigenvalues, and real linearly independent eigenvectors $\forall \xi = (\xi_1,\xi_2) \in \mathbb{R}^2$. The eigenvalues of K_ξ are

$$\lambda_1 = \vec{u}\cdot\vec{\xi}, \quad \lambda_{2,3} = \lambda_1 \pm a\|\vec{\xi}\|, \tag{11}$$

with $a = \sqrt{gH}$. The local *Froude* number defined by the ratio

$$\mathrm{Fr} = \frac{\|\vec{u}\|}{a}, \tag{12}$$

plays the same role as the Mach number in gas dynamics.

2.3. Exact solutions

To simplify the results section, we recall here a number of analytical solutions of the shallow water equations.

Lake at rest solution. This solution is easily obtained assuming $u = v = 0$ and integrating (1) and (2) over an arbitrary control volume $\mathcal{V}$ obtaining

$$\int_{\mathcal{V}} \frac{\partial H}{\partial t}\,dx\,dy = -\oint_{\partial\mathcal{V}} H\vec{u}\cdot\vec{n}\,dl = 0,$$

and similarly

$$\int_{\mathcal{V}} \frac{\partial (H\vec{u})}{\partial t}\,dx\,dy = -\int_{\mathcal{V}} gH\nabla H_{tot}\,dx\,dy.$$

If $H_{tot}(x,y,t=0) = H_0, \forall(x,y) \in \Omega$, from the arbitrariness of $\mathcal{V}$, one gets the exact solution

$$[H_{tot}(x,y,t), u(x,y,t), v(x,y,t)] = [H_0,0,0] \; \forall(x,y) \in \Omega \quad \text{and} \quad t \geqslant 0. \tag{13}$$

Note that this is independent on the shape of $B(x,y)$, as long as ∇H_{tot} is integrable.

A class of 2D potential solutions. In [41] the authors have presented a class of analytical solutions obtained by choosing the velocity vector to be given by a potential ψ: $(-v,u) = \nabla\psi$. A simple steady state for the water height is obtained provided that

$$\Delta\psi = 0 \quad \text{and} \quad u\frac{\partial H}{\partial x} + v\frac{\partial H}{\partial y} = \frac{\partial \psi}{\partial y}\frac{\partial H}{\partial x} - \frac{\partial \psi}{\partial x}\frac{\partial H}{\partial y} = 0.$$

A simple solution is $H = \psi + \alpha$, with α constant. The discharge equation is then exactly satisfied if [41]:

$$B = g^{-1}\left(C - \frac{\|\nabla\psi\|^2}{2}\right) - \psi - \alpha,$$

where C is another constant. We can choose ψ as the real part of a function f, $\psi(x,y) = \operatorname{Re} f(z)$, where f is holomorphic in $z = x + iy$. The reader is referred to [41] and to the results section for an example.

Travelling vortex solutions. In the case $B(x,y) = 0$, another class of solutions can be obtained by setting $\vec{u} = \vec{u}_\infty + \vec{u}'$, with $\vec{u}_\infty$ constant. If, in cylindrical coordinates (r,θ), we set $\vec{u}' = (u'_r, u'_\theta) = (0, u'_\theta(r))$, then:

$$\nabla \cdot \vec{u} = \nabla \cdot \vec{u}' = \frac{1}{r}\frac{\partial(ru'_r)}{\partial r} + \frac{1}{r}\frac{\partial u'_\theta(r)}{\partial \theta} = 0,$$

so that the water height equation becomes

$$\frac{\partial H}{\partial t} + \vec{u}_\infty \cdot \nabla H = 0,$$

with solution $H(x,y,t) = H_0(\xi(x,y,t))$ with $\xi = (x,y) - \vec{u}_\infty t$, and with $H_0(x,y)$ the initial water height profile. For the velocity we have instead

$$\frac{\partial \vec{u}'}{\partial t} + (\vec{u}_\infty \cdot \nabla)\vec{u}' + (\vec{u}' \cdot \nabla)\vec{u}' + g\nabla H_0(\xi) = 0.$$

Clearly, one possible solution is $\vec{u}'(x,y,t) = \vec{u}'(\xi(x,y,t))$, such that $\forall \xi \in \mathbb{R}^2$

$$(\vec{u}'(\xi) \cdot \nabla)\vec{u}'(\xi) + g\nabla H_0(\xi) = 0.$$

Ultimately, we have to choose the initial states of the water height and of the tangential velocity, such that the last equality is verified. A travelling vortex is obtained if one sets (in cartesian coordinates, see also [23]):

$$\vec{u}' = \begin{cases} \Gamma(1 + \cos(\omega r_c))(y_c - y, x - x_c) & \text{if} \quad \omega r_c \leqslant \pi \\ 0 & \text{otherwise} \end{cases},$$

with Γ the vortex intensity parameter, (x_c, y_c) the coordinates of the vortex core, and r_c the distance from the vortex core, and ω an angular wave frequency determining the width of the vortex. By integrating the velocity equation in the radial direction, we obtain for the water height:

$$H_0(r_c) = H_\infty + \begin{cases} \frac{1}{g}\left(\frac{\Gamma}{\omega}\right)^2 (h(\omega r_c) - h(\pi)) & \text{if} \quad \omega r_c \leqslant \pi \\ 0 & \text{otherwise} \end{cases},$$

with

$$h(x) = 2\cos(x) + 2x\sin(x) + \frac{1}{8}\cos(2x) + \frac{x}{4}\sin(2x) + \frac{12}{16}x^2.$$

Other vortex shapes can be obtained by using different definitions of $\vec{u}'$.

Thacker's 2D periodic oscillations. In [52] two classes of exact solutions have been shown, corresponding to nonlinear oscillations in a basin with paraboloid shape:

$$B(x,y) = B(r_c) = -H_0\left(1 - \frac{r_c^2}{a^2}\right),$$

with r_c the distance from the center of the paraboloid, H_0 the height of the center of the basin, and a, a parameter. Two unsteady analytical solutions exist, for which $H(x,y,t) = \max(f(r_c,t) - B(r_c), 0)$.

The first class of solutions describes the oscillations of a planar free surface level for which:

$$f(x,y,t) = \frac{\eta H_0}{a^2}(-\eta + 2(x - x_c)\cos(\omega t) + 2(y - y_c)\sin(\omega t)),$$

with $\omega = \sqrt{2gH_0/a^2}$ the frequency, and η a parameter.

Another set of periodic solutions describes the curved oscillations of the free surface. In this case:

$$f(r_c,t) = H_0\left(-1 + \frac{\sqrt{1-A^2}}{1 - A\cos(\omega t)} - \frac{r_c^2}{a^2}\left(-1 + \frac{1-A^2}{(1 - A\cos(\omega t))^2}\right)\right),$$

with $\omega = \sqrt{8gH_0/a^2}$ the frequency, and, given $r_0 > 0$, A is the shape parameter

$$A = \frac{a^2 - r_0^2}{a^2 + r_0^2}.$$

3. Numerical discretization

We consider here the approximation of weak solutions of the two dimensional system of conservation laws

$$\frac{\partial \mathbf{u}}{\partial t} + \nabla \cdot \mathcal{F}(\mathbf{u}) = 0 \quad \text{on} \quad \Omega \times [0, t_f] \subset \mathbb{R}^2 \times \mathbb{R}, \tag{14}$$

with $\mathbf{u} : \Omega_T \to \mathbb{R}^m$ the vector of conserved quantities, and $\mathcal{F} = [\mathcal{F}_1, \mathcal{F}_2] : \mathbb{R}^m \to \mathbb{R}^{m\times 2}$ the conservative fluxes. The system is supposed to be *hyperbolic*, hence for any $\vec{\xi} = (\xi_1, \xi_2) \in \mathbb{R}^2, \vec{\xi} \neq \vec{0}$ the matrix

$$K(\vec{\xi}, \mathbf{u}) = \frac{\partial \mathcal{F}_1(\mathbf{u})}{\partial \mathbf{u}} \xi_1 + \frac{\partial \mathcal{F}_2(\mathbf{u})}{\partial \mathbf{u}} \xi_2 = \frac{\partial \mathcal{F}(\mathbf{u})}{\partial \mathbf{u}} \cdot \vec{\xi}, \tag{15}$$

admits a full set of real eigenvalues and linearly independent eigenvectors. When necessary to simplify the discussion, we will also consider the scalar advection problem.

$$\frac{\partial u}{\partial t} + \vec{a} \cdot \nabla u = 0. \tag{16}$$

Throughout the text we will make use of bold characters whenever we refer to vector quantities (unknowns, fluxes, etc.), while in the scalar case we shall use small italic symbols.

We discretize the spatial domain Ω by means of an unstructured triangulation denoted by $\mathcal{T}_h$, the parameter h being a reference grid spacing. In each triangle $T \in \mathcal{T}_h$, we denote by $\vec{n}_j$ the local inward normal to the edge facing node j, scaled by the length of the edge (see Fig. 2). The local normals verify

$$\sum\nolimits_{j \in T} \vec{n}_j = 0. \tag{17}$$

On $\mathcal{T}_h$, we will mainly make use of a continuous piecewise linear representation. For example, given the nodal values $\mathbf{u}_i = \mathbf{u}(x_i, y_i)$ we set

$$\mathbf{u}_h(x, y, t) = \sum_{i \in \mathcal{T}_h} \psi_i(x, y) \mathbf{u}_i(t), \tag{18}$$

with

$$\psi_i(x_j.y_j) = \delta_{ij}, \quad \nabla \psi_i|_T = \frac{\vec{n}_i}{2|T|}, \quad \sum_{j \in T} \psi_j = 1, \tag{19}$$

where δ_{ij} is Kronecker's delta. Often, this representation is also used for the flux $\mathcal{F}$.

In time dependent simulations, we break the temporal domain $[0, t_f]$ in a series of discrete intervals $\{[t^n, t^{n+1}]\}_{n=0}^{N-1}$, with $t^0 = 0$ and $t^N = t_f$. We denote by Δt the time step $\Delta t = t^{n+1} - t^n$.

3.1. Residual distribution in the steady case

When seeking steady state solutions of (14), the schemes we consider are based on the computation and splitting of the *local element residuals* $\boldsymbol{\phi}^T$ defined as

$$\boldsymbol{\phi}^T(\mathbf{u}_h) = \oint_{\partial T} \mathcal{F}_h(\mathbf{u}_h) \cdot \hat{n}\, dl \quad \forall T \in \mathcal{T}_h, \tag{20}$$

where $\hat{n}$ is the outward unit normal to ∂T, and with $\mathcal{F}_h$ a continuous discrete approximation of the flux. The nodal values of the unknown are obtained by iterating until steady state the pseudo-time explicit update

$$\mathbf{u}_i^{n+1} = \mathbf{u}_i^n - \frac{\Delta t}{|S_i|} \sum_{T \in D_i} \boldsymbol{\phi}_i^T(\mathbf{u}_h^n), \tag{21}$$

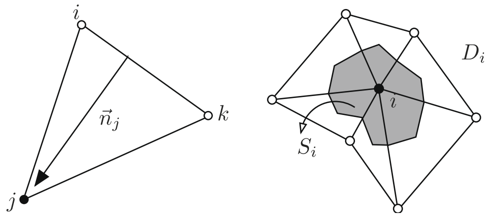


Fig. 2. Nodal normals (left), and median dual cell S_i and nodal stencil D_i (right).

with D_i the stencil of node i (Fig. 2), $|S_i|$ the area of the dual cell of the node i obtained as

$$|S_i| = \sum_{T \in D_i} \frac{|T|}{3}, \tag{22}$$

and where the *local nodal residuals* or *split residuals* $\boldsymbol{\phi}_j^T$ satisfy the "conservation equation"

$$\sum_{j \in T} \boldsymbol{\phi}_j^T(\mathbf{u}_h) = \boldsymbol{\phi}^T(\mathbf{u}_h), \quad \forall\, T \in \mathcal{T}_h. \tag{23}$$

Equivalently, denoting by $\boldsymbol{\beta}_j$ the distribution matrix of local node j, such that

$$\boldsymbol{\phi}_j^T = \boldsymbol{\beta}_j \boldsymbol{\phi}^T, \tag{24}$$

we must have

$$\sum_{j \in T} \boldsymbol{\beta}_j = \mathrm{Id}, \tag{25}$$

Id being the $m \times m$ identity matrix. Note that *the pseudo-time marching iterations* (21) *are ultimately a means of obtaining the approximate solution verifying*

$$\sum_{T \in D_i} \boldsymbol{\phi}_i^T(\mathbf{u}_h) = 0. \tag{26}$$

3.2. Residual distribution for time dependent problems

For time dependent computations, schemes like (21) are, in general, not suitable. Consistent formulations are suggested in [6,17,22,45]. Here, following [6,42], we compute a *space–time* residual defined as

$$\boldsymbol{\Phi}^T(\mathbf{u}_h) = \int_{t^n}^{t^{n+1}} \int_T \left(\frac{\partial \mathbf{u}_h}{\partial t} + \nabla \cdot \mathcal{F}_h \right) dx\,dy\,dt. \tag{27}$$

In particular, assuming a linear variation in time of the fluxes, we can write

$$\boldsymbol{\Phi}^T(\mathbf{u}_h) = \int_T (\mathbf{u}_h^{n+1} - \mathbf{u}_h^n)\,dx\,dy + \frac{\Delta t}{2}\left(\oint_{\partial T} \mathcal{F}_h^{n+1} \cdot \hat{n}\,dl + \oint_{\partial T} \mathcal{F}_h^n \cdot \hat{n}\,dl \right). \tag{28}$$

If we take $\mathbf{u}_h$ linear in space, we finally have

$$\boldsymbol{\Phi}^T = \frac{|T|}{3} \sum_{j \in T} (\mathbf{u}_j^{n+1} - \mathbf{u}_j^n) + \frac{\Delta t}{2}\left(\boldsymbol{\phi}^T(\mathbf{u}_h^{n+1}) + \boldsymbol{\phi}^T(\mathbf{u}_h^n) \right). \tag{29}$$

In every slab $\Omega \times [t^n, t^{n+1}]$, given $\mathbf{u}_h^n$ we compute the nodal values of $\mathbf{u}_h^{n+1}$ from the algebraic system

$$\sum_{T \in D_i} \boldsymbol{\Phi}_i^T(\mathbf{u}_h) = 0 \quad \forall T \in \mathcal{T}_h, \tag{30}$$

where the Φ_i^Ts define some splitting of $\Phi^T(\mathbf{u}_h)$, that is

$$\sum_{j \in T} \boldsymbol{\Phi}_j^T(\mathbf{u}_h) = \boldsymbol{\Phi}^T(\mathbf{u}_h). \tag{31}$$

As in the steady case, when possible, we denote by $\boldsymbol{\beta}_j$ the distribution matrix of local node j, such that

$$\boldsymbol{\Phi}_j^T = \boldsymbol{\beta}_j \boldsymbol{\Phi}^T, \quad \sum_{j \in T} \boldsymbol{\beta}_j = \mathrm{Id}. \tag{32}$$

3.3. Examples of linear first order positive schemes

3.3.1. The N scheme

When solving the steady limit of the scalar advection equation (16), one easily shows (with the notation of equation (17), see also Fig. 2)

$$\phi^T = \sum_{j \in T} k_j u_j, \quad k_j = \frac{\vec{a} \cdot \vec{n}_j}{2}.$$

The N scheme reads

$$\phi_i^{\mathrm{N}} = k_i^+(u_i - u_{in}), \tag{33}$$

with $u_{in} = -N\sum_{j\in T} k_j^- u_j$, being $N = 1/\sum_{j\in T} k_j^+$. The N scheme is a multidimensional upwind and positive scheme [20]. In particular, one can write

$$\phi_i^{\mathrm{N}} = -\sum_{\substack{j\in T\\ j\neq i}} k_i^+ N k_j^- (u_j - u_k) = \sum_{\substack{j\in T\\ j\neq i}} c_{ij}^T (u_i - u_j), \tag{34}$$

with $c_{ij}^T = -k_i^+ N k_j^- \geqslant 0$. When combined with a positivity preserving time integration scheme this leads to a discrete maximum principle for the numerical solution [20].

For hyperbolic systems, one can define a matrix N scheme [53] as follows:

$$\boldsymbol{\phi}_i^{\mathrm{N}} = K_i^+ (\mathbf{u}_i - \mathbf{u}_{in}), \tag{35}$$

where the matrix K_i is defined as

$$K_i = \frac{1}{2}\frac{\partial \boldsymbol{\mathcal{F}}(\bar{\mathbf{u}})}{\partial \mathbf{u}} \cdot \vec{n}_i, \tag{36}$$

for some (arbitrary) locally linearized state $\bar{\mathbf{u}}$, and with K_i^+ computed as usual via eigenvalue decomposition. Here, following [18], the inflow state vector $\mathbf{u}_{in}$ is computed such that the scheme is always conservative:

$$\mathbf{u}_{in} = -N\left(\sum_{j\in T} K_j^+ \mathbf{u}_j - \boldsymbol{\phi}^T\right), \tag{37}$$

with the matrix N defined as

$$N = \left(\sum_{j\in T} K_j^+\right)^{-1}. \tag{38}$$

For linear symmetric systems, the N scheme is energy stable [3]. Concerning its positivity, a simple-wave analysis has been proposed in [7] to justify the absence of spurious numerical oscillations when using the matrix N scheme.

For time dependent problems, as in [41,42], we use, in conjunction with splitting (35) of the spatial residual, the following splitting of (28)

$$\boldsymbol{\Phi}_i^{\mathrm{N}} = \frac{|T|}{3}(\mathbf{u}_i^{n+1} - \mathbf{u}_i^n) + \frac{\Delta t}{2}\left(\boldsymbol{\phi}_i^{\mathrm{N}}(\mathbf{u}_h^{n+1}) + \boldsymbol{\phi}_i^{\mathrm{N}}(\mathbf{u}_h^n)\right). \tag{39}$$

As remarked in [41], this corresponds to the combination of the N scheme in space with second order trapezium rule integration in time. In the case of linear scalar advection, the solution obtained with this scheme verifies a discrete maximum principle under a constraint on the size of the time step [6,20].

The matrix N scheme is at most first order accurate and yields non-oscillatory solutions in all practical applications.

3.3.2. The Lax–Friedrichs scheme

For steady scalar advection, the Lax–Friedrichs scheme is defined as

$$\phi_i^{\mathrm{LF}} = \frac{1}{3}\phi^T + \frac{1}{3}\alpha \sum_{\substack{j\in T\\ i\neq j}} (u_i - u_j). \tag{40}$$

This scheme is a two dimensional generalization of the 1D Lax–Friedrichs one. We have

$$\phi_i^{\mathrm{LF}} = \frac{1}{3}\sum_{j\in T} k_j u_j + \frac{1}{3}\alpha \sum_{\substack{j\in T\\ k\neq i}} (u_i - u_j) = \sum_{\substack{j\in T\\ j\neq i}} c_{ij}^T (u_i - u_j),$$

with $c_{ij}^T = (\alpha - k_j)/3 \geqslant 0$, provided that $\alpha \geqslant \max_{j\in T}|k_j| > 0$. In this case, when combined with a positivity preserving time integration scheme, the LF scheme leads to solutions enjoying a discrete maximum principle [20]. For hyperbolic systems, the LF scheme reads

$$\boldsymbol{\phi}_i^{\mathrm{LF}} = \frac{1}{3}\boldsymbol{\phi}^T + \frac{1}{3}\alpha \sum_{\substack{j\in T\\ j\neq i}} (\mathbf{u}_i - \mathbf{u}_j), \tag{41}$$

where if $\rho(\cdot)$ denotes the spectral radius of a matrix, α is normally chosen as [2]

$$\alpha \geqslant \max_{j\in T} \rho(K_j). \tag{42}$$

As for the N scheme, in the time dependent case we define a LF splitting given by

$$\boldsymbol{\Phi}_i^{\mathrm{LF}} = \frac{|T|}{3}(\mathbf{u}_i^{n+1} - \mathbf{u}_i^n) + \frac{\Delta t}{2}\left(\boldsymbol{\phi}_i^{\mathrm{LF}}(\mathbf{u}_h^{n+1}) + \boldsymbol{\phi}_i^{\mathrm{LF}}(\mathbf{u}_h^n)\right), \tag{43}$$

corresponding to the combination of the LF scheme in space with second order trapezium rule integration in time. In the case of linear scalar advection, the solution obtained with this scheme verifies a discrete maximum principle under a constraint on the size of the time step [20].

The LF scheme is at most first order. In the case of the SWE, we will show later that it preserves the positivity of the water height.

3.3.3. Limited nonlinear schemes

Nonlinear schemes are needed to combine a non-oscillatory behavior of the discrete solution (and eventually positivity preservation) and higher accuracy. In the $\mathcal{RD}$ framework, the accuracy of the discretization can be formally characterized by means of a truncation error analysis initially proposed in [1]. We omit here the details of this analysis, for which we refer to the above mentioned reference and to [8,20,41,44], and limit ourselves to mention the two results that are of interest here.

The first is that a truncation error of the type $\mathcal{E} = Ch^2$ will be obtained provided that the split residuals satisfy an estimate of the type (in two space dimensions)

$$\|\boldsymbol{\phi}_i^T\| \leqslant Ch^3 \quad \forall i \quad \text{and} \quad \forall T,$$

in the steady case, and

$$\|\boldsymbol{\Phi}_i^T\| \leqslant Ch^4 \quad \forall i \quad \text{and} \quad \forall T \quad \text{and} \quad \forall [t^n, t^{n+1}],$$

in the time dependent case. In particular, these estimates are always verified by Petrov–Galerkin schemes in the form

$$\boldsymbol{\phi}_i^T = \int_T \omega_i^T \nabla \cdot \boldsymbol{\mathcal{F}}_h(\mathbf{u}_h)\, dx\, dy,$$

and

$$\boldsymbol{\Phi}_i^T = \int_{t^n}^{t^{n+1}} \int_T \omega_i^T \left(\frac{\partial \mathbf{u}_h}{\partial t} + \nabla \cdot \boldsymbol{\mathcal{F}}_h(\mathbf{u}_h) \right) dx\, dy\, dt,$$

in the time dependent case, where the ω_i^Ts are a set of uniformly bounded test functions such that

$$\sum_{j \in T} \omega_j^T = \text{Id}.$$

As a particular case, schemes that read

$$\boldsymbol{\phi}_i^T = \boldsymbol{\beta}_i^T \boldsymbol{\phi}^T \quad (\boldsymbol{\Phi}_i^T = \boldsymbol{\beta}_i^T \boldsymbol{\Phi}^T \text{ in the time dependent case}),$$

with *uniformly bounded distribution matrices* $\boldsymbol{\beta}_i^T$, are formally second order accurate. Schemes that belong to this class are often said to be *linearity preserving* or $\mathcal{LP}$.

In this study, we adopt the construction proposed in [7,6]. The idea is to start with a linear first order positivity preserving scheme, and to devise a way to map its local residuals onto a set of nonlinear *positivity and linearity preserving* residuals. Note that by positivity preserving scheme here we intend one that, for scalar advection and in the steady case, can be written as

$$\phi_i^T = \sum_{j \in T} c_{ij}^T (u_i - u_j) \quad \text{with} \quad c_{ij} \geqslant 0,$$

and similarly in the time dependent case (see [20] for more details). The reader is referred to [1,22,23,33] and references therein for a review of other techniques.

In the scalar case, the *limiting technique* consists in mapping the distribution coefficients β_j^{P} of a positivity preserving scheme onto nonlinear bounded distribution coefficients β_j^*. The nonlinear mapping should verify the following properties:

$$|\beta_j^*| < C < \infty, \quad j = 1, \ldots, 3 \tag{44}$$

$$\beta_j^{\mathrm{P}} = 0 \Rightarrow \beta_j^* = 0, \quad j = 1, \ldots, 3 \tag{45}$$

$$\beta_j^{\mathrm{P}} \beta_j^* \geqslant 0, \quad j = 1, \ldots, 3 \tag{46}$$

$$\sum_{j \in T} \beta_j^* = 1 \tag{47}$$

Clearly, (44) is the $\mathcal{LP}$ condition, while (45) and (46) ensure the positivity preservation property. For example, in the steady case we can write

$$\phi_i^* = \frac{\phi_i^*}{\phi} \frac{\phi}{\phi_i^{\mathrm{P}}} \phi_i^{\mathrm{P}} = \frac{\beta_i^*}{\beta_i^{\mathrm{P}}} \sum_{\substack{j \in T \\ j \neq i}} c_{ij} (u_i - u_j) = \sum_{\substack{j \in T \\ j \neq i}} c_{ij}^* (u_i - u_j),$$

with $c_{ij}^* = \frac{\beta_i^*}{\beta_i^{\mathrm{P}}} c_{ij}$. Since $c_{ij} \geqslant 0$, then (46) guarantees $c_{ij}^* \geqslant 0$, hence the positivity of the resulting scheme.

Concerning the mapping, a common choice, which is adopted here too, is to use the so-called PSI limiter

$$\beta_i^* = \frac{\max(\beta_i^{\mathrm{P}}, 0)}{\sum_{j \in T} \max(\beta_j^{\mathrm{P}}, 0)}. \tag{48}$$

For scalar steady problems, this limiter has been very successful when applied to the N scheme, in which case we recover the PSI scheme of Struijs [48,38]. In this case, here we will speak of the limited N (LN) scheme. We will instead refer to the limited Lax–Friedrichs (LLF) scheme, as to the one obtained by limiting (40). Note that, when using (48), if $\beta_j^{\mathrm{P}} \leqslant 0$, then $\beta_j^* = 0$, otherwise one easily shows that $\beta_j^* \leqslant \beta_j^{\mathrm{P}}$, hence

$$\gamma_j^* = \frac{\beta_j^*}{\beta_j^{\mathrm{P}}} \in [0, 1]. \tag{49}$$

For systems, as in [7,6], we decompose the residual on the basis of the solution space given by the eigenvectors of the flux Jacobian. On each triangle, fixed a direction $\vec{\xi}$, we compute $\{\mathbf{l}_\sigma\}_{\sigma=1}^m$ and $\{\mathbf{r}_\sigma\}_{\sigma=1}^m$, the left and right eigenvectors of $K_j(\vec{\xi}, \bar{\mathbf{u}})$ (cf. Eq. (15)), evaluated on $\bar{\mathbf{u}}$, the locally averaged state of $\mathbf{u}_h$. We then project the residuals on $\{\mathbf{l}_\sigma\}_{\sigma=1}^m$:

$$\varphi_\sigma^T = \mathbf{l}_\sigma \boldsymbol{\phi}^T \quad \text{and} \quad \varphi_{\sigma,j}^{\mathrm{P}} = \mathbf{l}_\sigma \boldsymbol{\phi}_j^{\mathrm{P}}. \tag{50}$$

Each component σ is now treated as a scalar residual. We compute $\beta_{\sigma,j}^{\mathrm{P}} = \varphi_{\sigma,j}^{\mathrm{P}} / \varphi_\sigma^T$, and use (48) to get nonlinear coefficients $\beta_{\sigma,j}^*$. Finally, we set

$$\varphi_{\sigma,j}^* = \beta_{\sigma,j}^* \phi_\sigma^T \quad \text{and} \quad \boldsymbol{\phi}_j^* = \sum_\sigma \varphi_{\sigma,j}^* \mathbf{r}_\sigma. \tag{51}$$

We use the same construction in the time dependent case, obviously replacing $\boldsymbol{\phi}^T$, and $\boldsymbol{\phi}_j^{\mathrm{P}}$ by $\boldsymbol{\Phi}^T$, and $\boldsymbol{\Phi}_j^{\mathrm{P}}$.

The resulting scheme is $\mathcal{LP}$ by construction. Its stability on simple-waves is studied in [7]. We will show later that, in the case of the SWE, this construction allows to build discretizations preserving the positivity of the relative water height.

In this paper, *the direction $\vec{\xi}$ needed for this construction is always taken to be the direction of the local velocity vector $\vec{u}$, computed in each element from a locally averaged state of the conservative variables* $\mathbf{u}$.

3.3.4. Stability and dissipation

Stability and convergence proofs are often based on bounds on the L^2 norm of the solution (linear problems), or on entropy inequalities (such as (4)) [12,28,50]. The attempt to formulate discrete variants of these stability criteria for $\mathcal{RD}$ has not been very fruitful up to now, and few results exist (see *e.g.* [3]).

The relevant issue is really making sure of the existence and uniqueness of the discrete solution, and of its convergence with the mesh parameter h. This is ultimately linked to the properties of the discrete algebraic Eq. (26) (or (30)). Consider for example the case of the steady scalar advection equation. One way to look at the problem is that if we can rewrite the steady discrete Eq. (26) as linear algebraic system[1]

$$A_h u_h = f, \tag{52}$$

we should be sure that the matrix A_h is invertible. This issue is studied in [2]. For positivity preserving multidimensional upwind $\mathcal{RD}$ schemes such as the N and the LN schemes, in the reference it is shown that the matrix A_h admits a block triangular decomposition, and that each block $A_{\xi\eta}$ is invertible. To generalize the analysis to non-upwind discretizations such as for example the LLF scheme, in [2] it is suggested to replace (52) with the iterative update

$$u^{n+1} = u^n - \omega(A_h u^n - f) \quad \text{with} \quad \omega \in \mathbb{R}^+.$$

This procedure will converge if, for some $0 < r < 1$, one has

$$\|(Id - \omega A_h)v\|^2 \leqslant r\|v\|^2,$$

for any arbitrary $v \in \mathbb{R}^M$, and M denoting the total number of unknowns. Developing the last expression, one ends up with the requirement

$$v^t A_h v \geqslant \frac{1-r}{2\omega}\|v\|^2 + \frac{\omega}{2}\|A_h v\|^2 > C_h\|v\|^2 \geqslant 0,$$

which brings us back to the necessity showing the coercivity of the discretization, and/or of establishing a L^2 stability estimate of the type $v^t A_h v > 0$.

For linear first order $\mathcal{RD}$ schemes such as the N and LF ones, this estimate can be easily demonstrated [3,20]. The situation is less clear for the limited schemes. In [11], for example, some sources of instability (in the L^2 sense seen above) related to the limiting process are pointed out for the scalar LN scheme. However, this scheme yields in practice good iterative and

[1] At least for smooth solutions, eventually by means of a linearization of the nonlinear algebraic system.

grid convergence. This is in line with the result of [2]: scalar multidimensional upwind schemes lead to well posed algebraic equations.

The LLF scheme, instead, shows in practice poor convergence. This is observed especially on smooth problems, where one obtains wiggly numerical solutions, symptom of the presence of undamped spurious modes.

To cure this problem, we follow [2]. The idea is to add a stabilizing upwind bias to the discretization, by means of a streamline dissipation term:

$$\phi_i^{*s} = \beta_i^*\phi^T + \theta(\mathcal{T}_h, u_h)\int_T (\vec{a}\cdot\nabla\psi_i)(\vec{a}\cdot\nabla u)\,dx\,dy = \beta_i^*\phi^T + \frac{\theta(\mathcal{T}_h, u_h)}{|T|}k_i\phi^T, \tag{53}$$

where the additional superscript s stands for stabilized. The rationale for this modification is that the energy production associated to the "stabilized" scheme reads now (cf. analysis above, and see [2,20] for details)

$$\begin{aligned} v^tAv &= \sum_{T\in\mathcal{T}_h}\sum_{j\in T} v_j\left(\beta_j^*\phi^T + \theta(\mathcal{T}_h, v_h)\int_T (\vec{a}\cdot\nabla\psi_j)(\vec{a}\cdot\nabla v_h)\,dx\,dy\right) \\ &= \int_{\partial\Omega}\frac{v_h^2}{2}|\vec{a}\cdot\hat{n}|\,dl + \sum_{T\in\mathcal{T}_h}\int_T (h_T(\vec{\xi}^*\cdot\nabla v_h)(\vec{a}\cdot\nabla v_h) + \theta(\mathcal{T}_h, v_h)(\vec{a}\cdot\nabla v_h)^2), \end{aligned} \tag{54}$$

where as we shall see shortly, the vector $\vec{\xi}^*$ depends on the distance between the nodes of T and its gravity center, and on the β_j^*s. Independently on this, the important point is that now, for some definition of $\theta(\mathcal{T}_h, u_h) \geqslant 0$, the condition $v^tAv > 0$ will be verified.

The parameter $\theta(\mathcal{T}_h, u_h)$ is introduced for two reasons. One is to provide a correct scaling of the streamline dissipation term with respect to mesh size and advection speed, so that the additional term has the same dimensions as the element residual ϕ^T (and equivalently of the limited nonlinear residual $\beta_i^*\phi^T$). The other is to make sure that the additional term is only added in correspondence of smooth regions of the solution. For this reason, from now on we shall write

$$\theta(\mathcal{T}_h, u_h) = \tau(\mathcal{T}_h)\epsilon(u_h),$$

where $\tau(\mathcal{T}_h)$ is basically the standard streamline dissipation stabilization parameter, and $\epsilon(u_h)$ is the smoothness sensor. Definitions for these parameters will be given in Section 5.

In the time dependent case, following the initial developments of [40], we use a similar technique. To illustrate how we proceed, let us assume to be only interested in smooth solutions, so that we can work with a locally linearized constant coefficients quasi-linear problem. To devise a consistent modification of the mass matrix, we make use of a well-known analogy between $\mathcal{RD}$ and Petrov–Galerkin. Other formulations might however be thought of, as the geometrical analysis of [22] shows. Different ways exist to present the analogy between $\mathcal{RD}$ and finite elements (see *e.g.* [35,24,6,43]). In the simplest setting, one recasts a linearity preserving scheme as a perturbation of the Galerkin finite element scheme:

$$\sum_{T\in D_i}\beta_i\phi^T = \int_\Omega \psi_i\vec{a}\cdot\nabla u_h\,dx\,dy + \sum_{T\in\mathcal{T}_h}\int_T \delta_i^T\vec{a}\cdot\nabla u_h\,dx\,dy = 0,$$

with $\delta_i^T = \beta_i - 1/3$ if $i\in T$, $\delta_i^T = 0$ otherwise. This corresponds to chose as a test function the quantity

$$\omega_i^T = \psi_i + \delta_i^T,$$

with ψ_i the basis function (19). Whenever $\beta_i \geqslant 0\ \forall i\in T$, due to the properties of the linear basis functions (19), one can find a unique point $P_T\in T$ such that $\forall j\in T$ we have $\psi_j(P_T) = \beta_j$. Denoting by G_T the gravity center of the element, we can use the linearity of ψ_i to re-write the Petrov–Galerkin scheme as

$$\int_\Omega \psi_i\vec{a}\cdot\nabla u_h\,dx\,dy + \sum_{T\in\mathcal{T}_h} h_T\int_T \vec{\xi}\cdot\nabla\psi_i\vec{a}\cdot\nabla u_h\,dx\,dy = 0,$$

with h_T a local mesh size, and $h_T\vec{\xi} = P_T - G_T$. This is equivalent to rewriting the test function as

$$\omega_i^T = \psi_i + h_T\vec{\xi}\cdot\nabla\psi_i,$$

which now closely reminds of the SUPG test function. However, let now $\vec{\xi}^*$ be the direction giving the distribution bias corresponding to a limited scheme (cf. Eq. (54)). Since the limiting process does not guarantee any control over the location of P_T, the vector $\vec{\xi}^*$ generally does not introduce a bias in the streamline direction, that is $\vec{\xi}^*$ is not necessarily in the direction of the propagation speed $\vec{a}$. In the case of the LLF scheme, $\vec{\xi}^*$ is likely to have the direction of the largest component of the solution gradient, that is (for steady advection) the cross-wind direction. It is not unlikely, however, that $\vec{\xi}^*$ might even point upstream. This leads to poor stability. The cure proposed in [2] restores the correct direction in the bias of the discretization.

To go to the time dependent case, we first discretize in time with the trapezium scheme:

$$\frac{2}{\Delta t}u^{n+1} + \vec{a}\cdot\nabla u^{n+1} = \frac{2}{\Delta t}u^n - \vec{a}\cdot\nabla u^n.$$

Being u^n known, we can work with the model equation

$$\gamma u + \vec{a}\cdot\nabla u = S(x, y),$$

with $\gamma \geqslant 0$. We apply the Petrov–Galerkin formulation to this non-homogeneous advection–reaction problem to get, neglecting the contribution of the source term on the right hand side, and after some manipulations:

$$\sum_{T \in D_i} \Phi_i = 0,$$

where

$$\Phi_i = \beta_i \int_T (\gamma u_h + \vec{a} \cdot \nabla u_h) dx\, dy + \int_T \left(\psi_i - \frac{1}{3}\right) \gamma u_h \, dx\, dy + \int_T \left(\psi_i - \frac{1}{3}\right) \vec{a} \cdot \nabla u_h \, dx\, dy.$$

However, for a constant $\vec{a}$, the last term vanishes due to the linear variation of u_h, and to the relation

$$\int_T \left(\psi_i - \frac{1}{3}\right) dx\, dy = 0. \tag{55}$$

Ultimately we get

$$\Phi_i = \beta_i \int_T (\gamma u_h + \vec{a} \cdot \nabla u_h) dx\, dy + \int_T \left(\psi_i - \frac{1}{3}\right) \gamma u_h \, dx\, dy.$$

Taking now $\beta_i = \beta_i^*$, given by a limited scheme, and adding the streamline stabilization term, one ends with

$$\Phi_i^{*s} = \left(\beta_i^* + \frac{\tau(\mathcal{T}_h)\epsilon(u_h)}{|T|} k_i\right) \int_T (\gamma u_h + \vec{a} \cdot \nabla u_h) dx\, dy + \int_T \left(\psi_i - \frac{1}{3}\right) \gamma u_h \, dx\, dy.$$

In the last expression, we can identify the contribution of the nonlinear limited scheme, plus the streamline upwind bias, plus the last term that can be recast as

$$\int_T \left(\psi_i - \frac{1}{3}\right) \gamma u_h \, dx\, dy = \gamma \frac{|T|}{36} \sum_{j \in T} D_{ij} u_j,$$

where the matrix D is symmetric positive semi-definite and given by

$$D = \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix}.$$

The additional energy production due to this term reads (cf. analysis above, and see [2,20] for details)

$$v^t A_D v = \sum_{T \in \mathcal{T}_h} \gamma \frac{|T|}{36} \sum_{j \in T} v_j \sum_{k \in T} D_{jk} v_k = \sum_{T \in \mathcal{T}_h} \gamma \frac{|T|}{36} \frac{1}{2} \sum_{\substack{i,j \in T \\ i \neq j}} (v_i - v_j)^2 \geqslant 0,$$

which shows its dissipative character.

We now go back to our original purpose of stabilizing the nonlinear limited scheme in the time dependent case. Firstly we note that the previous analysis only applies in smooth areas of the solution. This means that the additional stabilization operator should also be premultiplied by $\epsilon(u_h)$. Repeating the development for the time dependent advection equation, and including this last remark, we end up with

$$\Phi_i^{*s} = \beta_i^* \Phi^T + \epsilon(u_h) \left(\frac{k_i}{|T|} \tau(\mathcal{T}_h) \Phi^T + \frac{|T|}{36} \sum_{j \in T} D_{ij} (u_j^{n+1} - u_j^n) \right), \tag{56}$$

where for $\epsilon(u_h) = 1$ he get back exactly the Petrov–Galerkin scheme applied to the time dependent equation. In general, $\epsilon(u_h)$ and $\tau(\mathcal{T}_h)$ being constant in each element, we can write for a locally linear (or linearized) problem

$$\Phi_i^{*s} = \int_{t^n}^{t^{n+1}} \int_T \omega_i^T \left(\frac{\partial u_h}{\partial t} + \vec{a} \cdot \nabla u_h \right) dx\, dy\, dt \quad \text{with} \quad \omega_i^T = \beta_i^* + \epsilon(u_h) \left(\vec{a} \cdot \nabla \psi_i \tau(\mathcal{T}_h) + \psi_i - \frac{1}{3} \right), \tag{57}$$

which, even though not in the form of a linearity preserving scheme, does verify the second-order of accuracy conditions for time dependent problems [41,44] thanks to the uniform boundedness of ω_i^T (cf. Section 3.3.3).

Concerning the case of a hyperbolic system, to be rigorous the additional terms should be evaluated in terms of the quasi-linear form in entropy variables in order to actually yield a meaningful dissipative operator (cf. Section 2.2 and [12]). However, in our case these terms are actually active only in smooth regions of the solution. Moreover, their definition is such that they *do not* influence the conservative character of the discretization. Indeed, using (19) and (17), we immediately get:

$$\sum_{j \in T} \int_T \vec{a} \cdot \nabla \psi_j \tau(\mathcal{T}_h) \left(\frac{\partial u}{\partial t} + \vec{a} \cdot \nabla u_h \right) dx\, dy = 0,$$

and

$$\sum_{j\in T}\int_T\left(\psi_j-\frac{1}{3}\right)\left(\frac{\partial u}{\partial t}+\vec{a}\cdot\nabla u_h\right)dx\,dy=0.$$

This gives some freedom in the choice of the approximation of these terms for systems. In particular, we can evaluate them on a locally linearized quasi-linear form of the equations, in a chosen set of variables (see also the discussion in [30,29]). This corresponds to the evaluation of the streamline dissipation integral with a one point quadrature formula, which, as the analysis made in [5] shows, is sufficient to yield a well posed set of algebraic equations in the case of second order schemes. With this choice we obtain

$$\epsilon(\mathbf{u}_h)\int_T\frac{\partial\mathcal{F}(\bar{\mathbf{u}})}{\partial\mathbf{u}}\cdot\nabla\psi_j\;\boldsymbol{\tau}(\mathcal{T}_h)\frac{\partial\mathcal{F}(\bar{\mathbf{u}})}{\partial\mathbf{u}}\cdot\nabla\mathbf{u}_h\,dx\,dy=\epsilon(\mathbf{u}_h)\frac{K_i}{|T|}\boldsymbol{\tau}(\mathcal{T}_h)\boldsymbol{\phi}^{\mathrm{nc}},$$

in the steady case, while in the time dependent case we have

$$\begin{aligned}\epsilon(\mathbf{u}_h)\quad&\int_{t^n}^{t^{n+1}}\int_T\left(\tfrac{\partial\mathcal{F}(\bar{\mathbf{u}})}{\partial\mathbf{u}}\cdot\nabla\psi_j\boldsymbol{\tau}(\mathcal{T}_h)+\psi_i\mathbf{I}-\tfrac{\mathbf{I}}{3}\right)\left(\tfrac{\partial\mathbf{u}_h}{\partial t}+\tfrac{\partial\mathcal{F}(\bar{\mathbf{u}})}{\partial\mathbf{u}}\cdot\nabla\mathbf{u}_h\right)dx\,dy\,dt=\\ \epsilon(\mathbf{u}_h)\quad&\left(\tfrac{K_i}{|T|}\boldsymbol{\tau}(\mathcal{T}_h)\boldsymbol{\Phi}^{\mathrm{nc}}+\tfrac{|T|}{36}\sum_{j\in T}D_{ij}(\mathbf{u}_j^{n+1}-\mathbf{u}_j^n)\right),\end{aligned}$$

with the non-conservative residuals

$$\begin{aligned}\boldsymbol{\phi}^{\mathrm{nc}}=&\quad\int_T\tfrac{\partial\mathcal{F}(\bar{\mathbf{u}})}{\partial\mathbf{u}}\cdot\nabla\mathbf{u}_h\,dx\,dy=\sum_{j\in T}K_j(\bar{\mathbf{u}})\mathbf{u}_j\\ \boldsymbol{\Phi}^{\mathrm{nc}}=&\quad\int_{t^n}^{t^{n+1}}\int_T\left(\tfrac{\partial\mathbf{u}_h}{\partial t}+\tfrac{\partial\mathcal{F}(\bar{\mathbf{u}})}{\partial\mathbf{u}}\cdot\nabla\mathbf{u}_h\right)dx\,dy=\tfrac{|T|}{3}\sum_{j\in T}(\mathbf{u}_j^{n+1}-\mathbf{u}_j^n)+\tfrac{\Delta t}{2}(\boldsymbol{\phi}^{\mathrm{nc}}(\mathbf{u}_h^n)+\boldsymbol{\phi}^{\mathrm{nc}}(\mathbf{u}_h^{n+1})),\end{aligned}$$

$\bar{\mathbf{u}}$ being a local (constant) average state of the state vector $\mathbf{u}$. Note that now $\boldsymbol{\tau}(\mathcal{T}_h)$ is in general a stabilization matrix. In practice, the additional terms become active only in smooth regions of the solution, we have simplified things by replacing the non-conservative residuals $\boldsymbol{\phi}^{\mathrm{nc}}$ with the conservative local approximations $\boldsymbol{\phi}^T$. In the steady case, this leads to the stabilized nonlinear schemes defined by:

$$\boldsymbol{\phi}_i^{*s}=\boldsymbol{\beta}_i^*\boldsymbol{\phi}^T+\epsilon(\mathbf{u}_h)\frac{K_i}{|T|}\boldsymbol{\tau}(\mathcal{T}_h)\boldsymbol{\phi}^T,\tag{58}$$

where the superscript s stands for stabilized. For time dependent simulations we use instead:

$$\boldsymbol{\Phi}_i^{*s}=\boldsymbol{\beta}_i^*\boldsymbol{\Phi}^T+\epsilon(\mathbf{u}_h)\left(\frac{K_i}{|T|}\boldsymbol{\tau}(\mathcal{T}_h)\boldsymbol{\Phi}^T+\frac{|T|}{36}\sum_{j\in T}D_{ij}(\mathbf{u}_j^{n+1}-\mathbf{u}_j^n)\right).\tag{59}$$

Once more we recall that in these formulas the stabilization terms arise from the application of the Petrov–Galerkin analogy (57) to a locally linearized (constant coefficients) quasi-linear form of the system.

Note that, as shown in [2], and as we shall see in some of the numerical tests, for systems also the matrix limited N scheme is subject to the same stability problems of the LLF scheme, and needs the addition of the stabilization operators. In the following, *we will refer to the limited and stabilized schemes used in this paper as to the LNs and LLFs schemes.*

As a last remark, we note that including the stabilization terms leads to the loss of formal monotonicity (*viz.* positivity preservation). Numerical results show an essentially non-oscillatory behavior, with very small oscillations across discontinuities [2,40]. To minimize this side effect, the parameter $\epsilon(\mathbf{u}_h)$ is chosen such that $\epsilon(\mathbf{u}_h)\approx h$ in discontinuities, while $\epsilon(\mathbf{u}_h)\approx 1$ elsewhere. Details of how to compute $\epsilon(\mathbf{u}_h)$ and $\boldsymbol{\tau}(\mathcal{T}_h)$ will be given in Section 5. To simplify the notation, in the following we will omit the dependence of these parameters on solution and mesh, simply writing $\epsilon(\mathbf{u}_h)=\epsilon_h$, and $\boldsymbol{\tau}(\mathcal{T}_h)=\boldsymbol{\tau}_h$.

3.3.5. *Obtaining the solution*

For the positivity analysis of Section 4.2, it is useful to introduce now the technique used to get the nodal values of the discrete solution. As anticipated in Section 3.1, we employ an explicit pseudo-time iterative technique. In the steady case, given the nodal values of the initial solution $\mathbf{u}_i^0$, we set $\forall p\geqslant 0$

$$\boldsymbol{\phi}^p=\boldsymbol{\phi}^T(\mathbf{u}^p),$$

with p the pseudo-time step number, and then we repeat

$$\mathbf{u}_i^{p+1}=\mathbf{u}_i^p-\frac{\Delta s}{|S_i|}\sum_{T\in D_i}\boldsymbol{\phi}_i^p,\tag{60}$$

where Δs is the pseudo time step, and the $\boldsymbol{\phi}_i^p$s representing the local splitting of $\boldsymbol{\phi}^p$. We continue this procedure until we have convergence in some norm of the residual.

In the time dependent case, we use the same technique. At each time step n, we set $\mathbf{u}_i^{n+1,0} = \mathbf{u}_i^n$ and

$$\mathbf{\Phi}^p = \mathbf{\Phi}^T(\mathbf{u}^{n+1,p}, \mathbf{u}^n),$$

with p the pseudo time step number. Now we repeat

$$\mathbf{u}_i^{n+1,p+1} = \mathbf{u}_i^{n+1,p} - \frac{\Delta s}{|S_i|} \sum_{T \in D_i} \mathbf{\Phi}_i^p, \tag{61}$$

where Δs is the pseudo time step, and the $\mathbf{\Phi}_i^p$s representing the local splitting of $\mathbf{\Phi}^p$. We continue this procedure until we have convergence in some norm of the residual (see Section 5 for more details).

4. Application to the SWE

To apply the schemes discussed in Section 3 to the SWE we have to take into account the source term modeling the variation of the bathymetry. The analysis reported in [20,41] shows that, in general, central and pointwise treatments of source terms lead to a loss of accuracy. Consistent discretizations are instead obtained by introducing the source term in the definition of the cell-residual. So, when seeking a steady solution of the SWE, we define the element residual $\boldsymbol{\phi}^T$ as

$$\boldsymbol{\phi}^T(\mathbf{u}_h) = \int_T (\nabla \cdot \mathcal{F}_h(\mathbf{u}_h) - \mathcal{S}_h(\mathbf{u}_h, x, y)) dx\, dy = \oint_{\partial T} \mathcal{F}_h(\mathbf{u}_h) \cdot \hat{n}\, dl - \int_T \mathcal{S}_h(\mathbf{u}_h, x, y) dx\, dy. \tag{62}$$

Similarly, in the time dependent case, the *space–time* residual $\Phi^T(\mathbf{u}_h)$ is computed as

$$\mathbf{\Phi}^T(\mathbf{u}_h) = \int_{t^n}^{t^{n+1}} \int_T \left(\frac{\partial \mathbf{u}_h}{\partial t} + \nabla \cdot \mathcal{F}_h - \mathcal{S}_h(\mathbf{u}_h, x, y) \right) dx\, dy\, dt = \frac{|T|}{3} \sum_{j \in T} (\mathbf{u}_j^{n+1} - \mathbf{u}_j^n) + \frac{\Delta t}{2} (\boldsymbol{\phi}^T(\mathbf{u}_h^{n+1}) + \boldsymbol{\phi}^T(\mathbf{u}_h^n)), \tag{63}$$

with $\boldsymbol{\phi}^T$ as in (62). Starting from these definitions, we proceed exactly as illustrated in Section 3. In particular, we note that the definitions of the N scheme and of the Lax–Friedrichs scheme remain unchanged, the presence of the source term being taken into account directly in the definition of the residual. The limiting, and the stabilization steps (see Sections 3.3.3, 3.3.4) remain unchanged, except at the wet/dry interface which will be discussed shortly.

In the following sections, we study the properties of the $\mathcal{RD}$ discretizations considered in this paper, with respect to their application to the SWE. Three main aspects are discussed:

- Well-balancedness,
- preservation of the positivity of the water height H,
- treatment of the wetting/drying interface.

4.1. Well-balancedness

When solving the SWE the numerical balance between flux divergence and the source term modeling the bed slope variation is very important. The respect of this balance is known in literature as the *well-balancedness* of the discretization. The analysis of the accuracy of $\mathcal{RD}$ discretizations in presence of source terms reported in [41,20] shows that, for a linear spatial approximation, and as long as we use a linearity preserving scheme, combined with definition (62) of the residual, we retain the $\mathcal{O}(h^2)$ truncation error when approximating a regular solution.

In addition to this, in [20] it has been shown that if the source term is discretized independently of the fluxes, second order of accuracy is in general lost, with the unique exception of central schemes. As in the homogeneous case, a good design criterion is to look for linearity preserving discretizations having bounded distribution coefficients. However, the presence of the additional stabilization terms requires a slight generalization of the discussion made in [41,20], concerning the SWE.

We start here from the actual evaluation of the spatial residual $\boldsymbol{\phi}^T$. As in [41], we make the hypothesis that for the SWE *the integrals in (62) are evaluated exactly with respect to the linear approximation of the water height H_h*. This leads to the following formulas for the spatial residual (cf. Eq. (2)):

$$\begin{aligned}
\boldsymbol{\phi}^T &= \oint_{\partial T} \begin{bmatrix} H_h(\vec{u}_h \cdot \vec{n}) \\ H_h\vec{u}_h(\vec{u}_h \cdot \vec{n}) \end{bmatrix} dl + \frac{1}{2} \oint_{\partial T} \begin{bmatrix} 0 \\ gH_h^2 \cdot \vec{n} \end{bmatrix} dl + \int_T H_h \begin{bmatrix} 0 \\ g\nabla B_h \end{bmatrix} dx\, dy \\
&= \oint_{\partial T} \begin{bmatrix} H_h(\vec{u}_h \cdot \vec{n}) \\ H_h\vec{u}_h(\vec{u}_h \cdot \vec{n}) \end{bmatrix} dl + \int_T gH_h \begin{bmatrix} 0 \\ \nabla H_h \end{bmatrix} dx\, dy + \int_T gH_h \begin{bmatrix} 0 \\ \nabla B_h \end{bmatrix} dx\, dy \\
&= \oint_{\partial T} \begin{bmatrix} H_h(\vec{u}_h \cdot \vec{n}) \\ H_h\vec{u}_h(\vec{u}_h \cdot \vec{n}) \end{bmatrix} dl + \frac{g\overline{H}}{2} \sum_{j \in T} \begin{bmatrix} 0 \\ (H_j + B_j)\vec{n}_j \end{bmatrix},
\end{aligned} \tag{64}$$

where Gauss–Green's formula has been used to pass from the first to the second line, and exact integration with respect to the linear approximation (18) (cf. also equation (19)) to get to the final expression, in which

$$\overline{H} = \frac{1}{3}\sum_{j\in T} H_j.$$

The interesting result is that, due to (17), when evaluated with (64), the residual $\boldsymbol{\phi}^T(\mathbf{u}_h)$ vanishes identically when $\mathbf{u}_h$ is the lake at rest solution, that is when $\vec{u}_h = 0$ and $H(x,y) + B(x,y) = H_0$, with H_0 a constant. This leads to the proposition (see [41] for more):

Proposition 4.1. *Provided that the same numerical representation is used for the water height and for the local height of the bottom $B(x,y)$, and that the local residual is evaluated exactly with respect to this numerical representation of H and B, linearity preserving $\mathcal{RD}$ schemes preserve exactly the lake at rest solution, independently on topology of the mesh, character of $B(x,y)$ and polynomial degree of the approximation.*

For steady calculations, last proposition applies to the stabilized schemes of type (58). Scheme (56), however, is not in the form of a linearity preserving distribution, so we need to generalize Proposition 4.1.

We start by a remark concerning the form of the stabilization used for the SWE. As in the general case, this is done using a locally linearized (constant coefficients) form of the quasi-linear form. In particular, let $\mathcal{F}^c$ be the convective part of the flux

$$\mathcal{F}^c = \begin{bmatrix} H\vec{u} \\ H\vec{u}\otimes\vec{u} \end{bmatrix}.$$

On an element T, given the local average $\bar{\mathbf{u}}$, we consider the linearized quasi-linear form of the SWE:

$$\frac{\partial}{\partial t}\begin{bmatrix} H \\ H\vec{u} \end{bmatrix} + \frac{\partial \mathcal{F}^c(\bar{\mathbf{u}})}{\partial \mathbf{u}} \cdot \nabla \begin{bmatrix} H \\ H\vec{u} \end{bmatrix} + \begin{pmatrix} 0 & 0 \\ g\overline{H} & 0 \end{pmatrix} \cdot \nabla \begin{bmatrix} H_{tot} \\ H\vec{u} \end{bmatrix} = 0.$$

For a linear representation of H, B and $H\vec{u}$, the application of the Petrov–Galerkin formula (57) yields

$$\boldsymbol{\Phi}_i^{*s} = \boldsymbol{\beta}_i^{*}\boldsymbol{\Phi}^{nc} + \epsilon_h\left(\frac{K_i}{|T|}\boldsymbol{\tau}_h\boldsymbol{\Phi}^{nc} + \frac{|T|}{36}\sum_{j\in T} D_{ij}(\mathbf{u}_j^{n+1} - \mathbf{u}_j^{n})\right),$$

where

$$\boldsymbol{\Phi}^{nc} = \int_{t^n}^{t^{n+1}}\int_T \left(\frac{\partial}{\partial t}\begin{bmatrix} H \\ H\vec{u} \end{bmatrix}_h + \frac{\partial \mathcal{F}^c(\bar{\mathbf{u}})}{\partial \mathbf{u}} \cdot \nabla \begin{bmatrix} H \\ H\vec{u} \end{bmatrix}_h + \begin{pmatrix} 0 & 0 \\ g\overline{H} & 0 \end{pmatrix} \cdot \nabla \begin{bmatrix} H_{tot} \\ H\vec{u} \end{bmatrix}_h\right) dx\,dy\,dt.$$

As done before, we now replace the non-conservative residual by its conservative approximation, so that we end up again with

$$\boldsymbol{\Phi}_i^{*s} = \boldsymbol{\beta}_i^{*}\boldsymbol{\Phi}^{T} + \epsilon_h\left(\frac{K_i}{|T|}\boldsymbol{\tau}_h\boldsymbol{\Phi}^{T} + \frac{|T|}{36}\sum_{j\in T} D_{ij}(\mathbf{u}_j^{n+1} - \mathbf{u}_j^{n})\right), \tag{65}$$

where

$$\boldsymbol{\Phi}^{T} = \frac{|T|}{3}\sum_{j\in T}(\mathbf{u}_j^{n+1} - \mathbf{u}_j^{n}) + \frac{\Delta t}{2}\left(\boldsymbol{\phi}^T(\mathbf{u}_h^n) + \boldsymbol{\phi}^T(\mathbf{u}_h^{n+1})\right),$$

with $\boldsymbol{\phi}^T$ as in (64). Note in particular, that for the SWE *the difference between the non-conservative approximation of the residual and its conservative one is in the evaluation of the integral of the convective flux $\mathcal{F}^c$ (cf. equation (64)). The gravitational terms are evaluated in the exact same way in the two cases.*

Concerning the properties of the scheme, as recalled in Section 3.3.3, the analysis of [41] shows that also for non-homogeneous problems schemes that can be recast as

$$\boldsymbol{\Phi}_i^{T} = \int_{t^n}^{t^{n+1}}\int_T \omega_i^T\left(\frac{\partial \mathbf{u}_h}{\partial t} + \nabla\cdot\mathcal{F}^h(\mathbf{u}_h) - \mathcal{S}_h(\mathbf{u}_h, x, y)\right) dx\,dy\,dt,$$

verify the formal condition for having a $\mathcal{O}(h^2)$ truncation error (hence second-order of accuracy), *provided that ω_i^T is uniformly bounded* [41,20]. Additionally, we immediately see that if $\mathbf{u}_h$ is the lake at rest solution, that is if $\vec{u}_h = 0$, and $H_h(x,y) + B_h(x,y) = H_0$, with H_0 constant, then scheme (65) leads to the discrete equation

$$\boldsymbol{\mathcal{M}}(\boldsymbol{\beta}_i, \boldsymbol{\mathcal{T}}_h, \mathbf{u}_h)(\mathbf{U}^{n+1} - \mathbf{U}^{n}) = 0, \tag{66}$$

where $\mathbf{U}_i = \mathbf{u}_i$. So, provided that the mass matrix $\boldsymbol{\mathcal{M}}(\boldsymbol{\beta}_i, \boldsymbol{\mathcal{T}}_h, \mathbf{u}_h)$ is invertible, scheme (65) will preserve this solution. In a more general way, note that, provided that ω_i^T is uniformly bounded and locally differentiable, and that we integrate the equations exactly with respect to the linear variation of H_h and B_h, on the lake at rest solution we have

$$\begin{aligned}
&\int_{t^n}^{t^{n+1}}\int_T \omega_i^T\nabla\cdot\mathcal{F}^c(\mathbf{u}_h)dx\,dy\,dt + \int_{t^n}^{t^{n+1}}\int_T gH_h\omega_i^T\begin{bmatrix} 0 \\ \nabla(H_h + B_h) \end{bmatrix} dx\,dy\,dt \\
&\quad = \int_{t^n}^{t^{n+1}}\oint_{\partial T}\omega_i^T\mathcal{F}^c(\mathbf{u}_h)\cdot\vec{n}\,dl\,dt - \int_{t^n}^{t^{n+1}}\int_T \mathcal{F}^c(\mathbf{u}_h)\cdot\nabla\omega_i^T dx\,dy\,dt + \int_{t^n}^{t^{n+1}}\int_T gH_h\omega_i^T\begin{bmatrix} 0 \\ \nabla H_0 \end{bmatrix} dx\,dy\,dt = 0,
\end{aligned}$$

since $\mathcal{F}^c(\mathbf{u}_h)$ vanishes on the lake at rest state, and since any consistent interpolation gives

$$\nabla H_0 = H_0 \sum_{j\in T} \nabla\psi_j = 0.$$

This leads again to a discrete equation of the type (66), and *provided that the mass matrix is invertible*, to the preservation of the lake at rest state. This proves the following generalization of proposition 4.1.

Proposition 4.2. *Provided that the same numerical representation is used for the water height and for the local height of the bottom $B(x,y)$, and that all integrals are evaluated exactly with respect to this numerical representation of H and B, schemes that can be recast as*

$$\int_0^{t_f}\int_\Omega \omega_i\left(\frac{\partial u_h}{\partial t} + \nabla\cdot\mathcal{F}_h(\mathbf{u}_h) - \mathcal{S}_h(\mathbf{u}_h,x,y)\right)dx\,dy\,dt,$$

preserve exactly the lake at rest solution, independently on topology of the mesh, character of $B(x,y)$ and polynomial degree of the approximation, provided that

$$\omega_i = \sum_{T\in\mathcal{T}_h} \omega_i^T$$

is uniformly bounded and locally differentiable, and that the associated mass matrix is invertible.

The purpose of the stabilization is precisely to ensure that the mass matrix of the $\mathcal{RD}$ discretization is invertible. Even though no analytical proof of this fact is given in this paper (we refer however the reader to [13] for the analysis of the mass matrix of the SUPG scheme which has close resemblance to the one of scheme (65)), the numerical results will show that proposition 4.2 is indeed verified by our schemes, and that, moreover, we achieve grid convergence with the expected rate.

4.2. Positivity of the water height

We consider now the preservation of the positivity of the water height. We analyze the Lax–Friedrichs scheme to show that, under a constraint on the size of the time step, it can indeed ensure this property. The hypotheses under which this is true for the LLF scheme are given.

In the following analysis, we look for the conditions under which, starting from non-negative nodal values of the water height, we get a positive value of H in each node of the mesh, when using the explicit iteration schemes (21) or (61).

Explicit Euler update: LF scheme. We start with the explicit update (21). Using (41), we have for the water height:

$$H_i^{n+1} = H_i^n - \frac{\Delta t}{|S_i|}\sum_{T\in D_i}\left(\frac{1}{3}\phi^T_{H^n} + \frac{1}{3}\alpha_T\sum_{\substack{j\in T\\ j\neq i}}(H_i^n - H_j^n)\right), \tag{67}$$

where the sub-script $_T$ has been added to the LF dissipation coefficient to make the exposition clearer. For a linear approximation of $\mathbf{u}_h$ we have

$$\phi_H^T = \oint_{\partial T} H\vec{u}\cdot\hat{n}\,dl = \frac{1}{2}\sum_{j\in T} H_j\vec{u}_j\cdot\vec{n}_j. \tag{68}$$

Using this expression the update (67) becomes

$$H_i^{n+1} = \left(1 - \frac{\Delta t}{|S_i|}\frac{1}{3}\sum_{T\in D_i}(\frac{\vec{u}_i\cdot\vec{n}_i}{2} + 2\alpha_T)\right)H_i^n + \frac{1}{3}\frac{\Delta t}{|S_i|}\sum_{T\in D_i}\sum_{\substack{j\in T\\ j\neq i}}\left(\alpha_T - \frac{\vec{u}_j\cdot\vec{n}_j}{2}\right)H_j^n. \tag{69}$$

Due to definition of α_T in (42) (see also (11)) we have

$$\alpha_T - \frac{\vec{u}_j\cdot\vec{n}_j}{2} > 0, \tag{70}$$

so the quantity in parentheses in the second term of (69) is positive. For the coefficient of H_i^n we can write

$$1 - \frac{\Delta t}{|S_i|}\frac{1}{3}\sum_{T\in D_i}\left(\frac{\vec{u}_i\cdot\vec{n}_i}{2} + 2\alpha_T\right) \geqslant 1 - \frac{\Delta t}{|S_i|}\sum_{T\in D_i}\alpha_T, \tag{71}$$

so that if

$$\Delta t < \frac{|S_i|}{\sum_{T\in D_i}\alpha_T}, \tag{72}$$

then given positive nodal values of H at time t^n, H_i^{n+1} will be positive $\forall i$. This shows that the LF scheme preserves the positivity of the water height. We set

$$\Delta t^{EE} = \frac{|S_i|}{\sum_{T\in D_i}\alpha_T}, \tag{73}$$

where the superscript EE stands for Explicit Euler. In practice, we compute the time step as

$$\Delta t = \nu\Delta t^{EE}, \quad \nu < 1. \tag{74}$$

Explicit Euler update: LLF scheme. Concerning the nonlinear LLF scheme, one can easily see that *provided that the limiting is performed directly on the water height equation, and if* (74) is satisfied, the LLF scheme preserves the positivity of the water height. This is basically a consequence of property (49). In fact, proceeding as before, we have for the LLF scheme

$$\begin{aligned} H_i^{n+1} = &\left(1 - \frac{\Delta t}{|S_i|}\sum_{T\in D_i}\gamma_i^*\sum_{\substack{j\in T\\ j\neq i}} c_{ij}^{LF}\right)H_i^n + \frac{\Delta t}{|S_i|}\sum_{T\in D_i}\gamma_i^*\sum_{\substack{j\in T\\ j\neq i}} c_{ij}^{LF}H_j^n \\ \geqslant &\left(1 - \frac{\Delta t}{|S_i|}\sum_{T\in D_i}\gamma_i^*\alpha_T\right)H_i^n + \frac{\Delta t}{|S_i|}\sum_{T\in D_i}\gamma_i^*\sum_{\substack{j\in T\\ j\neq i}} c_{ij}^{LF}H_j^n. \end{aligned} \tag{75}$$

Clearly, if (74)is verified, and using (49), we have

$$\Delta t < \frac{|S_i|}{\sum_{T\in D_i}\alpha_T} \leqslant \frac{|S_i|}{\sum_{T\in D_i}\gamma_i^*\alpha_T}.$$

Hence, condition (74) is enough to guarantee the positivity of the water height when using the LLF scheme.

Trapezium time integration: LF scheme. We will now try to carry over our analysis to the time dependent case. The time discretizations used in this case are implicit. However, we can exploit the explicit pseudo-time stepping solution strategy (61) to simplify the analysis. With the notation of §3.3.5, we have

$$\begin{aligned} H_i^{n+1,p+1} &= H_i^{n+1,p} - \Delta s(H_i^{n+1,p} - H_i^n) \\ &\quad - \frac{\Delta s\Delta t}{2|S_i|}\sum_{T\in D_i}\left(\frac{1}{3}\phi^T_{H^{n+1,p}} + \frac{1}{3}\alpha_T\sum_{\substack{j\in T\\ j\neq i}}(H_i^{n+1,p} - H_j^{n+1,p})\right) \\ &\quad - \frac{\Delta s\Delta t}{2|S_i|}\sum_{T\in D_i}\left(\frac{1}{3}\phi^T_{H^n} + \frac{1}{3}\alpha_T\sum_{\substack{j\in T\\ j\neq i}}(H_i^n - H_j^n)\right). \end{aligned} \tag{76}$$

Using again (68) and rearranging terms we obtain

$$\begin{aligned} H_i^{n+1,p+1} = &\left\{1 - \Delta s\left[1 + \frac{1}{3}\frac{\Delta t}{2|S_i|}\sum_{T\in D_i}\left(\frac{\vec{u}_i^{n+1,p}\cdot\vec{n}_i}{2} + 2\alpha_T\right)\right]\right\}H_i^{n+1,p} + \Delta s\left[1 - \frac{1}{3}\frac{\Delta t}{2|S_i|}\sum_{T\in D_i}\left(\frac{\vec{u}_i^{n}\cdot\vec{n}_i}{2} + 2\alpha_T\right)\right]H_i^n + \frac{1}{3} \\ &\times \frac{\Delta s\Delta t}{2|S_i|}\sum_{T\in D_i}\sum_{\substack{j\in T\\ j\neq i}}\left(\alpha_T - \frac{\vec{u}_j^{n+1,p}\cdot\vec{n}_j}{2}\right)H_j^{n+1,p} + \frac{1}{3}\frac{\Delta s\Delta t}{2|S_i|}\sum_{T\in D_i}\sum_{\substack{j\in T\\ j\neq i}}\left(\alpha_T - \frac{\vec{u}_j^{n}\cdot\vec{n}_j}{2}\right)H_j^n. \end{aligned} \tag{77}$$

If we define α_T according to (42), and we take the maximum of its values at times t^n and t^{n+1}, we are sure of the positivity of the last two terms (provided that we have no negative water heights at time t^n, and at the pth iteration). Concerning the terms containing H_i^n, we proceed as before to get

$$1 - \frac{1}{3}\frac{\Delta t}{2|S_i|}\sum_{T\in D_i}\left(\frac{\vec{u}_i^n\cdot\vec{n}_i}{2} + 2\alpha_T\right) \geqslant 1 - \frac{\Delta t}{2|S_i|}\sum_{T\in D_i}\alpha_T.$$

Provided that

$$\Delta t < \frac{2|S_i|}{\sum_{T\in D_i}\alpha_T} = 2\Delta t^{EE},$$

these terms will give a positive contribution. Note that, not surprisingly for the trapezium scheme, last condition is twice as large as the one given by the explicit Euler discretization (Eq. (74)). In practice, we set

$$\Delta t = 2\nu\Delta t^{EE}, \quad \nu < 1. \tag{78}$$

At last, we consider the terms containing $H_i^{n+1,p}$. Again we use the estimate

$$1-\Delta s\left[1+\frac{1}{3}\frac{\Delta t}{2|S_i|}\sum_{T\in D_i}\left(\frac{\vec{u}_i^{n+1,p}\cdot\vec{n}_i}{2}+2\alpha_T\right)\right]\geqslant 1-\Delta s\left[1+\frac{\Delta t}{2|S_i|}\sum_{T\in D_i}\alpha_T\right].$$

So that the positivity of the LF scheme with trapezium time integration requires the satisfaction of the pseudo-time step constraint

$$\Delta s<\frac{1}{1+\frac{\Delta t}{2|S_i|}\sum_{T\in D_i}\alpha_T}=\frac{1}{1+\nu}. \tag{79}$$

Trapezium time integration: LLF scheme (space–time). Consider now the space time nonlinear discretization obtained by limiting the LF scheme with trapezium time integration (cf. section 3.4). As before, we can show that *provided that the limiting is performed directly on the water height equation, the positivity of the LLF scheme is a consequence of property* (49). *In particular, after a few manipulations, we can rewrite the update for the LLF scheme with trapezium time integration as*

$$\begin{aligned}H_i^{n+1,p+1}=&\left\{1-\Delta s\sum_{T\in D_i}\frac{\gamma_i^*}{3|S_i|}\left[|T|+\frac{\Delta t}{2}\left(\frac{\vec{u}_i^{n+1,p}\cdot\vec{n}_i}{2}+2\alpha_T\right)\right]\right\}H_i^{n+1,p}+\Delta s\sum_{T\in D_i}\frac{\gamma_i^*}{3|S_i|}\left[|T|-\frac{\Delta t}{2}\left(\frac{\vec{u}_i^{n}\cdot\vec{n}_i}{2}+2\alpha_T\right)\right]H_i^n\\&+\frac{1}{3}\frac{\Delta s\Delta t}{2|S_i|}\sum_{T\in D_i}\gamma_i^*\sum_{\substack{j\in T\\ j\neq i}}\left(\alpha_T-\frac{\vec{u}_j^{n+1,p}\cdot\vec{n}_j}{2}\right)H_j^{n+1,p}+\frac{1}{3}\frac{\Delta s\Delta t}{2|S_i|}\sum_{T\in D_i}\gamma_i^*\sum_{\substack{j\in T\\ j\neq i}}\left(\alpha_T-\frac{\vec{u}_j^{n}\cdot\vec{n}_j}{2}\right)H_j^n.\end{aligned} \tag{80}$$

Obviously, the important terms to look at are the ones multiplying H_i, the definition of α_T assuring the positivity of the "off-diagonal" terms. Due to the local character of the limiting procedure, we are obliged to proceed in an element-wise fashion. In particular, proceeding as before, one immediately shows that the term multiplying H_i^n is positive if in every element we make sure that

$$\Delta t\leqslant\frac{2|T|}{3\alpha_T},$$

leading to the modified time step constraint

$$\Delta t=2\nu\min_{T\in\mathcal{T}_h}\frac{|T|}{3\alpha_T}\quad \nu<1. \tag{81}$$

Note that this constraint corresponds to the local positivity of the LF scheme with trapezium time integration [20]. Concerning the term multiplying $H_i^{n+1,p}$, setting in each element

$$\nu^T=\frac{3\alpha_T\Delta t}{2|T|},$$

we immediately see that the positivity of H will be preserved provided that we can ensure that

$$1-\Delta s\sum_{T\in D_i}\frac{\gamma_i^*|T|}{3|S_i|}\left[1+\frac{\nu^T}{3\alpha_T}(\frac{\vec{u}_i^{n+1,p}\cdot\vec{n}_i}{2}+2\alpha_T)\right]\geqslant 0.$$

Since $\alpha_T\geqslant\vec{u}_i^{n+1,p}\cdot\vec{n}_i/2$, and $\gamma_i^*\leqslant 1$, we have

$$\frac{\gamma_i^*|T|}{3|S_i|}\left[1+\frac{\nu^T}{3\alpha_T}\left(\frac{\vec{u}_i^{n+1,p}\cdot\vec{n}_i}{2}+2\alpha_T\right)\right]\leqslant\frac{\gamma_i^*|T|}{3|S_i|}(1+\nu^T)\leqslant\frac{|T|}{3|S_i|}\min_{T\in D_i}(1+\nu^T).$$

Finally, using (22) we end up with the result that the LLF scheme with trapezium time integration will preserve the positivity of the water height provided that for each node

$$\Delta s\leqslant \mathit{min}_{T\in D_i}\frac{1}{1+\nu^T}. \tag{82}$$

Note that, due to the locality of the analysis and to the regular use of the definition of α_T to bound terms, the conditions obtained (Eqs. (81) and (82)) are probably over-constraining.

4.3. Handling dry and partially dry cells

This section is devoted to the description of the treatment of dry and partially dry cells. A dry cell is one in which $H_j=0,\forall j\in T$. When this happens, it is easy to see that the residual and the split residuals are zero, since solution vector, fluxes, and source term are identically zero over the element. Hence, completely dry elements pose no particular problem.

Most numerical problems arise in partially dry cells. These are cells in which $H_j = 0$ for some of the nodes and $H_k > 0$ at least in one vertex. We have to distinguish between three situations. Cells T with constant bed height $B_j = B\ \forall j$ are called *flat*. If instead

$$H_j = 0, \quad H_k > 0 \quad \text{and} \quad B_j > B_k \quad j, k \in T,$$

we speak of cells with *adverse slope*, whereas if

$$H_j = 0, \quad H_k > 0 \quad \text{and} \quad B_j < B_k \quad j, k \in T,$$

we speak of *downward slope*. The issues we have to solve in these front cells are the following:

1. *Detection of dry nodes.* In practice, we encounter cells in which some nodes have heights $H_j \neq 0$, however $H_j \ll 1$. We should make sure that these cells are treated properly.
2. *Positivity.* We have to ensure that our schemes always keep positive values of the water height.
3. *Artificial velocities in front elements.* Along the wetting–drying front one typically observes two behaviors. For flat cells or downward slopes, the schemes behave like one expects: the water runs in direction of the dry nodes. For adverse slopes, one faces a problem. To see this, let us have a look at the representation in the top pictures of Fig. 3. In the left picture we can see the lake at rest situation in a front element, and in the right one we see its numerical representation for a linear variation of the water and bottom heights. Clearly, a non-physical slope in the water free surface is introduced by the linear numerical representation. This leads to artificial velocities in the downward direction, so that eventually the lake at rest state is not preserved numerically.
4. *Undefined velocities*. Even if conserved quantities are well defined, for (nearly) dry nodes, the velocity $\vec{u}$ is not, that is, the quotient $u = Hu/H$ is not necessarily bounded.

We illustrate hereafter the way in which these issues have been handled.

4.3.1. Dry nodes detection

In cells with very low water heights, the element residual $\boldsymbol{\phi}^T$ (or $\boldsymbol{\Phi}^T$) should be small, and so should be the split residuals. In practice, we have to deal with undefined velocity vectors. This is in part due to the limited machine precision, and can lead to unphysical values of the residuals. The solution adopted here is quite common in shallow water simulations [32,10,47] and consists in introducing a cut-off value of the water height, say $C_{\vec{u}} \ll h$, below which we set the velocity to zero, that is:

$$\vec{u} = \begin{cases} \frac{H\vec{u}}{H} & \text{if } H \geqslant C_{\vec{u}} \\ 0 & \text{if } H < C_{\vec{u}} \end{cases}.$$

Note that we *do not introduce any cut-off value on the water height itself*. The positivity of H is dealt with differently as we shall see shortly. Note that the choice of $C_{\vec{u}}$ is not necessarily trivial. For some test problems, we will analyze its influence on the numerical output. What we found out is that, probably due to the very small discharge in vicinity of the wetting/drying front, the value of this constant affects little the numerical solution. A possible criterion to choose $C_{\vec{u}}$ is given in Section 5.

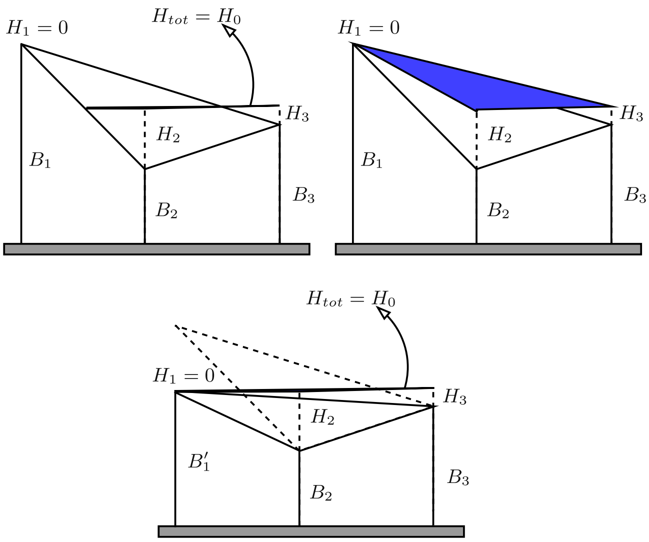


Fig. 3. Lake at rest in front cells with adverse slope. Top left: real situation. Top-right: artificial gradient of total water height due to linear approximation. Bottom: recovery of constant total water height through redefinition of the bottom height.

4.3.2. Water height positivity: nonlinear schemes in front cells

In Section 4.2 we have shown that, under a time step constraint, the LLF scheme can preserve the positivity of the water height. However, this is only true if the limiting is performed directly on the water height equation. The eigenvector projection described in Section 3.3.3, while improving the convergence of the overall algorithm [4], *a priori* does not guarantee the preservation of the positivity of H.

To enforce positivity, we proceed as follows. Let us consider for example the time dependent case. We start by noting that, when projecting back to conservative variables (cf. Eq. (51)), we get for the water height residual sent to node i:

$$\boldsymbol{\Phi}^*_{1,i} = \varphi^*_{1,i}\mathbf{r}_{1,1} + \varphi^*_{2,i}\mathbf{r}_{2,1} + \varphi^*_{3,i}\mathbf{r}_{3,1}.$$

Using the expression of the eigenvectors given in the Appendix A.1, the fact that the projection is performed using the eigenvectors in the velocity direction $\vec{\xi} = \vec{u}/\|\vec{u}\|$, and that $\varphi^*_{\sigma,i} = \beta^*_{\sigma,i}\varphi^T_\sigma$ with $\beta^*_{\sigma,i} \leqslant 1$, we get

$$\boldsymbol{\Phi}^*_{1,i} = \varphi^*_{2,i} + \varphi^*_{3,i} \leqslant |\varphi^T_2| + |\varphi^T_3|.$$

When we make explicit use of the expression of the left eigenvectors (cf. Appendix A.1) in the velocity direction, we can further bound the previous expression using the components of the element residual in conservative variables. In particular, one easily shows that

$$\boldsymbol{\Phi}^*_{1,i} \leqslant \frac{1}{2}(|1-\mathrm{Fr}| + 1 + \mathrm{Fr})|\boldsymbol{\Phi}^T_1| + \frac{1}{a}(|\boldsymbol{\Phi}^T_2| + |\boldsymbol{\Phi}^T_3|) = \Phi_{\mathrm{lim}},$$

where Fr denotes the Froude number introduced in Section 2.2, evaluated using local mean values of the unknowns. The last inequality is valid for all the nodes of the element. Based on the explicit update (61), and on (22), we define the following local *worst guess* for the value of the minimal water height at the new pseudo-time iteration:

$$H_{\mathrm{lim}} = H_{\mathrm{min}} - \frac{3\Delta s}{|T|}\Phi_{\mathrm{lim}},$$

with

$$H_{\mathrm{min}} = \min_{j\in T}(\min(H^n_j, H^{n+1,p}_j)). \tag{83}$$

We finally modify the limiting procedure by setting in equation (50)

$$\varphi^T_\sigma = \begin{cases} \mathbf{l}_\sigma\boldsymbol{\Phi}^T & \text{if } H_{\mathrm{lim}} \geqslant 0 \\ (\boldsymbol{\Phi}^T)_\sigma & \text{if } H_{\mathrm{lim}} < 0 \end{cases},$$

and similarly for the split residuals. In addition to this, we smoothly switch off the stabilization in front cells, so that in these elements we recover the positivity preserving limited scheme. This is controlled by the definition of $\epsilon(\mathbf{u}_h)$ given in Section 5.

4.3.3. Bed slope in front elements

To solve the problem illustrated in Fig. 3, we follow [16]. The observation is that for the lake at rest state we have

$$H_{tot,j} = H_0, \quad \forall j \in T \Rightarrow \nabla H_h = -\nabla B_h. \tag{84}$$

In cells with adverse slope this is not the case anymore. For example, in the situation of Fig. 3 we have $B_1 > H_0 = H_{tot,2} = H_{tot,3}$. This leads to different gradients ∇H_h and $-\nabla B_h$, so that $\boldsymbol{\phi}^T \neq 0$, and spurious velocities are generated. As in [16], the problem is cured by using, in front cells with adverse slope, a modified value of the bathymetry in the dry nodes. In particular, in the computation of the spatial residual, for all nodes j with $H_j = 0$ we replace B_j by H_{max}, with

$$H_{\mathrm{max}} = \max_{\substack{j\in T\\ H_j>0}}(H_j + B_j). \tag{85}$$

This approach cures the problem in the lake at rest case, for which we obtain again $\boldsymbol{\phi}^T = 0$ (cf. Fig. 3).

Note that the nodes in which we redefine the bathymetry all verify $H = 0$ and $\vec{u} = 0$. Hence, in front cells which are not at rest, this modification does not alter the total mass of water contained in the cell, and does not alter the mass flux, at least not directly. However, this procedure does alter the bed slope seen by the flow in these cells, leading to a reduction of the slope. We have tried to correct this flaw by pre-multiplying the term ∇H_{tot} by a factor of the type

$$\left|\frac{B_{\mathrm{max}} - H_{\mathrm{min}}}{H_{\mathrm{max}} - H_{\mathrm{min}}}\right|,$$

with $B_{\max}$ the max value of the bed height in the cell, and $H_{\min}$ the minimum of the total water height over the cell. We have observed no influence of this further modification in our calculations. The results presented later are obtained without this correction. This issue deserves further attention and will be further studied in the future.

5. Implementation details

Before discussing the numerical tests, we give some details on the practical implementation of the schemes. Important points are the computation of the physical variables from the conserved ones, the computation of the residual, and the definition of ϵ_h and $\boldsymbol{\tau}_h$ in the stabilization term. A summary is given hereafter.

Physical variables. We recompute the vector of physical variables $[H\,u\,v]$ after each iteration (21) (or (61) in the unsteady case). When doing this, we apply the velocity cut-off described in Section 4.3.1. This allows to have always at each iteration current nodal values of the conserved and of the physical variables. Concerning the definition of the cut-off value $C_{\vec{u}}$, we tried different solutions. The important issue is to make sure that this value is sufficiently small compared to the mesh size. We found out experimentally that a good definition of this constant is

$$C_{\vec{u}} = \left(\frac{h}{L_{\text{ref}}}\right)^2, \tag{86}$$

where L_{ref} is a reference geometrical dimension of the spatial domain computed as

$$L_{\text{ref}} = \underset{i,j\in\mathcal{T}_h}{max}\|\vec{x}_i - \vec{x}_j\|_\infty.$$

Residual evaluation. As already seen in Section 4.1, we compute the spatial residual as in (64). The contour integral is evaluated with a 2 points Gaussian formula on each edge of T. In doing this, we use a linear variation of the conserved variables $\mathbf{u}_h$, and interpolate the nodal velocities (recomputed after each update) to obtain the $\vec{u}_h\cdot\hat{n}$ terms in each Gauss point.

Stabilization term. The role of the parameter $\boldsymbol{\tau}_h$ in the stabilization terms is to provide the correct scaling w.r.t. mesh size and wave speeds. Ultimately, $\boldsymbol{\tau}_h$ must guarantee that the additional term has the "dimensions" of the cell-residual. It is easily seen that this is achieved through a definition guaranteeing that $\boldsymbol{\tau}_h = \mathcal{O}(h)\times\mathcal{O}(\|\vec{u}\|^{-1})$. After a review of the literature on SUPG and Least-Squares stabilized Galerkin schemes, we decided to test the following two formulations (cf. equation (38) and see [12,51,34] and references therein)

$$\boldsymbol{\tau}_1 = |T|\left(\sum_{j\in T}|K_j|\right)^{-1} = \frac{|T|}{2}N, \tag{87}$$

and (see Eq. (11) for the notation)

$$\boldsymbol{\tau}_2 = \frac{2h}{2\|\vec{u}\| + a}, \tag{88}$$

where all quantities are evaluated using mean values of the variables. Neither of these definitions are optimal. The study of better formulations is under way, in the spirit of [29,30,39].The smoothness monitor ϵ_h should instead guarantee that the extra terms are active mainly in smooth parts of the solution. Here we follow [2], and use a definition based on the entropy residual which we locally approximate as

$$\varphi_E = \bar{\mathbf{v}}^t\boldsymbol{\Phi}^T,$$

where $\bar{\mathbf{v}}$ is an average over the element of the vector of entropy variables (6). Following the definitions given in [2], we use as a smoothness sensor the quantity

$$\epsilon_h = \min\left(1, \frac{\overline{E}\|\vec{u}\|^T_{L^\infty}|T|}{L_{\text{ref}}|\varphi_E|}\right)e^{-\frac{h}{L_{\text{ref}}}\frac{H_{\text{ref}}}{H_{\min}}}. \tag{89}$$

In the last expression we make use of the following local quantities: $\overline{E}$ which is an elementwise reference value of the energy (5), computed using local averages of the variables; $\|\vec{u}\|^T_{L^\infty}$ which is the largest component of the local averaged speed; $H_{\min}$ given by (83). We also use the global values of L_{ref}, the reference length introduced in (86), and H_{ref} taken as the maximum value of H in the initial solution.The exponential factor is added to take into account the occurrence of dry areas. Please note that this factor is always of order one, except for small values of $H_{\min}$ in correspondence of which it quickly tends toward zero. *In the time dependent case we use of the same formulas, except that* $\varphi_E = \bar{\mathbf{v}}^t\boldsymbol{\Phi}^T/\Delta t$, *and* $h = (|T|\Delta t)^{1/3}$ *in* (88).

Explicit pseudo-time stepping. All the results presented are obtained by solving the nonlinear algebraic equations by means of an explicit pseudo-time stepping procedure. In particular, in the steady case we perform our iterations as

$$\mathbf{u}_i^{n+1} = \mathbf{u}_i^n - \frac{\Delta t_i}{|S_i|}\sum_{T\in D_i}\boldsymbol{\phi}_i(\mathbf{u}_h^n),$$

where Δt_i is the local time step respecting (74) with $\nu = 0.7$. We consider the solution as being converged when the norm of the water height residual drops to a sufficiently low value, that is

$$\frac{\|R^n\|_{L^1}}{\|R^0\|_{L^1}} < m \quad \text{with} \quad (R^n)_i = \sum_{T\in D_i} (\phi_i(\mathbf{u}_h^n))_H.$$

In steady computations we normally take $m = 10^{-8} - 10^{-10}$.Similarly, in time dependent computations the solution is updated as (cf. Section 3.4.1)

$$\mathbf{u}_i^{n+1,p+1} = \mathbf{u}_i^{n+1,p} - \frac{\Delta s_i}{|S_i|} \sum_{T\in D_i} \mathbf{\Phi}_i(\mathbf{u}_h^p),$$

where while the (physical) time step is computed always according to (81), the local pseudo-time step Δs_i is computed for each node from (82). In all computations we have set $\nu = 0.7$. The solution at time t^{n+1} is assumed to be converged when

$$\frac{\|R^p\|_{L^1}}{\|R^0\|_{L^1}} < m \quad \text{with} \quad (R^p)_i = \sum_{T\in D_i} (\mathbf{\Phi}_i(\mathbf{u}_h^p))_H.$$

In the time dependent case, we normally take $m = 10^{-3} - 10^{-4}$.

6. Numerical results

This section presents an extensive evaluation of the discretization proposed. Our objectives are first to show that the simpler LLFs scheme yields results comparable to the ones obtained with the LNs scheme, and then to evaluate its behavior on test-cases involving the formation and movement of wetting/drying fronts.

6.1. Hydraulic jump over a wedge

To verify the monotonicity of the schemes proposed, as well as to estimate the influence of different definitions of the stabilization matrix, we consider the approximation of a hydraulic jump over a wedge [31,41]. A sketch of the initial solution as a close-up view of the mesh are given in Fig. 4. The incoming flow is super-critical with Fr= 2.74, the wedge angle is 8.95°, and $B(x, y) = 0$ everywhere. The mesh size is $h \approx 1/20$.

In Fig. 5 we report the iterative convergence histories of all the nonlinear discretizations. The basic limited schemes present an erratic convergence, the residual not reaching machine zero. The stabilized ones show instead a (quasi-)monotonic convergence to machine accuracy. As it could be expected, the convergence of the LNs schemes is faster, due to their upwind character. We remark however that for the LLFs schemes the computation of the residual is much cheaper, the number of matrix operations being considerably reduced. This is especially true when using the stabilization parameter (88). Concerning the differences between the two definitions of the stabilization parameters, for the LNs scheme the cheaper choice $\boldsymbol{\tau} = \boldsymbol{\tau}_2$ leads to a slightly increased number of iterations. The inverse is observed for the LLFs scheme.

To visualize the resolution of the hydraulic jump, we report in Figs. 6 and 7 a 3D view of the water height (flow direction right-to-left in the figures). The pictures show a sharp capturing of the discontinuity, spread over 2–3 cells. No oscillations are present, *also in the case of the stabilized schemes, and whatever the definition of the stabilization parameter*. In the case of the LLFs schemes the stabilization even improves the resolution of the shock which is sharper. The transition to the post-shock state is also smoother. These observations are confirmed by the plots in Figs. 8 and 9, where we report the distribution of H at the outlet boundary ($x = 4$), and along the line $y = 1$, respectively.

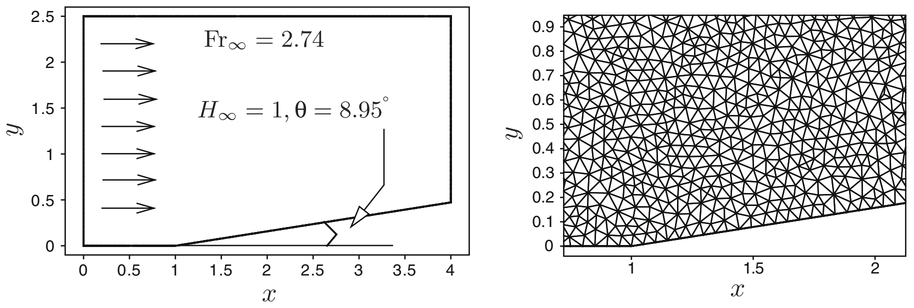

$\mathrm{Fr}_\infty = 2.74$
$H_\infty = 1, \theta = 8.95^\circ$
y
x
0
0.5
1
1.5
2
2.5
3
3.5
4
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

Fig. 4. Hydraulic jump. Problem description and mesh.

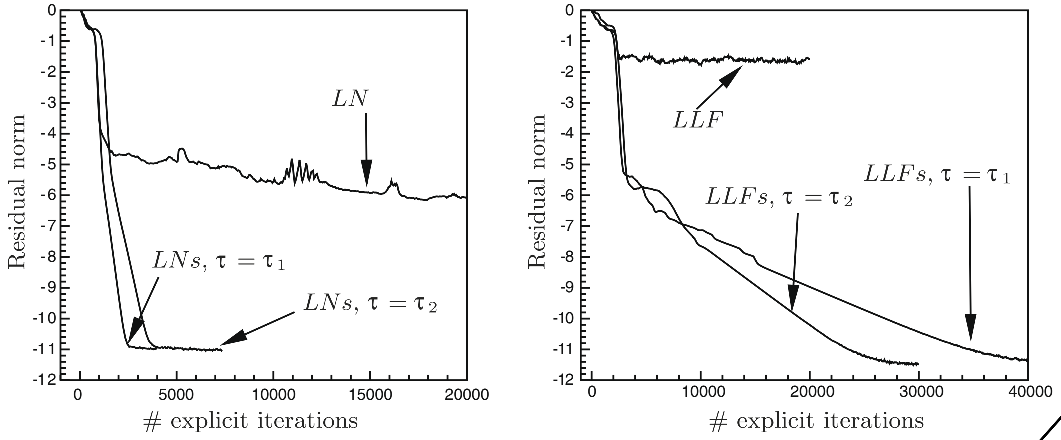


Fig. 5. Hydraulic jump. Iterative convergence. Left: LN and LNs schemes. Right: LLF and LLFs schemes.

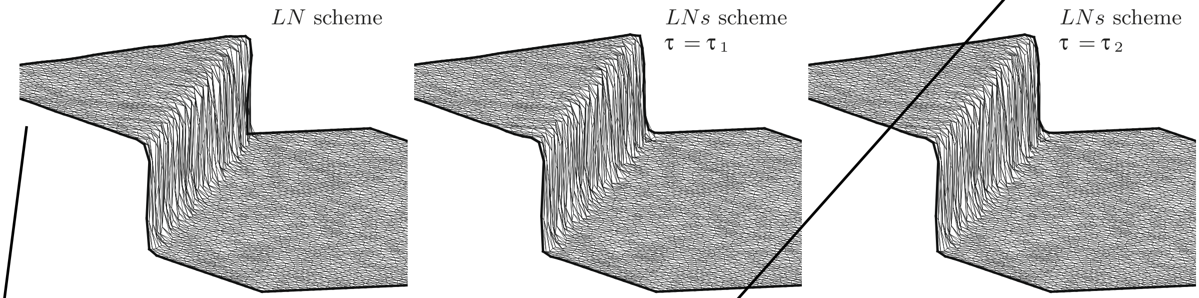
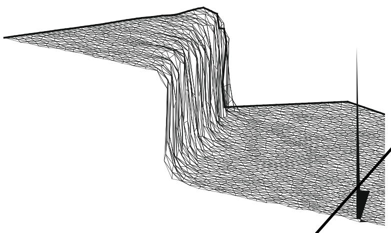


Fig. 6. Hydraulic jump. 3D plot of the water height (flow from right to left). Left: LN scheme. Center: LNs scheme with $\boldsymbol{\tau} = \boldsymbol{\tau}_1$ (Eq. (87)). Right: LNs scheme with $\boldsymbol{\tau} = \boldsymbol{\tau}_2$ (eq. (88)).



6.2. Smooth 2D potential solution

We consider now the approximation of a particular member of the family of 2D exact potential solutions of Section 2.3. On the spatial domain $[-1,1]^2$ we consider the solution corresponding to the choice $\psi = xy$, $\alpha = 1.5$, $C = 30$, and $g = 10$. With these choices top and bottom boundaries are sub-critical inlets, while the left and right boundaries are sub-critical outlets. The Froude number never exceeds one in the domain. Further details can be found in [41]. We start the computations from the exact solution and monitor their convergence to the steady numerical solution The same experiment was already per-

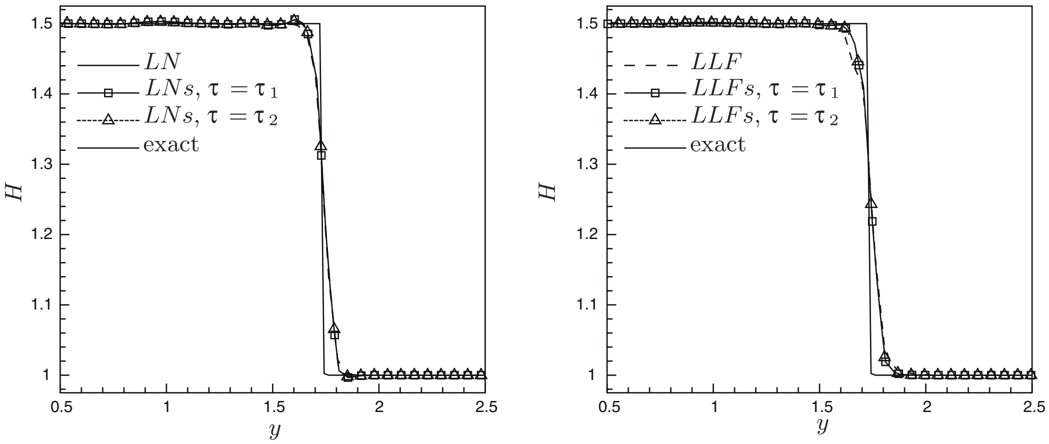
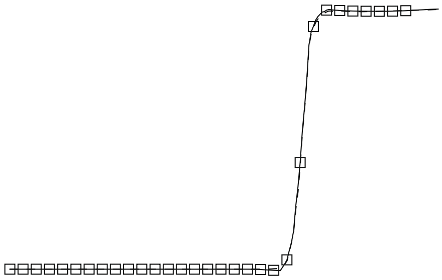


Fig. 8. Hydraulic jump. Water height at outlet boundary ($x = 4$). Left: LN and LNs schemes. Right: LLF and LLFs schemes.



formed in [41] with the LNs scheme, hence here we focus on the LLFs scheme. Our objective is to show the effect of the stabilization when approximating smooth solutions.

We compute the solution on a series of 4 unstructured irregular triangulations, similar to the one in Fig. 4. To have the same local irregularity of the grid in every computation, the meshes are generated independently. Weak boundary conditions are used everywhere.

We start by comparing the contours of the total water height H_{tot} obtained with the LLF and LLFs schemes. We discuss the results obtained with the simplified (and more interesting) formulation $\boldsymbol{\tau} = \boldsymbol{\tau}_2$ (Eq. (88)). The results obtained with the more expensive stabilization parameter of equation (87) are nearly identical. The results obtained on the finest mesh are reported in Fig. 10. The contour plots on the left and in the middle pictures clearly show the beneficial effect of the stabilization in "killing" the spurious modes present in the LLF solution. This is also visible from the picture on the right in the same figure showing the data along the diagonal (line $y = x, y \geqslant 0$). The high frequency oscillations of the LLF scheme are clearly visible in the line plot.

Lastly, in Fig. 11, we report on the left the typical iterative convergence histories obtained with the LLFs scheme, and on the right the grid convergence plot, showing that indeed we achieve second order of accuracy. This is in line with the results obtained with the LNs scheme in [41].

6.3. Pseudo-1D transonic flow over smooth bed

We consider now a 1D flow in a channel with the following variation of the bottom [46,26,21]

$$B(x,y) = B(x) = \begin{cases} 0.2 - 0.05(x-10)^2 & \text{if} \quad 8 \leqslant x \leqslant 12 \\ 0 & \text{otherwise} \end{cases}. \tag{90}$$

Different steady solutions can be computed depending on the choice of boundary conditions. We evaluate the influence of the stabilization on the trans-critical case.

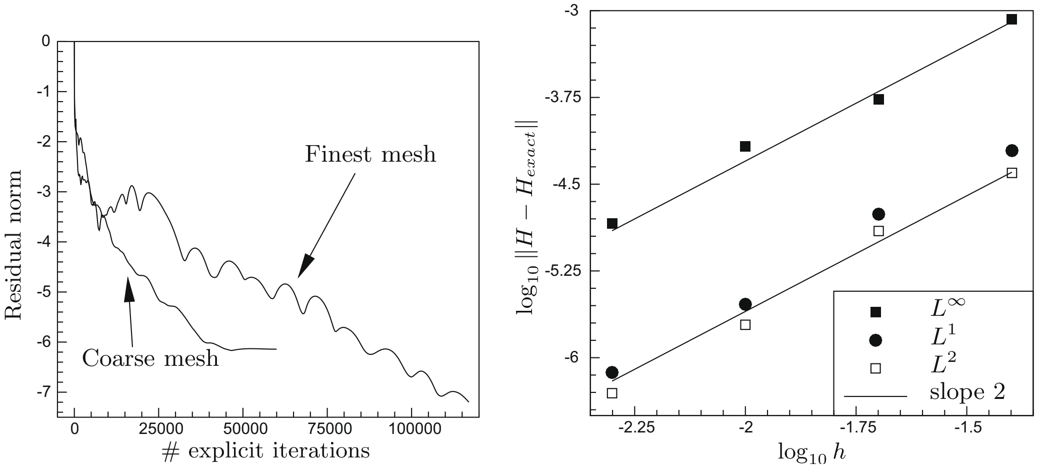


Fig. 11. 2D potential solution. Iterative convergence (left) and grid convergence of the water height H (right). LLFs scheme with stabilization parameter $\boldsymbol{\tau} = \boldsymbol{\tau}_2$ (equation (88)).

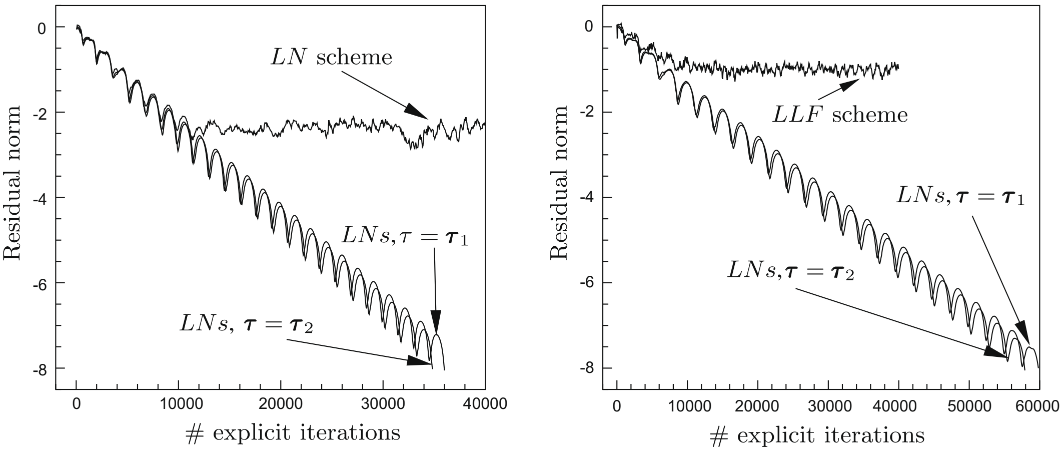


Fig. 12. Pseudo-1D transonic flow: iterative convergence. Left: LN and LNs schemes. Right: LLF and LLFs schemes.

The SWE are solved on the spatial domain $[0,20] \times [0,0.5]$ on an irregular unstructured mesh similar to the one of Fig. 4. The reference mesh size is $h = 1/10$. Periodic boundary conditions are applied along the y-direction. Weak boundary conditions are used at the left and right boundaries, imposing on the left $Hu = 0.18$ and zero v velocity, and on the right the water height $H = 0.33$.

The convergence histories are reported in Fig. 12. The LN and LLF schemes present a very irregular iterative convergence, with a stall after a decrease of one or two orders of magnitude in the water height residual. The stall takes place earlier for the LLF scheme. Conversely, all the stabilized schemes converge (eventually to machine accuracy). Due to the nature of the problem, the convergence is non-monotone, though smooth. As for the hydraulic jump computation, the LNs schemes converge a bit faster. However, in this case the difference is less pronounced. The two definitions of the stabilization parameter $\boldsymbol{\tau}$ give nearly identical results, and convergence histories.

In Fig. 13 we show the distribution of the total water height H_{tot} along the line $y = 0.25$. We compare the basic nonlinear schemes with the stabilized ones obtained with $\boldsymbol{\tau} = \boldsymbol{\tau}_2$. In the case of the LN and LNs schemes, the results are almost identical. No oscillations are present. Differences are instead visible between the LLF and LLFs schemes solutions. In particular, weak spurious oscillations are visible in the LLF solution. This is not a shock capturing problem: it is related to the presence of mild spurious modes, as illustrated on the previous test case. This is confirmed by the absence of these oscillations in the solution of the LLFs scheme, which is clearly monotone.

Last, we look at the errors in the discharge along the x-direction, which should be constant. In Figs. 14 and 15, we report the error distributions on the lower boundary of the computational domain (where periodic boundary conditions are imposed), and in the middle of the domain ($y = 0.25$), for the limited schemes (pictures on the left), and for the limited and stabilized schemes (pictures on the right). All the plots show an error peak in correspondence of the shock. This is related

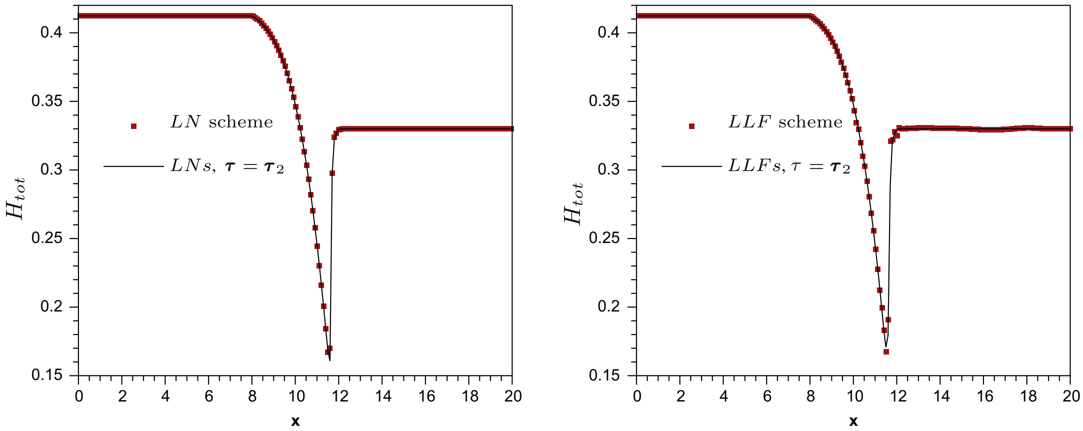


Fig. 13. Pseudo-1D transonic flow: total water height, plot of the data along the line $y = 0.25$. Left: LN and LNs scheme ($\boldsymbol{\tau} = \boldsymbol{\tau}_2$). Right: LLF and LLFs scheme ($\boldsymbol{\tau} = \boldsymbol{\tau}_2$).

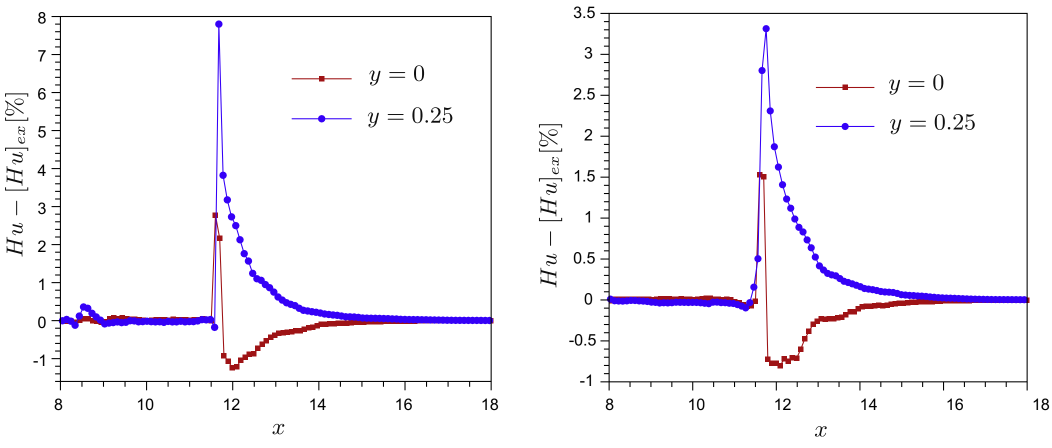


Fig. 14. Pseudo-1D transonic flow: error in the discharge. Left: LN scheme (max ≈ 7.8%). Right: LNs scheme (max ≈ 3.4%, $\boldsymbol{\tau} = \boldsymbol{\tau}_2$). Note that the left and right pictures have different error axis limits.

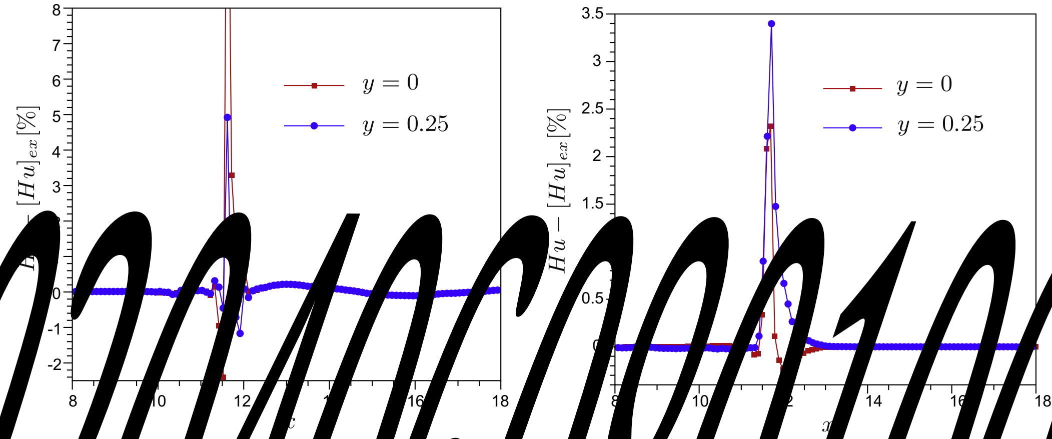


Fig. 15. Pseudo-1D transonic flow error in the discharge. Left: LNs scheme (max ≈ 15.[illegible]%). Right: LLFs scheme (max ≈ 3[illegible]%, $\tau = \tau_2$

to the fact that across the discontinuity the direction of the velocity vector is not well-defined, locally giving place to large errors in its components. Outside the shock, the errors are always below 5%, which is not bad considering that *the problem has been solved using a 2D irregular mesh instead of a 1D one*. What is impressive is the error reduction brought by the stabilization. The errors are reduced roughly by half for the LNs schemes, and to about one fourth for the LLFs scheme. For the latter, in particular, errors are present in a narrower region close to the shock, their absolute value being even smaller than in the case of the LNs scheme.

6.4. Circular dam break

We simulate the break of a circular dam separating two basins with water levels $H = 10$ and $H = 0.5$. The radius of the initial discontinuity is $r = 60$. A sketch of the initial solution is given in Fig. 16. We model only one quarter of the dam, using symmetry boundary conditions along the $y = 0$ and $x = 0$ axes. Our computational domain is hence the square $[0, 100]^2$. The unstructured triangulation used for the simulations is shown in Fig. 16. The reference grid size is $h = 2$. We have run the simulations up to time $t = 3$, and compared the LN, LNs, LLF and LLFs schemes (cf. Sections 3.3.1, 3.3.2, 3.3.3, and 3.3.4) with different definitions of the stabilization parameter $\boldsymbol{\tau}$.

We report a 3D visualization of the water height obtained with the nonlinear stabilized schemes in Figs. 17 and 19. We can see that the outward moving bore is computed without any spurious oscillations.

In a second set of pictures, we visualize the capabilities of the stabilization terms of dissipating the spurious high frequency modes. In our experience, the velocities are the variables more heavily affected by these modes. For this reason,

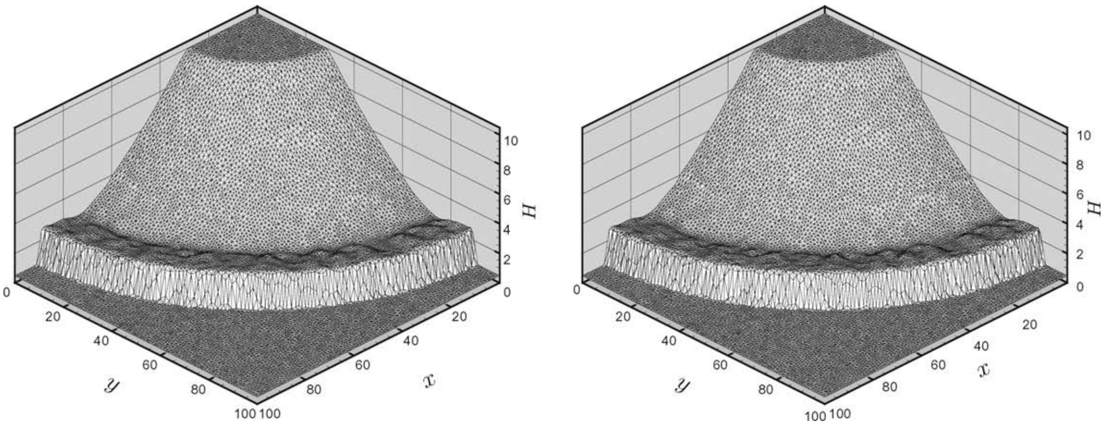
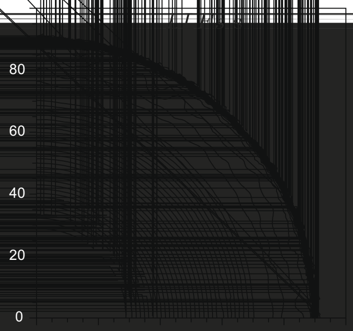


Fig. 17. Circular dam break. 3D plot of the water height. Left: LNs scheme with stabilization parameter $\boldsymbol{\tau}_1$ (equation (87)). Right: LNs scheme with stabilization parameter $\boldsymbol{\tau}_2$ (Eq. (88)).

we evaluate the effect of the stabilization by visualizing in Figs. 18 and 20 the contour plots of the norm of the velocity vector.

The spurious modes in the LN scheme solution are visible in the left picture in Fig. 18. The stabilization term helps somewhat in removing these modes when $\boldsymbol{\tau} = \boldsymbol{\tau}_1$, as visible in the picture in the center on the same figure. To our surprise, the


80
60
40
20
0

a dedicated machine (an Intel(R) Xeon(R) CPU 5160 3.00 GHz processor), and measured the CPU times needed to obtain the final solution.

We present the results in Table 1 for the hydraulic jump problem, in Table 2 for the transonic flow, and in Table 3 for the dam break problem. In the first two tables we report, for all the schemes, the number of iterations needed to achieve the steady state solution for different cut-off values of the residual norm

$$e_H = \log\frac{\|R_H\|_{L^1}}{\|R_H^0\|_{L^1}},$$

having denoted by R_H the array of the water height residuals, and by R_H^0 the array of the water height residuals at the first iteration. Note that the same quantity is used in all the iterative convergence plots reported earlier. In all the tables, we normalize the computational times with respect to the CPU time of the LLFs scheme with $\boldsymbol{\tau} = \boldsymbol{\tau}_2$ (cf. Eq. (88)), which is supposed to be the cheapest of all the schemes. In particular, in the tables we denote by T* the normalized times, and by T the actual (dimensional) CPU time. In conclusion, when larger than unity, T* tells us how many times a given scheme is slower than the LLFs scheme with $\boldsymbol{\tau} = \boldsymbol{\tau}_2$. The same notation is used in Table 3. However, since the results refer in this case to a time dependent problem, we only report the normalized CPU times, and the total number of (physical) time iterations.

A quick look at the tables immediately reveals that the LLFs scheme with $\boldsymbol{\tau} = \boldsymbol{\tau}_2$ is indeed the fastest of all the schemes. In average, the LNs scheme require about 1.4 more CPU time to get the solution in the steady case, and about 1.65 more time in te time dependent case. An exception is obtained at the finest level of convergence of the hydraulic jump, where the supercritical nature of the flow favours the LNs schemes which have a more marked upwind character. The LLFs scheme itself is about twice as slow when using $\boldsymbol{\tau} = \boldsymbol{\tau}_1$.

In particular, concerning the steady state computations, a quick calculation shows that, in comparison with the LLFs scheme with $\boldsymbol{\tau} = \boldsymbol{\tau}_2$, one iteration of the LLFs scheme with $\boldsymbol{\tau} = \boldsymbol{\tau}_1$ is about 1.97 times slower for the hydraulic jump, and about 1.85 times slower for the pseudo one-dimensional test case. This considerable difference is mainly due to the need of assembling and inverting the absolute value Jacobians needed to evaluate (87). Similarly, one iteration with the LNs scheme with $\boldsymbol{\tau} = \boldsymbol{\tau}_2$ is about 2.13 times and 2 times slower for the hydraulic jump and the pseudo one-dimensional problem respectively. Lastly, one iteration of the LNs scheme with $\boldsymbol{\tau} = \boldsymbol{\tau}_1$ is roughly 2.2 and 2.1 times slower for the steady problems considered.

Note that, if the explicit solver were replaced by Newton iterations, these figures might become even more favourable for the LLFs scheme with $\boldsymbol{\tau} = \boldsymbol{\tau}_2$, due to the little matrix algebra needed for the assembly of the residual Jacobian needed for the Newton loop. However, this has to be verified yet.

Concerning the time dependent circular dam break, the figures reported in Table 3 favour even more the LLFs scheme with $\boldsymbol{\tau} = \boldsymbol{\tau}_2$. In particular, the same scheme with $\boldsymbol{\tau} = \boldsymbol{\tau}_1$ gives the most expensive discretization, one (physical) time iteration being roughly 1.72 times slower. Lastly, one time iteration with one of the two LNs schemes is roughly 1.6 times slower.

6.6. *Travelling vortex: grid convergence*

We evaluate the accuracy of the LLFs scheme in time dependent computations on the travelling vortex test case described in Section 2.3. The parameters used in the computations are: $\Gamma = 15$, $\omega = 4\pi$, $\vec{u}_\infty = (6, 0)$, and $g = 1$. The problem is solved on the domain $[0, 1]^2$ with $(x_c, y_c) = (0.5, 0.5)$. In order to follow the movement of the vortex, we apply periodic boundary conditions on the left and right ends of the domain. Weak far field conditions are set on the top and bottom boundaries. We compute the solution up to time $t = 1/6$ when the vortex is back in its initial position.

We start by visualizing the effect of the stabilization in Fig. 21. In the pictures we report, on an unstructured mesh with the same topology of the one in Fig. 16 and $h = 1/80$, the contours of the exact solution (left picture), of the solution obtained with the LLFs scheme with $\boldsymbol{\tau} = \boldsymbol{\tau}_2$ (picture in the center), and with the LLF scheme (rightmost picture). It is evident that the stabilization is successful in removing the spurious modes present in the LLF solution. In the left picture in Fig. 22, we compare the data extracted on the line $y = 0.5$ for the exact solution, and for the LLFs and LLF numerical solutions. The improvement in accuracy brought by the stabilization is quite impressive.

Finally, we report a grid convergence study for the LLFs scheme (with $\boldsymbol{\tau} = \boldsymbol{\tau}_2$), in the right picture in Fig. 22. The results show that we achieve the expected second order of accuracy.

6.7. *Lake at rest tests for the LLFs scheme*

To assess the well-balancedness of the LLFs scheme, we consider here two tests involving a smooth variation of the bathymetry. On the domain $[0, 2] \times [0, 1]$, we consider the following shape of the bed [41,46]:

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

We discretize the spatial domain with an unstructured triangulation with the same topology of the one in Fig. 16, and with $h = 1/100$. As a first test, we impose as initial solution the lake at rest state $[H_{tot}, u, v] = [1, 0, 0]$, and let the time dependent version of the LLFs scheme (with $\boldsymbol{\tau} = \boldsymbol{\tau}_2$, and cf. Sections 3.2 and 3.3) evolve the solution until $t = 0.5$. We then compute the

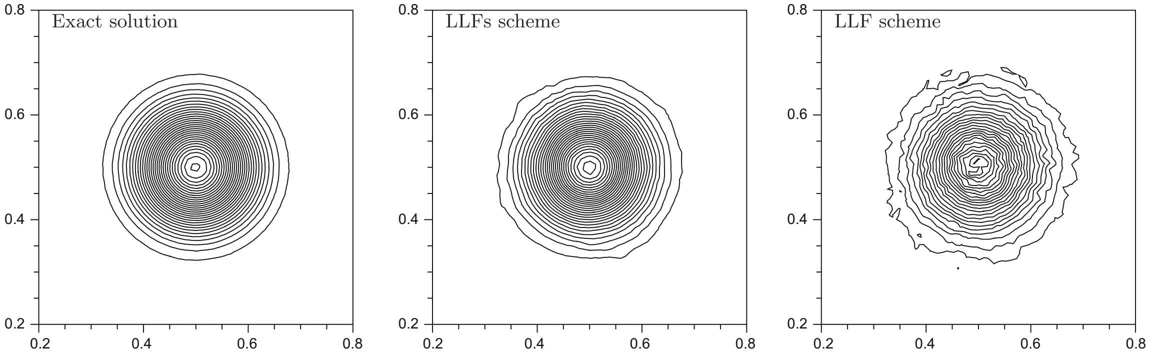


Fig. 21. Vortex advection. Contour plots of the solution after one period. Left: exact/initial solution. Center: LLFs scheme with $\boldsymbol{\tau} = \boldsymbol{\tau}_2$. Right: LLF scheme.

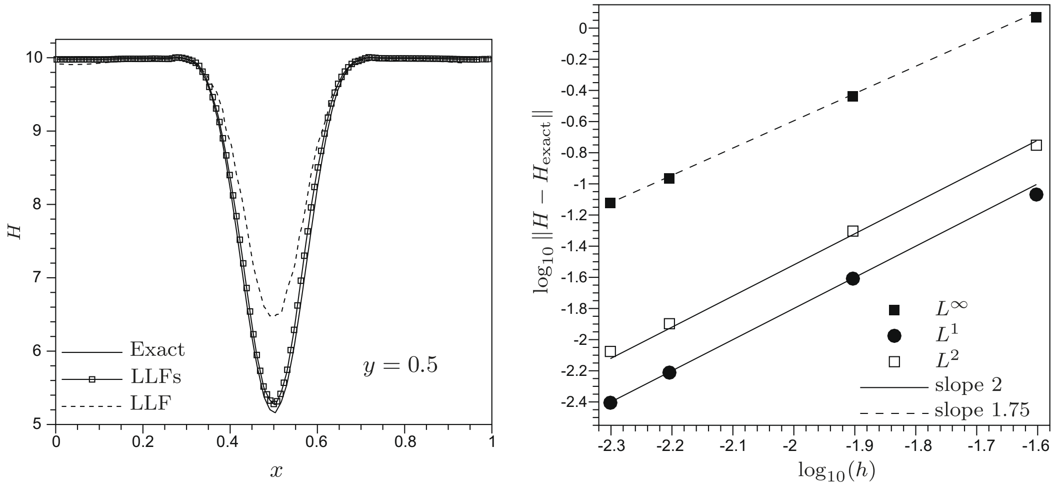


Fig. 22. Vortex advection. Left: solutions along the line $y = 0.5$. Right: grid convergence for the LLFs scheme with $\boldsymbol{\tau} = \boldsymbol{\tau}_2$.

Table 4
Lake at rest solution: errors at time $t = 0.5$, LLFs scheme with $\boldsymbol{\tau} = \boldsymbol{\tau}_2$.

	L^∞	L^1	L^2
$e_{H_{tot}}$	8.955510e−17	2.605999e-17	3.183067e−17
e_u	1.567940e−18	2.485329e−19	3.201703e−19
e_v	1.432740e−18	1.789517e−19	2.327169e−19

norm of the errors $e_{H_{tot}} = \|H_{tot} - 1\|$, $e_u = \|u\|$, and $e_v = \|v\|$. These errors are reported on Table 4, for a computation run in double precision. We preserve the initial solution, down to the machine zero. A similar result was shown in [41] for the LN scheme.

Next, we perturb the initial steady state by setting

$$H_{tot} = \begin{cases} 1.01 & \text{if} \quad 0.05 < x < 0.15 \\ 1 & \text{otherwise} \end{cases}.$$

We compute the solution with the LLFs scheme until time $t = 0.48$, using $g = 9.8182$ for the gravity acceleration. We report snapshots of the solutions obtained at times $t = 0.12$, $t = 0.24$, $t = 0.36$, and $t = 0.48$ in Figs. 23–26, respectively. In the figures, on the left we plot 30 contours of the water height (from 0.92 to 1.011 in all the pictures), while on the right we report the data extracted along the line $y = 0.5$.[2]

[2] The bed height has been rescaled for plotting purposes.

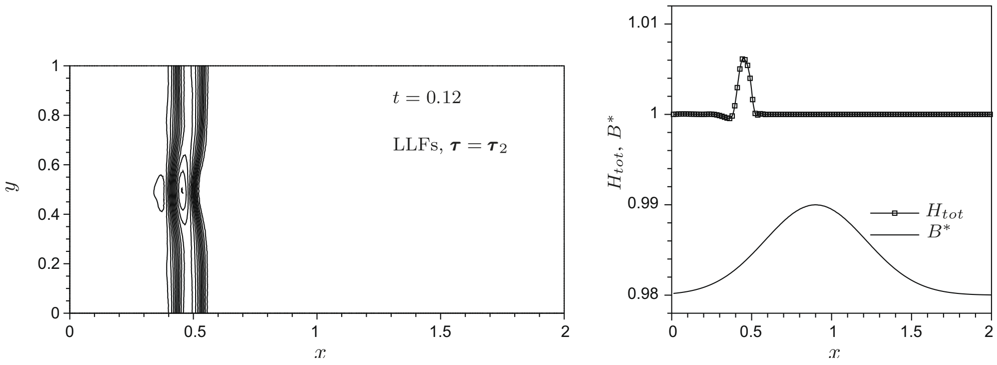


Fig. 23. Perturbation of the steady lake at rest solution, $t = 0.12$. LLFs scheme with $\boldsymbol{\tau} = \boldsymbol{\tau}_2$. Left: contour plot of the total water height. Right: solution along the line $y = 0.5$. Note: $B^* = B/80 + 0.98$.

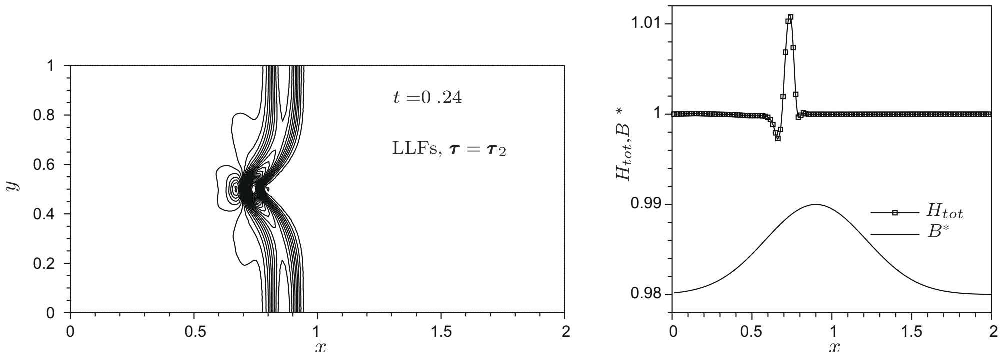
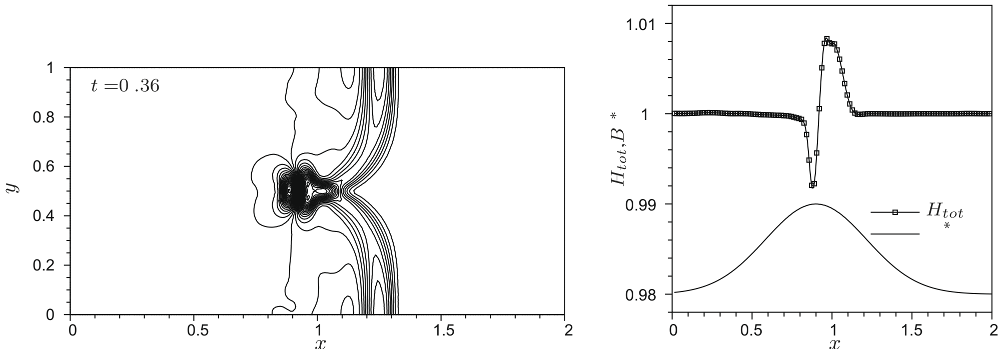


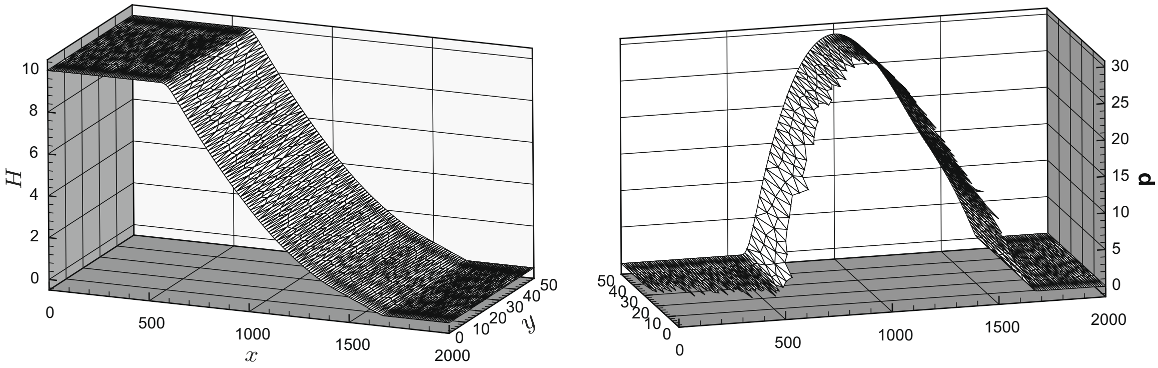
Fig. 24. Perturbation of the steady lake at rest solution, $t = 0.24$. LLFs scheme with $\boldsymbol{\tau} = \boldsymbol{\tau}_2$. Left: contour plot of the total water height. Right: solution along the line $y = 0.5$. Note: $B^* = B/80 + 0.98$.



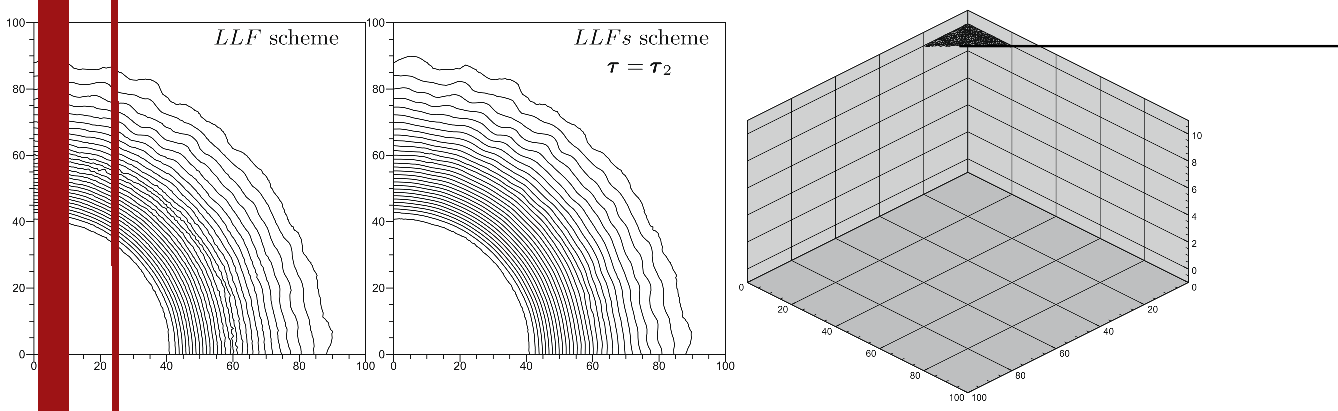
Our results compare very favorably with the ones reported in [41,46,54,55]. In particular, while being clearly well-balanced and preserving the lake at rest state before the perturbation, the stabilized LLFs scheme yields a nice reproduction of the interaction of the incoming wave with the non-flat bottom.

6.8. *Pseudo-1D dam break on dry bed*

We consider now a one-dimensional dam break on dry bed [46]. The test involves the break of a dam separating a basin containing 10 meters of water from a dry region. The bed slope is zero everywhere. We solved the problem on the two dimensional domain $[0, 2000] \times [0, 50]$, imposing periodic boundary conditions in the y-direction. The triangular mesh used is similar to the one in Fig. 4. We computed the solution up to time $t = 40$ with the LLFs scheme with $\boldsymbol{\tau} = \boldsymbol{\tau}_2$.



We report in Fig. 27 a three dimensional visualization of the solution obtained with the LLFs scheme setting the cut-off constant to $C_{\ddot{u}} = 10^{-12}$ (cf. Section 4.3.1). One-dimensional plots of the data extracted in the middle of the domain (line $y = 25$) are instead shown in Fig. 28. In the pictures, the influence of the choice of the cut-off constant $C_{\ddot{u}}$ is also shown. The results are practically independent on the value of this parameter. Differences in the distribution of the discharge Hu



are only observed when $C_{\vec{u}} > 10^{-5}$, while we had to go above $C_{\vec{u}} = 10^{-4}$ to be able to see differences in the water height distribution. Independently of the value of this parameter, the schemes keep the water height positive, confirming the analysis of Section 4.2. The smooth variation of the water height and of the discharge are well reproduced, thanks to the stabilization.

6.9. Circular dam break on dry bed

This test is a two dimensional variant of the previous one. Computational domain, mesh, and initial solution are as in the circular dam break test of Section 6.4, except that now the "left and right" states are $H = 10$ and $H = 0$. The final time of the simulation is $t = 1.75$.

The results are visualized in Fig. 29. In the left and middle pictures, the beneficial effect of the stabilization term is visible. In these pictures, we report a contour plot of the water height for the LLF scheme and the LLFs one, with $\boldsymbol{\tau} = \boldsymbol{\tau}_2$. The stabilized scheme gives nice smooth and nicely circular contours. A three dimensional view of the LLFs scheme solution is reported on the right in the same figure. The picture shows a nice reproduction of the wetting of the dry area. No oscillations are present. The value of the cut-off constant $C_{\vec{u}}$ used to obtain these results is $C_{\vec{u}} = 10^{-12}$. The degree of circular symmetry of the solution can be seen in Fig. 30, where we have reported plots of the data along the lines $y = 0$, $y = x$, and $x = 0$ for the LLFs solution.

Finally, the influence of the cut-off constant $C_{\vec{u}}$ is visualized in the line plots of Fig. 31. The solutions obtained with the different definitions of $\boldsymbol{\tau}$ are virtually identical. For this reason, all the following computations have been run using Eq. (86) to estimate the value of this parameter.

Note that our results compare very favorably with the ones of [9,46].

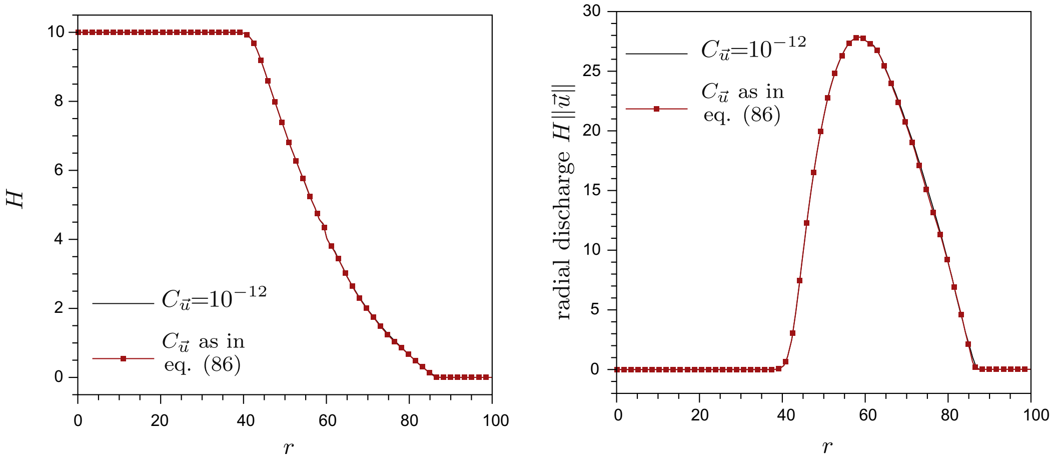
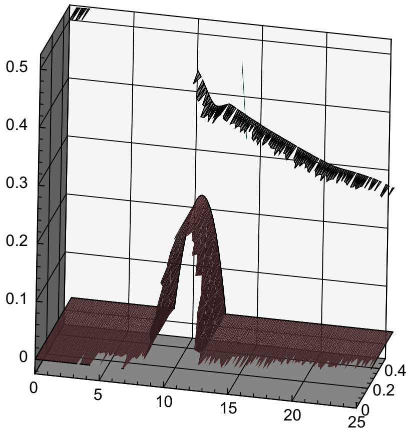


Fig. 31. Circular dam break over dry bed. LLFs scheme, and $\boldsymbol{\tau} = \boldsymbol{\tau}_2$. Influence of the cut-off $C_{\vec{u}}$. Data extracted from the line $y = x$. Left: water height. Right: radial discharge $H\|\vec{u}\|$.



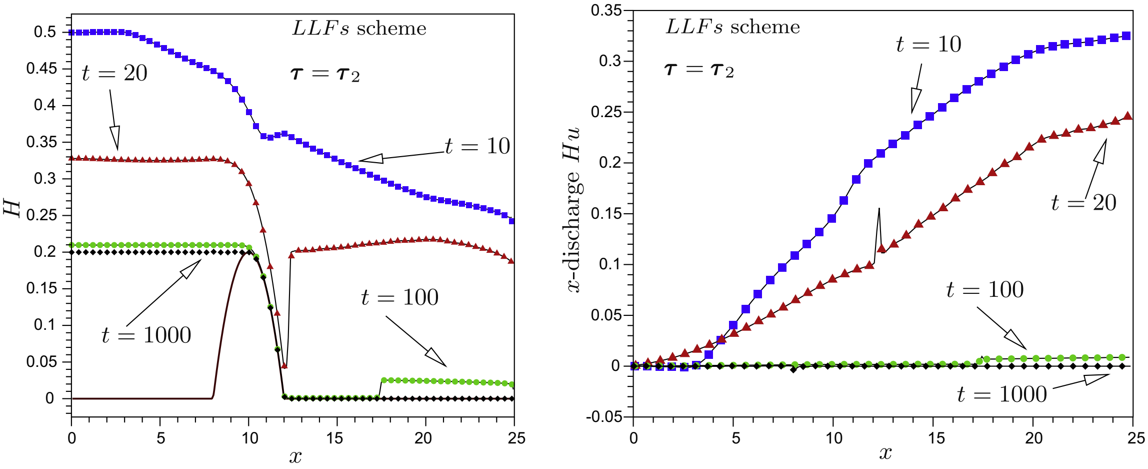


Fig. 34. Pseudo-1D drain on smooth hump. LLFs scheme with $\boldsymbol{\tau} = \boldsymbol{\tau}_2$ (Eq. (88)), and $C_{\vec{u}}$ as in (86). Data extracted from the lines $y = 0$ (solid) and $y = 0.25$ (symbols). Left: water height. Right: discharge.

6.10. Pseudo-1D draining over smooth bed

This is another classical test [46,26]. It involves the computation of the drying of a one-dimensional channel of length 25. The variation of the bed height is given by (90). At time $t = 0$, one has $H_{tot} = 0.5$ everywhere, with $\vec{u} = 0$. As time advances, one observes the water flowing out of the domain, until most of it is dry, with the exception of the region ahead of the hump in bed height.

We have solved this problem on the two dimensional domain $[0, 25] \times [0, 0.5]$, with periodic boundary conditions in the y direction. The mesh is an unstructured triangulation similar to the ones used for the other tests. The reference mesh size is $h = 1/200$. We show the results obtained with the LLFs scheme, with $\boldsymbol{\tau} = \boldsymbol{\tau}_2$, and $C_{\vec{u}}$ as in (86). Weak characteristic boundary conditions are used on the left and on the right ends of the domain using the initial solution as reference state on the left, and the dry state $[H, u, v] = [0, 0, 0]$ on the right.

We report in Figs. 32 and 33 a three dimensional visualization of the evolution in time of the water height. Line plots of the data extracted along the lines $y = 0$ and $y = 0.25$ are reported in Fig. 34. In the last figure, the solid lines represent the solution at $y = 0$, while the symbols represent the data at $y = 0.25$. Despite of the fact that we solved the problem on a two dimensional unstructured mesh, our results are in excellent agreement with the ones presented in published literature [46,26].

6.11. Pseudo-1D wetting/drying on sloping shore

This test is taken from [37] (see also [49,36]). It involves the interaction of a solitary wave with a sloping shore. A sketch of the initial solution is reported in Fig. 35. The initial solution is a wave described by the analytical profile $H_0(x) = \max(0, F_H - B)$, and $\vec{u}_0 = (u_0(x), 0)$, with.

$$F_H = D + \delta_H \text{sech}^2(\gamma(x - x_1)), \quad u_0(x) = \sqrt{\frac{g}{D}} H_0(x),$$

and, as in [37], we set $D = 1$, $\delta_H = 0.019$, and

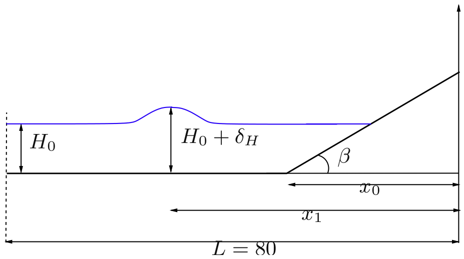


Fig. 35. Pseudo-1D wetting/drying on sloping shore. Sketch of the initial solution.

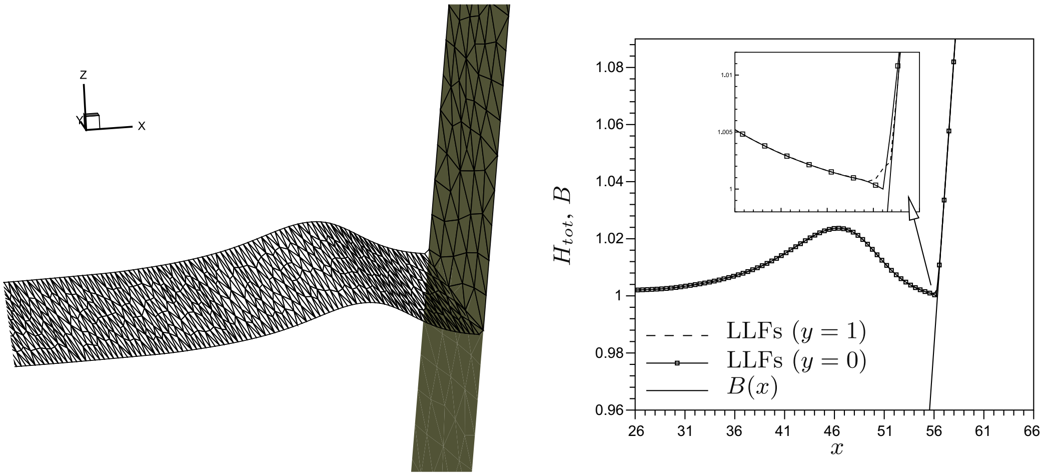


Fig. 36. Pseudo-1D wetting/drying on sloping shore: time $t = 9$. Left : 3D view of the solution. Right: line plots of the data extracted at $y = 0$ and $y = 1$. Solution obtained with the LLFs scheme.

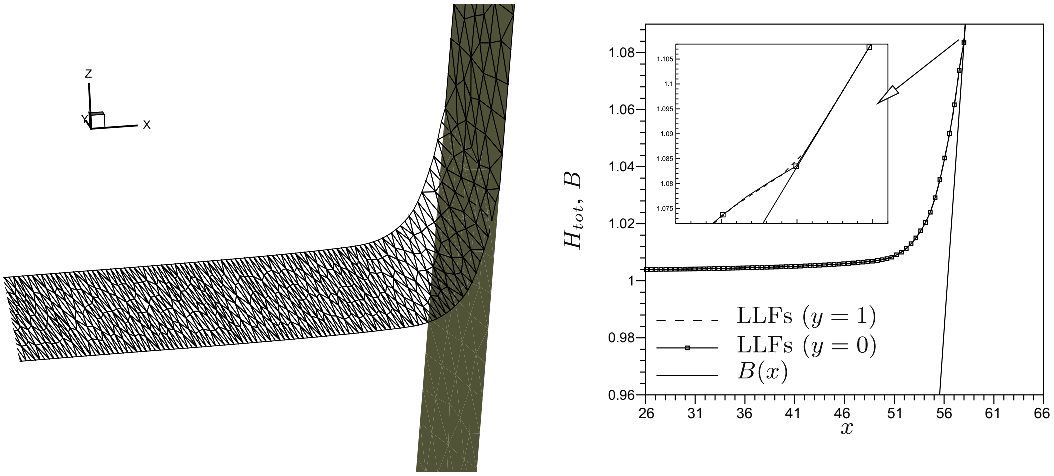


Fig. 37. Pseudo-1D wetting/drying on sloping shore: time $t = 17$. Left : 3D view of the solution. Right: line plots of the data extracted at $y = 0$ and $y = 1$. Solution obtained with the LLFs scheme.

$$\gamma = \sqrt{\frac{3\delta_H}{4D}}, \quad x_1 = \sqrt{\frac{4D}{3\delta_H}}\text{arcosh}\left(\sqrt{\frac{1}{0.05}}\right).$$

The simulations have been run on the domain $[0,80] \times [0,2]$ with the LLFs scheme, and using an unstructured triangulation similar to the one in Fig. 16, and with $h = 0.4$. Periodic boundary conditions have been used in the y-direction, and a weak far field condition has been imposed on the left boundary. In Figs. 36–41 we present the results in terms of a three dimensional

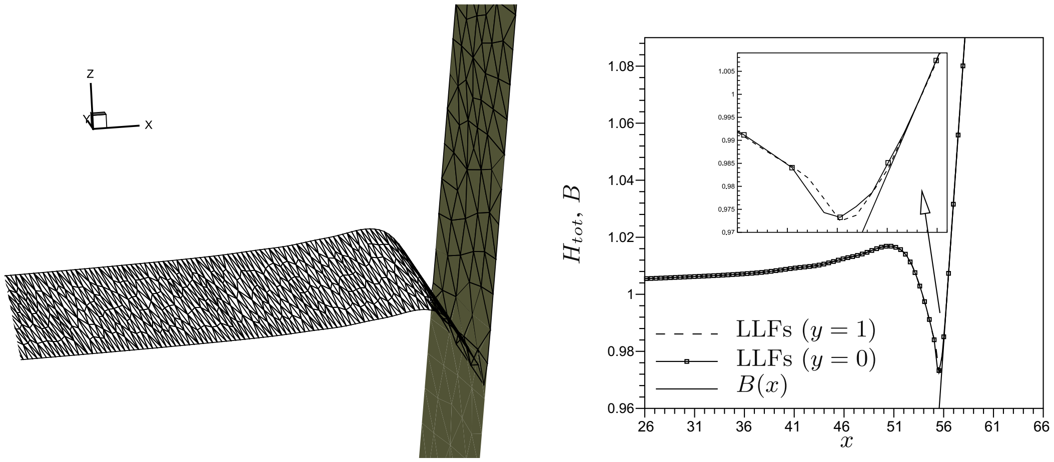


Fig. 38. Pseudo-1D wetting/drying on sloping shore: time $t = 23$. Left : 3D view of the solution. Right: line plots of the data extracted at $y = 0$ and $y = 1$. Solution obtained with the LLFs scheme.

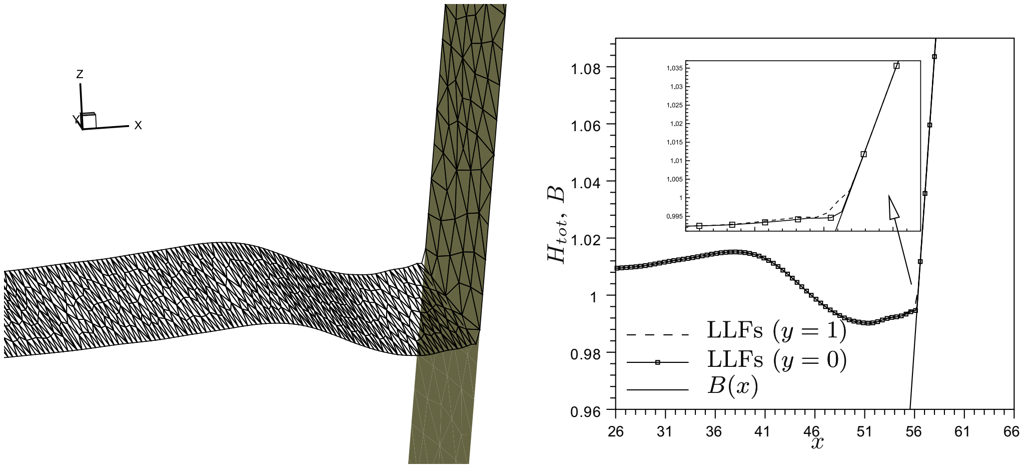


Fig. 39. Pseudo-1D wetting/drying on sloping shore: time $t = 28$. Left : 3D view of the solution. Right: line plots of the data extracted at $y = 0$ and $y = 1$. Solution obtained with the LLFs scheme.

view of the water height profile at different times, and of its one-dimensional distribution on the periodic boundary, and at $y = 1$.

Our results are in very good agreement with the ones presented in [37]. The wetting/drying process is well reproduced, as shown clearly by the one-dimensional plots. At time $t = 75$ the solution has almost reached a stationary state, as confirmed by the fact that we have

$$\left\| \frac{d\mathbf{u}_H}{dt} \right\|_{L^1(\Omega)} < 10^{-6}, \quad \text{where} \quad \left(\frac{d\mathbf{u}_H}{dt} \right)_i = \frac{H_i^{n+1} - H_i^n}{\Delta t}.$$

6.12. Thacker's periodic solutions

We report here the results obtained for the periodic oscillations of Thacker, described in Section 2.3. We have run the simulations for several oscillation periods on the domain $[-2, 2]^2$, using an unstructured grid with the same topology as the one of Fig. 16, and mesh size $h = 1/25$. As in the last test, we used the LLFs scheme with $\boldsymbol{\tau} = \boldsymbol{\tau}_2$, and $C_{\vec{u}}$ as in (86).

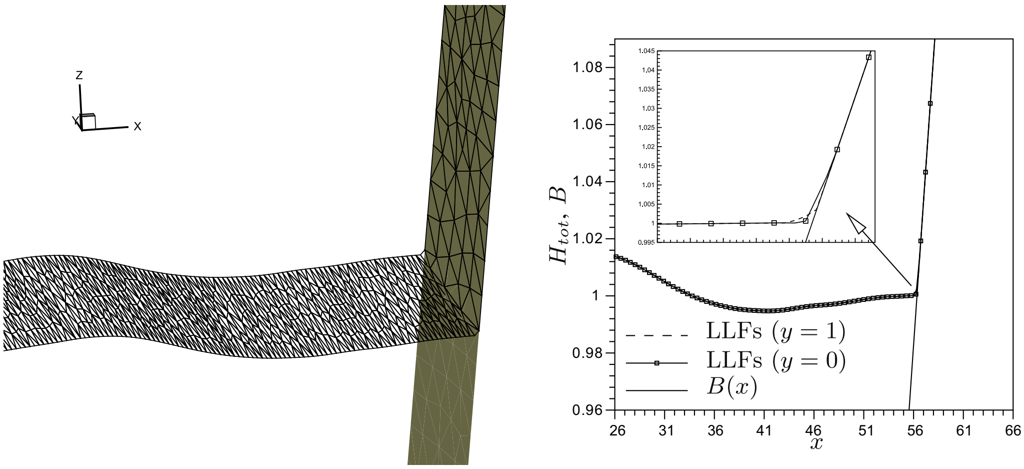


Fig. 40. Pseudo-1D wetting/drying on sloping shore: time $t = 32.5$. Left: 3D view of the solution. Right: line plots of the data extracted at $y = 0$ and $y = 1$. Solution obtained with the LLFs scheme.

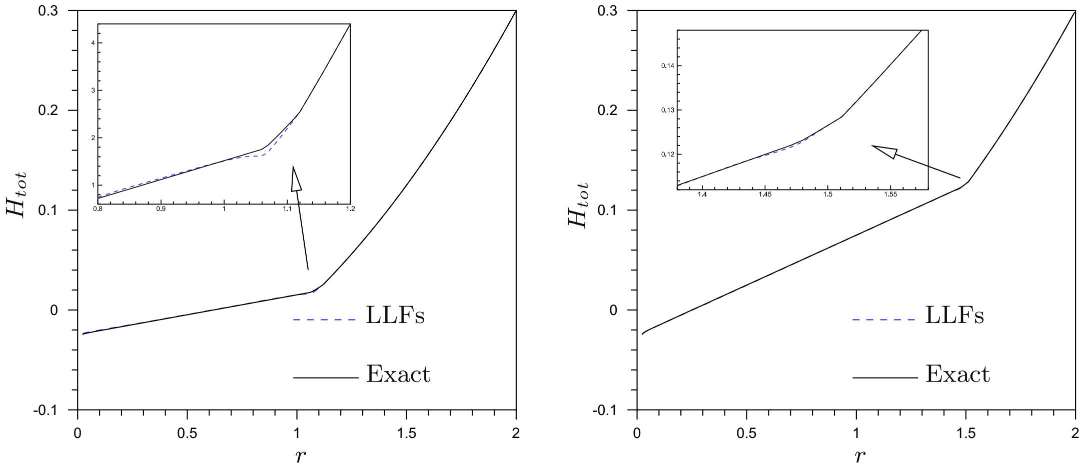


Fig. 43. Thacker's periodic planar solution. Left: time $t = 2T + T/2 + T/3$. Right: time $t = 3T$. Solutions obtained with the LLFs scheme.

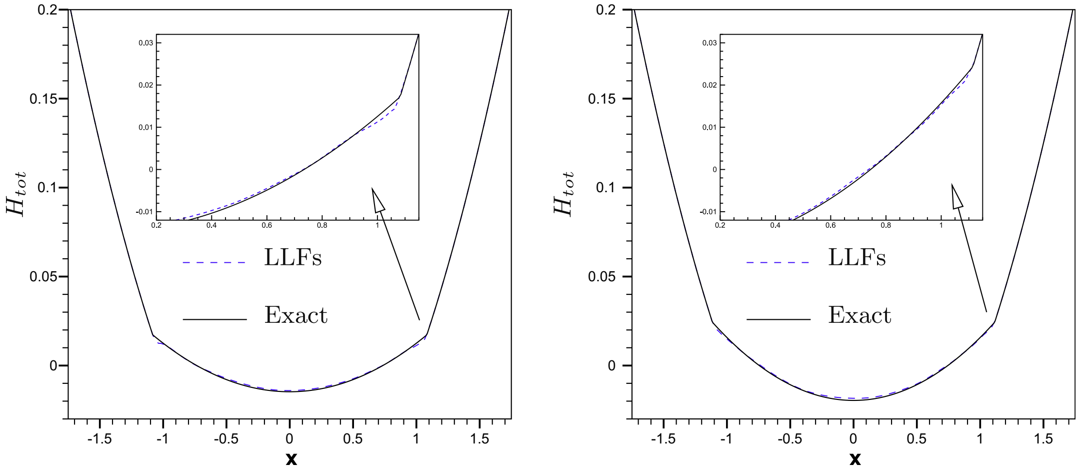


Fig. 44. Thacker's periodic curved solution. Left: time $t = 2T + T/3$. Right: time $t = 2T + T/2$. Solutions obtained with the LLFs scheme.

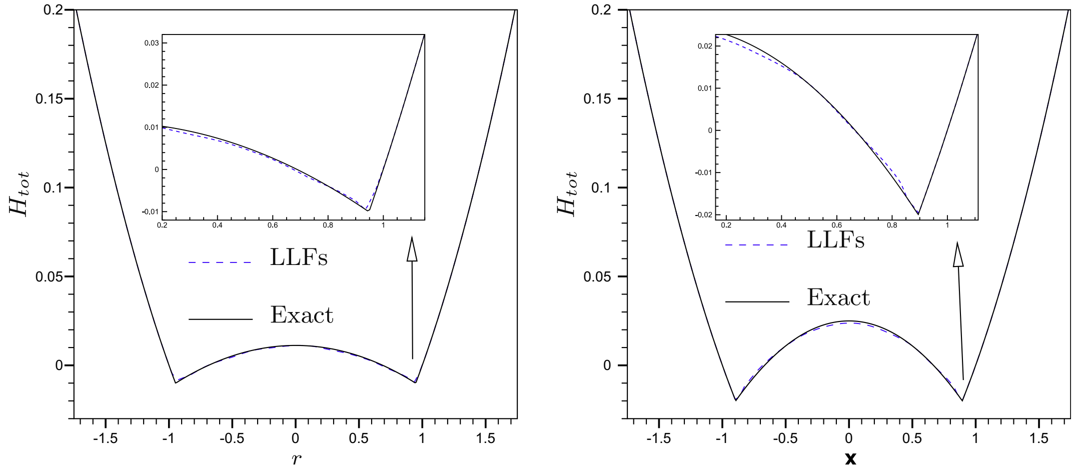


Fig. 45. Thacker's periodic curved solution. Left: time $t = 2T + T/2 + T/3$. Right: time $t = 3T$. Solutions obtained with the LLFs scheme.

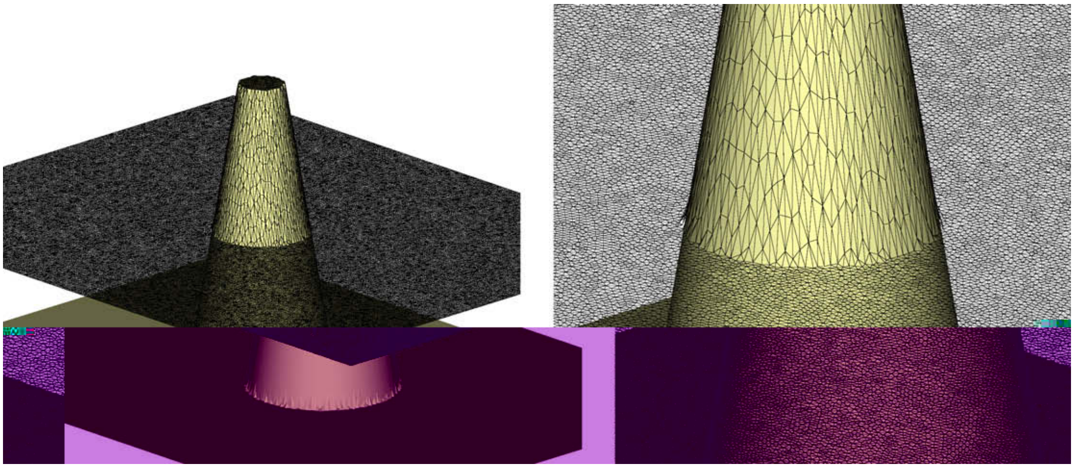


Fig. 46. Run-up on a circular island. 3D visualization of the total water height: lake at rest state.

Table 5
Lake at rest solution: errors at time $t = 5$, LLFs scheme with $\boldsymbol{\tau} = \boldsymbol{\tau}_2$.

	L^∞	L^1	L^2
$e_{H_{tot}}$	2.775558e−17	1.532978e−19	1.643908e−18
e_u	2.221603e−18	7.987680e−21	5.578502e−20
e_v	1.252903e−18	6.400735e−21	4.081257e−20

For the planar solution we set (cf. Section 2.3) $a = 1$, $H_0 = 0.1$, and $\eta = 0.5$. This gives an oscillation period $T \approx 4.44$. For the curved oscillations we set instead $a = 1$, $H_0 = 0.1$, and $r_0 = 0.8$, leading to a period $T \approx 2.22$.

We report the results in Figs. 42 and 43, for the planar oscillations, and in Figs. 44 and 45, for the curved ones. For the planar solution we plot the data extracted along the line $y = 0$ for $x > 0$, while for the curved oscillations we show the data sampled all along the line $y = 0$.

The agreement with the exact solution is excellent. Note that the numerical result is still almost perfectly periodic after three periods. The wetting/drying front is approximated without any spurious oscillations.

6.13. Wave run-up on a conical island

As a last test we consider the runup of a solitary wave over a conical island. We refer to [36,15,32] and references therein for details concerning the bathymetry defining the island.

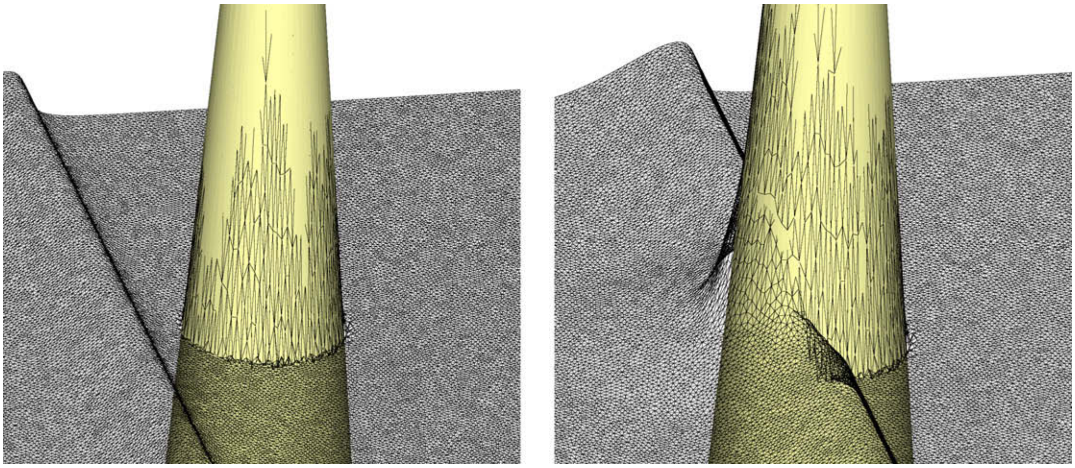


Fig. 47. Run-up on a circular island. 3D visualization of the total water height: front side and lateral runup, formation of symmetric waves.

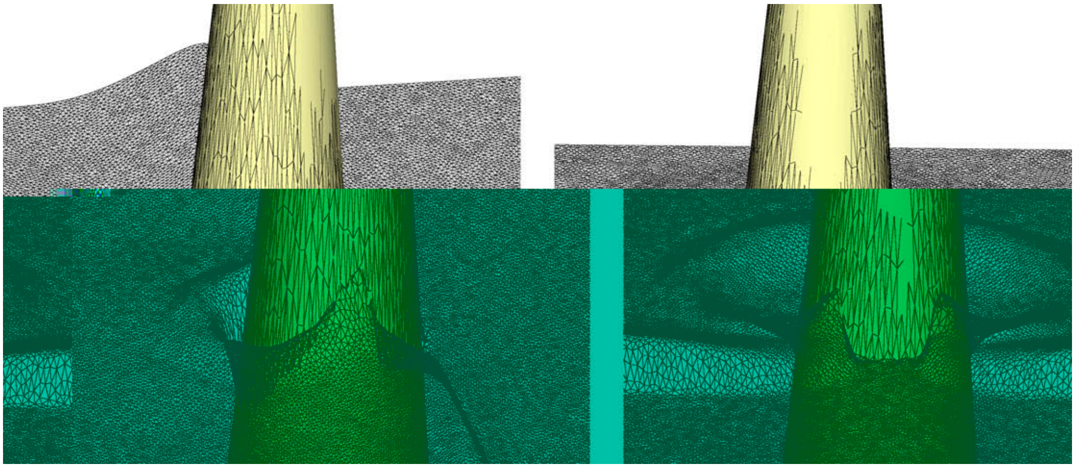


Fig. 48. Run-up on a circular island. 3D visualization of the total water height: lateral runup, propagation of symmetric waves.

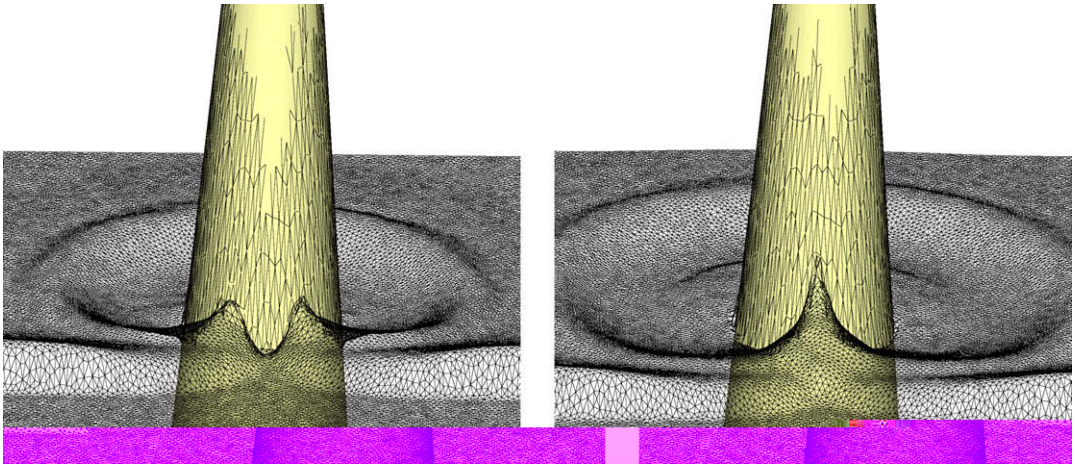


Fig. 49. Run-up on a circular island. 3D visualization of the total water height: symmetric waves joining on the rear side, and rear side runup.

As in [36], the simulations have been run on the computational domain $[0,25] \times [0,30]$. Following [32], we choose a coordinate system with origin located at the right end of the wavemaker. The conical island is then centered at $(x_c, y_c) = (12.96, 13.80)$. We impose far field characteristic boundary conditions on all the boundaries. The reference mesh size of the unstructured triangulation used for the simulations is $h = 0.25$ All the computations have been run with the LLFs scheme, $\boldsymbol{\tau} = \boldsymbol{\tau}_2$, and $C_{\vec{u}}$ given by (86).

Before presenting the simulation of the runup, as for the test of Section 6.6, we evaluate the well balancedness of the LLFs scheme, by initializing the solution with the lake at rest state obtained by setting

$$[H_{tot}, u, v] = [B(x,y) + \max(0, H_0 - B(x,y)), 0, 0],$$

with $H_0 = 0.32$ [36]. A visualization of the overall geometry, and of this initial solution is given in Fig. 46. We report in Table 5 the errors with respect to this stationary state obtained at time $t = 5$ with the LLFs scheme. As in Section 6.6, we find that the lake at rest is preserved down to machine zero, even in presence of dry areas.

Next we perturb the steady state by imposing a solitary wave defined by a hyperbolic secant. Details can be found in [36,37,32]. In Figs. 47–49 a three dimensional visualization of the runup process is shown.

The following remarks can be made. All the results are quite clean. No oscillations are visible. The preservation of the lake at rest state before the wave reaches the island can be clearly seen on the left in Fig. 47. The run up of the front side of the island is visible on the right in the same figure. In Figs. 48 and 49, we can observe the formation of the two waves running around the island and joining on its rear side. This eventually leads to the run-up on the rear side of the island, which is clearly visible in Fig. 49. Very good agreement with the results of [36] can be observed.

At last, we visualize in Fig. 50 the time evolution of the free surface parameter

$$\eta = H_{tot}(t) - H_{tot}(t = 0),$$

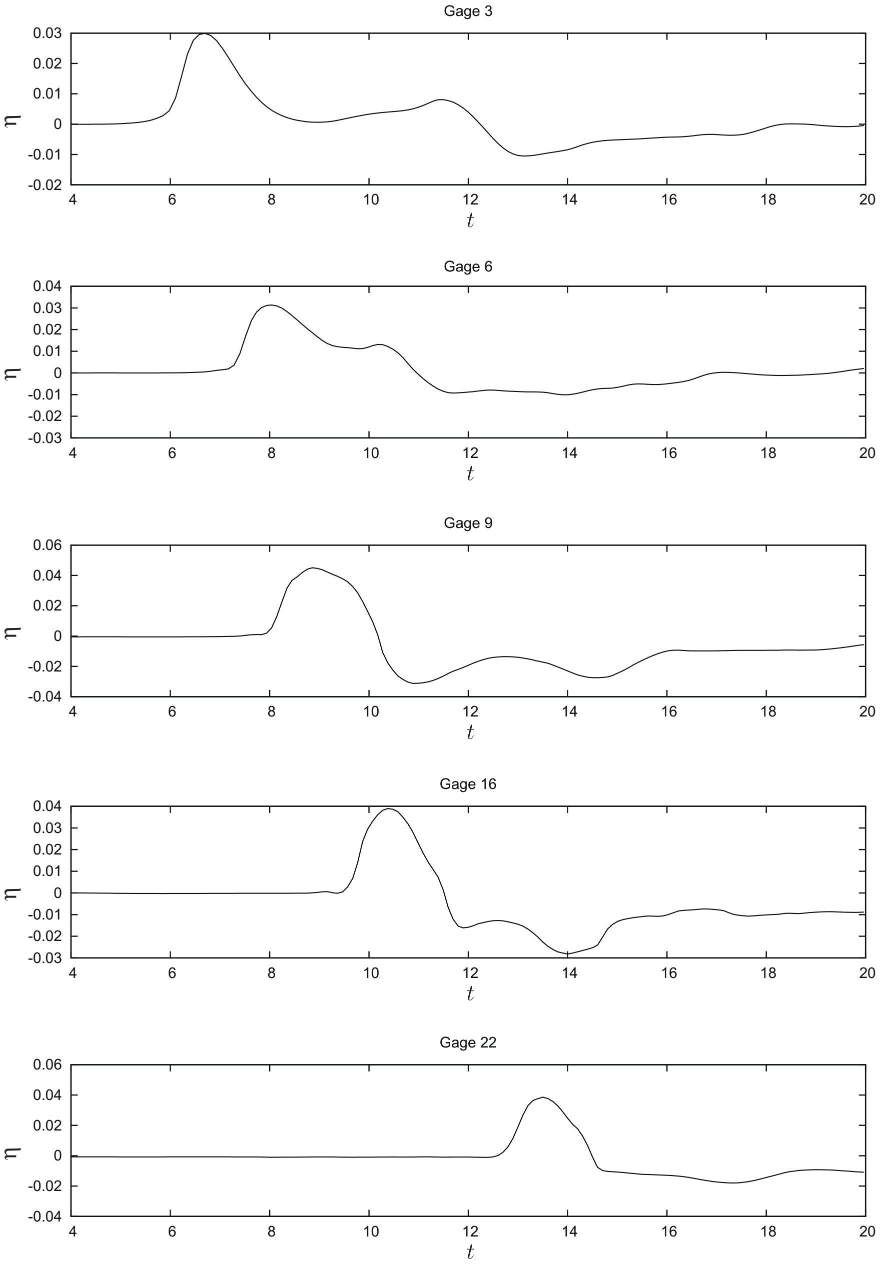


Fig. 50. Run-up on a circular island. Time variation of the free surface $\eta = H_{tot} - H_{tot}(t = 0)$ at wave gages 6, 9, 16, and 22 of benchmark problem of [32].

in the gage points given in [32]. Due to the unstructured nature of the mesh, we do have a slight difference in the position of these probes. In particular, the location of the gages for our computations is the following:

$$\begin{aligned}
&\text{Gage3}: \quad x = 6.796, \quad y = 13.045\\
&\text{Gage6}: \quad x = 9.273, \quad y = 13.722\\
&\text{Gage9}: \quad x = 10.365, \quad y = 13.789\\
&\text{Gage16}: \quad x = 12.930, \quad y = 11.213\\
&\text{Gage22}: \quad x = 15.560, \quad y = 13.800,
\end{aligned}$$

which are very close to the ones used in the reference. The time evolution of η obtained here agrees very well with the results reported in [32], especially for the first three gages, which are located on the front of the conical island. The results of the last two probes, located on the rear of the island, are still quite close to those of the reference. However a phase shift is present which might be explained by the fact that we neglected the friction term in the discharge equations, which is instead considered in [32]. We judge these results very encouraging.

7. Conclusions

We have discussed an approach to discretize the shallow water equations on unstructured grids, including dry areas. The technique proposed is based on a stabilized nonlinear variant of a multidimensional Lax–Friedrichs scheme, obtained adapting the ideas of [2,40].

The schemes proposed are conservative, well balanced, and second order accurate. The numerical results clearly demonstrate their capabilities in handling discontinuous flows and dry geometries without spurious oscillations. The preservation of the positivity of the water height is guaranteed under a time step constraint, without the need of a cutoff on the water height itself.

Our results are certainly comparable with the ones obtained with state of the art Finite Volume Godunov discretizations for Shallow Water simulations (*e.g.* [46,32,9,36,26,25] and references therein). However, the residual approach at the basis of our work allows is easily generalized to very high order of accuracy on unstructured grids, without losing any of its basic properties such as compactness, non-oscillatory behavior, and positivity of the water height. How to do this for steady state problems is discussed for example in [5].

There are however a number of important issues to be investigated and improved. From a practical point of view, it is certainly necessary to replace the explicit iteration procedure with a Newton solver. From the point of view of efficiency, however, in our view the most disappointing fact is to have an implicit highly nonlinear discretization, and still being subject to a time step constraint in unsteady simulations. The existence of an upper bound for the size of the time step is a known condition for the preservation of the positivity of the solution when integrating time dependent conservation laws [14]. However, we still think the fact that an implicit discretization needing a nonlinear iterative solution process should allow the use of large time steps. Residual distribution schemes with this property have been proposed in the past [6,19], based on a two-layer solution procedure. This might be a solution. Another possible route could be the use of a space–time framework with a discontinuous representation in time. This will definitely be the subject of research in the near future.

An additional important point will be to find formulas for the stabilization matrix $\boldsymbol{\tau}$ allowing an optimal scaling of the stabilization, without the need of performing any matrix inversion (as in (87)). Definition (88) is an attempt but it is far from optimal. The work published in [29,30,39] gives possible guidelines to achieve this goal.

Lastly, concerning the wetting/drying process, we are sure that improvements are still possible. As in some published Finite Volume solvers (see *e.g.* [25] and references therein), we might for example exploit exact linearized solvers at the wet/dry interface to devise a distribution strategy not only based on the positivity requirement, as we do now. Lastly, it would be interesting to find an improved procedure allowing the preservation of steady lake at rest states in front cells at the same time not perturbing the slope in cells with upward moving flow.

Acknowledgments

We are grateful to R. Abgrall. His numerous remarks made our work substantially easier. S. Noelle is also acknowledged for many useful discussions on shallow water simulations.

Appendix A

A.1. Jacobian eigenvectors

Given $\vec{\xi} = (\xi_x, \xi_y) \in \mathbb{R}^2$, with $\|\xi\| = 1$, the Jacobian $K(\vec{\xi}, \mathbf{u})$ (Eq. (15)) for the SWE is

$$K(\vec{\xi}, \mathbf{u}) = \begin{bmatrix} 0 & \xi_x & \xi_y \\ a^2\xi_x - u\vec{u}\cdot\vec{\xi} & u\xi_x + \vec{u}\cdot\vec{\xi} & u\xi_y \\ a^2\xi_y - v\vec{u}\cdot\vec{\xi} & v\xi_x & v\xi_y + \vec{u}\cdot\vec{\xi}, \end{bmatrix}$$

where $\vec{u} = (u, v)$ is the flow speed, and with the celerity $a = \sqrt{gH}$. The eigenvalues of $K(\vec{\xi}, \mathbf{u})$ are obtained by (11), setting $\|\xi\| = 1$. The corresponding right eigenvectors are given by

$$\mathbf{r}_1 = \begin{bmatrix} 0 \\ -a\xi_x \\ a\xi_y \end{bmatrix}, \quad \mathbf{r}_2 = \begin{bmatrix} 1 \\ u + a\xi_x \\ v + a\xi_y \end{bmatrix}, \quad \mathbf{r}_3 = \begin{bmatrix} 1 \\ u - a\xi_x \\ v - a\xi_y \end{bmatrix}$$

The expression for the left eigenvectors is the following:

$$\mathbf{l}_1 = \frac{1}{a}\left[u\xi_y - v\xi_x, -\xi_y, \xi_x\right]$$
$$\mathbf{l}_2 = \frac{1}{2a}\left[a - \vec{u}\cdot\vec{\xi}, \xi_x, \xi_y\right]$$
$$\mathbf{l}_3 = \frac{1}{2a}\left[a + \vec{u}\cdot\vec{\xi}, -\xi_x, -\xi_y\right]$$

References

[1] R. Abgrall, Toward the ultimate conservative scheme: following the quest, J. Comput. Phys. 167 (2) (2001) 277–315.

[2] R. Abgrall, Essentially non oscillatory residual distribution schemes for hyperbolic problems, J. Comput. Phys. 214 (2) (2006) 773–808.

[3] R. Abgrall, T.J. Barth, Residual distribution schemes for conservation laws via adaptive quadrature, Siam. J. Sci. Comput. 24 (3) (2002) 732–769.

[4] R. Abgrall, A. Larat, M. Ricchiuto, Construction of high order residual distribution schemes, in: 35th CFD VKI/ADIGMA Course on High Order Discretization Methods, 2008.

[5] R. Abgrall, A. Larat, M. Ricchiuto, C. Tavé. Simplified stabilisation procedures for residual distribution schemes. Comput. Fluids, in press, doi:10.1016/j.compfluid.2008.01.031.

[6] R. Abgrall, M. Mezine, Construction of second order accurate monotone and stable residual distribution schemes for unsteady flow problems, J. Comput. Phys. 188 (2003) 16–55.

[7] R. Abgrall, M. Mezine, Construction of second-order accurate monotone and stable residual distribution schemes for steady flow problems, J. Comput. Phys. 195 (2004) 474–507.

[8] R. Abgrall, P.L. Roe, High order fluctuation schemes on triangular meshes, Siam. J. Sci. Comput. 19 (3) (2003) 3–36.

[9] F. Alcrudo, P.G. Navarro, A high resolution Godunov-type scheme in finite volumes for the 2d shallow water equations, Int. J. Numer. Meth. Fluids 16 (1993) 489–505.

[10] E. Audusse, Personal communication, 2006.

[11] T.J. Barth, An energy look at the N scheme, Working Notes (1996).

[12] T.J. Barth, Numerical methods for gasdynamic systems on unstructured meshes, in: D. Kröner, M. Ohlberger, C. Rohde (Eds.), An Introduction to Recent Developments in Theory and Numerics for Conservation Laws, of Lecture Notes in Computational Science and Engineering, vol. 5, Springer-Verlag, Heidelberg, 1998, pp. 195–285.

[13] P.B. Bochev, M.D. Gunzburger, J.N. Shadid, Stability of the SUPG finite element method for transient advection–diffusion problems, Comp. Meth. Appl. Mech. Eng. 193 (23–26) (2004) 2301–2323.

[14] C. Bolley, M. Crouzeix, Conservation de la positivité lors de la discétization des problèmes d'èvolution paraboliques, R.A.I.R.O. Analyse Numérique 12 (1978) 237–254.

[15] M.J. Briggs, C.E. Synolakis, G.S. Harkins, D.R. Green, Laboratory experiments of tsunami runup on a circular island, Pure Appl. Geophys. 144 (3/4) (1995) 569–593.

[16] P. Brufau, P. García-Navarro, Unsteady free surface flow simulation over complex topography with a multidimensional upwind technique, J. Comput. Phys. 186 (2) (2003) 503–526.

[17] D. Caraeni, L. Fuchs, Compact third-order multidimensional upwind discretization for steady and unsteady flow simulations, Comput. Fluids 34 (4-5) (2005) 419–441.

[18] Á. Csík, M. Ricchiuto, H. Deconinck, A conservative formulation of the multidimensional upwind residual distribution schemes for general nonlinear conservation laws, J. Comput. Phys. 179 (2) (2002) 286–312.

[19] Á. Csík, M. Ricchiuto, H. Deconinck, S. Poedts, Space–time residual distribution schemes for hyperbolic conservation laws, in: 15th AIAA Computational Fluid Dynamics Conference, Anahein, CA, USA, June 2001.

[20] H. Deconinck, M. Ricchiuto, Residual distribution schemes: foundation and analysis, in: E. Stein, R. de Borst, T.J.R. Hughes (Eds.), Encyclopedia of Computational Mechanics, John Wiley & Sons Ltd., 2007. doi:10.1002/0470091355.ecm054.

[21] A.I. Delis, Th. Katsaounis, Relaxation schemes for the shallow water equations, Int. J. Numer. Meth. Fluids 41 (2003) 695–719.

[22] P. De Palma, G. Pascazio, G. Rossiello, M. Napolitano, A second-order accurate monotone implicit fluctuation splitting scheme for unsteady problems, J. Comput. Phys. 208 (1) (2005) 1–33.

[23] J. Dobes, H. Deconinck, Second order blended multidimensional upwind residual distribution scheme for steady and unsteady computations, J. Comput. Appl. Math. 215 (1) (2006) 378–389.

[24] A. Ferrante, H. Deconinck, Solution of the unsteady Euler equations using residual distribution and flux corrected transport, Technical Report VKI-PR 97-08, von Karman Institute for Fluid Dynamics, 1997.

[25] J.M. Gallardo, C. Parés, Manuel Castro, On a well-balanced high-order finite volume scheme for shallow water equations with topography and dry areas, J. Comput. Phys. 227 (1) (2007) 574–601.

[26] T. Gallouët, J.-M. Hérard, N. Seguin, Some approximate Godunov schemes to compute shallow-water equations with topography, Comput. Fluids 32 (2003) 479–513.

[27] A. Harten, On the symmetric form of systems of conservation laws with entropy, J. Comput. Phys. 49 (1) (1983) 151–164.

[28] G. Hauke, A symmetric formulation for computing transient shallow water flows, Comp. Meth. Appl. Mech. Eng. 163 (1998) 111–122.

[29] G. Hauke, Simple stabilizing matrices for the computation of compressible flows in primitive variables, Comp. Meth. Appl. Mech. Eng. 190 (2001) 6881–6893.

[30] G. Hauke, T.J.R. Hughes, A comparative study of different sets of variables for solving compressible and incompressible flows, Comp. Meth. Appl. Mech. Eng. 153 (1998) 1–44.

[31] M.E. Hubbard, M. J Baines, Conservative multidimensional upwinding for the steady two-dimensional shallow water equations, J. Comput. Phys. 138 (1997) 419–448.

[32] M.E. Hubbard, N. Dodd, A 2d numerical model of wave run-up and overtopping, Coast. Eng. 47 (1) (2002) 1–26.

[33] M.E. Hubbard, P.L. Roe, Compact high resolution algorithms for time dependent advection problems on unstructured grids, Int. J. Numer. Meth. Fluids 33 (5) (2000) 711–736.

[34] G.J. LeBeau, S.E. Ray, S.K. Aliabadi, T.E. Tezduyar, SUPG finite element computations of compressible flows with the entropy formulation and conservation variables formulations, Comp. Meth. Appl. Mech. Eng. 104 (1993) 422–497.

[35] J. Maerz, G. Degrez, Improving time accuracy of residual distribution schemes, Technical Report VKI-PR 96-17, von Karman Institute for Fluid Dynamics, 1996.

[36] F. Marche. Theoretical and numerical study of shallow water models, Applications to nearshore hydrodynamics, PhD thesis, Université de Bordeaux I, 2005. Available online: <www.math.u-bordeaux1.fr/~marche/THESE_Marche.pdf>.

[37] F. Marche, P. Bonneton, P. Fabrie, N. Seguin, Evaluation of well-balanced bore-capturing schemes for 2D wetting and drying processes, Int. J. Numer. Meth. Fluids 53 (2007) 867–894.

[38] H. Paillère, H. Deconinck, Compact cell vertex convection schemes on unstructured meshes, in: H. Deconinck, B. Koren (Eds.), Euler and Navier–Stokes Solvers Using Multi-Dimensional Upwind Schemes and Multigrid Acceleration, of Notes on Numerical Fluid Mechanics, vol. 57, Vieweg, Braunchweig, 1997.

[39] M. Polner, L. Pesch, J.J.W. van der Vegt, Construction of stabilization operators for Galerkin least-squares discretizations of compressible and incompressible flows, Comp. Meth. Appl. Mech. Eng. 196 (2007) 2431–2448.

[40] M. Ricchiuto, R. Abgrall, Stable and convergent residual distribution for time dependent conservation laws, in: ICCFD4 Proceedings, Springer-Verlag, 2006.

[41] M. Ricchiuto, R. Abgrall, H. Deconinck, Application of conservative residual distribution schemes to the solution of the shallow water equations on unstructured meshes, J. Comput. Phys. 222 (2007) 287–331.

[42] M. Ricchiuto, Á. Csík, H. Deconinck, Residual distribution for general time dependent conservation laws, J. Comput. Phys. 209 (1) (2005) 249–289.

[43] M. Ricchiuto, N. Villedieu, R. Abgrall, H. Deconinck, On uniformly high-order accurate residual distribution schemes for advection–diffusion, J. Comput. Appl. Math. 215 (2) (2008) 378–389.

[44] G. Rossiello, P. De Palma, G. Pascazio, M. Napolitano, Third-order-accurate fluctuation splitting schemes for unsteady hyperbolic problems, J. Comput. Phys. 222 (1) (2007) 332–352.

[45] G. Rossiello, P. De Palma, G. Pascazio, M. Napolitano, Second-order-accurate explicit fluctuation splitting schemes for unsteady problems. Comput. Fluids, in press, doi:10.1016/j.compfluid.2008.01.021.

[46] M. Seaïd, Non-oscillatory relaxation methods for the shallow-water equations in one and two space dimensions, Int. J. Numer. Meth. Fluids 46 (2004) 457–484.

[47] M. Seaïd, Personal communication, 2006.

[48] R. Struijs, H. Deckoninck, P.L. Roe, Fluctuation splitting schemes for the 2d Euler equations VKI LS 1991-01, Comput. Fluid Dyn. (1991).

[49] C.E. Synolakis, The runup of solitary waves, J. Fluid Mech. 185 (1987) 523–545.

[50] E. Tadmor, Skew-self-adjoint form for systems of conservation laws, J. Math. Anal. Appl. 103 (1984) 428–442.

[51] T.E. Tezduyar, M. Senga, Stabilization and shock-capturing parameters in SUPG formulation of compressible flows, Comp. Meth. Appl. Mech. Eng. 195 (2006) 621–1632.

[52] W.C. Thacker, Some exact solutions to the nonlinear shallow-water wave equations, J. Fluid Mech. 107 (1981) 499–508.

[53] E. van der Weide, H. Deconinck, Positive matrix distribution schemes for hyperbolic systems, in: Computational Fluid Dynamics, Wiley, New York, 1996, pp. 747–753.

[54] Y. Xing, C.-W. Shu, High-order finite difference WENO schemes with the exact conservation property for the shallow-water equations, J. Comput. Phys. 208 (1) (2005) 206–227.

[55] Y. Xing, C.-W. Shu, High-order well-balanced finite volume WENO schemes and discontinuous Galerkin methods for a class of hyperbolic systems with source terms, J. Comput. Phys. 214 (2) (2006) 567–598.