Available online at www.sciencedirect.com
ScienceDirect
Journal of Computational Physics 222 (2007) 287–331
JOURNAL OF COMPUTATIONAL PHYSICS
www.elsevier.com/locate/jcp

Application of conservative residual distribution schemes to the solution of the shallow water equations on unstructured meshes

M. Ricchiuto [a,*], R. Abgrall [a,b], H. Deconinck [c]

[a] *INRIA Futurs, Projet Scalapplix and Mathématiques Appliquées de Bordeaux, Université Bordeaux 1, 351, Course de la Libération, 33405 Talence Cedex, France*
[b] *Institut Universitaire de France, Université Bordeaux 1, 351, Course de la Libération, 33405 Talence Cedex, France*
[c] *von Karman Institute for Fluid Dynamics, 72, Chaussée de Waterloo, B-1640 Rhode-St-Genèse, Belgium*

Received 3 November 2005; received in revised form 29 May 2006; accepted 8 June 2006
Available online 2 August 2006

Abstract

We consider the numerical solution of the shallow water equations on unstructured grids. We focus on flows over *wet* areas. The extension to the case of *dry* bed will be reported elsewhere. The shallow water equations fall into the category of systems of conservation laws which can be symmetrized thanks to the existence of a mathematical entropy coinciding, in this case, with the total energy. Our aim is to show the application of a particular class of conservative residual distribution ($\mathscr{RD}$) schemes to the discretization of the shallow water equations and to analyze their discrete accuracy and stability properties. We give a review of conservative $\mathscr{RD}$ schemes showing relations between different approaches previously published, and recall L^∞ stability and accuracy criteria characterizing the schemes. In particular, the accuracy of the $\mathscr{RD}$ method in presence of source terms is analyzed, and conditions to construct rth order discretizations on irregular triangular grids are proved. It is shown that the $\mathscr{RD}$ approach gives a natural way of obtaining high order discretizations which, moreover, preserves *exactly* the steady *lake at rest* solution independently on mesh topology, nature of the variation of the bottom and polynomial order of interpolation used for the unknowns. We also consider more general analytical solutions which are less investigated from the numerical view point. On irregular grids, linearity preserving $\mathscr{RD}$ schemes yield a truly second order approximation. We also sketch a strategy to achieve discretizations which preserve *exactly* some of these solutions. Numerical results on steady and time-dependent problems involving smooth and non-smooth variations of the bottom topology show very promising features of the approach.

Keywords: Conservative schemes; Residual distribution; Shallow water equations; Lake at rest solution; Truly 2D analytical solution; High order accuracy; Unstructured grids

* Corresponding author. Tel.: +33 540003794; fax: +33 540003895.
E-mail address: Mario.Ricchiuto@inria.fr (M. Ricchiuto).

0021-9991/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcp.2006.06.024

1. Introduction

This paper considers the solution of the two-dimensional shallow water equations on unstructured meshes. The shallow water system describes the motion of shallow free surface flows under the action of gravity force. We consider the case of frictionless flow over *wet* non-flat bed. The extension to *dry* bed including bed friction will be reported elsewhere. The equations constitute a non-homogeneous system of conservation laws in terms of local water height and discharge (or momentum), where the source term models the effects of the variation of the bottom on the flow. It is known that this system admits a mathematical entropy, in the sense of Harten [1], which symmetrizes the system and coincides with the total energy in the flow [2]. A review of the nonlinear stability principle associated to the total energy can be found in [3] and will only be briefly recalled in this paper. Moreover, the shallow water system has a number of exact solutions with a known simple analytical form. Among these we consider the approximation of the so called *lake at rest* solution consisting of still flow over a non-flat bottom. Independently on the shape of the bottom, the system admits the exact steady solution consisting of zero flow speed and constant total height of the water. *We also review additional truly two-dimensional steady solutions which are less investigated from the numerical view point.*

We are interested in the application of the family of conservative residual distribution ($\mathscr{RD}$) schemes proposed in [4–8] to the discretization of the shallow water equations. Due to the lack of a multidimensional conservative linearization of the flux Jacobians of the system [9], standard matrix $\mathscr{RD}$ schemes cannot be applied. The class of schemes of [4–8] gives a simple solution to this problem. Previous investigations considering the application of residual distribution to the shallow water equations have been published in [10–12]. The schemes considered in the references are however non-conservative or based on *ad hoc* conservative corrections and do not propose a general framework for the discretization of systems of conservation laws. Moreover, the references do not contain any theoretical basis for the analysis of the accuracy and stability of the discretizations and consider only steady-state computations or time-dependent computations based on first order inconsistent discrete approximations. Here, we show how the schemes of [4–8] can be used to approximate steady and time-dependent solutions of the shallow water equations and, following [6,7,13,14], we propose nonlinear schemes which are formally second order accurate and which provide a non-oscillatory approximation of discontinuities. The reader may consult [15] for an alternative construction of second order monotone $\mathscr{RD}$ discretizations for time dependent problems.

The structure of the paper is the following. We start by recalling the shallow water equations, their symmetric form, the associated energy stability principle and *several exact steady-state solutions including* the lake at rest solution. In Section 3 we then present the conservative schemes of [4–7] for steady and unsteady problems. We show how to extend these scheme to problems with source terms. We analyze the accuracy of $\mathscr{RD}$ schemes in presence of source terms. Second order $\mathscr{RD}$ schemes are proved to guarantee the preservation of the exact *lake at rest* solution independently on the topology of the mesh, on the variation of the bottom and on the order of interpolation of the unknowns. Differently from other numerical techniques [16–19], this is achieved very naturally thanks to the truly residual character of the schemes. Finally, we present and discuss the numerical results obtained on a number of representative steady and time-dependent test cases in Section 4. Some comments related to the future development of the method are proposed in the conclusion.

2. The shallow water system and its properties

2.1. Conservation law form of the equations

Frictionless shallow free surface flows under the action of gravity force can be modeled by the following system of the shallow water equations:

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

where $\Omega \times [0, t_f]$ is the space-time domain over which solutions are sought, and the vector of *conserved variables* and fluxes are given by

$$\mathbf{u} = \begin{bmatrix} H \\ Hu \\ Hv \end{bmatrix}, \quad \mathscr{F}(\mathbf{u}) = [\,\mathscr{F}_1 \quad \mathscr{F}_2\,] = \begin{bmatrix} Hu & Hv \\ Hu^2 + g\frac{H^2}{2} & Huv \\ Huv & Hv^2 + g\frac{H^2}{2} \end{bmatrix} \tag{2}$$

with H the *local relative water height*, $\vec{u} = (u, v)$ the flow speed and g the gravity acceleration. The source term $\mathscr{S}(\mathbf{u})$ models the effects of the shape of the bottom on the flow and is given by

$$\mathscr{S}(\mathbf{u}, x, y) = -\begin{bmatrix} 0 & gH\frac{\partial B(x,y)}{\partial x} & gH\frac{\partial B(x,y)}{\partial y} \end{bmatrix}^{\mathrm{T}}, \tag{3}$$

where the superscript $^{\mathrm{T}}$ denotes the transpose of a vector (or of a matrix) and $B(x, y)$ is the local height of the bottom. We define $H_{\mathrm{tot}} = H + B$, the total water height (Fig. 1). We consider only flows over *wet* bed ($H > \varepsilon > 0$). The extension to *dry* flows ($H \geqslant 0$) will be reported elsewhere. In the simplified case of flat bottom, in which we will assume $B = 0\ \forall (x, y) \in \Omega$, system (1) expresses the conservation of the total water height $H_{\mathrm{tot}} = H$ and of the discharge $H\vec{u}$ in presence of the gravity force. Note that the system indeed admits physical discontinuous solutions (hydraulic jumps). As a consequence, the use of conservative numerical discretizations is necessary for a correct approximation of these features. Moreover, the schemes used to approximate (1) should also have some form of L_∞ stability in order to provide a non-oscillatory numerical solution in correspondence of discontinuities.

2.2. Symmetric quasi-linear form and total energy equation

One of the most interesting properties of system (1) is that the total energy in the flow

$$E(\mathbf{u}) = H\left(\frac{1}{2}gH + gB + \frac{\vec{u}\cdot\vec{u}}{2}\right), \tag{4}$$

respects the inequality [2,3]

$$\frac{\partial E}{\partial t} + \nabla\cdot(\vec{u}E) + \nabla\cdot\left(\vec{u}\frac{gH^2}{2}\right) \leqslant 0. \tag{5}$$

The energy E, which is convex in $\mathbf{u}$, acts for the system as a mathematical entropy, in the sense of Harten [1]. In particular, introducing the vector of *symmetrizing variables* $\mathbf{v}$ given by [3]

$$\mathbf{v} = \frac{\partial E(\mathbf{u})}{\partial \mathbf{u}}^{\mathrm{T}} = \begin{bmatrix} p \\ u \\ v \end{bmatrix}, \quad p = gH - \frac{\vec{u}\cdot\vec{u}}{2}, \tag{6}$$

the system can be written in the symmetric quasi-linear form

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

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

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

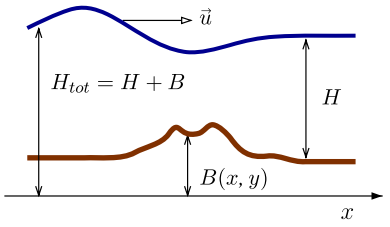

Fig. 1. Shallow water equations: basic unknowns.

Note that the convexity of $E(\mathbf{u})$ also implies that A_0 is positive definite. The total energy equation is recovered multiplying on the left (7) by $\mathbf{v}^{\mathrm{T}}$, obtaining [3]

$$\mathbf{v}^{\mathrm{T}}A_0\frac{\partial \mathbf{v}}{\partial t}+\mathbf{v}^{\mathrm{T}}A_1\frac{\partial \mathbf{v}}{\partial x}+\mathbf{v}^{\mathrm{T}}A_2\frac{\partial \mathbf{v}}{\partial y}-\mathbf{v}^{\mathrm{T}}\mathscr{S}(\mathbf{v},x,y)=\mathbf{v}^{\mathrm{T}}\left(\frac{\partial \mathbf{u}(\mathbf{v})}{\partial t}+\nabla\cdot\mathscr{F}(\mathbf{v})-\mathscr{S}(\mathbf{v},x,y)\right)\leqslant 0. \tag{9}$$

The symmetry of the flux Jacobians guarantees that the matrix

$$K=A_1\xi_1+A_2\xi_2$$

has real eigenvalues and eigenvectors $\forall\vec{\xi}=(\xi_1,\xi_2)\in\mathbb{R}^2$, hence the system admits *simple wave*-like solutions traveling with speeds associated with the eigenvalues of the flux Jacobians. For completeness, we report the eigenvalues of $K=A_1\xi_1+A_2\xi_2$, given by

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

with a the speed of propagation $a=\sqrt{gH}$. It is also useful to introduce the Froude number

$$Fr=\frac{\sqrt{\vec{u}\cdot\vec{u}}}{a}. \tag{10}$$

2.3. Steady solutions

Developing the second and third lines of (1) and (2), we get

$$u\left[\frac{\partial(Hu)}{\partial x}+\frac{\partial(Hv)}{\partial y}\right]+H\left[u\frac{\partial u}{\partial x}+v\frac{\partial u}{\partial y}+g\frac{\partial(H+B)}{\partial x}\right]=0$$

and

$$v\left[\frac{\partial(Hu)}{\partial x}+\frac{\partial(Hv)}{\partial y}\right]+H\left[u\frac{\partial v}{\partial x}+v\frac{\partial v}{\partial y}+g\frac{\partial(H+B)}{\partial y}\right]=0.$$

Using the notation

$$\mathscr{I}=\frac{u^2+v^2}{2}+g(H+B)=\frac{u^2+v^2}{2}+gH_{\mathrm{tot}},$$

we get

$$H\left[u\frac{\partial u}{\partial x}+v\frac{\partial u}{\partial y}+g\frac{\partial(H+B)}{\partial x}\right]=H\frac{\partial\mathscr{I}}{\partial x}+Hv\left(\frac{\partial u}{\partial y}-\frac{\partial v}{\partial x}\right)$$

and

$$H\left[u\frac{\partial v}{\partial x}+v\frac{\partial v}{\partial y}+g\frac{\partial(H+B)}{\partial y}\right]=H\frac{\partial\mathscr{I}}{\partial y}+Hu\left(\frac{\partial v}{\partial x}-\frac{\partial u}{\partial y}\right).$$

The steady shallow water equations are equivalent to:

$$\frac{\partial(Hu)}{\partial x}+\frac{\partial(Hv)}{\partial y}=0, \tag{11a}$$

$$\left(\frac{\partial(Hu)}{\partial x}+\frac{\partial(Hv)}{\partial y}\right)\vec{u}+H\nabla\mathscr{I}+H\,\mathrm{curl}\,\vec{u}\begin{pmatrix}-v\\ u\end{pmatrix}=0. \tag{11b}$$

We look for elementary solutions of (11a) and (11b).

2.3.1. The lake at rest solution

We recall here the so called *lake at rest* exact solution of the equations. This solution is easily obtained assuming $u=v=0$ and integrating (1) over an arbitrary control volume $\mathscr{V}$ obtaining

$$\int_{\mathscr{V}} \frac{\partial H}{\partial t}\,\mathrm{d}x\,\mathrm{d}y = -\oint_{\partial\mathscr{V}} H\vec{u}\cdot\vec{n}\,\mathrm{d}l = 0$$

and

$$\int_{\mathscr{V}} \frac{\partial (H\vec{u})}{\partial t}\,\mathrm{d}x\,\mathrm{d}y = -\int_{\mathscr{V}} gH\nabla H_{\mathrm{tot}}\,\mathrm{d}x\,\mathrm{d}y.$$

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

$$\begin{aligned} &H_{\mathrm{tot}}(x,y,t) = H_{\mathrm{tot}}(x,y,t=0) = H_0 \quad \forall (x,y)\in\Omega \text{ and } t\geqslant 0,\\ &u = v = 0 \quad \forall (x,y)\in\Omega \text{ and } t\geqslant 0. \end{aligned} \tag{12}$$

Note that this is independent on the shape of $B(x,y)$, as long as ∇H_{tot} is integrable. Later on we will show a class of schemes preserving *exactly* this solution independently on the form of $B(x,y)$, mesh topology and degree of the discrete polynomial approximation of the unknowns.

2.3.2. Two-dimensional solutions

In order to mimic geophysical flows, we look for solutions such that

$$\nabla\cdot\vec{u} = 0.$$

This divergence-free condition is satisfied if

$$u = \frac{\partial\psi}{\partial y}, \quad v = -\frac{\partial\psi}{\partial x}$$

for some given function ψ. Then (11a) becomes

$$u\frac{\partial H}{\partial x} + v\frac{\partial H}{\partial y} + H\nabla\cdot\vec{u} = 0,$$

that is

$$u\frac{\partial H}{\partial x} + v\frac{\partial H}{\partial y} = \frac{\partial\psi}{\partial y}\frac{\partial H}{\partial x} - \frac{\partial\psi}{\partial x}\frac{\partial H}{\partial x} = \det(\nabla\psi,\nabla H) = 0,$$

so that a simple choice is $H = \psi + \alpha$ where α is a constant.

We now examine (11b). We have

$$\mathrm{curl}\,\vec{u} = -\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} = -\Delta\psi,$$

so that a curl-free solution is obtained provided $\Delta\psi = 0$ and then

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

where C is another constant.

We now have to choose ψ. It is well known that harmonic functions are real parts of holomorphic functions, $\psi(x,y) = \mathrm{Re}f(z)$ where f is holomorphic in $z = x + \mathrm{i}y$. The function f is arbitrary. Examples will be given in the results section.

3. Conservative $\mathcal{RD}$ schemes

3.1. Basic formulation

Consider an unstructured discretization of Ω composed by non-overlapping triangles. We denote the mesh by $\mathscr{T}_h$, with h a reference global element size (e.g. largest element diameter) and by E the generic element in the

grid. Given a vector of state variables $\mathbf{w}(\mathbf{u})$, on $\mathscr{T}_h$, we denote the piecewise linear continuous interpolation of the nodal values $\mathbf{w}_i = \mathbf{w}(\mathbf{u}_i)$ by $\mathbf{w}_h$. We consider schemes of the form

$$|S_i|\frac{\mathrm{d}\mathbf{u}_i}{\mathrm{d}t} + \sum_{E\in\mathscr{D}_i}\phi_i(\mathbf{w}_h) = 0 \quad \forall i \in \tau_h \tag{13}$$

with $\{\mathbf{u}_i^0\}_{i\in\mathscr{T}_h} = \{\mathbf{u}_0(x_i, y_i, t=0)\}_{i\in\mathscr{T}_h}$, being $\mathbf{u}_0(x,y)$ the initial solution, $|S_i|$ the area of the median dual surface of node i and $\mathscr{D}_i$ the stencil of node i, i.e., the set of triangles containing i as a node (see Fig. 2). In the homogeneous case $\mathscr{S} = 0$, the quantities $\phi_i(\mathbf{w}_h)$ determine some form of the splitting

$$\sum_{j\in E}\phi_j(\mathbf{w}_h) = \phi^h(\mathbf{w}_h) = \oint_{\partial E}\mathscr{F}(\mathbf{w}_h)\cdot\hat{n}\,\mathrm{d}l \quad \forall E \in \mathscr{T}_h. \tag{14}$$

The quantity $\phi^h(\mathbf{w}_h)$ is called the *local element residual* while the ϕ_js are referred to as the *local nodal residuals* or *split residuals*. If not stated otherwise, we assume the residual ϕ^h to be computed with exact integration. As remarked in [4–8], the direct use of the *integral* formulation of the problem, guarantees that, as long as the consistency relation (14) is satisfied, scheme (13) reduces to a discrete approximation of the Rankine–Hugoniot jump conditions across a discontinuity, hence the schemes are *conservative by construction*. Indeed, when seeking a discontinuous steady-state solution, scheme (13) can be seen as an iterative procedure to reduce the *conservation defect* represented by the residual. Clearly, the definition of the splitting (14) determines the final properties of the discretization. However, before giving some design criteria for the ϕ_js and introducing the schemes used in this paper, we will add some remarks concerning the computation of the residual. In particular, we remark that using Gauss' theorem one can write (see Fig. 3)

$$\phi^h(\mathbf{w}_h) = \int_E \nabla\cdot\mathscr{F}(\mathbf{w}_h)\,\mathrm{d}x\,\mathrm{d}y = \int_E \frac{\partial\mathscr{F}(\mathbf{w}_h)}{\partial\mathbf{w}}\cdot\nabla\mathbf{w}_h\,\mathrm{d}x\,\mathrm{d}y$$

and finally for a piecewise linear $\mathbf{w}_h$

$$\phi^h = \sum_{j\in E}\widetilde{K}_j\mathbf{w}_j, \quad \widetilde{K}_j = \frac{1}{2}\left(\int_E \frac{\partial\mathscr{F}(\mathbf{w}_h)}{\partial\mathbf{w}}\right)\cdot\vec{n}_j \tag{15}$$

with $\vec{n}_j$ the inward pointing vector normal to the edge of E facing node j, scaled by the length of the edge. In (15) the $\widetilde{K}_j$ matrices represent the projection of an *exact mean-value* flux Jacobian along $\vec{n}_j$. As we will see later, the sign of these Jacobians, defined in the usual matrix sense via eigenvalue decomposition, can be used to devise upwind discretizations. Here we limit ourselves to the following remarks.

(i) The computation of the *exact* mean-value Jacobians, needed to evaluate the residual as in (15) and for the distribution, is difficult or even impossible, especially if one seeks a closed form analytical expressions. To partially cure this, in [20] the authors propose to replace the exact mean-value Jacobians with approximate ones, obtained through volume (surface in 2D) Gaussian integration of the quasi-linear form. In the reference, the authors prove a Lax–Wendroff theorem for the $\mathscr{RD}$ schemes obtained in this

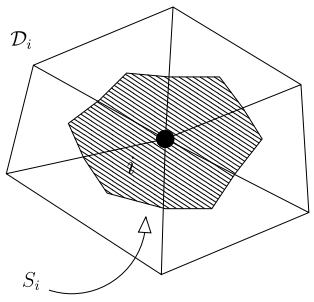

Fig. 2. Median dual cell S_i and *stencil* $\mathscr{D}_i$.

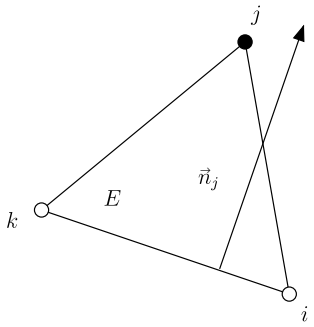

Fig. 3. Nodal normal $\vec{n}_j$.

way, guaranteeing, under standard hypotheses and *provided that the number of Gaussian points is large enough*, the convergence to the correct weak solutions. Unfortunately, the cost of this technique is high, even with the adaptive quadrature strategy proposed in the reference.

(ii) The simpler approach proposed in [4–8] and used here relies on a direct approximation of the contour integral in (14), thus ensuring the conservative character of the schemes. Arbitrary linearizations can be used to evaluate the flux Jacobians, needed in the distribution step. To distinguish this case from the one of Eq. (15), we shall denote by K_j the projection of the flux Jacobians along $\vec{n}_j$, when an arbitrary (inexact) linearization is used.

(iii) In the non-homogeneous case (1), the schemes are still defined by (13), except that now $\phi_i(\mathbf{w}_h)$ determine some form of the splitting

$$\sum_{j\in E}\phi_j(\mathbf{w}_h)=\phi^h(\mathbf{w}_h)=\oint_{\partial E}\mathscr{F}(\mathbf{w}_h)\cdot\hat{n}\,\mathrm{d}l-\int_E\mathscr{S}(\mathbf{w}_h,x,y)\,\mathrm{d}x\,\mathrm{d}y. \tag{16}$$

As defined by (16), the residual still represents a *degree of non-equilibrium*, this time not only related to conservation but to the balance between the local net flux through the element and the forcing terms. The integrals (16) are computed with quadrature formulas of the same order (see Section 4 for details).

3.2. Fully discrete schemes

3.2.1. Integration of the ODE

When steady solutions are sought, it is customary to integrate the system of ODEs (13) with a properly chosen time discretization. Denoting by $\mathbf{w}_h^n$ and $\mathbf{w}_h^{n+1}$ the piecewise continuous linear discrete interpolation of the nodal values $\{\mathbf{w}_i(t^n)\}_{i\in\mathscr{T}_h}$ and $\{\mathbf{w}_i(t^{n+1})\}_{i\in\mathscr{T}_h}$, the following schemes will be considered.

Explicit Euler scheme

$$\mathbf{u}_i^{n+1}=\mathbf{u}_i^n-\frac{\Delta t}{|S_i|}\sum_{E\in\mathscr{D}_i}\phi_i(\mathbf{w}_h^n)\quad\forall i\in\mathscr{T}_h \tag{17}$$

with $\Delta t=t^{n+1}-t^n$ the time-step.

Crank–Nicholson ($\mathscr{CN}$) scheme. Two different formulations of the $\mathscr{CN}$ scheme are considered. The first is given by

$$|S_i|\frac{\delta\mathbf{u}_i}{\Delta t}=-\sum_{E\in\mathscr{D}_i}\frac{\phi_i(\mathbf{w}_h^{n+1})+\phi_i(\mathbf{w}_h^n)}{2}\quad\forall i\in\mathscr{T}_h \tag{18}$$

with $\delta\mathbf{u}_i=\mathbf{u}_i^{n+1}-\mathbf{u}_i^n$. We also consider the $\mathscr{CN}$ scheme

$$|S_i|\frac{\delta\mathbf{u}_i}{\Delta t}=-\sum_{E\in\mathscr{D}_i}\phi_i(\mathbf{w}_h^{n+1/2})\quad\forall i\in\mathscr{T}_h \tag{19}$$

with $\mathbf{w}_h^{n+1/2} = \mathbf{w}(\mathbf{u}_h^{n+1/2})$ and

$$\mathbf{u}_h^{n+1/2} = \frac{1}{2}(\mathbf{u}_h^n + \mathbf{u}_h^{n+1}).$$

The $\mathscr{CN}$ scheme necessitates at each iteration the solution of a nonlinear system of algebraic equations.

3.2.2. Space-time schemes

When considering time-dependent flows, it is known that the formulation (13) is inconsistent and limits the accuracy *in space* to first order [14]. For this reason, in the time-dependent case we make use of the *space-time* formulation obtained by replacing (13) with the algebraic system of nonlinear equations

$$\sum_{E\in\mathscr{D}_i} \Phi_i(\mathbf{w}_h) = 0 \quad \forall i \in \mathscr{T}_h, \tag{20}$$

where now in each element of $\mathscr{T}_h$, the Φ_js define some form of the splitting of the space-time residual $\Phi^h(\mathbf{w}_h)$

$$\sum_{j\in E} \Phi_j(\mathbf{w}_h) = \Phi^h(\mathbf{w}_h) = \int_{t^n}^{t^{n+1}} \int_E \left(\frac{\partial \mathbf{u}(\mathbf{w}_h)}{\partial t} + \nabla\cdot \mathscr{F}(\mathbf{w}_h) - \mathscr{S}(\mathbf{w}_h, x, y) \right) \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t. \tag{21}$$

The conservative schemes proposed in [6–8] are obtained by approximating (21) as

$$\Phi^h = \frac{|E|}{3}\sum_{j\in E}\delta\mathbf{u}_j + \frac{\Delta t}{2}\oint_{\partial E}(\mathscr{F}^n + \mathscr{F}^{n+1})\cdot\hat{n}\,\mathrm{d}l - \frac{\Delta t}{2}\int_E (\mathscr{S}^n + \mathscr{S}^{n+1})\,\mathrm{d}x\,\mathrm{d}y \tag{22}$$

with $\mathscr{F}^n = \mathscr{F}(\mathbf{w}_h^n)$, $\mathscr{S}^n = \mathscr{S}(\mathbf{w}_h^n, x, y)$ and similarly for $\mathscr{F}^{n+1}$ and S^{n+1}. Using (16), last expression can be conveniently recast as

$$\Phi^h = \frac{|E|}{3}\sum_{j\in E}\delta\mathbf{u}_j + \frac{\Delta t}{2}\phi^h(\mathbf{w}_h^{n+1}) + \frac{\Delta t}{2}\phi^h(\mathbf{w}_h^n). \tag{23}$$

A trivial splitting of Φ^h is obtained by setting

$$\Phi_i = \frac{|E|}{3}\delta\mathbf{u}_i + \frac{\Delta t}{2}(\phi_i(\mathbf{w}_h^{n+1}) + \phi_i(\mathbf{w}_h^n))$$

in (20), with ϕ_i a given splitting of the spatial residual ϕ^h. Summing up over the elements, one ends with

$$|S_i|\delta\mathbf{u}_i + \frac{\Delta t}{2}\sum_{E\in\mathscr{D}_i}(\phi_i(\mathbf{w}_h^{n+1}) + \phi_i(\mathbf{w}_h^n)) = 0,$$

which is nothing else that (18). This shows that first-order and inconsistent schemes can be recast as space-time schemes when combined with the $\mathscr{CN}$ time integration (18).

Another possibility is to approximate (21) as

$$\Phi^h = \frac{|E|}{3}\sum_{j\in E}\delta\mathbf{u}_j + \Delta t\oint_{\partial E}\mathscr{F}(\mathbf{w}_h^{n+1/2})\cdot\hat{n}\,\mathrm{d}l - \Delta t\int_E \mathscr{S}(\mathbf{w}_h^{n+1/2}, x, y)\,\mathrm{d}x\,\mathrm{d}y \tag{24}$$

with $\mathbf{w}_h^{n+1/2}$ as in (19). We remark that, in smooth regions, (22) and (24) are equally accurate approximations of (21). As before, we can recast the last expression as

$$\Phi^h = \frac{|E|}{3}\sum_{j\in E}\delta\mathbf{u}_j + \Delta t\phi^h(\mathbf{w}_h^{n+1/2}), \tag{25}$$

which can be used to show that first-order and inconsistent $\mathscr{RD}$ schemes can be recast as space-time schemes of this type when combined with the $\mathscr{CN}$ time integration (19). For a given splitting of the spatial residual ϕ_i, this is achieved by setting

$$\Phi_i = \frac{|E|}{3}\delta\mathbf{u}_i + \Delta t\phi_i(\mathbf{w}_h^{n+1/2}).$$

3.3. Basic properties

We recall here some design criteria that can be used for the definition of the split residuals ϕ_j, and Φ_j. The issue of conservation has been discussed in the previous subsection and will not be mentioned. We just assume that all the schemes respect (14) (or (20)), and hence are conservative due to the definition of ϕ^h (respectively, Φ^h). Note that under this hypothesis and under a continuity assumption on the split residuals, a scheme of the form (13) can be shown to respect a Lax–Wendroff theorem ensuring that *provided some standard stability assumptions are met*, if convergent (with h) conservative $\mathscr{RD}$ schemes convergence to a weak solution of the problem. The proof of this theorem can be found in [21] for the case of exact evaluation of ϕ^h and in [20] for the case of approximate quadrature. We remark that no general convergence proof exist up to now for $\mathscr{RD}$, except for some scalar schemes. For the schemes considered in this paper, the practical experience shows that they are stable and convergent. In the following, we consider the issues of the non-oscillatory character of the solution, and of the accuracy of the method.

3.3.1. L_∞ stability criterion

When approximating discontinuous solutions, the non-oscillatory character of the discrete approximation is an important property. For scalar problems, a LED (local extremum diminishing) Principle on unstructured meshes [22,23] can be used to characterize the stability of (13). This ultimately translates into a *positivity* criterion when one of the time integration schemes presented in the previous subsection is applied and for the space-time schemes as well [22,14,23,6–8]. In the simplest case of a linear (or linearized) scalar problem, this positivity criterion can be explicitly stated by rewriting the fully discrete scheme as

$$\mathscr{A}U^{n+1} = \mathscr{B}U^n$$

with U^{n+1} and U^n the arrays of the nodal unknowns. A scheme is positive if $\mathscr{A}$ is a monotone ($\mathscr{A}_{ii} \geqslant 0$, $\mathscr{A}_{ij} \leqslant 0$) diagonally dominant matrix and $\mathscr{B}$ is a positive matrix ($\mathscr{B}_{ij} \geqslant 0\ \forall i,j$). For a homogeneous problem, this condition can be shown to guarantee a *discrete maximum principle* [22,14,23,8]. An extension of this criterion to linear systems has been recently proposed in [13]. Consider for example the symmetric system

$$\frac{\partial \mathbf{v}}{\partial t} + A_1 \frac{\partial \mathbf{v}}{\partial t} + A_2 \frac{\partial \mathbf{v}}{\partial y} = 0. \tag{26}$$

The basic idea behind the analysis presented in the reference is that for $\mathscr{RD}$ schemes the variation of the solution can be expressed as the convex combination of *local signals*

$$\frac{\mathrm{d}\mathbf{v}_i}{\mathrm{d}t} = \sum_{E\in\mathscr{D}_i} \gamma_E \frac{\mathrm{d}\widetilde{\mathbf{v}}_i}{\mathrm{d}t}; \quad \gamma_E |S_i| \frac{\mathrm{d}\widetilde{\mathbf{v}}_i}{\mathrm{d}t} = -\phi_i\ \forall E \in \mathscr{T}_h \tag{27}$$

with the scalars γ_E respecting

$$\gamma_E \in (0,1], \quad \sum_{E\in\mathscr{D}_i} \gamma_E = 1$$

and in the case of the schemes considered here given by

$$\gamma_E = \frac{|E|}{3|S_i|}.$$

Denoting by $\langle,\rangle$ the standard Euclidean product, in the case of a linear system, it is shown in [13] that the solution can *locally* be decomposed in simple waves

$$\mathbf{v}_h(x,y) = \sum_\sigma \sum_{j\in E} \varphi_j^\sigma \boldsymbol{r}_\sigma, \quad \varphi_\sigma = \langle \mathbf{v}_j, \boldsymbol{r}_\sigma \rangle = C_j^\sigma + \alpha_j^\sigma \vec{\xi}\cdot\vec{x} \text{ on } E$$

with $C_j^\sigma, \alpha_j^\sigma \in \mathbb{R}$, $\vec{x} = (x,y)$ the position vector and $\boldsymbol{r}_\sigma$ an eigenvector of $K = A_1'\xi_1 + A_2'\xi_2$. If a linear scheme is used to discretize the equations, the wave decomposition leads to [13]

$$\phi_i = \sum_\sigma \sum_{j\in E} c_{ij}^\sigma(\varphi_j^\sigma - \varphi_i^\sigma)\mathbf{r}_\sigma,$$

implying that also the ϕ_js are sum of simple waves. Due to the linearity of the scheme, one can then consider the evolution due to each simple wave, obtaining

$$\gamma_E|S_i|\frac{\mathrm{d}\widetilde{\mathbf{v}}_i}{\mathrm{d}t} = -\sum_{j\in E} c_{ij}^\sigma(\varphi_j^\sigma - \varphi_i^\sigma)\mathbf{r}_\sigma.$$

The last relation finally leads to

$$\left\langle \gamma_E|S_i|\frac{\mathrm{d}\widetilde{\mathbf{v}}_i}{\mathrm{d}t}, \mathbf{r}_\sigma \right\rangle = -\sum_{j\in E} c_{ij}^\sigma(\varphi_j^\sigma - \varphi_i^\sigma)\|\mathbf{r}_\sigma\|^2,$$

which allows to extend the scalar LED principle to the simple wave σ. Note that this *does not* imply that if $\mathbf{v}_h^n$ is a simple wave, then so is $\mathbf{v}_h^{n+1}$, however, it gives a means of explaining the monotone behavior exhibited by some first order linear schemes. Indeed, when this analysis is performed for a fully discrete scheme and in the case $\mathbf{v}_h^n$ is a simple wave, a scheme is said to be stable if it is possible to show a direction $\vec{\xi}$ for which

$$\|\widetilde{\mathbf{v}}_i\| \leqslant \max_{j\in E} \|\varphi_j^\sigma \mathbf{r}_\sigma\|. \tag{28}$$

Since, as remarked before, $\mathbf{v}_i^{n+1}$ can be written as a convex combination of the $\widetilde{\mathbf{v}}_i$ values, the bounds (28) imply a local stability of the discrete solution.

Remark 3.1. The extension of this criterion to the non-homogeneous case, and hence its application to the shallow water equations, is still unclear. For linear symmetric systems with a source term $\mathscr{S}(x,y)$ *independent on the solution*, some very local results can be obtained following [13]. In particular, for some schemes one can show that the solution respects bounds of the type

$$\min_{j\in\mathscr{D}_i} \varphi_j^\sigma + \alpha \min_{j\in\mathscr{D}_i} \varphi_j^{\mathscr{S},\sigma} \leqslant \langle \widetilde{\mathbf{v}}_i, \mathbf{r}_\sigma\rangle \leqslant \max_{j\in\mathscr{D}_i} \varphi_j^\sigma + \alpha \max_{j\in\mathscr{D}_i} \varphi_j^{\mathscr{S},\sigma} \tag{29}$$

with $\alpha \geqslant 0$, and where the component $\varphi_j^{\mathscr{S},\sigma}$ is of the form

$$\varphi_j^{\mathscr{S}\sigma} = \langle \mathscr{S}(x_j,y_j), \mathbf{r}_\sigma\rangle. \tag{30}$$

However, it is difficult to go further with (29).

3.3.2. Linearity preservation and residual schemes

Case of steady homogeneous systems. The resolution of complex structures is another essential issue when approximating solutions of conservation laws. The accuracy of the schemes considered here can be characterized by a residual property.

This property is best presented by resorting to the following analysis. Let $\mathbf{w}$ be an exact smooth solution verifying $\nabla\cdot\mathscr{F}(\mathbf{w}) = 0$. Let $\mathbf{w}_h$ be a rth order accurate continuous piecewise polynomial approximation of nodal values of $\mathbf{w}$, $\{\mathbf{w}_i = \mathbf{w}(x_i,y_i)\}_{i\in\mathscr{T}_h}$. Let $\mathscr{F}_h$ be a continuous rth order accurate approximation on $\mathscr{T}_h$ of $\mathscr{F}(\mathbf{w})$. Define the error

$$\mathscr{E}(\mathbf{w}_h) := \sum_{i\in\mathscr{T}_h} \varphi_i \sum_{E\in\mathscr{D}_i} \phi_i(\mathbf{w}_h)$$

with φ_i the nodal values of a smooth function $\varphi \in C_0^1(\Omega)$. Since $\mathbf{w}_h$ is not the numerical solution given by the $\mathscr{RD}$ scheme but the rth order continuous piecewise polynomial approximation of the smooth exact solution $\mathbf{w}$, in general $\mathscr{E}(\mathbf{w}_h) \neq 0$. The magnitude of this error gives an estimate on the accuracy of the schemes. It has been shown [24,23,25] that, provided that the mesh satisfies the constraint

$$C_1 \leqslant \sup_{E\in\mathscr{T}_h} \frac{h}{|E|} \leqslant C_2, \quad C_1, C_2 \in \mathbb{R}^+,$$

at steady-state a $\mathscr{RD}$ scheme respects the error estimate $\mathscr{E}(\mathbf{w}_h) = \mathscr{O}(h^r)$ provided that

$$\phi_i = \mathcal{O}(h^{r+1}), \tag{31}$$

see Appendices A and B for more details. Hence, in the case of the piecewise linear approximation considered here, the condition $\phi_i = \mathcal{O}(h^3)$, represents the *condition for a $\mathscr{RD}$ scheme to be second order accurate*.

Remark 3.2. Note that the condition $\phi_i = \mathcal{O}(h^{r+1})$ only guarantees that the truncation error of the scheme is $\mathcal{O}(h^r)$. However, in no way this guarantees that the scheme actually does converge with the rate r. Some additional stability properties must be verified to achieve this. As a counter-example, we mention that the Galerkin scheme does verify the accuracy condition, however the error blows up under mesh refinement due to the unstable character of the scheme.

The important fact is the following estimate (now $\mathbf{w}_h$ is the second order continuous piecewise polynomial approximation of the smooth exact solution $\mathbf{w}$, and $\boldsymbol{\mathscr{F}}_h$ is a second order approximation of $\boldsymbol{\mathscr{F}}(\mathbf{w})$)

$$\phi^h = \int_E \nabla\cdot\boldsymbol{\mathscr{F}}_h(\mathbf{w}_h)\,\mathrm{d}x\,\mathrm{d}y = \int_E \nabla\cdot(\boldsymbol{\mathscr{F}}_h(\mathbf{w}_h) - \boldsymbol{\mathscr{F}}(\mathbf{w}))\,\mathrm{d}x\,\mathrm{d}y = \oint_{\partial E}(\boldsymbol{\mathscr{F}}_h(\mathbf{w}_h) - \boldsymbol{\mathscr{F}}(\mathbf{w}))\cdot\hat{n}\,\mathrm{d}l = \mathcal{O}(h^3),$$

since $\boldsymbol{\mathscr{F}}_h$ is a second order accurate. Last estimate allows to give the following characterization.

Definition 3.3 (*Linearity preserving scheme*). A $\mathscr{RD}$ scheme is *linearity preserving* ($\mathscr{LP}$) if

$$\phi_i(\mathbf{w}_h) = \beta_i\phi^h(\mathbf{w}_h)$$

with uniformly bounded matrices β_i:

$$\max_{E\in\mathscr{T}_h}\max_{j\in E}\|\beta_j\| < C < \infty \quad \forall\phi^h, \mathbf{w}_h, \mathbf{u}_h^0, h, \ldots$$

If a continuous rth order approximation is used in the discretization, $\mathscr{LP}$ schemes satisfy *by construction* the accuracy condition $\phi_i = \mathcal{O}(h^{r+1})$.

Ultimately, this definition gives a criterion for the design of high order accurate schemes.

Case of unsteady homogeneous systems. The analysis of the time dependent case is quite similar. Details are reported in Appendix A. The idea is, given a smooth solution of the time dependent problem (denoted by $\mathbf{u}$) and continuous rth order accurate approximations of $\mathbf{u}$ and $\boldsymbol{\mathscr{F}}(\mathbf{u})$ (denoted by $\mathbf{u}_h$ and $\boldsymbol{\mathscr{F}}_h$) to estimate the error

$$\mathscr{E}(\mathbf{u}_h, t_f) := \sum_{n=0}^{N}\sum_{i\in\mathscr{T}_h}\varphi_i^{n+1}\sum_{E\in\mathscr{D}_i}\Phi_i(\mathbf{u}_h)$$

with φ_i^{n+1} the nodal value at time t^{n+1} of a smooth function $\varphi\in C^1(\Omega\times[0,t_f])$ with compact support. As before, since $\mathbf{u}_h$ *is not the numerical solution given by the $\mathscr{RD}$ scheme but the rth order continuous piecewise polynomial approximation of the smooth exact solution* $\mathbf{u}$, in general $\mathscr{E}(\mathbf{u}_h, t_f)\neq 0$. Its magnitude gives an estimate of the accuracy of the scheme.

The analysis of Appendix A applies both when the discretization in time is done via a finite difference scheme (or any classical high order time integration method), and when using a space-time approach (see, e.g. [26,27]). In particular, in the appendix we show that, provided that the approximation in time is at least rth order accurate, if

$$\Phi_i = \mathcal{O}(h^{r+2}),$$

then the scheme respects an error estimate of the type $\mathscr{E}(\mathbf{u}_h, t_f) = \mathcal{O}(h^r)$.[1] In the case of a piecewise linear approximation, the condition reduces to $\Phi_i = \mathcal{O}(h^4)$.

As in the steady case, the important fact is that simple manipulations show that $\Phi^h = \mathcal{O}(h^{r+2})$ ($\Phi^h = \mathcal{O}(h^4)$ for a linear approximation). Hence, the use of a $\mathscr{LP}$ scheme, defined by

[1] The condition $\Phi_i = \mathcal{O}(h^{r+2})$ is not the same as for steady-state, in which case one must have $\phi_i = \mathcal{O}(h^{r+1})$. This is due to the definition of the residuals which are obtained by integration in space and time, giving and extra h in the scaling of Φ_i.

$$\Phi_i = \beta_i \Phi^h$$

with β_i uniformly bounded (see Definition 3.3), is sufficient to guarantee the formal satisfaction of a $\mathcal{O}(h^r)$ (respectively, $\mathcal{O}(h^2)$) error bound (however, see Remarks 3.2 and A.9).

Extension to non-homogeneous systems. We extend the analysis of [23–26,28] to non-homogeneous systems in Appendix B. More precisely, in the appendix we prove the conditions for the $\mathscr{RD}$ schemes considered here to respect an error estimate of the type $\mathscr{E} = \mathcal{O}(h^r)$.

The details of the proof are given in Appendix B. As in all the other cases, the important fact to recall is that the condition allowing to have an $\mathcal{O}(h^r)$ error bound is (in the steady case)

$$\phi_i(\mathbf{w}_h) = \mathcal{O}(h^{r+1})$$

with $\mathbf{w}_h$ a continuous rth order approximation of a smooth exact solution of the problem $\mathbf{w}$. Since, by simple manipulations one can show that $\phi^h(\mathbf{w}_h) = \mathcal{O}(h^{r+1})$, we conclude that, once more, *a valuable criterion for the construction of a rth order scheme is given by* Definition 3.3, *that is by the use of a linearity preserving scheme*. Similar conclusions are obtained in the time dependent non-homogeneous case (see Appendix B for details).

3.3.3. Linearity preservation and the lake at rest solution

The meaning of the $\mathscr{LP}$ condition is that, for a piecewise linear approximation, provided that $\phi_i = \beta_i \phi^h$ ($\Phi_i = \beta_i \Phi^h$ in the time dependent case) with β_i uniformly bounded, within $\mathcal{O}(h^2)$ exact solutions of the continuous problem are also solutions of the discrete equations (once more we refer to Appendices A and B for more details). For the shallow water equations, this has a very interesting application. We can easily construct schemes that preserve exactly some steady solutions.

Proposition 3.4. *Denote by* $\mathbf{w}$ *the set of quantities chosen as primary variables in the numerical approximation of* (1). *Provided that the same numerical representation is used for* $\mathbf{w}$ *and for the local height of the bottom* $B(x,y)$, *and that the local residual is evaluated exactly with respect to* H *and* B, *linearity preserving* $\mathscr{RD}$ *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 the following three different choices of* $\mathbf{w}$:

1. *conservative variables* $\mathbf{u}$, *Eq.* (2);
2. *symmetrizing variables* $\mathbf{v}$, *Eq.* (6);
3. *primitive variables* $\mathbf{p} = [H \quad u \quad v]^{\mathrm{T}}$.

Proof. The proof is obtained quite easily by noting that on the lake at rest solution, for all the choices of variables, the vector of unknowns reduces to $\mathbf{w} = [H \quad 0 \quad 0]^{\mathrm{T}}$. Suppose now to discretize the spatial domain Ω with a mesh $\mathscr{T}_h$ composed of non-overlapping elements E. On $\mathscr{T}_h$ denote by $\mathbf{w}_h$ and B_h continuous piecewise polynomial approximations of $\mathbf{w}$ and B that can be written as

$$\mathbf{w}_h = \sum_{i \in \mathscr{T}_h} \psi_i \mathbf{w}_i, \quad B_h = \sum_{i \in \mathscr{T}_h} \psi_i B_i$$

with i the generic node of the mesh and with the *shape functions* $\psi_i(x,y)$ respecting the obvious consistency constraint

$$\sum_{j \in E} \psi_j(x,y) = 1. \tag{32}$$

Consider now the semi-discrete scheme (13) and compute the spatial residual

$$\phi^h = \int_E (\nabla \cdot \mathscr{F}_h - \mathscr{S}_h)\,\mathrm{d}x\,\mathrm{d}y.$$

The first component of ϕ^h is

$$\int_E \nabla \cdot (H^h \vec{u}^h)\,\mathrm{d}x\,\mathrm{d}y = \oint_{\partial E} H^h (\vec{u}^h \cdot \hat{n})\,\mathrm{d}l = 0,$$

since $\vec{u}_j = 0$ in all the nodes of E and so will be any consistent interpolation $\vec{u}^h$. Second and third components can be written together in vector form as

$$\int_E (\nabla \cdot (H^h \vec{u}^h \otimes \vec{u}^h) + g H^h \nabla (H^h + B_h)) \,\mathrm{d}x\,\mathrm{d}y = \oint_{\partial E} H^h \vec{u}^h (\vec{u}^h \cdot \hat{n}) \,\mathrm{d}l + g \int_E H^h \nabla H^h_{\mathrm{tot}} \,\mathrm{d}x\,\mathrm{d}y.$$

Since $\vec{u}^h = 0$, these components of the residuals reduce to

$$g \int_E H^h \nabla H^h_{\mathrm{tot}} \,\mathrm{d}x\,\mathrm{d}y = g \int_E H^h \sum_{j \in E} H^h_{\mathrm{tot}_j} \nabla \psi_j \,\mathrm{d}x\,\mathrm{d}y = g H_0 \int_E H^h \sum_{j \in E} \nabla \psi_j \,\mathrm{d}x\,\mathrm{d}y = 0,$$

since on the lake at rest solution $H^h_{\mathrm{tot}_j} = H_0 \ \forall j \in E$ and using condition (32). This shows that $\forall E \in \mathscr{T}_h$, $\phi^h = 0$ on the lake at rest solution, hence for any $\mathscr{LP}$ scheme

$$|S_i| \frac{\mathrm{d}\mathbf{u}_i}{\mathrm{d}t} = 0,$$

which achieves the proof for scheme (13). In the case of the space-time schemes, a similar reasoning leads to the conclusion that on the lake at rest solution the nonlinear system (20) reduces to the identity $0 = 0$. This shows that the proposition holds also for these schemes. □

Remark 3.5. Note that the hypothesis that ϕ^h must be computed exactly with respect to the discrete approximations of H and B, does not mean that it has to be exact with respect to the whole vector $\mathbf{w}_h$, the fluxes being polynomials containing monomials of different degrees in the different components of $\mathbf{w}_h$. Moreover, the result not only applies to the (physical) choices of variables listed in the proposition but to any set of variables reducing to $\mathbf{w} = [H \quad 0 \quad 0]^{\mathrm{T}}$ on this exact solution.

Remark 3.6. In theory, we can do similar things for other two-dimensional solutions. The problem reduces to finding proper approximations $\mathscr{F}_h$ and $\mathscr{S}_h$, such that for a steady flow the residual is exactly zero. For example, assume that we start with a linear approximation of the primitive variables $\mathbf{p}_h = [H_h, \vec{u}_h]$ of Proposition 3.4. Suppose to be looking for a solution on which $\nabla \cdot (H\vec{u}) = 0$. This is a strong hypothesis, however it is true for problems with constant discharge. The big advantage of the conservative $\mathscr{RD}$ approach considered in this paper is that it allows to mimic the behavior of the continuous solutions. In particular, as long as exact integration is used, hence preserving conservation, we are allowed to play with the differential form of the equations. In particular, we can write for the second and third components of the residual

$$\int_E \left[\frac{\partial}{\partial x} \begin{pmatrix} H_h u_h^2 \\ H_h u_h v_h \end{pmatrix} + \frac{\partial}{\partial y} \begin{pmatrix} H_h u_h v_h \\ H_h v_h^2 \end{pmatrix} + g H_h \nabla B_h \right] = \int_E \left[\nabla \cdot \vec{q}(\mathbf{p}_h) \begin{pmatrix} u_h \\ v_h \end{pmatrix} + H_h \nabla \mathscr{I}(\mathbf{p}_h) + H_h \operatorname{curl} \vec{u}_h \begin{pmatrix} -v_h \\ u_h \end{pmatrix} \right], \tag{33}$$

where $\vec{q}(\mathbf{p}_h) = H_h \vec{u}_h$ and $\mathscr{I}(\mathbf{p}_h) = \vec{u}_h^2/2 + g(H^h + B^h)$ are numerical approximations of the local discharge and energy, consistent with $\mathbf{p}_h$. The idea is then to add to the natural approximations $\mathscr{F}(\mathbf{p}_h)$ and $\mathscr{S}(\mathbf{p}_h)$ terms of the type

$$\nabla(\mathscr{I}_h - \mathscr{I}(\mathbf{p}_h)) = \nabla \left(\sum_j \mathscr{I}_j \psi_j - \mathscr{I}(\mathbf{p}_h) \right) \tag{34}$$

and

$$\nabla \cdot (\vec{q}_h - \vec{q}(\mathbf{p}_h)) = \nabla \cdot \left(\sum_j \vec{q}_j \psi_j - \vec{q}(\mathbf{p}_h) \right) \tag{35}$$

with ψ_j the standard linear basis functions. On piecewise linear elements, when included in the residual these corrections give lead to terms of $\mathscr{O}(h^3)$, which are within the truncation error. Hence these terms do not spoil the accuracy of $\mathscr{LP}$ schemes. Moreover, one easily checks that on an initial solution with constant discharge $\vec{q} = \vec{q}_0$ and energy $\mathscr{I} = \mathscr{I}_0$, and assuming that the rotational term is zero, one gets

$$\phi^h = \begin{bmatrix} \int_E \nabla \cdot \vec{q}_h = \vec{q}_0 \cdot \sum_j \vec{n}_j/2 \\ \int_E (\vec{u}_h \nabla \cdot \vec{q}_h + H_h \nabla \mathscr{I}_h) = \bar{u}\vec{q}_0 \cdot \sum_j \vec{n}_j/2 + \overline{H}\mathscr{I}_0 \sum_j \vec{n}_j/2 \end{bmatrix} = 0$$

with $\bar{u}$ and $\overline{H}$ average velocity and water height. We see that these corrections guarantee that the residual vanishes *exactly* on solutions with constant discharge and energy, thus allowing an *exact preservation* of this class of analytical solutions. Unfortunately, the hypothesis of a vanishing rotational terms is extremely strong. Indeed, in general we do not know yet how to handle this term. The only case that we know of in which we can guarantee that $\operatorname{curl}\vec{u}_h$ vanishes identically is when computing pseudo-1d solutions with grid aligned velocity, on structured grids. As an example, one easily checks that for the grid on the left in Fig. 4 one has (with the notation and local node numbering of the right picture in the same figure)

$$\left(\frac{\partial v_h}{\partial x} - \frac{\partial u_h}{\partial y}\right)_A = \left(\frac{\partial v_h}{\partial x} - \frac{\partial u_h}{\partial y}\right)_B = \sum_j v_j \frac{n_{jx}}{2} - \sum_j u_j \frac{n_{jy}}{2} = -u_0 \frac{n_{1y} + n_{3y}}{2} = 0$$

using the fact that the flow is quasi-1d ($u_3 = u_1 = u_0$), that it is aligned with the grid ($v = 0$), and that $0 = \sum_j n_{jy} = n_{1y} + n_{3y}$, since $n_{2y} = 0$. In practice, on this type of grids, one verifies that pseudo-1d solutions are preserved up to machine accuracy by $\mathscr{LP}$ schemes, when introducing the corrections (34) and (35). At present, we do not know how to generalize this technique to more general situations.

3.4. Examples of linear conservative $\mathscr{RD}$ schemes

We give some examples of $\mathscr{RD}$ schemes for (1). We focus our attention on truly upwind schemes. The generalization of the ideas of the paper to non-upwind schemes will be reported in a forthcoming publication.

3.4.1. Multidimensional upwind schemes

Perhaps one of the biggest advantages of the $\mathscr{RD}$ approach is its ability to incorporate true multidimensional upwinding into the discretization. A multidimensional upwind ($\mathscr{MU}$) scheme is defined as one for which [29] (see Eq. (15))

$$\widetilde{K}_i^+ = 0 \Rightarrow \phi_i = 0 \quad \text{on any } E \in \mathscr{T}_h$$

with $\widetilde{K}_i^{\pm}$ defined in the usual matrix sense, using the eigenvalue decomposition of $\widetilde{K}_i$. Note that, for a scalar problem this corresponds to splitting the element residual only to the nodes situated downstream in E with respect to the local multidimensional wave speed. In this paper we consider the following $\mathscr{MU}$ schemes.

The LDA scheme. The LDA scheme is perhaps the most successful linear linearity preserving $\mathscr{RD}$ scheme. It is defined by the splitting

$$\phi_i^{\mathrm{LDA}}(\mathbf{w}_h) = \beta_i^{\mathrm{LDA}} \phi^h(\mathbf{w}_h) = K_i^+ N \phi^h, \quad N = \left(\sum_{j \in E} K_j^+\right)^{-1}. \tag{36}$$

First of all, note that due to the definition of β_i^{LDA}, the use of exact mean-value Jacobians $\widetilde{K}_i$ or of approximate ones, denoted by K_i, does not alter the conservative character of the scheme, which respects (16) by con-

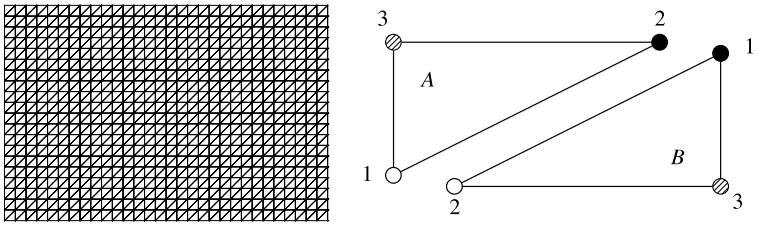

Fig. 4. Regular grid of Remark 3.6.

struction. For this reason, from now on we assume that (36) is obtained with an approximate linearization, which is the simplest possible approach. *Note also that since the shallow water system (1) and (2) is symetrizable, the matrices $K_i^+ N$ are always defined, see [24]. In fact, since the Jacobian matrices of the system have no common eigenvalues, the matrix $\sum_{j\in E} K_j^+$ is invertible, see again [24] for details*. For steady computations, fully discrete versions of the scheme can be obtained by choosing one of the time integration strategies of Section 3.2. For time-dependent computations a second order scheme is obtained by using the space-time formulation of Section 3.2.2 (however, see [30,15]), still distributing the (space-time) residual with β_i^{LDA} defined by (36). A different definition of the space-time LDA scheme is also possible. The scheme is obtained noting that, in the case of a homogeneous linear system such as (26), the residual (21) becomes

$$\Phi^h = \sum_{j\in E} (C_j^{n+1}\mathbf{v}_j^{n+1} + C_j^n \mathbf{v}_j^n)$$

with

$$C_j^{n+1} = \frac{\Delta t}{2} K_j + \frac{|E|}{3}, \quad C_j^n = \frac{\Delta t}{2} K_j - \frac{|E|}{3}.$$

In [8,31] it is shown that the C_j matrices are true generalizations to space-time prismatic elements of the K_j Jacobians. Under a proper time-step restriction [8,31] (see also Section 4), the eigenvalues of all the C_j^n matrices are negative. A conservative and space-time $\mathscr{M}\mathscr{U}$ scheme is then obtained as

$$\Phi_i^{\text{ST-LDA}}(\mathbf{w}_h) = \beta_i^{\text{ST-LDA}} \Phi^h(\mathbf{w}_h) = C_i^+ M \Phi^h, \quad M = \left(\sum_{j\in E} C_j^+\right)^{-1} \tag{37}$$

with $C_j^+ = C_j^{n+1,+}$, and Φ^h an approximation of (21) [6–8].

The N scheme. The N scheme is the *optimal* upwind first order scheme, i.e., the one with the least amount of *cross-wind* numerical dissipation. For steady problems, if the residual is computed as in (15), it is defined by the splitting

$$\phi_i^{\text{N}}(\mathbf{w}_h) = \widetilde{K}_i^+ (\mathbf{w}_i - \mathbf{w}_{in}), \quad \mathbf{w}_{in} = -\widetilde{N} \sum_{j\in E} \widetilde{K}_j^- \mathbf{w}_j. \tag{38}$$

A more general formulation is obtained by manipulating the last expression as follows [8]:

$$\phi_i^{\text{N}} = \widetilde{K}_i^+ \mathbf{w}_i + \widetilde{K}_i^+ \widetilde{N} \sum_{j\in E} \widetilde{K}_j^- \mathbf{w}_j = \widetilde{K}_i^+ \mathbf{w}_i + \overbrace{\widetilde{K}_i^+ N \sum_{j\in E} \widetilde{K}_j \mathbf{w}_j}^{\phi_i^{\text{LDA}}} - \widetilde{K}_j^+ \widetilde{N} \sum_{j\in E} \widetilde{K}_j^+ \mathbf{w}_j$$

having used the relation $\widetilde{K}_j^- = \widetilde{K}_j - \widetilde{K}_j^+$. Finally one has, dropping the $\widetilde{}$ over the flux Jacobians

$$\phi_i^{\text{N}}(\mathbf{w}_h) = \phi_i^{\text{LDA}}(\mathbf{w}_h) + d_i^{\text{N}}(\mathbf{w}_h), \quad d_i^{\text{N}}(\mathbf{w}_h) = \sum_{j\in E} K_i^+ N K_j^+ (\mathbf{w}_i - \mathbf{w}_j). \tag{39}$$

First of all we remark that due to the relations

$$\sum_{j\in E} \phi_j^{\text{LDA}} = \phi^h, \quad \sum_{j\in E} d_j^{\text{N}} = 0$$

as written in (39), the N scheme is conservative *independently on the linearization used for the flux Jacobians, and of the definition of ϕ^h*. In particular, the conservative scheme considered in this paper is obtained by computing the residual as in (16), while a simple average is used to evaluate the K_js (see Section 4). One easily shows that the d_j^{N}s are local dissipation terms (see, e.g. [24,23,8]). Hence, the conservative N scheme can be written as the LDA scheme plus some anisotropic dissipation. This approach is quickly shown to be equivalent to the formulation of [4] which is obtained by redefining $\mathbf{w}_{in}$ in (38) in a way that forces the validity of (16)[2] (see [8] for more details).

[2] $\mathbf{w}_{in} = -N\left(\phi^h - \sum_{j\in E} K_j^+ \mathbf{w}_j\right)$, which reduces to (38) when using an exact mean-value linearization.

Concerning the time-dependent case, the N scheme can be extended to the space-time framework of Section 3.2.2 in two ways. One way is to use the Crank–Nicholson time-integrators (18) or (19), in which case we will refer to the scheme simply as to the N scheme, with

$$\Phi_i^{\mathrm{N}} = \frac{|E|}{3}\delta\mathbf{u}_i + \frac{\Delta t}{2}\left(\phi_i^{\mathrm{N}}(\mathbf{w}_h^{n+1}) + \phi_i^{\mathrm{N}}(\mathbf{w}_h^{n})\right) \quad \text{or} \quad \Phi_i^{\mathrm{N}} = \frac{|E|}{3}\delta\mathbf{u}_i + \Delta t\phi_i^{\mathrm{N}}(\mathbf{w}_h^{n+1/2}). \tag{40}$$

Alternatively, one can use the formulation of [31]. In this case, we refer to the space-time ST-N scheme as the one defined by the local nodal residuals (see Eq. (37) for the notation)

$$\Phi_i(\mathbf{w}_h)^{\text{ST-N}} = \beta_i^{\text{ST-LDA}}\Phi^h(\mathbf{w}_h) + d_i^{\text{ST-N}}, \quad d_i^{\text{ST-N}} = \sum_{j\in E} C_i^+ M C_j^+ (\mathbf{w}_i^{n+1} - \mathbf{w}_j^{n+1}) \tag{41}$$

with $\Phi^h(\mathbf{w}_h)$ as in (22) or (24). In the linear case, the ST-N scheme reduces to a truly space-time variant of the N scheme (see [7,8] for more details). *In the linear case both the N and the ST-N scheme are L^∞ and L^2 stable.*

3.5. Nonlinear schemes

To complete the review of $\mathscr{RD}$ schemes, we briefly recall the construction of the nonlinear schemes. Several approaches are possible but we will only discuss two of them, one of which is actually used in the computations presented later.

3.5.1. Blended schemes

The first approach we consider is the nonlinear *blending* of a $\mathscr{LP}$ scheme with a monotone one. As noted in [13,24] the choice of the two underlying linear schemes is constrained by the need of some compatibility between the two, in the sense that this technique works best when the linear schemes are based on similar distribution strategies. For example, the blending between the N and the LDA scheme can be performed and is well posed (and is the one most often used in practice [23,24,32,4,33–35]). Blending central discretizations, such as the SUPG scheme [36,32], with the N scheme can lead to ill-posed situations in which the blending parameter cannot be computed in a way that guarantees the positivity (or the accuracy) of the scheme. A general guideline could be that central monotone schemes can only be blended with central $\mathscr{LP}$ schemes and similarly for $\mathscr{MU}$ schemes. In general a blended scheme can be written as

$$\phi_i = (\boldsymbol{I} - \theta(\mathbf{v}_h))\phi_i^{\mathscr{LP}}(\mathbf{v}_h) + \theta(\mathbf{v}_h)\phi_i^{\mathscr{M}}(\mathbf{v}_h),$$

where $\mathscr{M}$ stands for monotone, and the blending matrix $\theta(\mathbf{v}_h)$ is computed in a way that ensures local positivity (or L_∞ stability) and linearity preservation in smooth regions. Possible choices for this matrix are described in [23,24,32,4,33] and will not be discussed here. In the case of the blended LDA/N scheme, relation (39) allows to give a different interpretation of the blending. Indeed one easily shows that

$$\phi_i^{\text{LDA/N}} = \phi_i^{\text{LDA}}(\mathbf{v}_h) + \theta(\mathbf{v}_h)d_i^{\mathrm{N}}(\mathbf{v}_h). \tag{42}$$

Hence, the blending can be seen as the addition to the LDA scheme of a *properly scaled* dissipation guaranteeing non-oscillatory approximations of discontinuities and second-order of accuracy. In the case of the blending between the N and the LDA scheme, this was already recognized in [34,35]. This fact has been used in [28] to construct very high order non-oscillatory schemes for scalar conservation laws.

3.5.2. Limited schemes

The drawback of the nonlinear blending is that the definition of $\theta(\mathbf{v}_h)$ is generally difficult. Moreover, blended schemes often lack robustness. An approach that has been shown to lead to more robust nonlinear schemes is the *limiting* of a monotone scheme. This technique is thoroughly discussed in [13,23] and finds its theoretical justification in the L_∞ stability criterion introduced in the same references. For a linear system of the type (26), the idea behind this approach is, given a monotone scheme with nodal residuals $\phi_i^{\mathscr{M}}$ and a local direction $\vec{\xi}$, to decompose the nodal residuals as

$$\phi_i^{\mathscr{M}} = \sum_\sigma \langle \boldsymbol{l}^\sigma, \phi_i^{\mathscr{M}}\rangle \boldsymbol{r}^\sigma = \sum_\sigma \varphi_i^{\mathscr{M},\sigma}\boldsymbol{r}^\sigma,$$

being $\boldsymbol{l}^\sigma$ and $\boldsymbol{r}^\sigma$ the left and right eigenvectors of $A_1\xi_1 + A_2\xi_2$. Each $\varphi_i^{\mathscr{M},\sigma}$ can be seen as a scalar residual. As a result, the element residual is decomposed as:

$$\phi^h = \sum_\sigma \varphi^\sigma \boldsymbol{r}^\sigma, \quad \varphi^\sigma = \sum_{j\in E} \varphi_j^{\mathscr{M},\sigma}.$$

On each scalar wave one constructs nonlinear scalar residuals φ_i^σ respecting

$$\begin{cases} \varphi_i^\sigma \varphi_i^{\mathscr{M},\sigma} \geqslant 0, \\ \varphi_i^\sigma = \beta_i \varphi^\sigma, \quad |\beta_i| < \infty, \\ \sum\limits_{j\in E} \varphi_j^\sigma = \varphi^\sigma. \end{cases}$$

The first condition guarantees that, if the monotone scheme respects the L_∞ stability criterion, so does the nonlinear scheme. The proof is reported in [13] and will not be recalled here. Second and third conditions guarantee that the resulting scheme is consistent and linearity preserving. There are several ways of constructing mappings $\{\varphi_j^{\mathscr{M},\sigma}\}_{j\in E} \to \{\varphi_j^\sigma\}_{j\in E}$ respecting the three constraints above. A review can be found in [13,23,25], while conditions for the well-posedness of the whole procedure are given in [6–8]. The advantage of this approach, compared to the blending, is that it only needs the evaluation of one scheme. It also results in more robust schemes as shown in [13,23]. Linearity preservation is obtained by construction. Conversely, the L^2 stability properties of the scheme are not clear. This is confirmed by the poor iterative convergence that limited nonlinear $\mathscr{RD}$ schemes shown in practice. This issue is discussed in [37], and is only briefly recalled here, together with the solution adopted in the reference. The main problem of the limiting approach is that it produces schemes which are entirely based on L^∞ stability conditions, without taking into account neither the dissipative behavior of the resulting scheme, nor the directional propagation of the information, typical of many solution to (26). As shown in [37], this can lead to situations in which the final algebraic problem is very ill-conditioned. To cure this, in the reference it is proposed to rewrite the limited schemes as:

$$\phi_i(\mathbf{w}_h) = \beta_i^* \phi^h(\mathbf{w}_h) + h\theta(\mathbf{w}_h) \int_E \frac{\partial \mathscr{F}(\mathbf{w}_h)}{\partial \mathbf{w}} \cdot \nabla\psi_i \nabla \cdot \mathscr{F}(\mathbf{w}_h)\,\mathrm{d}x\,\mathrm{d}y = \left(\beta_i^* + \frac{\theta(\mathbf{w}_h)}{h} K_i\right)\phi^h(\mathbf{w}_h), \tag{43}$$

where $\theta(\mathbf{w}_h)$ guarantees that the least-squares type correction is only active in smooth regions, and β_i^* is the distribution matrix of the basic (non-stabilized) limited scheme. In [37] it is shown that this approach indeed cures the problem. Moreover, even though strict monotonicity is formally lost, a proper definition of $\theta(\mathbf{w}_h)$ yields in practice very little oscillations even for very difficult computations, as confirmed by the numerical results reported in [37]. We refer to the reference for further details concerning the extension to steady nonlinear problems. We also mention that the extension of the work of [37] to the time-dependent case is under development.

In this paper, we use nonlinear schemes obtained by limiting the N scheme using the mapping: $\varphi_i^\sigma = 0$ if $\varphi_i^{\mathscr{M},\sigma} = 0$ or $\varphi^\sigma = 0$, otherwise, $\varphi_i^\sigma = \beta_i \varphi^\sigma$ with

$$\beta_i = \frac{\max(0, \beta_i^{\mathscr{M}})}{\sum_{j\in E} \max(0, \beta_j^{\mathscr{M}})}, \quad \beta_j^{\mathscr{M}} = \frac{\varphi_i^{\mathscr{M},\sigma}}{\varphi^\sigma}.$$

In steady computations, the monotone scheme is given by the N scheme (39) with definition (16) of the residual, while in time-dependent computations we use the space-time N scheme corresponding to the $\mathscr{CN}$ time-stepping (19) with the residual computed according to (24) or the ST-N scheme (41) with the same definition of the residual. Being $\mathscr{LP}$ the nonlinear schemes respect Proposition 3.4. As already remarked, the extension of the work of [37] to the time-dependent case is under development. So, in time-dependent computations, the nonlinear schemes are obtained with the basic limiting strategy, as in [14,7].

4. Computational details and numerical results

In this section we discuss the results obtained on the shallow water equations. We consider a number of representative steady and time-dependent problems involving flows over flat and non-flat bottom. For the

computations we have used the conservative N scheme and its limited variant. In particular, the steady results have been computed with scheme (13) with explicit Euler *local* time-stepping. The distribution has been obtained according to (39) for the N scheme, or using the limiting procedure of Section 3.5.2 for the limited N scheme (LN scheme). The time-step has been computed in each node according to

$$\Delta t_i = \nu \frac{|S_i|}{\sum_{E\in\mathscr{D}_i} \max_k \lambda_i^{k,+}} \tag{44}$$

with λ_i^k is the kth eigenvalue of K_i. If not stated otherwise, we took $\nu = 0.8$. For flows over flat bed the residual has been computed according to (14) using 3 points Gaussian formulas on each edge of the element, yielding exact integration in terms of the linearly varying symmetrizing variables $\mathbf{v}_h$ (6). We also implemented version of the schemes based on linearly varying primitive (see Proposition 3.4) and conserved variables. We observed almost no changes in the numerical output. In presence of non-zero bed slope, the integral of the source term has been evaluated as

$$\int_E H\nabla B_h(x,y)\,\mathrm{d}x\,\mathrm{d}y = \sum_{j\in E} \overline{H} B_j \frac{\vec{n}_j}{2}$$

with the average relative water height $\overline{H}$ computed integrating exactly

$$\overline{H} = \frac{1}{|E|}\int_E H(\mathbf{v}_h)\,\mathrm{d}x\,\mathrm{d}y$$

when assuming a linear variation of $\mathbf{v}_h$. Note that $H(\mathbf{v}_h)$ can be easily shown to be quadratic in $\mathbf{v}_h$ so that a 3 points formula involving the mid-points of the edges is enough for this purpose. We have also implemented more involved formulas in which $H(\mathbf{v}_h)$ is written as the sum of a linear component (corresponding to the lake at rest solution) plus the difference due to the non-zero velocity and integrating the linear component as in the proof of Proposition 3.4 and the rest exactly (see also Remark 3.6). No differences have been observed in the numerical results. The use of the primitive or of the conservative variables as primary unknowns has no effect on the results obtained on non-flat bottom cases, as long as the hypotheses of Proposition 3.4 are respected.

In the time-dependent computations, we used either the space-time version of the N scheme corresponding to the $\mathscr{C}\mathscr{N}$ scheme (19) (referred to as N scheme in the results), or the ST-N scheme (41). In all the computations the time-step has been set to

$$\Delta t = 0.75 \min_{E\in\mathscr{T}_h} \Delta t_E, \quad \Delta t_E = \frac{2}{3}\min_{j\in E} \frac{|E|}{\max_k \lambda_j^{k,+}}$$

with Δt_E the value of the time-step guaranteeing [31,8,7] (see also Section 3)

$$\sum_{j\in E} d_j^{\text{ST-N}} = 0, \quad C_j^{n,+} = 0.$$

The space-time residual is given by (24) with flux and source-term integral computed as in the steady case only at $\mathbf{v}_h^{n+1/2}$. Both in steady and unsteady computations, the nonlinear scheme has been implemented as described in Section 3.5, setting $\vec{\xi} = \vec{u}$ (the velocity vector) for the wave decomposition. In the following, when talking about the limited N (LN) scheme we refer either the scheme obtained by limiting scheme (39) or the same scheme coupled with time-discretization (19) in the space-time framework. The scheme obtained by limiting the ST-N scheme is referred to limited ST-N (LST-N) scheme. The nonlinear system of algebraic equation (20) has been solved as in [6,7] with an explicit pseudo-time iterative technique. As in the reference, we observed that a number of iterations going from 20 to 40 is needed to converge the N scheme to machine accuracy in pseudo-time. For the nonlinear scheme, the same number of iterations leads to a decrease of the L_1 norm of the residual of 3–4 orders of magnitude. As observed by several authors, [14,13,6,7], machine accuracy is almost never reached for the limited scheme, except when adding the stabilization term (43). In this case, we will explicitly mention the use of this term in the text. We remark once again that the extension of the results of [37] to the time-dependent case are under development, so the nonlinear space-time limited schemes are obtained without any extra stabilization, as in [14,7]. Also the results obtained in time dependent

computations are very little affected by the choice of the primary set of unknowns and, in the case of the N scheme, by the use of (18) instead of (19).

Independently on the choice of primary unknowns $\mathbf{w}_h$ (conservative, entropy, or primitive variables), all the results shown in the paper have been obtained by evaluating on each element E the flux Jacobians needed in the K_j matrices (and for the computation of the time step) using the local arithmetic average of the nodal values of $\mathbf{w}_h$, which are the ones actually stored and evolved throughout the computation. The boundary conditions have been weakly enforced as in [38]. This is true for the reflective (inviscid wall) condition, for characteristic (far field) condition and for sub-critical inlet/outlet. An exception to this are the super-critical inlet condition, and periodicity, which have been imposed exactly, in a strong nodal sense.

4.1. Flows over flat bed

4.1.1. Hydraulic jump over a wedge

This problem has been considered to confirm the conservative and non-oscillatory character of the schemes. It consists of a super-critical $Fr = 2.74$ flow over a 8.95° wedge. We have run the test with the N and LN schemes. A sketch of the initial solution as a close-up view of the mesh are given in Fig. 5. The mesh size is $h = 1/20$. In the same figure, we report the convergence history of the explicit Euler time-integration in terms of the L_1 norm of the residual of the relative water height H. From the figure we see that, while the linear scheme converges to machine accuracy without any problem, the convergence of the limited scheme is somewhat erratic. This can be improved by adding the stabilization term (43), with however very little change in the solution (not shown). Indeed, it is in smooth regions that the extra stabilization is active, hence there is no great influence on the results for this case (we refer to [37] for more details on the matter). The results are visualized in terms of water height and Froude number. The contours of the computed relative height and a comparison of its outlet distribution with the exact solution are reported in Figs. 6 and 7 for both schemes. The following observations can be made. The discontinuity is captured monotonically, the LN scheme giving a much sharper approximation. This is particularly clear from Fig. 7, from which we also see that angle and

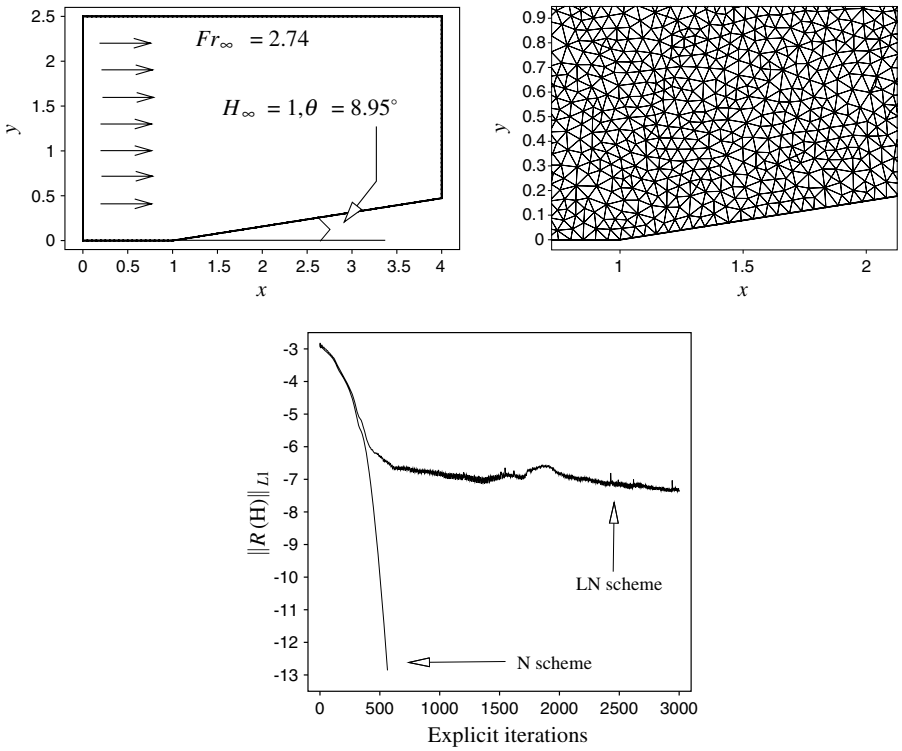

Fig. 5. Hydraulic jump over a wedge: sketch of the problem (top-left), mesh (top-right, $h = 1/20$) and convergence histories (bottom).

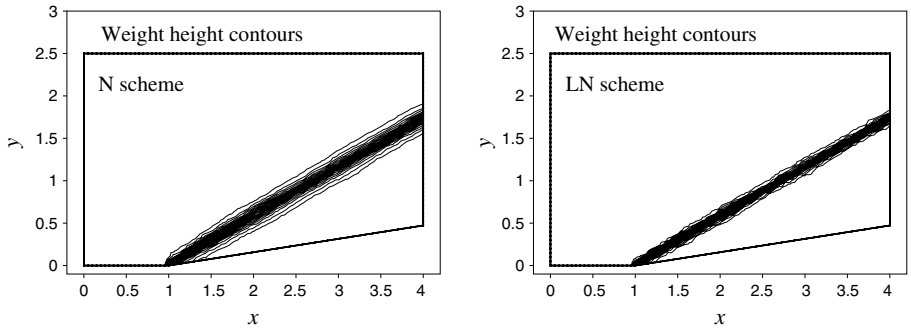

Fig. 6. Hydraulic jump over a wedge. Water height contour levels (20 levels between 0.9997 and 1.511). Left: N scheme. Right: LN scheme.

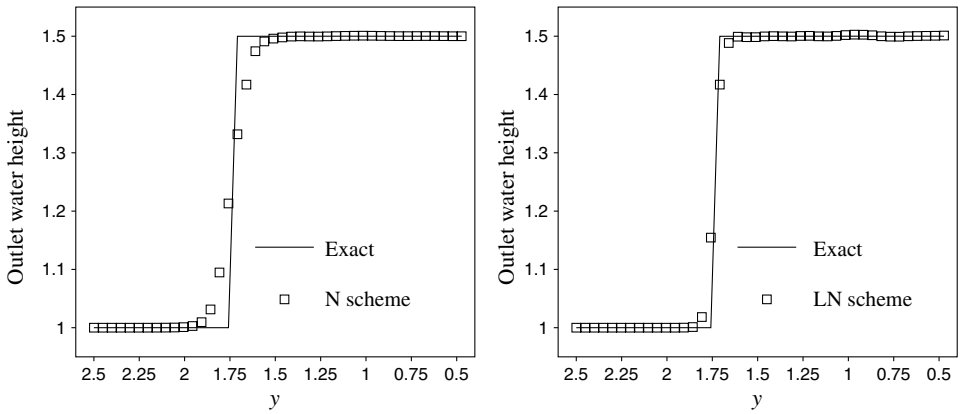

Fig. 7. Hydraulic jump over a wedge. Outlet water height. Left: N scheme. Right: LN scheme.

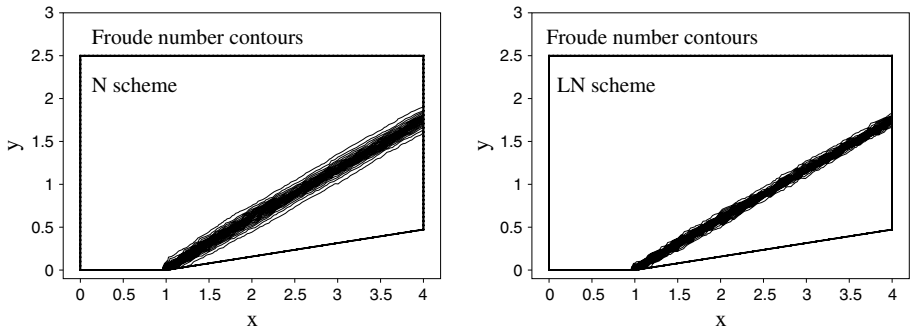

Fig. 8. Hydraulic jump over a wedge. Froude number contour levels (20 levels between 2.06 and 2.7405). Left: N scheme. Right: LN scheme.

strength of the jump are correctly reproduced, confirming the conservative character of the schemes. Similar comments can be made by looking at the contour lines of the computed Froude number, reported in Fig. 8, and at its comparison with the exact solution at the outlet, Fig. 9.

4.1.2. Trans-critical break of a circular dam

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 $r = 60$. Due to the difference in water height, the flow becomes rapidly

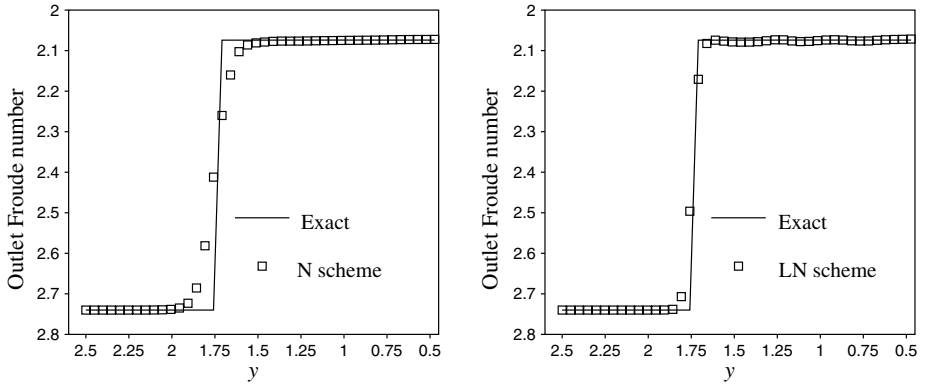

Fig. 9. Hydraulic jump over a wedge. Outlet Froude number. Left: N scheme. Right: LN scheme.

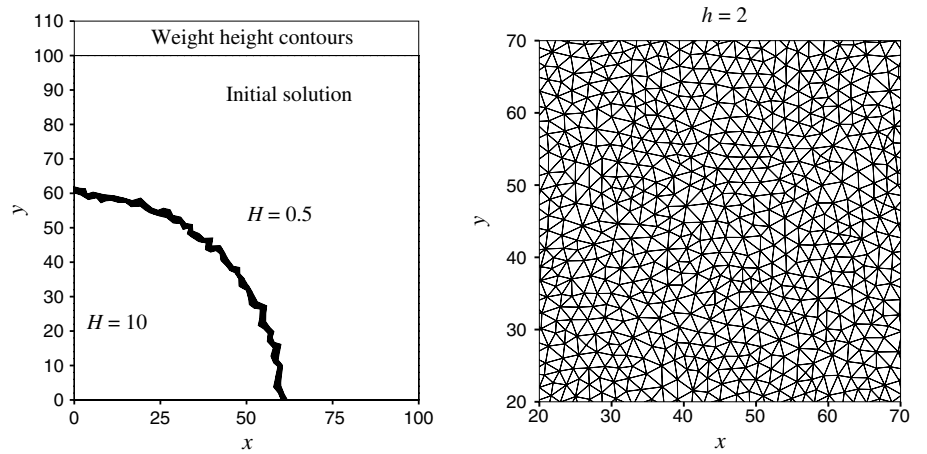

Fig. 10. Transcritical dam-break: Initial state (left) and mesh (right, $h = 2$).

trans-critical. A sketch of the initial solution and a zoom of the grid ($h = 2$) are reported in Fig. 10. Symmetry (reflective) boundary conditions have been used on the left and bottom boundaries, while on the top and on the right we imposed a characteristic far-field condition. The simulations have been run with the N scheme, with the ST-N scheme and with their limited variants until time $t = 3$. Contour plots of computed water height Froude number are given in Figs. 11 and 12. Even on this irregular grid, the flow acceleration and the right moving bore are very well reproduced. Even without extra stabilization of type (43), the LN scheme (second picture from the left) gives a quite smooth approximation of the flow acceleration. Concerning the LST-N scheme, the Froude number contours in the smooth part present small perturbations. At the origin of this feature might be the same mechanism acting when limiting centered monotone schemes, well described in [37]. However in this case the effects are less pronounced, and mainly visible in the velocity components (hence in the Froude number). Improvements are expected in the approximation of the smooth part, once the approach of [37] will be extended to the space-time schemes.

The capturing of the right moving water front, on the other hand, is monotone with all the schemes. The limited schemes yield a very sharp approximation of this feature. As remarked in [6,7], due to its upwind character in space *and time*, the ST-N scheme (third picture from the left) shows a considerably higher numerical dissipation compared to the N scheme (extreme left). Indeed, the water height waves (Fig. 11) are blurred into a unique smooth profile, while the right moving wave in the Froude contours (Fig. 12) is considerably smeared. Note that no problem whatsoever is encountered in the critical point. The computed water height and Froude number distributions along the line $y = x$ are shown in Figs. 13 and 14. The plots confirm our previous observations.

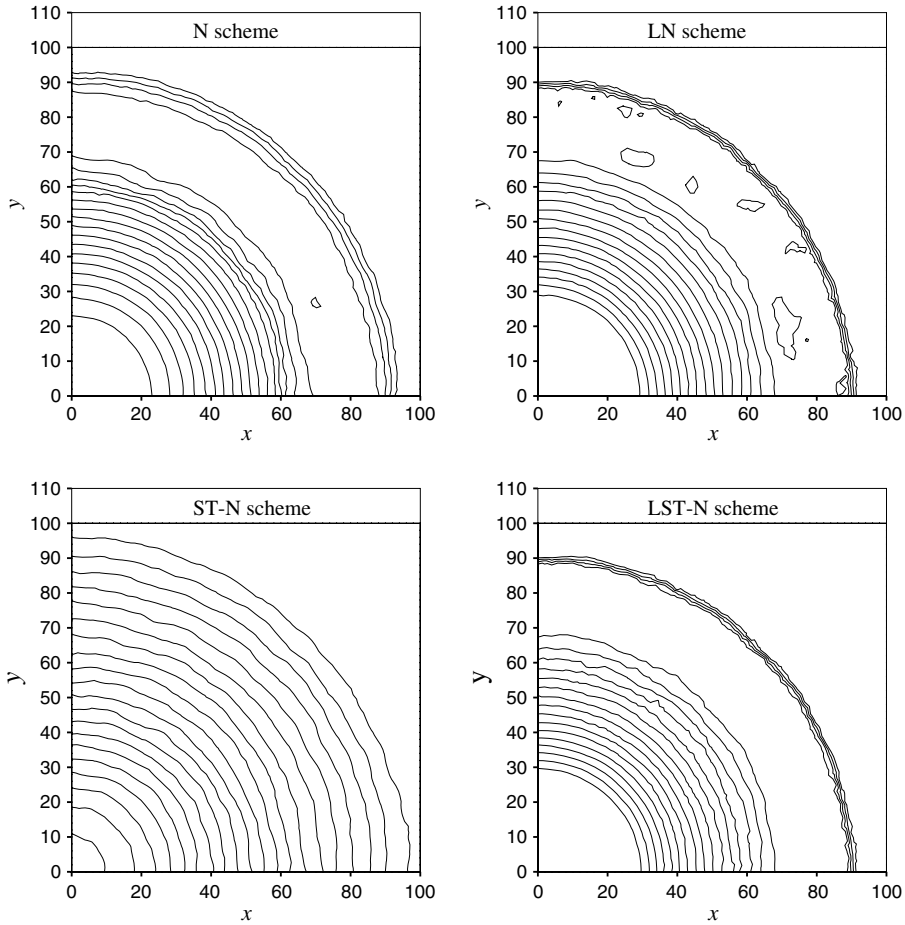

Fig. 11. Transcritical dam-break. Water height contours at $t = 3$ (20 levels between 0.498 and 10). Top row: N scheme (left) and LN scheme (right). Bottom row: ST-N scheme (left) and LST-N scheme (right).

4.1.3. Asymmetric break of a dam

This problem is taken from [39], and it is similar to the previous one, except that the geometry is more complex. We consider the sudden break of a dam separating two basins with water heights 5 and 10. The dam breaks asymmetrically at time $t = 0$ and we simulate the problem until time $t = 7.2$. A sketch of the geometry of the problem and a zoom of the unstructured mesh are given in Fig. 15. The reference mesh size if $h = 2$. The length of the breach is 75 and it starts at $y = 95$. The dam itself has a finite width of 10 and its left side is positioned at $x = 95$. Reflective boundary conditions are applied on all the edges of the domain. We show the results obtained with the nonlinear LN and LST-N schemes in terms of relative water height contours (Fig. 16), Froude number contours (Fig. 17) and distributions of water height and Froude number along the line $y = 160$ (Fig. 18). The contour plots show that both schemes compute in a very smooth way the water acceleration on the left of the dam, while the water wave moving to the right is very sharp and monotone in both the results. The reflection of this wave on the upper wall of the domain is clearly visible. The line plots of Fig. 18 confirm these observations.

4.2. Flows over non-flat bed

In this section we present some results obtained on non-flat, smooth and non-smooth bed shapes. The first test is a well-known 1D test for the shallow water equations on non-flat bed involving the steady trans-critical flow with a shock over a smooth hump. We then consider two tests involving the computation of the lake at

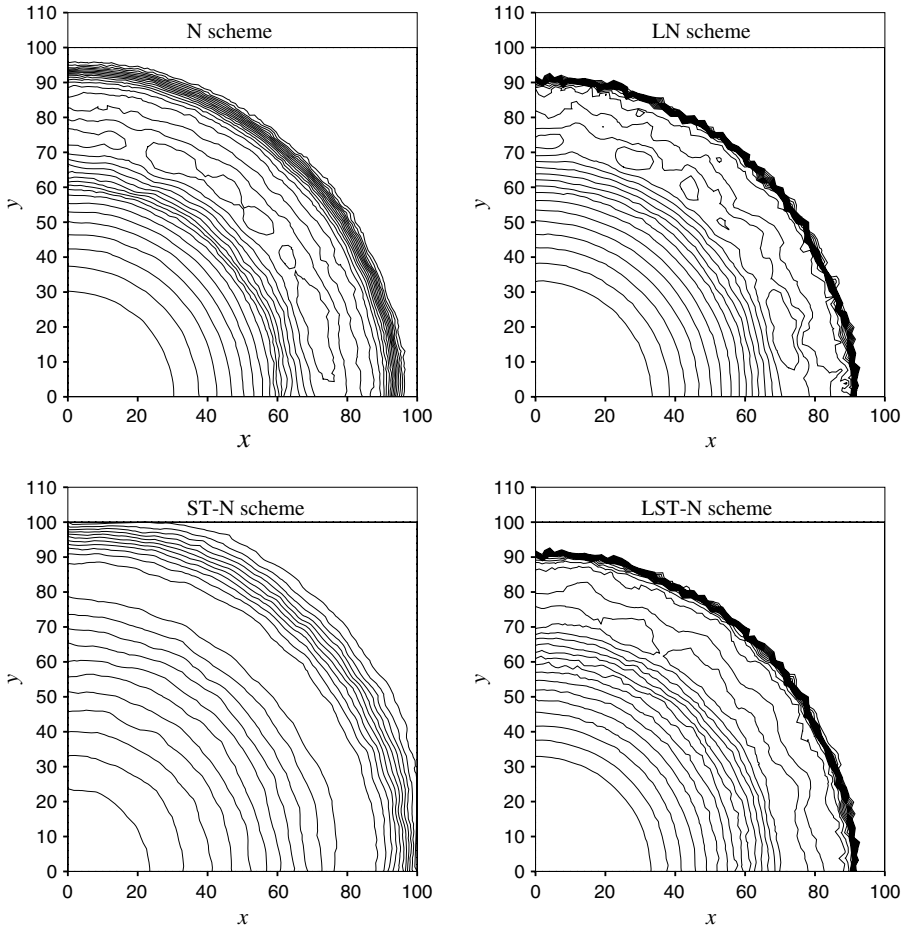

Fig. 12. Transcritical dam-break: Froude number contours at $t = 3$ (20 levels between 0.1 and 1.85). Top row: N scheme (left) and LN scheme (right). Bottom row: ST-N scheme (left) and LST-N scheme (right).

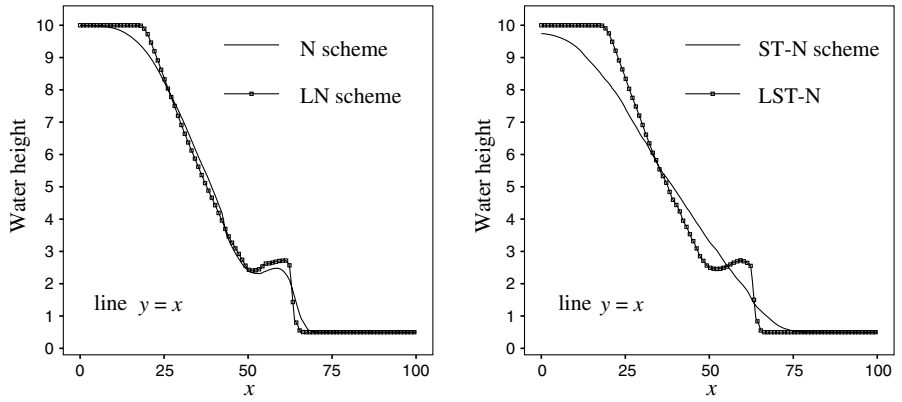

Fig. 13. Transcritical dam-break: water height distribution at $t = 3$ s on the line $y = x$ (symbols do not correspond to mesh points). Left: N and LN schemes. Right: ST-N and LST-N schemes.

rest solution. As a consequence of Proposition 3.4 the application of a linearity preserving scheme leads to an exact preservation of the analytical solution. This is indeed observed experimentally when using either scheme (13) in conjunction with the limited N scheme and the explicit Euler time-stepper, the limited N scheme or the

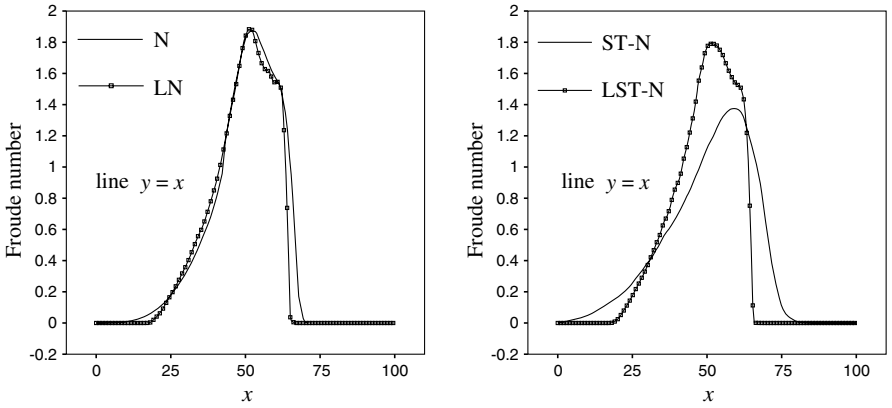

Fig. 14. Transcritical dam-break: Froude number distribution at $t = 3$ s on the line $y = x$ (symbols do not correspond to mesh points). Left: N and LN schemes. Right: ST-N and LST-N schemes.

limited ST-N scheme: the norm of the (spatial) residual stays at machine zero. In order to show the performances of the schemes in presence of variation of the bed height, we show the results obtained on problems involving perturbations of the exact lake at rest solution. The ST-N scheme has shown to be way too dissipative to be able to resolve this type of flows, hence no results with this scheme are shown. Moreover, the limited N scheme and the limited ST-N scheme have given nearly identical results. Only the ones obtained with the limited N scheme are shown. We test the schemes in two situations involving a smooth and a non-smooth variation of the bottom respectively. Lastly, we give an example of one of the truly two-dimensional solutions described in Section 2.3.2.

4.2.1. Trans-critical flow with a shock over a smooth hump

This is a particular (one-dimensional) case of the exact solutions of Section 2.3. It is obtained by assuming the following variation of the bottom [39–41]

$$B(x) = \begin{cases} 0.2 - 0.05(x-10)^2 & \text{if } 8 \leqslant x \leqslant 12, \\ 0 & \text{otherwise.} \end{cases} \tag{45}$$

Different steady solutions can be computed involving fully sub-critical, smooth trans-critical and trans-critical flow with a shock. In order to assess the shock capturing capabilities of the N scheme and of its limited variant in presence of non-flat bed, we consider here the case of steady trans-critical flow with a shock. We solve the

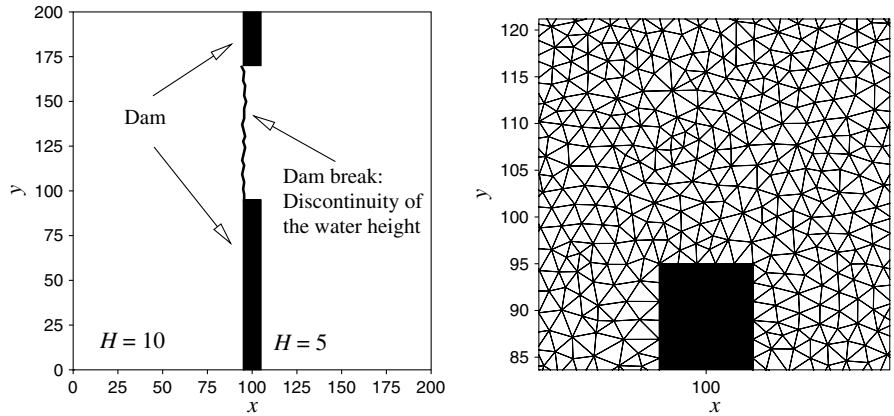

Fig. 15. Asymmetric dam break. Problem description (left) and zoom of the mesh (right).

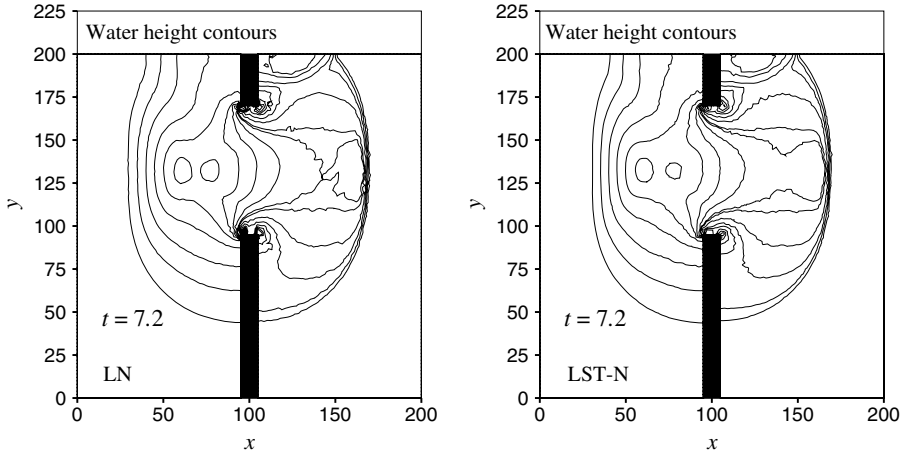

Fig. 16. Asymmetric dam break. Water height contours at time $t = 7.2$. 15 Contours between 4 and 9.95. LN scheme (left) and LST-N scheme (right).

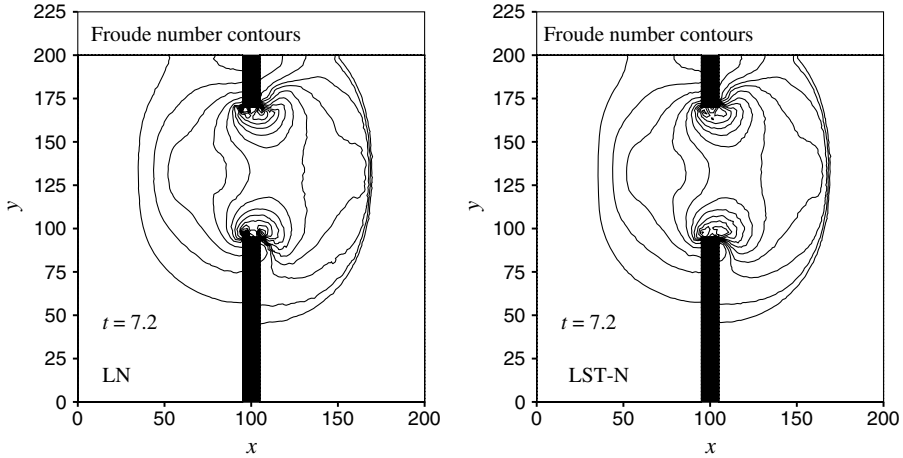

Fig. 17. Asymmetric dam break. Froude number contours at time $t = 7.2$. 15 Contours between 0.05 and 0.9. LN scheme (left) and LST-N scheme (right).

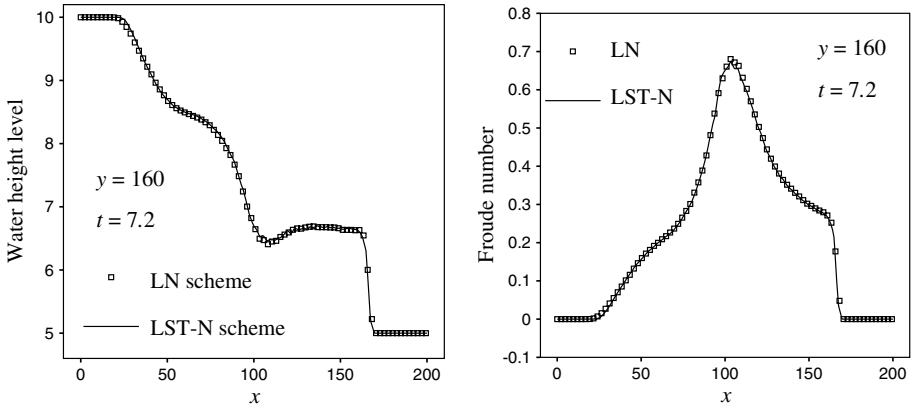

Fig. 18. Asymmetric dam break. Water height (left) and Froude number (right) distribution at $t = 7.2$ and $y = 160$. LN scheme (symbols) and LST-N scheme (line).

shallow-water equations on the spatial domain $[0,20]\times[0,0.5]$ on an irregular unstructured mesh similar to the ones used in the other computations. The reference mesh size is $h = 1/10$. We assume $B(x,y) = B(x)\ \forall y$ with $B(x)$ as in (45). Periodic boundary conditions are applied along the y direction. On the left boundary we assign as boundary condition the discharge $Hu = 0.18$ and zero v velocity, while on the right boundary we set $H = 0.33$. The steady-state solutions obtained with the N scheme and the LN scheme, using formulation (13) with explicit Euler time-stepping, are reported in Fig. 19. We can remark that the solutions are monotone and the shock approximation is very sharp, no problems are encountered in the critical point, the acceleration being smooth in both solutions, the approximation of the discharge (which should be constant and equal to 0.18 everywhere) is very good, *despite of the fact that the problem has been solved on a 2D irregular mesh instead that in 1D*. In particular, note that the peak of the error in the shock is due to the fact that across the discontinuity the direction of the velocity is not well-defined, locally giving place to large errors in the velocity components.

4.2.2. Lake at rest solution and $\mathscr{LP}$ schemes

We verify experimentally Proposition 3.4. On the domain $[0,1]^2$, we consider an initial state in which the velocity is zero and $H = 1 - B(x,y)$ with [17,42,43,39]

$$B(x,y) = 0.8\mathrm{e}^{-50((x-0.5)^2+(y-0.5)^2)}.$$

We compute the solution up to time $t = 0.5$ with the (space-time) limited N scheme on an irregular triangulation with the same topology as the one depicted in Fig. 10, and $h = 1/100$. In Table 1, we report the values (computation run in double precision) of the norms of the errors on water-height and velocity components. The results obviously confirm the theoretical result of the proposition. The numerical output is similar $\forall t > 0$, and independently on the choice of the primary unknowns. The results of the table have been obtained in symmetrizing variables.

4.2.3. Water height perturbation over 2D smooth bed

In this section we consider a test initially proposed in [17], and more recently used in [39,42,43] to assess the performances of *well-balanced* formulations of very high-order relaxation, finite difference and finite volume WENO, and discontinuous Galerkin discretizations. The spatial domain of the problem is $[0,2]\times[0,1]$. The following smooth bottom shape is assumed

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

corresponding to an ellipsoidal hump centered at $[0.9,0.5]$. The initial solution is obtained by perturbing the exact lake-at-rest state in the band $x \in [0.05,0.15]$: at $t = 0$, the velocity is set to zero everywhere, while the relative water height is set to

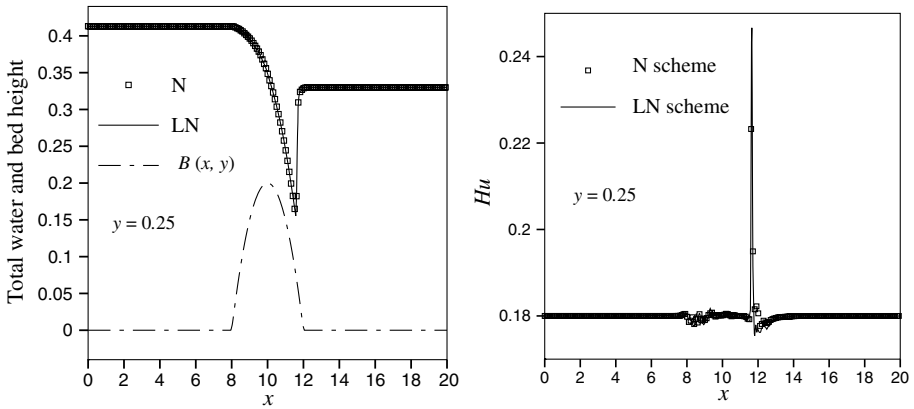

Fig. 19. Transcritical flow over a smooth hump. Total water height (left) and discharge (right). N scheme (symbols) and limited N scheme (solid line).

Table 1
Norm of the errors at time $t = 0.5$, LN scheme

	L^∞	L^1	L^2
H	7.491837e − 17	7.085969e − 17	7.107835e − 17
u	7.478237e − 17	7.161000e − 17	7.169336e − 17
v	7.478237e − 17	7.177553e − 17	7.177653e − 17

$$H = \begin{cases} 1.01 - B(x,y) & \text{if } 0.05 < x < 0.15, \\ 1 - B(x,y) & \text{otherwise.} \end{cases}$$

We solve the problem on an unstructured discretization of the domain with reference mesh size $h = 1/100$. As in [42,43] the gravity acceleration is set to $g = 9.812$. A contour plot of the total water height $H_{\text{tot}} = H + B$ at time $t = 0$, and a zoom of the grid are reported in Fig. 20. Characteristic BCs are imposed on the right and left end of the spatial domain, while the upper and lower boundaries are treated as symmetry lines. We consider the solution at four different times: $t = 0.12$, $t = 0.24$, $t = 0.36$ and $t = 0.48$. In Figs. 21 and 22 we visualize the results obtained with the space-time LN scheme. On the top rows we have reported the contours of H_{tot}, and on the bottom rows, we have reported the distribution of H_{tot} along the line $y = 0.5$. The scaling of the bed height used in the line plots is reported in the figures.

The following observations can be made. In the region ahead of the perturbation, the exact solution is *perfectly preserved* up to machine accuracy, as predicted by Proposition 3.4. In the region behind the perturbation, the solution quickly gets back to the lake-at-rest state with a small noise, probably due to grid irregularities. With respect to the results obtained, e.g. in [42], with a fifth-order finite difference WENO scheme on a structured mesh with $h = 1/100$, our results reproduce very well the interaction. The small structures of the solution shown in the reference are clearly visible in the results of the LN scheme. For this type of test, we expect perhaps some improvements in the approximation of the smooth part of the interaction, when the technique of [37] will be extended to the time-dependent case. Moreover, the use of a very high-order discretization (>2) is beneficial when approximating this type of problem, involving the propagation of a small perturbation. This is also a topic of future research. We recall, however, that Proposition 3.4 also applies to higher-order polynomial interpolations, as the ones used, e.g. in [26,27]. In this perspective, the results of this section are very encouraging: the LN scheme well reproduces the structure of the solution, while yielding a monotone approximation, and preserving exactly the lake-at-rest state in the unperturbed region. In particular, while this last property is a natural consequence of the residual approach used in this work, the well-balanced schemes of [42,43] are based on ad hoc constructions allowing to achieve, in the WENO and discontinuous Galerkin frameworks, the exact balance between flux divergence and source term.

For completeness, we report in Fig. 23 the total water height and the Froude number along the line $y = 0.5$, in the unperturbed region $x \in [0.5, 2]$. The results are the ones obtained with the N and LN schemes at time $t = 0.12$. The plots show the preservation of the exact solution up to machine accuracy obtained with the $\mathscr{LP}$ nonlinear scheme, and the very small deviation of the N scheme.

4.2.4. Water height perturbation over 2D non-smooth bed

We also consider a variant of the previous problem involving a non-smooth variation of the bed height. In particular, we set

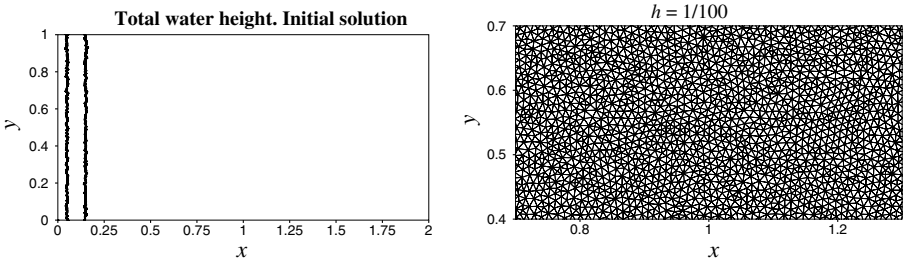

Fig. 20. Water height perturbation over smooth bed. Contours of total water height at time $t = 0$ (left) and zoom of the grid (right).

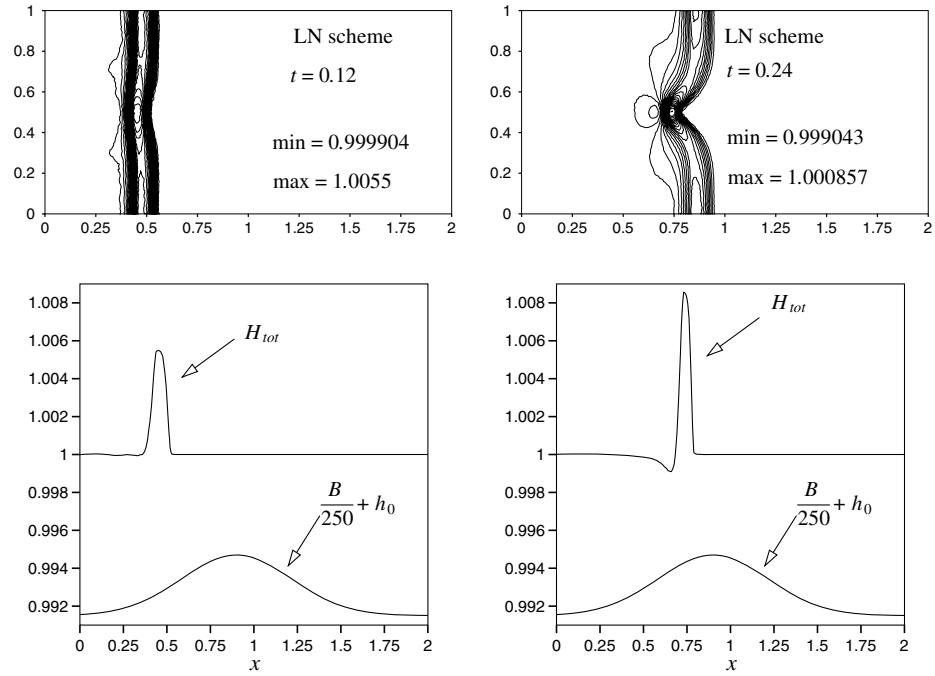

Fig. 21. Water height perturbation over smooth bed. Solution of the LN scheme at $t = 0.12$ (left) and $t = 0.24$ (right). Top: contour plot of total water height (20 contours). Bottom: distribution of H_{tot} at $y = 0.5$ ($h_0 = 0.9915$).

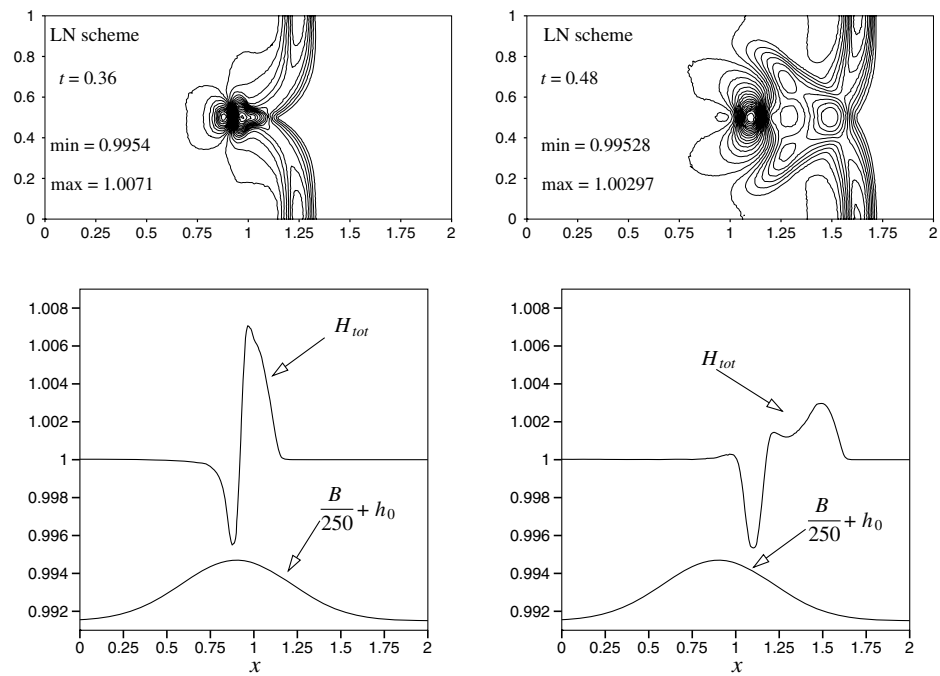

Fig. 22. Water height perturbation over smooth bed. Solution of the LN scheme at $t = 0.36$ (left) and $t = 0.48$ (right). Top: contour plot of total water height (20 contours). Bottom: distribution of H_{tot} at $y = 0.5$ ($h_0 = 0.9915$).

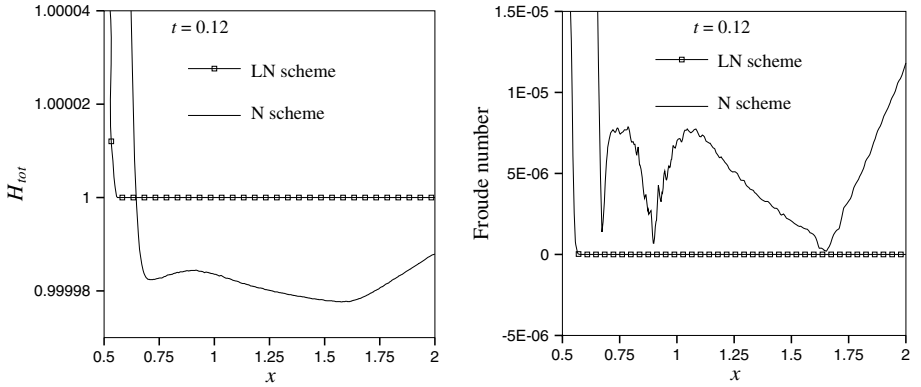

Fig. 23. Water height perturbation over smooth bed. Total water height (left) and Froude number (right) in the unperturbed region (line $y = 0.5$) at time $t = 0.12$. N scheme (line) and limited N scheme (line with symbols).

$$B(x,y) = 0.6e^{-\psi(x,y)}$$

with

$$\psi(x,y) = \begin{cases} \sqrt{(x-0.9)^2 + (y-0.5)^2} & \text{if } 0.3 \leqslant y \leqslant 0.7 \text{ and } 0.9 \leqslant x \leqslant 1.1, \\ 5(x-0.9)^2 + 50(y-0.5)^2 & \text{otherwise.} \end{cases}$$

In Fig. 24, we report a 3D view of bed shape B (properly scaled for plotting reasons). The computational set-up and the initial state are identical to the ones used in the previous test. The solution is qualitatively very close to the one obtained on the previous problem, until the perturbation reaches the discontinuity in B. We report in Fig. 25 the solution obtained with the LN scheme at times $t = 0.30$ and $t = 0.45$. As before, the scaling of the bed height used in the line plots is indicated in the pictures. The remarks made for the previous test apply also to these results. The lake-at-rest solution is preserved exactly in the unperturbed region, despite of the non-smoothness of the shape of the bottom, and of the irregular mesh. Similarly, the total water height behind the perturbation gets back to a constant value very close to one. The numerical solution of the nonlinear scheme is quite stable and monotone, even if the data of the problem are non-smooth, and the mesh quite irregular. Very small oscillations are present at later times of the simulation only in correspondence of the singular corners of $B(x,y)$, at $(x,y) = (0.9, 0.3)$, $(x,y) = (1.1, 0.3)$, $(x,y) = (0.9, 0.7)$ and $(x,y) = (1.1, 0.7)$. We believe that this is a consequence of the extremely low velocity and almost flat profile of H_{tot}, that these oscillations are not dissipated by the scheme. We expect this to be improved by the adaptation of the technique described in [37]. As done for the previous problem, in Fig. 26 we compare the solution of the N scheme with the one of LN scheme at time $t = 0.15$ in the unperturbed region, along the line $y = 0.5$: the LN scheme preserves the lake-at-rest state up to machine accuracy while very small perturbations are introduced by the first-order linear scheme.

4.2.5. Example of a truly 2D exact solution

We consider now the approximation of a particular member of the family of 2D exact solutions of Section 2.3. In particular, as described in Section 2.3.2, on the spatial domain $[-1,1]^2$ we consider a solution in which the velocity field is divergence-free, and obtained from the harmonic function

$$\psi = xy$$

as

$$u = \frac{\partial \psi}{\partial y} = x, \quad v = -\frac{\partial \psi}{\partial x} = -y.$$

The relative water height is taken as $H = 1.5 + \psi$, while the bed height is computed from

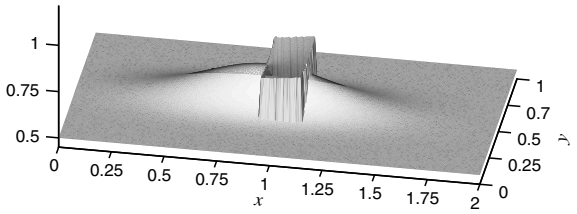

Fig. 24. 3D view of the scaled non-smooth bed.

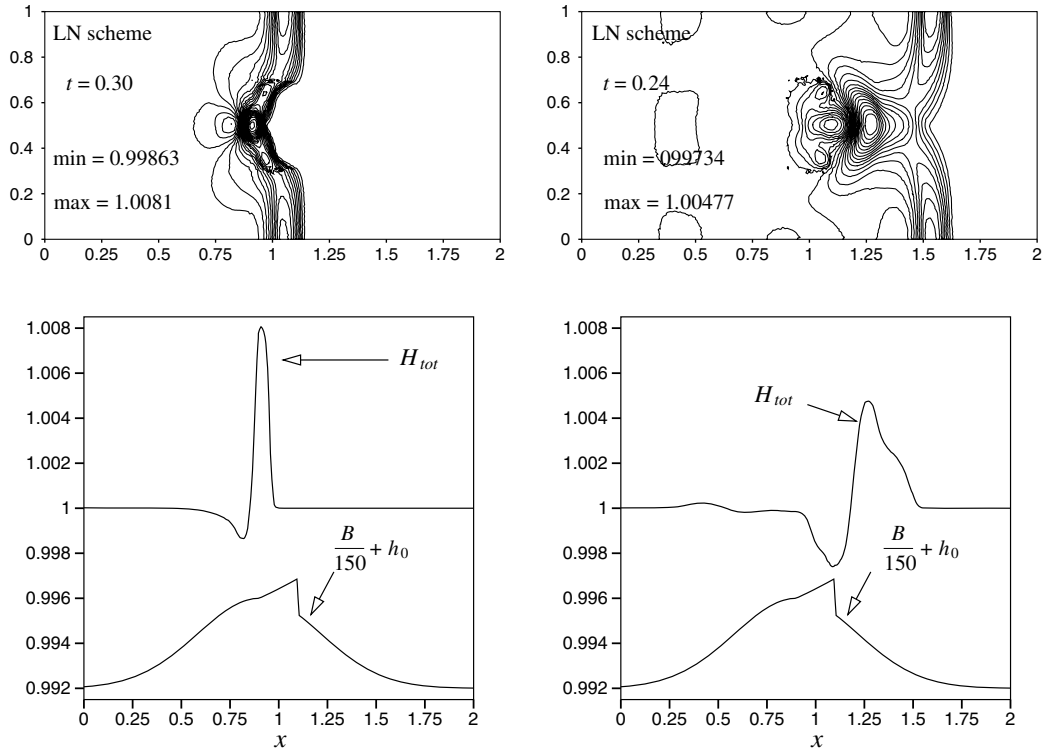

Fig. 25. Water height perturbation over non-smooth bed. Solution of the LN scheme at $t = 0.30$ (left) and $t = 0.45$ (right). Top: contour plot of total water height (20 contours). Bottom: distribution of H_{tot} at $y = 0.5$ ($h_0 = 0.992$).

$$g(1.5 + \psi) + gB(x, y) = 30 - \frac{x^2 + y^2}{2}$$

with the gravity acceleration taken to be $g = 10$. Some elements of the exact solution are visualized in Fig. 27. Note that the bottom and top boundaries are sub-critical inlets, while the left and right boundaries are sub-critical outlets. For this choice of parameters, the Froude number never exceeds one. Starting from the exact solution, we march toward steady-state using the nonlinear LN scheme (obtained by limiting (39)). The results are obtained by adding to the basic scheme the stabilization term proposed in [37], described in Section 3.5.2. The CFL parameter ν in (44) is set to $\nu = 0.1$. We compute the solution on a series of 4 unstructured irregular triangulations, similar to the ones used for the other problems. To have the same local irregularity of the grid in every computation, the meshes are generated independently. The reference mesh sizes are computed as $\tilde{h} = \sqrt{2 \times |[-1, 1]^2|/\#E} = \sqrt{8/\#E}$ and are given by $h_0 = 0.040112472606$, $h_1 = 0.020088585092$, $h_2 = 0.01007317032$, and $h_3 = 0.0050547294049$. Note that, due to the lack of locally refined regions, these reference lengths correspond almost exactly to the (uniform) mesh spacing along the boundaries, given by $\tilde{h}_0 = 1/25$, $\tilde{h}_1 = 1/50$, $\tilde{h}_2 = 1/100$, and $\tilde{h}_3 = 1/200$. Weak boundary conditions are used everywhere. In Fig. 28 we show the iterative convergence obtained on each calculation, in terms of the (properly shifted) L^1 norm of the water

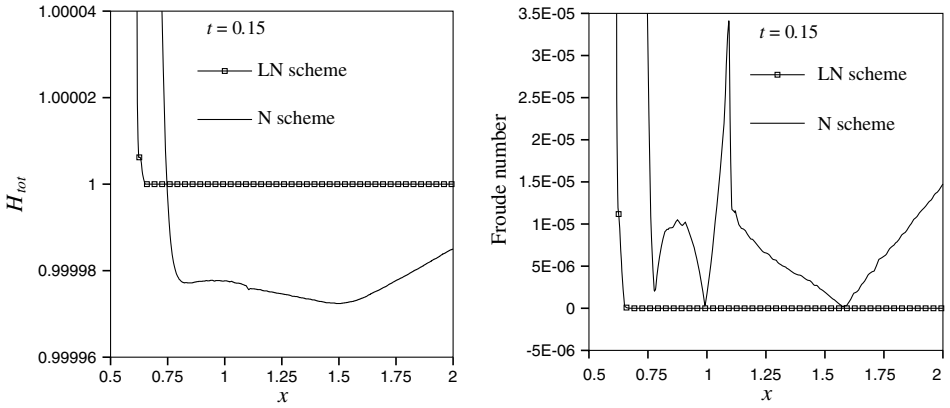

Fig. 26. Water height perturbation over non-smooth bed. Total water height (left) and Froude number (right) in the unperturbed region (line $y = 0.5$) at time $t = 0.15$. N scheme (line) and limited N scheme (line with symbols).

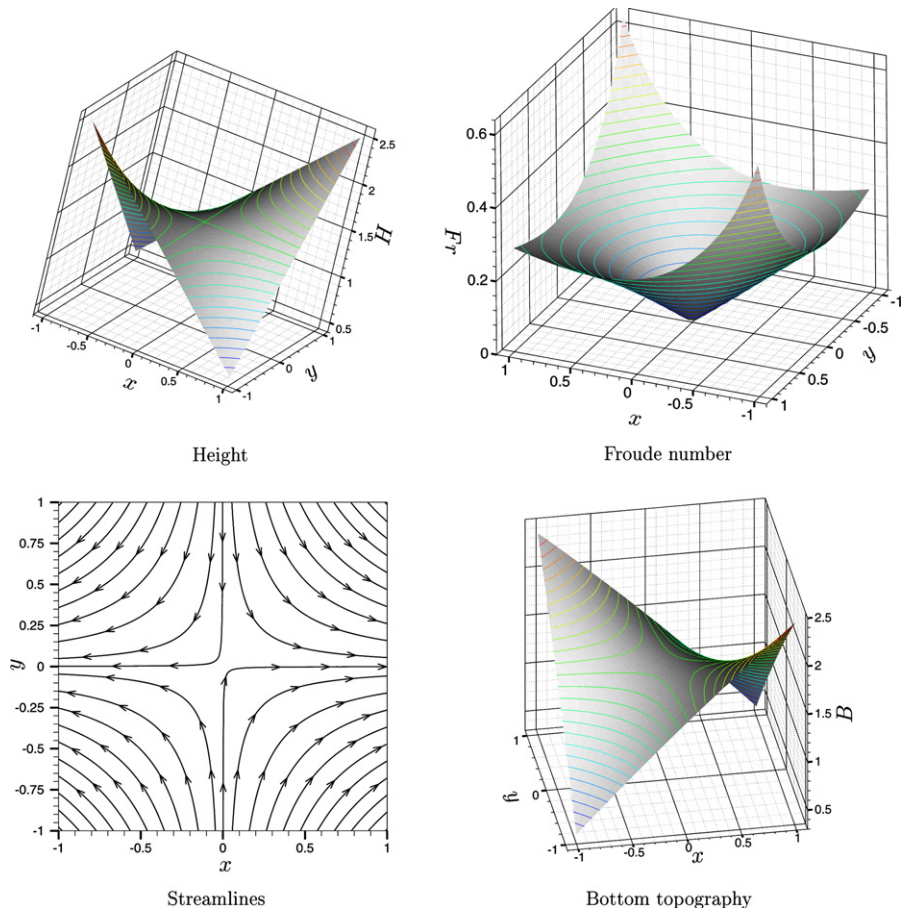

Fig. 27. Some description elements for the multidimensional steady solution.

height residual. When compared to the iterative convergence plot of Fig. 5, this result confirms the beneficial effect of the additional stabilization term of [37]. Next, we measure the rate of convergence toward the exact solution. The results obtained for the water height H are summarized in Table 2 and in Fig. 29, and confirm the second order of accuracy of the scheme.

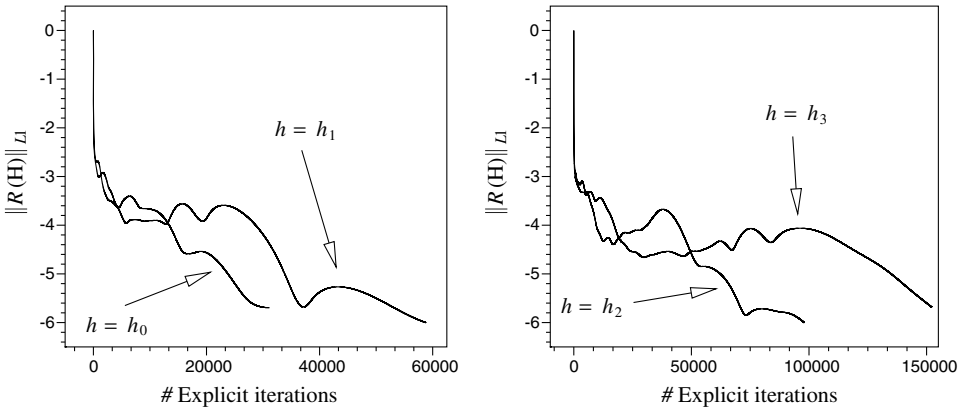

Fig. 28. Two-dimensional smooth solution: iterative convergence of the LN scheme with stabilization [37].

Table 2
2D exact solution

log(h)	$\log(\|\epsilon_H\|_{L^\infty})$	$\log(\|\epsilon_H\|_{L^1})$	$\log(\|\epsilon_H\|_{L^2})$
−1.39672	−3.50751	−3.54667	−3.52892
−1.69705	−4.35415	−4.35622	−4.35571
−1.99683	−4.92696	−4.92785	−4.92765
−2.29630	−5.44660	−5.44720	−5.44704

Grid convergence: relative water height H, limited N scheme.

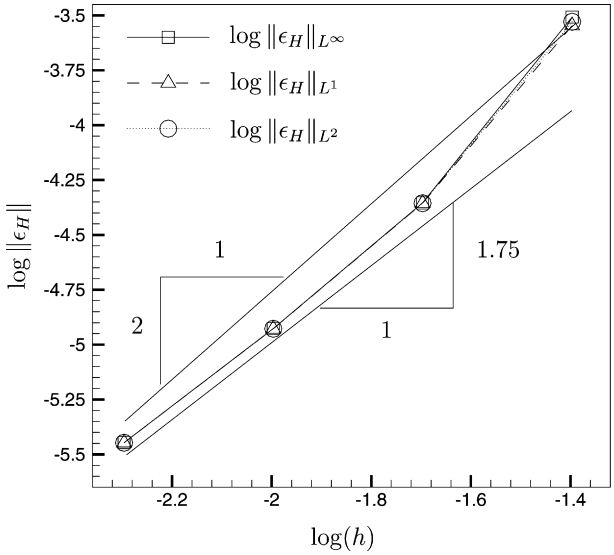

Fig. 29. 2D solution: grid convergence of the LN scheme. Relative water height H.

5. Conclusions

In this paper we have discussed the application of a family of conservative discretizations of the $\mathscr{RD}$ type to the solution of the shallow-water equations. The accuracy of the schemes has been characterized. We have given conditions to obtain schemes which are second order accurate in presence of source terms and, shown that the $\mathscr{RD}$ framework allows easily to construct schemes respecting these conditions by construction. These

schemes have also been proved to preserve *exactly* the lake at rest solution independently on the topology of the mesh, on the complexity of the bed shape and on the order of interpolation of the unknowns, which is very important for very high order extensions of the method which could be obtained following [25–27]. We also suggested a way to extend this property to more general analytical solutions. We have shown one technique to obtain nonlinear second order schemes with a strong L^∞-stable character which also preserve the lake at rest solution. An extensive numerical validation has been discussed. Results on a wide number of steady and time-dependent problems, involving the solutions of the shallow water equations on flat and non-flat bed, show the great potential of the approach.

The main challenges which remain to be faced are the extension of the schemes to the computations of dry areas, the inclusion of source terms containing stronger nonlinearities, such as the ones modeling bed friction, and the extension of the accuracy to more than second order. All these issues have as a common denominator: the improvement of the L^2 stability properties of the nonlinear schemes obtained with the limiting procedure. This will be initially achieved by adapting the technique proposed in [37]. Similar ideas can be coupled with higher order interpolation elements (see [25,23,27] for an overview) to obtain very high order schemes respecting an L^∞ stability criterion. These schemes are expected to be extremely competitive with state of the art discontinuous Galerkin discretizations.

Acknowledgments

This work was performed when the first author was still member of the Doctoral program of the von Karman Institute for Fluid Dynamics.

Appendix A. Truncation error and accuracy analysis for time dependent problems

In this appendix, we consider the solution of

$$\begin{aligned} &\frac{\partial \mathbf{u}}{\partial t} + \nabla \cdot \mathscr{F}(\mathbf{u}) = 0 \quad \text{on } \Omega \times [0, t_f], \\ &\mathbf{u}(x, y, t = 0) = \mathbf{u}^0(x, y). \end{aligned} \tag{46}$$

We will analyze the semi-discrete (discrete in time) equivalent of (46)

$$\frac{\partial \mathbf{u}}{\partial t} + \nabla \cdot \mathscr{F}(\mathbf{u}) \simeq \sum_{i=0}^{p} \frac{\alpha_i}{\Delta t_{n+1-i}} \delta \mathbf{u}^{n+1-i} + \sum_{j=0}^{q} \theta_j \nabla \cdot \mathscr{F}^{n+1-j} \tag{47}$$

having denoted $\delta \mathbf{u}^k = \mathbf{u}^k - \mathbf{u}^{k-1}$, with $\mathbf{u}^k = \mathbf{u}(x, y, t^k)$, $\mathscr{F}^{n+1-j} = \mathscr{F}(\mathbf{u}^{n+1-j})$, and with $\Delta t_k = t_k - t_{k-1}$ the (variable) time step.

Remark A.1. The α_i and θ_j coefficients can either be associated either to a finite difference formula used to approximate the time derivative, or to a high order time integration scheme, or, in the framework of a space-time approximation, to Gauss integration is time given a polynomial variation in time for $\mathbf{u}$ and $\mathscr{F}$ (see, e.g. [26,27]). The analysis of this appendix applies to all these cases.

The analysis can be done if the time step is not uniform, provided that the ratio between any consecutive time steps is uniformly bounded from below and above. For consistency, the coefficients α_i and θ_j in (47) verify

$$\sum_{i=0}^{p} \alpha_i = 1, \quad \sum_{j=0}^{q} \theta_j = 1. \tag{48}$$

In the following, in order to simplify the text, we assume a uniform time step, Δt.

We assume that (47) *is a* kth order approximation in time of (46), with $k \geqslant 1$. In particular, if the argument $\mathbf{u}$ in (47) has a smooth variation in time, then

$$\sum_{i=0}^{p} \alpha_i \frac{\delta \mathbf{u}^{n+1-i}}{\Delta t} + \sum_{j=0}^{q} \theta_j \nabla \cdot \mathscr{F}^{n+1-j} = \frac{\partial \mathbf{u}}{\partial t} + \nabla \cdot \mathscr{F}(\mathbf{u}) + \mathscr{O}(\Delta t^k). \tag{49}$$

For a given smooth $\mathbf{u}$, let $\mathbf{u}_h$ be a rth order accurate continuous piecewise smooth interpolant of the nodal values $\{\mathbf{u}_i\}_{i\in\mathscr{T}_h}$, with $r \geqslant 2$:

$$\mathbf{u}_h = \sum_{E\in\mathscr{T}_h}\sum_{j\in T}\psi_j\mathbf{u}_j \tag{50}$$

with, $\psi_j(x,y)$ continuous over the whole domain. We denote by $\mathscr{F}_h = \mathscr{F}_h(\mathbf{u}_h)$ a rth order accurate continuous piecewise smooth interpolant of the flux $\mathscr{F}(\mathbf{u})$. Without loss of generality, we assume that

$$\sum_{i=0}^{p}\alpha_i\frac{\delta\mathbf{u}_h^{n+1-i}}{\Delta t} + \sum_{j=0}^{q}\theta_j\nabla\cdot\mathscr{F}_h^{n+1-j} = \frac{\partial\mathbf{u}_h}{\partial t} + \nabla\cdot\mathscr{F}_h(\mathbf{u}_h) + \mathscr{O}(\Delta t^k) \tag{51}$$

at least within each element E.

Remark A.2. Given $\mathscr{T}_h$, the rth order interpolant $\mathbf{u}_h$ can be built in several ways. One, used in [25–28], is given by the use of $r-1$th Lagrangian triangular elements. However, other possibilities exist to build such continuous polynomials, e.g. based on the use of orthogonal polynomial and Gauss–Lobatto expansions. We refer to [44–46] and to the review [47] for further details. In any case, when $r \geqslant 3$ the interpolant (50) involves values of the solutions in nodes other than the vertices of the elements.

Finally, we note that if $\mathbf{u}$ is a smooth exact solution of (46), then

$$\sum_{i=0}^{p}\alpha_i\frac{\delta\mathbf{u}^{n+1-i}}{\Delta t} + \sum_{j=0}^{q}\theta_j\nabla\cdot\mathscr{F}^{n+1-j} = \frac{\partial\mathbf{u}}{\partial t} + \nabla\cdot\mathscr{F}(\mathbf{u}) + \mathscr{O}(\Delta t^k) = \mathscr{O}(\Delta t^k), \tag{52}$$

where the last equality is true since $\mathbf{u}$ verifies (46) in a pointwise manner.

An example of a second order discretization of type (47) is the Crank–Nicholson scheme (see also Eqs. (18) and (19))

$$\frac{\mathbf{u}^{n+1}-\mathbf{u}^n}{\Delta t} + \frac{1}{2}\left(\nabla\cdot\mathscr{F}^n + \nabla\cdot\mathscr{F}^{n+1}\right).$$

Other second order (as well as higher order) time integration methods such as the three points backwards scheme, the Adams–Bashfort scheme, etc., also fit in this framework.

We set, with clear notations,

$$\Phi^h = \int_{t^n}^{t^{n+1}}\int_E\left(\sum_{i=0}^{p}\alpha_i\frac{\delta\mathbf{u}_h^{n+1-i}}{\Delta t} + \sum_{j=0}^{q}\theta_j\nabla\cdot\mathscr{F}_h^{n+1-j}\right)\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t := \int_{t^n}^{t^{n+1}}\int_E\Psi^{n+1}(\mathbf{u}_h)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t. \tag{53}$$

Finally, we consider the $\mathscr{RD}$ schemes that write, at the time level t_{n+1},

$$\sum_{E\in\mathscr{D}_i}\Phi_i = 0, \tag{54}$$

where in all elements

$$\sum_{j\in E}\Phi_j = \Phi^h. \tag{55}$$

We look for the truncation error of the scheme. In particular, the purpose of this appendix is to describe what we intend by truncation error, and to precise conditions on the residuals Φ_i for which we have the best possible error.

Analysis. We follow the error analysis of [23–25], extending it to the time dependent case. The result obtained here improve the estimate initially given in [26]. *Using the same technique, similar results have been independently obtained in* [48].

The idea is to derive an estimate of how well an $\mathscr{RD}$ scheme reproduces the weak formulation of the problem, in correspondence of a smooth solution. This is obtained by *replacing the argument in the $\mathscr{RD}$ equations with a continuous interpolant of a smooth exact solution*. When doing this, the remainder gives us an estimate of the accuracy of the method.

Assumption A.3 (*Mesh regularity*). In the following, we consider triangulations $\mathscr{T}_h$ that are shape regular, i.e. there exists C, a finite constant independent of $\mathscr{T}_h$ such that in all the elements of $\mathscr{T}_h$ ratio of the inner diameter to the outer one is uniformly bounded by a constant C_{mesh}.

Assumption A.4 (*Time step scaling*). We assume the existence of two positive constants C_0 and C_1, eventually depending on the α_i and θ_j coefficients in (47), and on the definition of the distribution scheme (54), such that the time step verifies

$$C_0 \leqslant \frac{\Delta t}{h} \leqslant C_1 \tag{56}$$

uniformly with respect to h.

Let now $\varphi \in C^1(\Omega \times [0, t_f])$ with $\varphi(\cdot, t)$ having compact support on Ω. We set $\varphi_i^t := \varphi(x_i, y_i, t)$. We denote by φ_h the (r-order accurate) continuous piecewise polynomial spatial approximation of φ obtained as

$$\varphi_h^t = \varphi_h(x, y, t) = \sum_{i \in \mathscr{T}_h} \varphi_i^t \psi_i = \sum_{i \in \mathscr{T}_h} \varphi_i(x_i, y_i, t)\psi_i$$

with ψ_i rth order continuous piecewise polynomial basis functions on $\mathscr{T}_h$. Note that, due to the regularity of φ, we have

$$\left\| \frac{\partial \varphi}{\partial t} \right\|_{L^\infty(\Omega)} \leqslant C_2, \quad |\varphi_h^{t+\Delta t} - \varphi_h^t| \leqslant C_2 \Delta t \tag{57}$$

and

$$\|\varphi\|_{L^\infty(\Omega)} < \infty, \quad |\varphi_i - \varphi_j| \leqslant \|\nabla \varphi\|_{L^\infty(\Omega)} h \leqslant C_3 h \tag{58}$$

for some positive constants C_2 and C_3, and

$$\|\varphi_h\|_{L^\infty(\Omega)} \leqslant \|\varphi\|_{L^\infty(\Omega)}, \quad \|\nabla \varphi_h\|_{L^\infty(\Omega)} \leqslant C_2 \tag{59}$$

with C_4 depending on C_3 and the constant C_{mesh}. All the constants are uniform over $[0, t_f]$.

The global error is computed by multiplying (54) by $\varphi_i^{n+1} := \varphi_i^{t^{n+1}}$, and by adding for each mesh point and for each of the time levels between $t = 0$ to $t = N\Delta t = t_f$. We thus obtain the truncation error:

$$\mathscr{E}(\mathbf{u}_h, t_f) := \sum_{n=0}^{N} \left(\sum_{i \in \mathscr{T}_h} \varphi_i^{n+1} \sum_{E \in \mathscr{D}_i} \Phi_i(\mathbf{u}_h) \right) = \sum_{n=0}^{N} \left(\sum_{E \in \mathscr{T}_h} \sum_{i \in E} \varphi_i^{n+1} \Phi_i(\mathbf{u}_h) \right). \tag{60}$$

We recall that $\mathbf{u}_h$ *is not the solution obtained with the* $\mathscr{RD}$ *scheme but the continuous* r*th order piecewise polynomial interpolant of* $\mathbf{u}$, *smooth exact solution to* (46). Hence $\mathscr{E}(\mathbf{u}_h, t_f)$ is in general non-zero, and its magnitude is an indicator of the accuracy of the discretization.

Denoting by Φ_i^c the Galerkin residual

$$\Phi_i^c = \int_{t^n}^{t^{n+1}} \int_E \left(\sum_{i=0}^{p} \alpha_i \frac{\delta \mathbf{u}_h^{n+1-i}}{\Delta t} + \sum_{j=0}^{q} \theta_j \nabla \cdot \mathscr{F}_h^{n+1-j} \right) \psi_i \, \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}t,$$

we note that the identities

$$\sum_{i \in E} (\Phi_i - \Phi_i^c) = 0, \quad \sum_{i \in E} \varphi_j (\Phi_i - \Phi_i^c) = 0$$

allows to write (see Eq. (53))

$$\sum_{i \in E} \varphi_i^{n+1} \Phi_i = \int_{t^n}^{t^{n+1}} \int_E \varphi_h^{n+1} \Psi^{n+1}(\mathbf{u}_h) \, \mathrm{d}x \, \mathrm{d}y + \frac{1}{K} \sum_{i \in E} \sum_{j \in E} (\varphi_i^{n+1} - \varphi_j^{n+1})(\Phi_i - \Phi_i^c),$$

where $\varphi_h^{n+1} = \varphi_h^{t^{n+1}} = \varphi_h(x, y, t^{n+1})$, and K is the total number of nodes (i.e., degrees of freedom) of the element. Thus (60) rewrites

$$\mathscr{E}(\mathbf{u}_h, t_f) = \mathrm{I} + \mathrm{II} + \mathrm{III}$$

with

$$\mathrm{I} = \sum_{n=0}^{N} \int_{t^n}^{t^{n+1}} \int_{\mathbb{R}^2} \varphi_h^{n+1} \left(\sum_{i=0}^{p} \alpha_i \frac{\delta \mathbf{u}_h^{n+1-i}}{\Delta t} \right) \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t,$$

$$\mathrm{II} = \sum_{n=0}^{N} \int_{t^n}^{t^{n+1}} \int_{\mathbb{R}^2} \varphi_h^{n+1} \left(\sum_{j=0}^{q} \theta_j \nabla \cdot \mathscr{F}_h^{n+1-j} \right) \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t,$$

$$\mathrm{III} = \frac{1}{K} \sum_{n=0}^{N} \sum_{E \in \mathscr{T}_h} \sum_{i \in E} \sum_{j \in E} (\varphi_i^{n+1} - \varphi_j^{n+1})(\Phi_i - \Phi_i^c).$$

We rewrite $\mathrm{I} + \mathrm{II}$ as

$$\begin{aligned} \mathrm{I} + \mathrm{II} &= \sum_{n=0}^{N} \int_{t^n}^{t^{n+1}} \int_{\mathbb{R}^2} \varphi_h^{n+1} \Psi^{n+1} \,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t \\ &= \sum_{n=0}^{N} \int_{t^n}^{t^{n+1}} \int_{\mathbb{R}^2} \varphi_h(x,y,t) \Psi^{n+1} \,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t + \sum_{n=0}^{N} \int_{t^n}^{t^{n+1}} \int_{\mathbb{R}^2} (\varphi_h^{n+1} - \varphi_h(x,y,t)) \Psi^{n+1} \,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t. \end{aligned}$$

Now note that, due to (51) we have for the rth order approximation in space $\mathbf{u}_h$:

$$\begin{aligned} \int_0^{t_f} \varphi_h \left(\frac{\partial \mathbf{u}_h}{\partial t} + \nabla \cdot \mathscr{F}_h \right) \mathrm{d}t &= \sum_{n=0}^{N} \int_{t^n}^{t^{n+1}} \varphi_h \left(\frac{\partial \mathbf{u}_h}{\partial t} + \nabla \cdot \mathscr{F}_h \right) \mathrm{d}t \\ &= \sum_{n=0}^{N} \int_{t^n}^{t^{n+1}} \varphi_h \Psi^{n+1}(\mathbf{u}_h) \,\mathrm{d}t + \sum_{n=0}^{N} \int_{t^n}^{t^{n+1}} \varphi_h \mathscr{O}(\Delta t^k) \,\mathrm{d}t. \end{aligned}$$

Due to (58), the last term can be estimated to an order of

$$\mathscr{O}(\#\text{time steps}) \times \mathscr{O}(\Delta t) \times \mathscr{O}(\Delta t^k) = \mathscr{O}(\Delta t^{-1}) \times \mathscr{O}(\Delta t) \times \mathscr{O}(\Delta t^k) = \mathscr{O}(\Delta t^k).$$

The compactness of φ, hence of φ_h, finally allows to write

$$\sum_{n=0}^{N} \int_{t^n}^{t^{n+1}} \int_{\mathbb{R}^2} \varphi_h \Psi^{n+1}(\mathbf{u}_h) \,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t = \int_0^{t_f} \int_{\mathbb{R}^2} \varphi_h(x,y,t) \left(\frac{\partial \mathbf{u}_h}{\partial t} + \nabla \cdot \mathscr{F}_h \right) \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t + \mathscr{O}(\Delta t^k). \tag{61}$$

And finally we have for the error (see Eq. (60))

$$\begin{aligned} \mathscr{E}(\mathbf{u}_h, t_f) = &\int_0^{t_f} \int_{\mathbb{R}^2} \varphi_h(x,y,t) \left(\frac{\partial \mathbf{u}_h}{\partial t} + \nabla \cdot \mathscr{F}_h \right) \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t \\ &+ \sum_{n=0}^{N} \int_{t^n}^{t^{n+1}} \int_{\mathbb{R}^2} (\varphi_h^{n+1} - \varphi_h(x,y,t)) \left(\sum_{j=0}^{p} \alpha_i \frac{\delta \mathbf{u}_h^{n+1-i}}{\Delta t} \right) \\ &+ \sum_{n=0}^{N} \int_{t^n}^{t^{n+1}} \int_{\mathbb{R}^2} (\varphi_h^{n+1} - \varphi_h(x,y,t)) \left(\sum_{j=0}^{q} \theta_j \nabla \cdot \mathscr{F}_h^{n+1-j} \right) \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t \\ &+ \frac{1}{K} \sum_{n=0}^{N} \sum_{E \in \mathscr{T}_h} \sum_{j \in T} (\varphi_i - \varphi_j)(\Phi_i - \Phi_i^c) + \mathscr{O}(\Delta t^k). \end{aligned} \tag{62}$$

We look at each of the terms assuming that the exact solution $\mathbf{u}$ of (46) is smooth, so that (46) is satisfied in a pointwise manner. Under this assumption:

- First term

$$\int_0^{t_f}\int_{\mathbb{R}^2}\varphi_h\left(\frac{\partial \mathbf{u}_h}{\partial t}+\nabla\cdot\boldsymbol{\mathscr{F}}_h\right)\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t$$
$$=\int_0^{t_f}\int_{\mathbb{R}^2}\varphi_h\left(\frac{\partial(\mathbf{u}_h-\mathbf{u})}{\partial t}+\nabla\cdot(\boldsymbol{\mathscr{F}}_h-\boldsymbol{\mathscr{F}})\right)\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t$$
$$=-\int_0^{t_f}\int_{\mathbb{R}^2}\left((\mathbf{u}_h-\mathbf{u})\frac{\partial\varphi_h}{\partial t}+(\boldsymbol{\mathscr{F}}_h-\boldsymbol{\mathscr{F}})\cdot\nabla\varphi_h\right)\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t+\int_{\mathbb{R}^2}(\mathbf{u}_h^0-\mathbf{u}^0)\varphi_h\,\mathrm{d}x\,\mathrm{d}y.$$

Since both $\mathbf{u}_h$ and $\boldsymbol{\mathscr{F}}_h$ are rth order accurate in space, uniformly in time, and due to the compactness and regularity of φ (see Eqs. (57)–(59)), both terms are of $\mathcal{O}(h^r)$.

- Second and third terms. We make use of (52) to rewrite these terms as

$$\sum_{n=0}^{N}\int_{t^n}^{t^{n+1}}\int_{\mathbb{R}^2}(\varphi_h^{n+1}-\varphi_h(x,y,t))(\Psi^{n+1}(\mathbf{u}_h)-\Psi^{n+1}(\mathbf{u}))\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t$$
$$+\sum_{n=0}^{N}\int_{t^n}^{t^{n+1}}\int_{\mathbb{R}^2}(\varphi_h^{n+1}-\varphi_h(x,y,t))\mathcal{O}(\Delta t^k)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t=A+B. \tag{63}$$

Note now that, due to (57), within $[t^n,t^{n+1}]$ the difference $\varphi_h^{n+1}-\varphi_h(x,y,t)$ is $\mathcal{O}(\Delta t)$. Using the compactness of φ_h, and recalling that the number of time steps N is of $\mathcal{O}(\Delta t^{-1})$, the last term finally gives an $B=\mathcal{O}(\Delta t^{k+1})$. We now estimate A. We split $\Psi^{n+1}(\mathbf{u}_h)-\Psi^{n+1}(\mathbf{u})$ in the two contributions due to the time variation of the unknown, and the transport term. The first part gives:

$$\sum_{n=0}^{N}\int_{t^n}^{t^{n+1}}\int_{\mathbb{R}^2}(\varphi_h^{n+1}-\varphi_h)\left(\sum_{i=0}^{p}\alpha_i\frac{\delta(\mathbf{u}_h-\mathbf{u})^{n+1-i}}{\Delta t}\right)\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t.$$

As already remarked, $[t^n,t^{n+1}]$, the difference $\varphi_h^{n+1}-\varphi_h(x,y,t)$ is $\mathcal{O}(\Delta t)$, hence $(\varphi_h^{n+1}-\varphi_h(x,y,t))/\Delta t=\mathcal{O}(1)$. Since $N=\mathcal{O}(\Delta t^{-1})$, and due to the compactness and boundedness of φ, the whole integral is $\mathcal{O}(\mathbf{u}_h-\mathbf{u})=\mathcal{O}(h^r)$, since the spatial approximation is rth order accurate.

We now consider the term

$$\sum_{n=0}^{N}\int_{t^n}^{t^{n+1}}\int_{\mathbb{R}^2}(\varphi_h^{n+1}-\varphi_h(x,y,t))\left(\sum_{j=0}^{q}\theta_j\nabla\cdot(\boldsymbol{\mathscr{F}}_h-\boldsymbol{\mathscr{F}})^{n+1-j}\right)\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t$$
$$=-\sum_{n=0}^{N}\int_{t^n}^{t^{n+1}}\int_{\mathbb{R}^2}\sum_{j=0}^{q}\theta_j(\boldsymbol{\mathscr{F}}_h-\boldsymbol{\mathscr{F}})^{n+1-j}\cdot\nabla(\varphi_h^{n+1}-\varphi_h)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t$$

having integrated by parts, and having used the compactness of φ. The last term can be easily estimated by using again the compactness and regularity of φ (see Eqs. (57)–(59)), and using the fact that $\boldsymbol{\mathscr{F}}_h$ is a rth order accurate approximation of $\boldsymbol{\mathscr{F}}$, leading to:

$$-\sum_{n=0}^{N}\int_{t^n}^{t^{n+1}}\int_{\mathbb{R}^2}\sum_{j=0}^{q}\theta_j(\boldsymbol{\mathscr{F}}_h-\boldsymbol{\mathscr{F}})^{n+1-j}\cdot\nabla(\varphi_h^{n+1}-\varphi_h)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t$$
$$=\mathcal{O}(\Delta t^{-1})\times\mathcal{O}(\Delta t)\times\mathcal{O}(h^r)\times\mathcal{O}(\Delta t)=\mathcal{O}(\Delta t h^r).$$

So we finally get for the second and third terms in (62)

$$A+B=\mathcal{O}(h^r)+\mathcal{O}(\Delta t h^r)+\mathcal{O}(\Delta t^{k+1})=\mathcal{O}(h^r)+\Delta t(\mathcal{O}(h^r)+\mathcal{O}(\Delta t^k)).$$

- Last term. So far, we have obtained

$$\mathscr{E}(\mathbf{u}_h,t_f)=\mathcal{O}(\Delta t^k)+\mathcal{O}(h^r)+\Delta t(\mathcal{O}(h^r)+\mathcal{O}(\Delta t^k))+\frac{1}{K}\sum_{n=0}^{N}\sum_{E\in\mathscr{T}_h}\sum_{i\in E}\sum_{j\in E}(\varphi_i^{n+1}-\varphi_j^{n+1})(\Phi_i-\Phi_i^c).$$

Using (56), and isolating terms related to the accuracy of the temporal and spatial approximations (k and r, respectively), we rewrite the last estimate as

$$\mathscr{E}(\mathbf{u}_h, t_f) = \mathcal{O}(h^k) + \mathcal{O}(h^r) + \frac{1}{K}\sum_{n=0}^{N}\sum_{E\in\mathscr{T}_h}\sum_{i\in E}\sum_{j\in E}(\varphi_i^{n+1} - \varphi_j^{n+1})(\Phi_i - \Phi_i^c). \tag{64}$$

We finally analyze the last term, trying to deduce conditions for the scheme to be rth order accurate. First of all, note that, since $N = \mathcal{O}(\Delta t^{-1}) = \mathcal{O}(h^{-1})$, due to the second in (58), and since the total number of nodes in a bounded domain is of $\mathcal{O}(h^{-2})$, we can immediately say that

$$\frac{1}{K}\sum_{n=0}^{N}\sum_{E\in\mathscr{T}_h}\sum_{i\in E}\sum_{j\in E}(\varphi_i^{n+1} - \varphi_j^{n+1})(\Phi_i - \Phi_i^c) = \mathcal{O}(h^{-2})\mathcal{O}(\Phi_i - \Phi_i^c). \tag{65}$$

We start by analyzing Φ_i^c. Using (52) we easily get:

$$\begin{aligned}\Phi_i^c &= \int_{t^n}^{t^{n+1}}\int_E \psi_i\big(\Psi^{n+1}(\mathbf{u}_h) - \Psi^{n+1}(\mathbf{u})\big)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t + \int_{t^n}^{t^{n+1}}\int_E \mathcal{O}(\Delta t^k)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t\\ &= \int_{t^n}^{t^{n+1}}\int_E \psi_i\big(\Psi^{n+1}(\mathbf{u}_h) - \Psi^{n+1}(\mathbf{u})\big)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t + \mathcal{O}(h^{k+3})\end{aligned} \tag{66}$$

having used (56), and the fact that the area of element E is $|E| = \mathcal{O}(h^2)$.Next we evaluate the first integral. We immediately observe that

$$\int_{t^n}^{t^{n+1}}\int_E \psi_i \sum_{i=0}^{p}\alpha_i\frac{\delta(\mathbf{u}_h - \mathbf{u})^{n+1-i}}{\Delta t}\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t = \mathcal{O}(h^{r+2})$$

having used (56), the fact that ψ_i is uniformly bounded, that $|E| = \mathcal{O}(h^2)$, and that $\mathbf{u}_h$ is a rth order accurate approximation. Concerning the remaining term, we first rewrite it as

$$\begin{aligned}&\int_{t^n}^{t^{n+1}}\int_E \psi_i\sum_{j=0}^{q}\theta_j\nabla\cdot(\mathscr{F}_h - \mathscr{F})^{n+1-j}\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t\\ &\quad= \Delta t\oint_{\partial E}\psi_i\sum_{j=0}^{q}\theta_j(\mathscr{F}_h - \mathscr{F})^{n+1-j}\cdot\hat{n}\,\mathrm{d}l - \Delta t\int_E\sum_{j=0}^{q}\theta_j(\mathscr{F}_h - \mathscr{F})^{n+1-j}\cdot\nabla\psi_i\,\mathrm{d}x\,\mathrm{d}y.\end{aligned}$$

For the first part, considering that ψ_i is uniformly bounded, that $|\partial E| = \mathcal{O}(h)$, and that $\mathscr{F}_h$ is rth order accurate, we get:

$$\Delta t\oint_{\partial E}\psi_i\sum_{j=0}^{q}\theta_j(\mathscr{F}_h - \mathscr{F})^{n+1-j}\cdot\hat{n}\,\mathrm{d}l = \mathcal{O}(h^{r+2}).$$

Similarly, since $\mathscr{F}_h$ is rth order accurate, since $\nabla\psi_i = \mathcal{O}(h^{-1})$, while $|E| = \mathcal{O}(h^2)$, we have

$$-\Delta t\int_E\sum_{j=0}^{q}\theta_j(\mathscr{F}_h - \mathscr{F})^{n+1-j}\cdot\nabla\psi_i\,\mathrm{d}x\,\mathrm{d}y = \mathcal{O}(h^{r+2}).$$

Using (66), we conclude that $\Phi_i^c = \mathcal{O}(h^{r+2}) + \mathcal{O}(h^{k+3})$. Plugging this information into (64), and using (65), we get for the error

$$\begin{aligned}\mathscr{E}(\mathbf{u}_h, t_f) &= \mathcal{O}(h^k) + \mathcal{O}(h^r) + \mathcal{O}(h^{-2})\mathcal{O}(\Phi_i) + \mathcal{O}(h^{-2})\mathcal{O}(h^{r+2}) + \mathcal{O}(h^{-2})\mathcal{O}(h^{k+3})\\ &= \mathcal{O}(h^k) + \mathcal{O}(h^r) + \mathcal{O}(h^{-2})\mathcal{O}(\Phi_i).\end{aligned} \tag{67}$$

Last expression gives the *two conditions* for the method to be rth order accurate. These conditions are summarized in the following proposition.

Proposition A.5 (Truncation error and rth-order accuracy). *Given any smooth function $\varphi \in C^1(\Omega\times[0,t_f])$, satisfying the regularity assumptions (57)–(59), and with $\varphi(\cdot,t)$ with compact support on Ω. Given a triangulation satisfying the regularity assumption A.1. Given time step and mesh size uniformly of the same order as in assumption A.2. Given $\mathbf{u}_h$, the continuous rth order accurate piecewise polynomial interpolant of a smooth exact*

*solution to (46), and $\mathscr{F}_h$, a continuous r*th *order accurate piecewise polynomial approximation of $\mathscr{F}(\mathbf{u})$. The $\mathscr{RD}$ scheme (53)–(55) verifies the truncation error estimate*

$$\mathscr{E}(\mathbf{u}_h, t_f) := \sum_{n=0}^{N} \sum_{i \in \mathscr{T}_h} \varphi_i \sum_{E \in \mathscr{D}_i} \Phi_i(\mathbf{u}_h) = \mathscr{O}(h^r) \tag{68}$$

provided that the following conditions are met

1. *the time integration scheme (47) is at least r*th *order accurate, i.e., $k \geqslant r$;*
2. *$\Phi_i = \mathscr{O}(h^{r+2})$.*

Remark A.6. The weak form of the problem (46) is, for any φ, C^1 function with compact support on $\mathbb{R}^2$,

$$-\int_0^{t_f} \int_{\mathbb{R}^2} \left(\mathbf{u} \frac{\partial \varphi}{\partial t} + \mathscr{F} \cdot \nabla \varphi \right) \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t + \int_{\mathbb{R}^2} \mathbf{u}^0(x) \varphi(x, y, 0)\,\mathrm{d}x\,\mathrm{d}y = 0.$$

We have shown that under the assumptions of Proposition A.5, the truncation error is

$$\mathscr{E}(\mathbf{u}_h, t_f) = -\int_0^{t_f} \int_{\mathbb{R}^2} \left(\mathbf{u}_h \frac{\partial \varphi_h}{\partial t} + \mathscr{F}_h \cdot \nabla \varphi_h \right) \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t + \int_{\mathbb{R}^2} \mathbf{u}_h^0(x) \varphi_h(x, y, 0)\,\mathrm{d}x\,\mathrm{d}y \tag{69}$$

up to terms of order $\mathscr{O}(h^r)$. As similar analysis was done, e.g. in [49,50], following [51]. In particular, following [51], we say that the scheme is *r*th order accurate if

$$\|\mathscr{E}(\mathbf{u}_h, t_f)\| \leqslant C(u, h) h^r \left(\left\| \frac{\partial \varphi}{\partial t} \right\|_\infty + \|\nabla \varphi\|_\infty \right).$$

Here, this can be achieved if the conditions $k \geqslant r$ and $\Phi_i = \mathscr{O}(h^{r+2})$ are met.

The $\mathscr{LP}$ property. If the condition $k \geqslant r$ is met, if $\mathbf{u}_h$ is the *r*th order interpolant of a smooth exact solution $\mathbf{u}$, and given $\mathscr{F}_h$, the *r*th order approximation of $\mathscr{F}(\mathbf{u})$, a simple estimate of $\Phi^h(\mathbf{u}_h)$ gives

$$\begin{aligned}
\Phi^h &= \int_{t^n}^{t^{n+1}} \int_E \Psi^{n+1}(\mathbf{u}_h)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t = \int_{t^n}^{t^{n+1}} \int_E \left(\Psi^{n+1}(\mathbf{u}_h) - \Psi^{n+1}(\mathbf{u}) \right) \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t + \int_{t^n}^{t^{n+1}} \int_E \mathscr{O}(\Delta t^k)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t \\
&= \int_{t^n}^{t^{n+1}} \int_E \sum_{i=0}^{p} \alpha_i \frac{\delta(\mathbf{u}_h - \mathbf{u})^{n+1-i}}{\Delta t}\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t + \int_{t^n}^{t^{n+1}} \int_E \sum_{j=0}^{q} \theta_j \nabla \cdot (\mathscr{F}_h - \mathscr{F})^{n+1-j}\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t + \mathscr{O}(h^{k+3}) \\
&= \mathscr{O}(h^{r+2}) + \int_{t^n}^{t^{n+1}} \oint_{\partial E} \sum_{j=0}^{q} \theta_j (\mathscr{F}_h - \mathscr{F})^{n+1-j} \cdot \hat{n}\,\mathrm{d}l\,\mathrm{d}t = \mathscr{O}(h^{r+2}) + \mathscr{O}(h^{r+2}) = \mathscr{O}(h^{r+2}).
\end{aligned}$$

Having used (52) to obtain the second line; (56) and the fact that $|E| = \mathscr{O}(h^2)$ to obtain the third; the facts that $\mathbf{u}_h$ is a *r*th order approximation, that $|E| = \mathscr{O}(h^2)$, assumption (56), Green–Gauss theorem, and the hypothesis that $k \geqslant r$ to obtain the fourth; the *r*th order of accuracy of $\mathscr{F}_h$, and the fact that $|\partial E| = \mathscr{O}(h)$ to get to the final result. As a consequence of this estimate we can give the following characterization.

Definition A.7 (*Linearity preserving schemes*). A $\mathscr{RD}$ scheme for which $\Phi_i = \beta_i \Phi^h$, with β_i uniformly bounded, that is

$$\max_{E \in \mathscr{T}_h} \max_{j \in E} \|\beta_j\| < C < \infty \quad \forall \Phi^h, \mathbf{u}_h, \mathbf{u}_h^0, h, \Delta t, \ldots$$

is said to be *Linearity Preserving* ($\mathscr{LP}$). For a *r*th order variable and flux approximation, Linearity preserving residual distribution schemes verify by construction the truncation error estimate

$$\mathscr{E}(\mathbf{u}_h, t_f) = \mathscr{O}(h^r).$$

Remark A.8 (*Second order space-time schemes*). The analysis of this appendix extends trivially to the space-time schemes used in this paper. It suffices to use the Crank–Nicholson scheme in (47). In particular, the condition for second order of accuracy becomes in this case

$$\Phi_i = \mathscr{O}(h^4).$$

Remark A.9 (*Why $\Phi_i = \mathcal{O}(h^{r+2})$ is not sufficient*). We end this appendix by recalling the argument of Remark 3.2. The condition $\Phi_i = \mathcal{O}(h^{r+2})$ only guarantees that the truncation error of the scheme is $\mathcal{O}(h^r)$ (provided that $k \geqslant r$ is also verified). However, in no way it guarantees that the scheme actually does converge with the proper rate. Some additional stability properties must be enjoyed by the scheme in order to achieve this. As a counter-example, we mention that the Galerkin scheme does verify Proposition A.5, however the error blows up under mesh refinement due to the unstable character of the scheme.

Appendix B. Truncation error and accuracy analysis for non-homogeneous problems

In this appendix we analyze the accuracy of the $\mathscr{RD}$ approximation in presence of source terms, that is we consider the steady problem

$$\nabla \cdot \mathscr{F}(\mathbf{w}) - \mathscr{S}(\mathbf{w}, x, y) = 0. \tag{70}$$

The analysis extends to the inhomogeneous case the one of [23–25], and to systems the one of [28]. As in the previous appendix, the idea is to derive an estimate of how well an $\mathscr{RD}$ scheme reproduces the weak formulation of the problem, in correspondence of a smooth solution.

In the following, we consider grids verifying the regularity assumption A.1. We also consider a smooth function $\varphi \in C_0^1(\Omega)$ verifying (58) and (59).

Let $\mathbf{w}$ be a smooth exact solution, verifying (70) in a pointwise manner. Let $\mathbf{w}_h$ be its rth order accurate continuous piecewise polynomial approximation on $\mathscr{T}_h$ (see Remark A.2 concerning the choice of $\mathbf{w}_h$):

$$\mathbf{w}_h = \sum_{i \in \mathscr{T}_h} \psi_i \mathbf{w}_i = \sum_{i \in \mathscr{T}_h} \psi_i \mathbf{w}(x_i, y_i)$$

with ψ_i rth order continuous piecewise polynomial basis functions on $\mathscr{T}_h$. We consider now scheme (13) at steady-state

$$\sum_{E \in \mathscr{D}_i} \phi_i = 0 \quad \forall i \in \mathscr{T}_h \tag{71}$$

and analyze the error

$$\mathscr{E}(\mathbf{w}_h) := \sum_{i \in \mathscr{T}_h} \varphi_i \left(\sum_{E \in \mathscr{D}_i} \phi_i \right) \tag{72}$$

when the argument of the residuals ϕ_i is replaced by $\mathbf{w}_h$. Once more we underline that $\mathbf{w}_h$ is not the numerical solution given by (71), but the rth order continuous piecewise polynomial approximation of a smooth exact solution $\mathbf{w}$. Hence, in general $\mathscr{E}(\mathbf{w}_h) \neq 0$. The magnitude of this error gives an estimate on the accuracy of the schemes.

By inverting the two summations in (72) we have

$$\mathscr{E}(\mathbf{w}_h) = \sum_{E \in \mathscr{T}_h} \left(\sum_{i \in E} \varphi_i \phi_i(\mathbf{w}_h) \right). \tag{73}$$

We write the term between brackets as

$$\begin{aligned} \sum_{i \in E} \varphi_i \phi_i(\mathbf{w}_h) &= \sum_{i \in E} \varphi_i \phi_i^c(\mathbf{w}_h) + \sum_{i \in E} \varphi_i \big(\phi_i(\mathbf{w}_h) - \phi_i^c(\mathbf{w}_h) \big) \\ &= \int_E \varphi_h (\nabla \cdot \mathscr{F}_h(\mathbf{w}_h) - \mathscr{S}_h(\mathbf{w}_h, x, y)) \, \mathrm{d}x \, \mathrm{d}y + \sum_{i \in E} \varphi_i \big(\phi_i(\mathbf{w}_h) - \phi_i^c(\mathbf{w}_h) \big) \end{aligned}$$

with $\phi_i^c(\mathbf{w}_h)$ the Galerkin fluctuation

$$\phi_i^c(\mathbf{w}_h) = \int_E \psi_i (\nabla \cdot \mathscr{F}_h(\mathbf{w}_h) - \mathscr{S}_h(\mathbf{w}_h, x, y)) \, \mathrm{d}x \, \mathrm{d}y \tag{74}$$

and with $\mathscr{F}_h(\mathbf{w}_h)$ and $\mathscr{S}_h(\mathbf{w}_h,x,y)$ continuous rth order accurate numerical approximations of the flux and of the source term on $\mathscr{T}_h$.

Thus, Eq. (73) becomes

$$\mathscr{E}(\mathbf{w}_h)=\int_\Omega \varphi_h(\nabla\cdot\mathscr{F}_h(\mathbf{w}_h)-\mathscr{S}_h(\mathbf{w}_h,x,y))\,\mathrm{d}x\,\mathrm{d}y+\sum_{E\in\mathscr{T}_h}\sum_{i\in E}\varphi_i\big(\phi_i(\mathbf{w}_h)-\phi_i^c(\mathbf{w}_h)\big)=\mathrm{I}+\mathrm{II}.$$

Since by hypothesis $\mathbf{w}$ verifies (70) in a pointwise manner, we can estimate I by

$$\begin{aligned}&\int_\Omega \varphi_h(\nabla\cdot\mathscr{F}_h(\mathbf{w}_h)-\mathscr{S}_h(\mathbf{w}_h,x,y))\,\mathrm{d}x\,\mathrm{d}y\\&\quad=\int_\Omega \varphi_h(\nabla\cdot[\mathscr{F}_h(\mathbf{w}_h)-\mathscr{F}(\mathbf{w})]-[\mathscr{S}_h(\mathbf{w}_h,x,y)-\mathscr{S}(\mathbf{w},x,y)])\,\mathrm{d}x\,\mathrm{d}y.\end{aligned}$$

The next step is to use Green formula on each element and to sum up. Using the continuity of $\mathscr{F}_h(\mathbf{w}_h)$ across element edges, we get

$$\int_\Omega \varphi_h\quad\nabla\cdot(\mathscr{F}_h(\mathbf{w}_h)-\mathscr{F}(\mathbf{w}))\,\mathrm{d}x\,\mathrm{d}y=-\int_\Omega\nabla\varphi_h(\mathscr{F}_h(\mathbf{w}_h)-\mathscr{F}(\mathbf{w}))\,\mathrm{d}x\,\mathrm{d}y=\mathscr{O}(h^r),$$

$\mathscr{F}_h(\mathbf{w}_h)$ being an rth order accurate approximation of $\mathscr{F}(\mathbf{w})$, and thanks to (58). Since $\mathscr{S}_h$ is also an rth order accurate approximation of $\mathscr{S}$, (58) ensures that

$$\int_\Omega \varphi_h(\mathscr{S}_h(\mathbf{w}_h,x,y)-\mathscr{S}(\mathbf{w},x,y))\,\mathrm{d}x\,\mathrm{d}y=\mathscr{O}(h^r).\tag{75}$$

Hence we see that on a smooth solution, given rth order accurate flux and source term approximations, we have $\mathrm{I}=\mathscr{O}(h^r)$.

Then we estimate II. We start by estimating $\phi_i^c(\mathbf{w}_h)$:

$$\begin{aligned}\phi_i^c(\mathbf{w}_h)&=\int_E\psi_i\nabla\cdot[\mathscr{F}_h(\mathbf{w}_h)-\mathscr{F}(\mathbf{w})]\,\mathrm{d}x\,\mathrm{d}y-\int_E\psi_i(\mathscr{S}_h(\mathbf{w}_h,x,y)-\mathscr{S}(\mathbf{w},x,y))\,\mathrm{d}x\,\mathrm{d}y\\&=\oint_{\partial E}\psi_i(\mathscr{F}_h(\mathbf{w}_h)-\mathscr{F}(\mathbf{w}))\cdot\hat{n}\,\mathrm{d}l-\int_E(\mathscr{F}_h(\mathbf{w}_h)-\mathscr{F}(\mathbf{w}))\cdot\nabla\psi_i\,\mathrm{d}x\,\mathrm{d}y\\&\quad+\int_E\psi_j(\mathscr{S}_h(\mathbf{w}_h,x,y)-\mathscr{S}(\mathbf{w},x,y))\,\mathrm{d}x\,\mathrm{d}y\\&=\mathscr{O}(h^{r+1})+\mathscr{O}(h^{r+1})+\mathscr{O}(h^{r+2})=\mathscr{O}(h^{r+1})\end{aligned}\tag{76}$$

having used the boundedness of ψ_i, the fact that $\mathscr{F}_h$ and $\mathscr{S}_h$ are rth order accurate, the fact that $\nabla\psi_i=\mathscr{O}(h^{-1})$, and that $|\partial E|=\mathscr{O}(h)$ and $|E|=\mathscr{O}(h^2)$. The next remark is that, since

$$\sum_{i\in E}\phi_i=\int_E(\nabla\cdot\mathscr{F}_h(\mathbf{w}_h)-\mathscr{S}_h(\mathbf{w}_h,x,y))\,\mathrm{d}x\,\mathrm{d}y=\sum_{i\in E}\phi_i^c,$$

we have

$$\sum_{j\in E}\Big(\phi_j(\mathbf{w}_h)-\phi_j^c(\mathbf{w}_h)\Big)=0.$$

We can thus rewrite II as

$$\mathrm{II}=\sum_{E\in\mathscr{T}_h}\sum_{i\in E}\varphi_i(\phi_i-\phi_i^c)=\frac{1}{K}\sum_{E\in\mathscr{T}_h}\sum_{i\in E}\sum_{j\in E}(\varphi_i-\varphi_j)(\phi_i-\phi_i^c).$$

being K the total number of nodes (i.e. degrees of freedom) of the element. Using the estimate on I, we see that, to preserve the accuracy of the approximation of the flux and of the source term, we must make sure that $\mathrm{II}=\mathscr{O}(h^r)$. Last expression shows that

$$\mathrm{II}=\mathscr{O}(\#E)\times\mathscr{O}(\varphi_i-\varphi_j)\times\mathscr{O}(\phi_i-\phi_i^c)$$

using (59), and since $\#E=\mathscr{O}(h^{-2})$ for a shape regular mesh in two dimensions, we get

$$\mathrm{II} = \mathcal{O}(h^{-1}) \times \mathcal{O}(\phi_i - \phi_i^c),$$

so that, to satisfy $\mathrm{II} = \mathcal{O}(h^r)$, we must have $\phi_i - \phi_i^c = \mathcal{O}(h^{r+1})$. Finally, estimate (76) leads to the following necessary condition for second order of accuracy.

Proposition B.1 (Truncation error and rth-order accuracy). *Given a smooth function $\varphi \in C_0^1(\Omega)$, satisfying the regularity assumptions (58) and (59). Given a triangulation satisfying the regularity assumption A.1. Given $\mathbf{w}_h$, the rth order accurate continuous piecewise polynomial approximation of $\mathbf{w}$, a smooth exact solution to (70), and denoting by $\mathscr{F}_h$ and $\mathscr{S}_h$ continuous rth order accurate approximations to the exact flux and source term $\mathscr{F}(\mathbf{w})$, and $\mathscr{S}(\mathbf{w}, x, y)$ on $\mathscr{T}_h$. A $\mathscr{RD}$ scheme verifies the truncation error estimate*

$$\mathscr{E}(\mathbf{w}_h) := \sum_{i \in \mathscr{T}_h} \varphi_i \sum_{E \in \mathscr{D}_i} \phi_i(\mathbf{w}_h) = \mathcal{O}(h^r) \tag{77}$$

provided that the condition

$$\phi_i = \mathcal{O}(h^{r+1})$$

is met.

Remark B.2. The weak form of the problem (70) is, for any φ, C_0^1 function with compact support on $\mathbb{R}^2$,

$$-\int_{\mathbb{R}^2} \mathscr{F}(\mathbf{w}) \cdot \nabla \varphi \, \mathrm{d}x \, \mathrm{d}y + \int_{\mathbb{R}^2} \varphi \mathscr{S}(\mathbf{w}, x, y) \, \mathrm{d}x \, \mathrm{d}y = 0.$$

What we have shown is that under the assumptions of Proposition B.1, the truncation error is

$$\mathscr{E}(\mathbf{w}_h) = -\int_{\mathbb{R}^2} \mathscr{F}_h(\mathbf{w}_h) \cdot \nabla \varphi_h \, \mathrm{d}x \, \mathrm{d}y + \int_{\mathbb{R}^2} \varphi_h \mathscr{S}_h(\mathbf{w}, x, y) \, \mathrm{d}x \, \mathrm{d}y \tag{78}$$

up to terms of order $\mathcal{O}(h^r)$. As in Appendix A, we recall the analysis of [49–51]. In particular, following [51], we say that the scheme is rth order accurate if

$$\|\mathscr{E}(\mathbf{w}_h)\| \leqslant C(u, h) h^r \left(\|\varphi\|_\infty + \|\nabla \varphi\|_\infty \right).$$

Here, this can be achieved provided that the condition $\phi_i = \mathcal{O}(h^{r+1})$ is met (however, see Remarks 3.2 and A.9).

The $\mathscr{LP}$ property. Given continuous and rth order accurate flux and source terms approximations, $\mathscr{F}_h$ and $\mathscr{S}_h$, for a smooth exact solution one has

$$\phi^h(\mathbf{w}_h) = \oint_{\partial E} (\mathscr{F}_h - \mathscr{F}) \cdot \hat{n} \, \mathrm{d}l - \int_E (\mathscr{S}_h - \mathscr{S}) \, \mathrm{d}x \, \mathrm{d}y = \mathcal{O}(h^{r+1}) + \mathcal{O}(h^{r+2}) = \mathcal{O}(h^{r+1}).$$

As a consequence, we can give the following characterization.

Definition B.3 (*Linearity Preserving scheme*). A $\mathscr{RD}$ scheme is *linearity preserving* ($\mathscr{LP}$) if its distribution coefficients are uniformly bounded with respect to the solution and the data of the problem:

$$\max_{E \in \mathscr{T}_h} \max_{j \in E} \|\beta_j\| < C < \infty \quad \forall \phi^h, \mathbf{u}_h, \mathbf{u}_h^0, h \ldots$$

$\mathscr{LP}$ schemes satisfy by construction the error Eq. (77).

Remark B.4 (*Choice of $\mathscr{F}_h$ and $\mathscr{S}_h$*). The condition $\phi_i = \mathcal{O}(h^{r+1})$, only requires $\mathscr{F}_h$ to be rth order accurate and continuous, hence the use of the piecewise polynomial approximation

$$\mathscr{F}_h = \sum_{i \in \mathscr{T}_h} \mathscr{F}_i \psi_i$$

is sufficient, even though the choice $\mathscr{F}_h = \mathscr{F}(\mathbf{w}_h)$ is also possible. Conversely, since $|E| = \mathcal{O}(h^2)$, it might appear that a piecewise polynomial approximation of the source of degree $r - 2$ (hence $r - 1$th order accurate) is sufficient to guarantee $\phi^h = \mathcal{O}(h^{r+1})$. This is not true, since the analysis (in particular Eq. (75)) is only valid for

a rth order approximation $\mathscr{S}_h$. For example, for $r = 2$, an element-wise constant value of the source term can be used only if consistent with a second-order approximation in smooth regions. For example, one can use

$$\mathscr{S}_h|_E = \overline{\mathscr{S}} = \frac{1}{3}\sum_{j\in E}\mathscr{S}_j \quad \text{or} \quad \mathscr{S}_h|_E = \overline{\mathscr{S}} = \mathscr{S}(\bar{\mathbf{w}}, \bar{x}, \bar{y})$$

with $\bar{x}$ and $\bar{y}$ the coordinates of the gravity center of E, and $\bar{\mathbf{w}} = \mathbf{w}_h(\bar{x}, \bar{y})$.

Remark B.5 (*Extension to the time dependent case*). The extension of the analysis to the time dependent case

$$\frac{\partial \mathbf{u}}{\partial t} + \nabla\cdot\mathscr{F}(\mathbf{u}) - \mathscr{S}(\mathbf{u}, x, y) = 0 \quad \text{on } \Omega\times[0, t_f],$$
$$\mathbf{u}(x, y, t = 0) = \mathbf{u}^0(x, y)$$

is easily obtained, with minor modifications, from the analysis of Appendix A. In particular, it suffices to replace the temporal approximation (47) by (with obvious notation)

$$\frac{\partial \mathbf{u}}{\partial t} + \nabla\cdot\mathscr{F}(\mathbf{u}) - \mathscr{S}(\mathbf{u}, x, y) \simeq \sum_{i=0}^{p}\frac{\alpha_i}{\Delta t_{n+1-i}}\delta\mathbf{u}^{n+1-i} + \sum_{j=0}^{q}\theta_j(\nabla\cdot\mathscr{F}^{n+1-j} - \mathscr{S}^{n+1-j}).$$

The definition of the element residual (second in (53)) remains unchanged but now

$$\Psi(\mathbf{u}_h)^{n+1} = \sum_{i=0}^{p}\frac{\alpha_i}{\Delta t_{n+1-i}}\delta\mathbf{u}_h^{n+1-i} + \sum_{j=0}^{q}\theta_j(\nabla\cdot\mathscr{F}_h^{n+1-j} - \mathscr{S}_h^{n+1-j})$$

with $\mathbf{u}_h$, $\mathscr{F}_h$, and $\mathscr{S}_h$ continuous and rth order accurate piecewise polynomial approximations on $\mathscr{T}_h$ of $\mathbf{u}$, $\mathscr{F}(\mathbf{u})$, and $\mathscr{S}(\mathbf{u}, x, y)$, being $\mathbf{u}$ a smooth exact solution of the problem. The analysis remains unchanged, using the Galerkin residual given by

$$\Phi_i^c = \int_{t^n}^{t^{n+1}}\int_E \Psi(\mathbf{u}_h)^{n+1}\psi_i\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t.$$

The main difference is that, due to the new definition of $\Psi(\mathbf{u}_h)^{n+1}$, the first terms in Eq. (62) becomes now

$$\begin{aligned}
&\int_0^{t_f}\int_{\mathbb{R}^2}\varphi_h\left(\frac{\partial \mathbf{u}_h}{\partial t} + \nabla\cdot\mathscr{F}_h - \mathscr{S}_h\right)\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t\\
&\quad= \int_0^{t_f}\int_{\mathbb{R}^2}\varphi_h\left(\frac{\partial(\mathbf{u}_h - \mathbf{u})}{\partial t} + \nabla\cdot(\mathscr{F}_h - \mathscr{F}) - (\mathscr{S}_h - \mathscr{S})\right)\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t\\
&\quad= -\int_0^{t_f}\int_{\mathbb{R}^2}\left((\mathbf{u}_h - \mathbf{u})\frac{\partial\varphi_h}{\partial t} + (\mathscr{F}_h - \mathscr{F})\cdot\nabla\varphi_h + (\mathscr{S}_h - \mathscr{S})\varphi_h\right)\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t + \int_{\mathbb{R}^2}(\mathbf{u}_h^0 - \mathbf{u}^0)\varphi_h\,\mathrm{d}x\,\mathrm{d}y\\
&\quad= \mathscr{O}(h^r)
\end{aligned} \tag{79}$$

due to the regularity of φ, and to the fact that $\mathbf{u}_h$, $\mathscr{F}_h$, and $\mathscr{S}_h$ are rth order accurate. Eq. (62) also contains the additional term

$$-\sum_{n=0}^{N}\int_{t^n}^{t^{n+1}}\int_{\mathbb{R}^2}\left(\varphi_h^{n+1} - \varphi_h(x, y, t)\right)\left(\sum_{j=0}^{q}\theta_j\mathscr{S}_h^{n+1-j}\right)\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t.$$

However, the use of (63) shows that this term leads to an extra error

$$-\sum_{n=0}^{N}\int_{t^n}^{t^{n+1}}\int_{\mathbb{R}^2}\left(\varphi_h^{n+1} - \varphi_h(x, y, t)\right)\left(\sum_{j=0}^{q}\theta_j(\mathscr{S}_h - \mathscr{S})^{n+1-j}\right)\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t = \mathscr{O}(h^{r+3}),$$

since $\varphi_h^{n+1} - \varphi_h(x, y, t) = \mathscr{O}(\Delta t) = \mathscr{O}(h)$ in $[t^n, t^{n+1}]$, and since $\mathscr{S}_h$ is rth order accurate. This term is hence negligible with respect to the remaining ones.

The rest of the analysis is identical. One easily shows $\Phi_i^c = \mathcal{O}(h^{r+2}) + \mathcal{O}(h^{k+3})$, and ultimately the preliminary error estimate is given precisely (67), which leads to the same conditions for the an $\mathscr{RD}$ scheme to be rth order accurate, as in Proposition A.5.

In particular, if the time integration scheme is kth order accurate with $k \geqslant r$, and if $\Phi_i = \mathcal{O}(h^{r+2})$ one has

$$\mathscr{E}(\mathbf{u}_h, t_f) : \sum_{n=0}^{N} \sum_{i \in \mathscr{T}_h} \varphi_i \sum_{E \in \mathscr{D}_i} \Phi_i(\mathbf{u}_h) = \mathcal{O}(h^r).$$

Moreover, the error can be rewritten as

$$\mathscr{E}(u_h, t_f) = -\int_0^{t_f} \int_{\mathbb{R}^2} \left(\mathbf{u}_h \frac{\partial \varphi_h}{\partial t} + \mathscr{F}_h \cdot \nabla \varphi_h + \mathscr{S}_h \varphi_h \right) \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t + \int_{\mathbb{R}^2} \mathbf{u}_h^0(x) \varphi_h(x, y, 0)\,\mathrm{d}x\,\mathrm{d}y + \mathcal{O}(h^r)$$

and bounded as

$$\|\mathscr{E}(u_h, t_f)\| \leqslant C(u, h) h^r \left(\|\varphi\|_\infty + \|\frac{\partial \varphi}{\partial t}\|_\infty + \|\nabla \varphi\|_\infty \right).$$

As in the other cases, one easily shows that, provided that $k \geqslant r$, for a rth order approximation in space $\Phi^h = \mathcal{O}(h^{r+2})$. As a consequence, linearity preserving schemes still respect this error estimate by construction.

References

[1] A. Harten, On the symmetric form of systems of conservation laws with entropy, J. Comput. Phys. 49 (1983) 151–164.
[2] E. Tadmor, Skew-selfadjoint form for systems of conservation laws, J. Math. Anal. Appl. 103 (1984) 428–442.
[3] G. Hauke, A symmetric formulation for computing transient shallow water flows, Comp. Meth. Appl. Mech. Eng. 163 (1998) 111–122.
[4] 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.
[5] T. Quintino, M. Ricchiuto, Á. Csík, H. Deconinck, Conservative multidimensional upwind residual distribution schemes for arbitrary finite elements, in: ICCFD2 International Conference on Computational Fluid Dynamics 2, Sidney, Australia, July, 2002.
[6] M. Ricchiuto, Á. Csík, H. Deconinck, Conservative residual distribution schemes for general unsteady systems of conservation laws, in: ICCFD3 International Conference on Computational Fluid Dynamics 3, Toronto, Canada, July, 2004.
[7] M. Ricchiuto, Á. Csík, H. Deconinck, Residual distribution for general time dependent conservation laws, J. Comput. Phys. 209 (1) (2005) 249–289.
[8] M. Ricchiuto, Construction and analysis of compact residual discretizations for conservation laws on unstructured meshes, Ph.D. Thesis, von Karman Institute for Fluid Dynamics and Université Libre de Bruxelles, June, 2005.
[9] H. Deconinck, P.L. Roe, R. Struijs, A multidimensional generalization of Roe's difference splitter for the Euler equations, Comp. Fluids 22 (2/3) (1993) 215–222.
[10] P. Garcia-Navarro, M.E. Hubbard, A. Priestley, Genuinely multidimensional upwinding for the 2D shallow water equations, J. Comput. Phys. 121 (1) (1995) 79–93.
[11] M.E. Hubbard, M.J. Baines, Conservative multidimensional upwinding for the steady two-dimensional shallow water equations, J. Comput. Phys. 138 (1997) 419–448.
[12] H. Paillère, G. Degrez, H. Deconinck, Multidimensional upwind schemes for the shallow water equations, Int. J. Numer. Meth. Fluids 26 (1998) 987–1000.
[13] 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.
[14] 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.
[15] 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.
[16] J.M. Greenberg, A.-Y. Leroux, A well-balanced scheme for the numerical processing of source terms in hyperbolic systems, SIAM J. Numer. Anal. 33 (1996) 553–582.
[17] R.J. LeVeque, Balancing source terms and flux gradients in high-resolution Godunov method: the quasi-steady wave propagation algorithm, J. Comput. Phys. 146 (1998) 346–365.
[18] L. Gosse, A well-balanced flux-vector splitting scheme designed for hyperbolic systems of conservation laws with source terms, Comp. Math. Appl. 39 (2000) 135–159.
[19] M.E. Hubbard, P. Garcia-Navarro, Flux difference splitting and the balancing of source terms and flux gradients, J. Comput. Phys. 165 (1) (2000) 89–125.

[20] R. Abgrall, T.J. Barth, Residual distribution schemes for conservation laws via adaptive quadrature, SIAM J. Sci. Comput. 24 (3) (2002) 732–769.
[21] R. Abgrall, K. Mer, B. Nkonga, A Lax–Wendroff type theorem for residual schemes, in: M. Haffez, J.J. Chattot (Eds.), Innovative Methods for Numerical Solutions of Partial Differential Equations, World Scientific, Singapore, 2002, pp. 243–266.
[22] T.J. Barth, Numerical methods for conservation laws on structured and unstructured meshes, VKI LS 2003-05, 33rd Computational Fluid dynamics Course, von Karman Institute for Fluid Dynamics, 2003.
[23] R. Abgrall, M. Mezine, Residual distribution schemes for steady problems, VKI LS 2003-05, 33rd Computational Fluid dynamics Course, von Karman Institute for Fluid Dynamics, 2003.
[24] R. Abgrall, Toward the ultimate conservative scheme: following the quest, J. Comput. Phys. 167 (2) (2001) 277–315.
[25] R. Abgrall, P.L. Roe, High order fluctuation schemes on triangular meshes, J. Sci. Comput. 19 (3) (2003) 3–36.
[26] R. Abgrall, N. Andrianov, M. Mezine, Towards very high-order accurate schemes for unsteady convection problems on unstructured meshes, Int. J. Numer. Meth. Fluids 47 (8–9) (2005) 679–691.
[27] M. Ricchiuto, R. Abgrall, H. Deconinck, Construction of very high order residual distribution schemes for unsteady advection: preliminary results, VKI LS 2003-05, 3rd Computational Fluid dynamics Course, von Karman Institute for Fluid Dynamics, 2003.
[28] M. Ricchiuto, N. Villedieu, R. Abgrall, H. Deconinck, High order residual distribution schemes: discontinuity capturing crosswind dissipation and extension to advection diffusion, VKI LS, 34rd Computational Fluid dynamics Course, von Karman Institute for Fluid Dynamics, 2005.
[29] E. van der Weide, H. Deconinck, E. Issmann, G. Degrez, Fluctuation splitting schemes for multidimensional convection problems: an alternative to finite volume and finite element methods, Comput. Mech. 23 (2) (1999) 199–208.
[30] D. Caraeni, L. Fuchs, Compact third-order multidimensional upwind scheme for Navier–Stokes simulations, Theor. Comput. Fluid Dyn. 15 (2002) 373–401.
[31] Á. Csík, M. Ricchiuto, H. Deconinck, Space time residual distribution schemes for hyperbolic conservation laws over linear and bilinear elements, VKI LS 2003-05, 33rd Computational Fluid Dynamics Course, von Karman Institute for Fluid Dynamics, 2003.
[32] H. Deconinck, K. Sermeus, R. Abgrall, Status of multidimensional upwind residual distribution schemes and applications in aeronautics, AIAA Paper 2000-2328, June, 2000, AIAA CFD Conference, Denver, USA.
[33] K. Sermeus, H. Deconinck, An entropy fix for multidimensional upwind residual distribution schemes, Comp. Fluids 34 (4) (2005) 617–640.
[34] J.C.C. Henriques, L.M.C. Gato, Use of a residual distribution Euler solver to study the occurrence of transonic flow in wells turbine rotor blades, Comput. Mech. 29 (3) (2002) 243–253.
[35] J.C.C. Henriques, L.M.C. Gato, A multidimensional upwind matrix distribution scheme for conservative laws, Comp. Fluids 33 (2004) 755–769.
[36] T.J.R. Hughes, A. Brook, Streamline upwind Petrov–Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier–Stokes equations, Comp. Meth. Appl. Mech. Eng. 32 (1982) 199–259.
[37] R. Abgrall, Essentially non oscillatory residual distribution schemes for hyperbolic problems, J. Comput. Phys. 214 (2) (2006) 773–808.
[38] H. Paillère, Multidimensional upwind residual discretization schemes for the Euler and Navier–Stokes equations on unstructured meshes, Ph.D. Thesis, Université Libre de Bruxelles, 1995.
[39] M. Seaïd, Non-oscillatory relaxation methods for the shallow-water equations in one and two space dimensions, Int. J. Numer. Meth. Fluids 46 (2004) 57–484.
[40] T. Gallouët, J.-M. Hérard, N. Seguin, Some approximate Godunov schemes to compute shallow-water equations with topography, Comp. Fluids 32 (2003) 479–513.
[41] A.I. Delis, Th. Katsaounis, Relaxation schemes for the shallow water equations, Int. J. Numer. Meth. Fluids 41 (2003) 695–719.
[42] 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.
[43] 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.
[44] M. Dubiner, Spectral methods on triangles and other domains, J. Sci. Comput. 6 (4) (1991) 345–390.
[45] R.G. Owens, Spectral approximations on the triangle, Proc. R. Soc. Lond. A 454 (1998) 857–872.
[46] M. Blyth, C. Pozrikidis, A lobatto interpolation grid over the triangle, IMA J. Appl. Math. 71 (2006) 153–169.
[47] C. Eskilsonn, S. Sherwin, An introduction to spectral/HP finite elements methods for hyperbolic problems, VKI LS, 34rd Computational Fluid dynamics Course, von Karman Institute for Fluid Dynamics, 2005.
[48] G. Rossiello, P. De Palma, G. Pascazio, M. Napolitano, Third-order-accurate fluctuation splitting schemes for unsteady hyperbolic problems, J. Comput. Phys. (2006) (submitted for publication).
[49] S. Karni, A. Kurganov, G. Petrova, A smoothness indicator for adaptive algorithm for hyperbolic systems, J. Comput. Phys. 178 (2002) 323–341.
[50] S. Karni, A. Kurganov, Local error analysis for approximate solutions of hyperbolic conservation laws, Adv. Comput. Math. 22 (1) (2005) 79–99.
[51] E. Tadmor, Local error estimates for discontinuous solutions of non linear hyperbolic equations, SIAM J. Numer. Anal. 28 (1991) 891–906.