INTERNATIONAL JOURNAL FOR NUMERICAL METHODS IN FLUIDS
Int. J. Numer. Meth. Fluids 2007; **54**:543–590
Published online 16 January 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/fld.1417

The influence of source terms on stability, accuracy and conservation in two-dimensional shallow flow simulation using triangular finite volumes

J. Murillo[1], P. García-Navarro[1,*,†], J. Burguete[2] and P. Brufau[1]

[1]*Fluid Mechanics, CPS, University of Zaragoza, Spain*
[2]*E.E.A.D., C.S.I.C., Zaragoza, Spain*

SUMMARY

The two-dimensional shallow water model is a hyperbolic system of equations considered well suited to simulate unsteady phenomena related to some surface wave propagation. The development of numerical schemes to correctly solve that system of equations finds naturally an initial step in two-dimensional scalar equation, homogeneous or with source terms. We shall first provide a complete formulation of the second-order finite volume scheme for this equation, paying special attention to the reduction of the method to first order as a particular case.

The explicit first and second order in space upwind finite volume schemes are analysed to provide an understanding of the stability constraints, making emphasis in the numerical conservation and in the preservation of the positivity property of the solution when necessary in the presence of source terms. The time step requirements for stability are defined at the cell edges, related with the traditional Courant–Friedrichs–Lewy (CFL) condition. Copyright © 2007 John Wiley & Sons, Ltd.

Received 21 August 2006; Revised 7 November 2006; Accepted 10 November 2006

KEY WORDS: source terms; stability condition; wetting/drying; MUSCL-Hancock; unstructured grids; two-dimensional hydraulic model

1. INTRODUCTION

Numerical models to solve differential equations of applicability in hydraulic engineering have become a common tool and the interest in developing better, more accurate and robust methods has increased. In recent years, one of the topics of research in this field has been driven by the necessity to use the numerical methods in practical situations of growing complexity. Water flowing steadily or unsteadily over uneven, irregular and rough bed surfaces that can be flooded or emerge

[*]Correspondence to: P. García-Navarro, Fluid Mechanics, CPS, University of Zaragoza, Spain.
[†]E-mail: pigar@unizar.es

depending on the flow conditions and can, at the same time be the transport basis of a solute is a challenge for modellers and several contributions have been reported [1–8]. The question arises of what are the features to be sought in the numerical model for such kind of applications. Accuracy, conservation and numerical stability are central properties of a numerical method and must be carefully considered in first place trying to understand and to quantify as much as possible how they interfere with each other.

Begnudelli and Sanders [7] modelled shallow water flow and scalar transport over arbitrary topography involving wetting/drying fronts and reported that scalar predictions cannot be accurately predicted as undershoots and overshoots were generated even in cases with initial constant values of scalar concentration, requiring water depth and a scalar concentration tolerance to avoid an excessive mass error. In many applications, the stability region is not analysed taking into account both friction and bed level term influences and it is only admitted that, in certain situations, the time step size must be reduced to achieve better quality solutions.

In Murillo *et al.* [9] an upwind finite volume conservative model was presented to solve shallow water flow involving scalar transport in the presence of complex bottom variations using Roe's approximate solver to compute fluxes. In Murillo *et al.* [8], the numerical model was extended to include wetting/drying advance in complex geometries upholding conservation properties over the water mass and solute mass keeping accurately bounded the initial solute concentration values. The model was based in the redefinition of the eigenvectors of the system of equations to model correctly wetting fronts and required a redefinition of numerical fluxes along flood recession to prevent from excessive reduction of the time step according to the stability region. In this work, a MLG-Wierse limited cell gradient methods (MUSCL)-based extension to second order in space and time is presented and carefully studied.

The correct extension to second-order approach requires special attention in the discretization of the source terms. When applying the MUSCL or MUSCL-Hancock method it is necessary to properly discretize the source term. This can be done using a flux correction to ensure hydrostatic balance in still water conditions [5, 6]. In this paper a novel and elegant procedure to accurately and efficiently model second-order approach is derived constructing the numerical fluxes in a compact form where equilibrium is defined in a natural way.

The second-order reconstruction will be structured so that it is possible to identify the new stability region generated by the presence of source terms in the equations and those cases where the numerical scheme must be reduced to first-order approach. Also the accurate definition of the numerical fluxes in the presence of source terms ensuring equilibrium in steady state cases is formulated.

The explicit first and second order in space upwind finite volume schemes are analysed to provide an understanding of the stability constraints, making emphasis in the numerical conservation and in the preservation of the positivity property of the solution when necessary in the presence of source terms. The time step requirements for stability are defined at the cell edges, related with the traditional Courant–Friedrichs–Lewy (CFL) condition [10]. This will be first discussed for a scalar two-dimensional conservation equation with source terms. Two examples concerning the linear advection equation and a modification of the inviscid Burgers' equation will be used to test the performance of the numerical techniques. Then, the finite volume schemes will be formulated for systems of equations with source terms and the analysis of the stability, conservation and positivity constraints will be extended. The particular formulation in the case of the shallow water equations will be presented and several test cases of steady and unsteady shallow water flow with exact solution will be used to illustrate the relative performance of the schemes in each case.

2. SCALAR EQUATION WITH SOURCE TERMS

This is a conservation law expressing that a function u varies and is transported according to both the distribution of a flux function $\mathbf{f}$ and to a source term s, in the form:

$$\frac{\partial u(x,y)}{\partial t} + \nabla \mathbf{f} = s(u,x,y), \quad \mathbf{f} = (f_x, f_y) \tag{1}$$

At this point, the source term is assumed to follow $s = \nabla \boldsymbol{\tau}$, where $\boldsymbol{\tau}$ is a suitable vector, and the advection, or transport, velocity $\boldsymbol{\lambda}$ is

$$\boldsymbol{\lambda} = \frac{\mathrm{d}\mathbf{f}}{\mathrm{d}u} \tag{2}$$

To introduce the upwind finite volume scheme, (1) is integrated in a volume Ω:

$$\frac{\partial}{\partial t}\int_{\Omega} u(x,y)\,\mathrm{d}\Omega + \int_{\Omega} \nabla(\mathbf{f} - \boldsymbol{\tau})\,\mathrm{d}\Omega = 0 \tag{3}$$

In the (x, y) plane, the volumes are actually surfaces, $\mathrm{d}\Omega$ denotes the contour line and $\mathbf{n}$ is the unit outward normal vector to Ω. If Gauss's theorem is applied to the second integral in (3):

$$\frac{\partial}{\partial t}\int_{\Omega} u(x,y)\,\mathrm{d}\Omega + \oint_{\partial\Omega} (\mathbf{f} - \boldsymbol{\tau})\mathbf{n}\,\mathrm{d}l = 0 \tag{4}$$

If the domain is subdivided in cells Ω_i in a mesh fixed in time, (4) can also be applied to each cell. Calling $u_i(x, y)$ the discrete value of the function u at cell i and assuming that each cell is surrounded by a set of edges defined by the edge vertices e_k, as shown in Figure 1, (4) can be rewritten in a first approximation considering the fluxes affecting cell i as

$$\frac{\partial}{\partial t}\int_{\Omega_i} u_i(x,y)\,\mathrm{d}\Omega + \sum_{k=1}^{\mathrm{NE}} \int_{e_k}^{e_{k+1}} (\mathbf{f} - \boldsymbol{\tau})_j \mathbf{n}_k\,\mathrm{d}l = 0 \tag{5}$$

provided that the following condition applies over the cell edges [11]:

$$\sum_{k=1}^{\mathrm{NE}} \mathbf{n}_k l_k = 0 \tag{6}$$

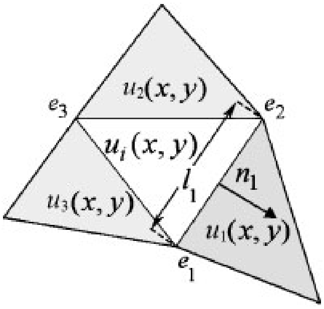

Figure 1. Cell parameters.

DOI: 10.1002/fld

and where $\mathbf{f}_j$ and $\boldsymbol{\tau}_j$ are the value of the functions $\mathbf{f}$ and $\boldsymbol{\tau}$, respectively, at the neighbour cell j connected to cell i through the edge k, $\mathbf{n}_k$ is the outward unit normal vector to the cell edge k, l_k is the corresponding edge length and NE is the number of edges in the cell. Condition (6) is necessary to ensure that, when a uniform flux crosses the cell domain, the cell state does not change in time (steady state).

The flux value inside cell i, $\mathbf{f}_i(x,y)$, can be added and subtracted from Equation (5) so that

$$\frac{\partial}{\partial t}\int_{\Omega_i} u_i(x,y)\,\mathrm{d}\Omega + \sum_{k=1}^{\mathrm{NE}}\int_{e_k}^{e_{k+1}} \delta(\mathbf{f}-\boldsymbol{\tau})_k\mathbf{n}_k\,\mathrm{d}l + \sum_{k=1}^{\mathrm{NE}}\int_{e_k}^{e_{k+1}} (\mathbf{f}-\boldsymbol{\tau})_i\mathbf{n}_k\,\mathrm{d}l = 0 \tag{7}$$

where $\delta\mathbf{f}_k = \mathbf{f}_j(x,y) - \mathbf{f}_i(x,y)$ and $\delta\boldsymbol{\tau}_k = \boldsymbol{\tau}_j(x,y) - \boldsymbol{\tau}_i(x,y)$. Depending on the spatial representation of $u(x,y)$, (7) will lead to different schemes.

2.1. *Explicit first-order upwind scheme for the scalar equation*

The simplest option is to represent u and $\boldsymbol{\tau}$ using piecewise constant values assigned to the centroid of the cells, (x_0, y_0), at a given time t, as Figure 2 shows. This is a first-order approximation in space [12].

The cell functions are thus constant, $u(x,y) = u_{i,0}^n$, $\boldsymbol{\tau}(x,y,t) = \boldsymbol{\tau}_{i,0}^n$ and the first integral in (7) can be approximated by the Euler approximation:

$$\frac{\partial}{\partial t}\int_{\Omega_i} u(x,y)\,\mathrm{d}\Omega \cong \frac{u_i^{n+1} - u_i^n}{\Delta t}A_i \tag{8}$$

where superscripts n and $n+1$ represent the solution at times t and $t+\Delta t$, respectively, Δt being the discrete time step and A_i the area of cell Ω_i. On the other hand, assuming piecewise constant values of the variables, the third integral of (7), using (6), vanishes leading to

$$\frac{u_i^{n+1} - u_i^n}{\Delta t}A_i + \sum_{k=1}^{\mathrm{NE}} (\delta(\mathbf{f}-\boldsymbol{\tau})_k\mathbf{n}_k) l_k = 0 \tag{9}$$

Using (2), the linearized advection velocity $(\widetilde{\boldsymbol{\lambda}\mathbf{n}})_k$ can be defined [13] as

$$\widetilde{\lambda}_k = (\widetilde{\boldsymbol{\lambda}\mathbf{n}})_k = \frac{(\mathbf{f}_{j,0} - \mathbf{f}_{i,0})}{(u_{j,0} - u_{i,0})}\mathbf{n}_k \tag{10}$$

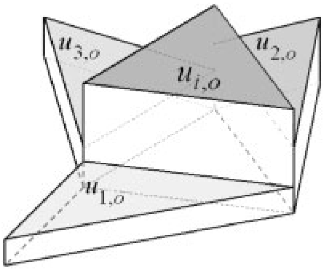

Figure 2. Piecewise constant representation of the variable u.

DOI: 10.1002/fld

Following the upwind philosophy, which discriminates the sense of propagation according to the sign of the advection velocity, the flux difference is split [13] as a sum of waves travelling in and out of a given cell

$$\delta\mathbf{f}_k\mathbf{n}_k = \delta\mathbf{f}_k^-\mathbf{n}_k + \delta\mathbf{f}_k^+\mathbf{n}_k = \widetilde{\lambda}_k^-\,\delta u_k + \widetilde{\lambda}_k^+\,\delta u_k \tag{11}$$

with $\widetilde{\lambda}^{\pm} = (\widetilde{\lambda} \pm |\widetilde{\lambda}|)/2$.

The edge source term $\delta\tau_k = (\delta\boldsymbol{\tau}\mathbf{n})_k$ is also split into in-going and out-going contributions

$$\delta\tau_k = (\delta\tau)_k^- + (\delta\tau)_k^+ \tag{12}$$

where

$$(\delta\tau)_k^{\pm} = \tfrac{1}{2}(1 \pm \operatorname{sgn}(\widetilde{\lambda}_k^{\pm}))(\delta\tau)_k \tag{13}$$

The updating scheme for cell i includes only the in-going contribution of flux and source term to that cell

$$u_i^{n+1} = u_i^n - \sum_{k=1}^{\mathrm{NE}} (\delta(\mathbf{f} - \boldsymbol{\tau})\mathbf{n})_k^- \frac{l_k}{A_i}\Delta t \tag{14}$$

Equation (14) is the later so-called unified discretization. This can also be expressed in a compact form as

$$u_i^{n+1} = u_i^n - \sum_{k=1}^{\mathrm{NE}} v_{i,k}^*\delta u_k, \quad v_{i,k}^* = \frac{\lambda_k^*}{(A_i/l_k)}\Delta t \tag{15}$$

with

$$\lambda_k^* = \widetilde{\lambda}_k^-\theta_k, \quad \theta_k = 1 - \left(\frac{\delta\tau^-}{\widetilde{\lambda}^-\,\delta u}\right)_k = 1 - \left(\frac{\delta\tau^-}{\delta f^-}\right)_k \tag{16}$$

Note that $v_{i,k}^*$ is a dimensionless quantity that plays the role of a local CFL number generalized to cases involving source terms. The coefficient θ_k expresses the discrete ratio of source term to flux differences. They will be key parameters in our discussion of stability conditions. It is worth noting here that definitions analogous to (16) are possible in cases in which the source term is not written as the divergence of a vector field.

2.1.1. *Influence of θ_k on the stability condition.*

In the homogeneous case $\theta_k = 1$. Then, numerical stability for scheme (15) is ensured if [9]

$$-1 \leqslant v_{i,k}^* \leqslant 0 \tag{17}$$

and, at the same time, the following condition on the monotonicity of the solution holds:

$$u^{\min} \leqslant u_{i,0}^{n+1} \leqslant u^{\max} \tag{18}$$

where $u_k^{\max} = \max\{u_{j,0}^n, u_{i,0}^n\}_k$ and $u_k^{\min} = \min\{u_{j,0}^n, u_{i,0}^n\}_k$.

The aim of the following analysis is to find the criterion that preserves (18) in presence of source terms. For that reason, first it is necessary to enforce (17) by requiring that the redefined eigenvalue is negative

$$\widetilde{\lambda}_k^* \leqslant 0 \tag{19}$$

and as, by definition, $\widetilde{\lambda}_k^- < 0$ the new coefficient must be positive:

$$\theta_k \geqslant 0 \tag{20}$$

Given a grid mesh and flow conditions, (17) is a limit on the value of the time step to meet the stability criterion. For the sake of simplicity, assume that all the $u_{j,0}$ values are uniform at the surrounding cells to cell i, but $u_{i,0} > u_{j,0}$. Condition (6) is the key to bound the size of the allowable incoming contributions to a cell [8] as

$$\left|\sum_k \lambda_k^* l_k \delta u_k \Delta t\right| \leqslant \max_k\{|\lambda_k^*| l_k\} \delta u_0 \Delta t = (|\lambda^*| l)_{k_{\max}} \delta u_0 \Delta t \tag{21}$$

where $\delta u_0 = u_{j,0} - u_{i,0} = \delta u_k$. As the updating flux crossing every edge k between cells i and j must be limited by the quantity that ensures that the final state at both cells is included between the initial values $\delta u_0 A_{\min}$, the following is also true:

$$(|\lambda^*| l)_{k_{\max}} \delta u_0 \Delta t \leqslant \delta u_0 A_{\min} \tag{22}$$

where $A_{\min} = \min\{A_i, A_j\}$ and A_i and A_j are the areas of cells i and j, respectively. In a more general case, $\delta u_0 = \max_k\{\delta u_k\}$.

Under these conditions (17) is reformulated as

$$\Delta t = \text{CFL}\, \Delta t_{\max}, \quad \text{CFL} \leqslant 1$$

and $\Delta t_{\max}$ given by the condition expressed in (22). It can also be formulated as a cell time step in terms of the edge-time steps as follows:

$$\Delta t_{\max} = \min\{\Delta t_k\}_{k=1,N\text{edge}}, \quad \Delta t_k = \frac{A_{\min,k}}{|\lambda_k^*| l_k}, \quad \theta_k \geqslant 0 \tag{23}$$

If θ_k is set equal to one in (23), the basic CFL stability condition for the homogeneous case is automatically recovered. Otherwise, (23) states a more general rule. At this point it is worth remarking the relevance of the source term discretization when analysing the stability region defined by (23). If a unified formulation has been used so that in equilibrium

$$(\delta(\mathbf{f} - \boldsymbol{\tau})\mathbf{n})_k^- = 0 \tag{24}$$

This means that $\theta_k = 0$, $\lambda_k^* = 0$ and the numerical scheme becomes unconditionally stable at steady state. Figure 3 represents the stability region of the scheme as a function of θ. The point $\theta = 1$ on the curve corresponds to the homogeneous case (no source term) and $\Delta t(\lambda_k^*)$ is the maximum time step compatible with stability in this case or CFL condition [10]. The rest of the curve corresponds to cases with source terms. In Figure 3 the sign relations among fluxes and source terms are displayed. The dashed zone ($0 \leqslant \theta_k \leqslant 1$) is the set of situations in which it could be possible to use larger time steps than the limit of the homogeneous stability: both fluxes and source terms have

DOI: 10.1002/fld

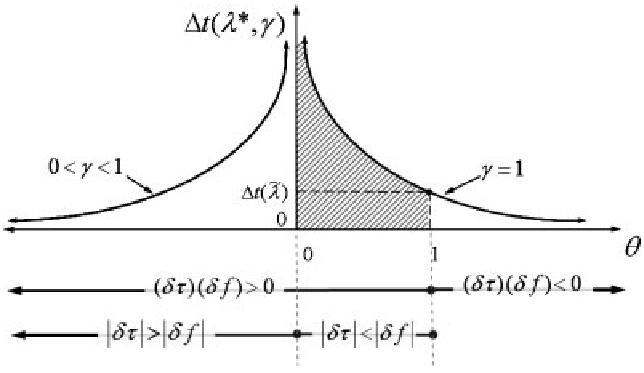
$\Delta t(\lambda^*,\gamma)$
$0<\gamma<1$
$\gamma=1$
$\Delta t(\tilde{\lambda})$
0
0
1
θ
$(\delta\tau)(\delta f)>0$
$(\delta\tau)(\delta f)<0$
$|\delta\tau|>|\delta f|$
$|\delta\tau|<|\delta f|$

Figure 3. Stability region. First-order scheme in the presence of source terms.

the same sign so the net contributions are reduced. In the case $\theta_k > 1$, stability requires a reduction of the time step size over that dictated by the CFL condition, as fluxes and source terms have an opposite sign, so the net contribution increases.

On the other hand, it is possible to find situations where $\theta_k < 0$. In that case the source term contribution dominates over the flux difference, $|\delta\tau^-|_k > |\delta f^-|_k$ and a different line of reasoning must be followed. Depending on the requirements over the solution different strategies can be tackled. If preservation of the sign of the variable is the main objective,

$$\begin{aligned} u_i^{n+1} &\leqslant 0 \quad \text{when } u_i^n, u_{j=1,2,3}^n \leqslant 0 \\ u_i^{n+1} &\geqslant 0 \quad \text{when } u_i^n, u_{j=1,2,3}^n \geqslant 0 \end{aligned} \tag{25}$$

the contributions between cells are also limited by the initial values stored in the cells so that the stability condition is determined also by the initial condition:

$$\Delta t_k = \gamma \frac{A_{\min,k}}{|\lambda_k^*| l_k}, \quad \gamma = \frac{\min\{|u_{i,0}|, |u_{j,0}|, |\delta u_k|\}}{|\delta u_k|}, \quad \theta_k < 0 \tag{26}$$

If u is a gradually varied function, the coefficient γ is 1 and the time step limit in (26) reduces to (23). Otherwise, $0 \leqslant \gamma < 1$ and an actual reduction in the time step is required. In the special case $\gamma = 0$ the local time step Δt_k would be zero according to (26). This is absurd and must be interpreted as condition of no information crossing that cell edge. In practice, a threshold value equal to the machine accuracy is defined for the minimum value of γ before imposing the condition that no information crosses the edge.

In cases in which the source term is not written as the divergence of a vector field, the above discussion is still valid if, at the discrete level, condition (24) can still be formulated. This is the case in the problems addressed in this work.

Finally, it is also possible to re-express (15) as

$$u_i^{n+1} = u_i^n - \sum_{k=1}^{\mathrm{NE}} \tilde{\lambda}_k^- \delta d_k \frac{l_k}{A_i} \Delta t, \quad \delta d_k = \delta u_k - \frac{\delta \tau_k}{\lambda^-} = \delta u_k \theta_k \tag{27}$$

that in steady state reduces to $\delta d_k = 0$. This formulation will be useful when analysing the second order in space approach.

2.2. Explicit second order in space upwind scheme for the scalar equation

The spatial discrete representation of the functions can be improved using information of the neighbour cells. The reconstruction functions can be defined as piecewise linear representations in the cells so that the scheme becomes a second order in space approximation [12]. As the cell representation function must be unique to preserve conservation, the techniques described in this section can only be applied to triangular cells because, as will be shown, the number of edges in which stability conditions are required cannot exceed the number of points used to define the representation function itself [11].

The piecewise linear reconstruction of a scalar variable u, over an element with centroid at (x_0, y_0) is expressed as

$$u_i(x, y) = u_i(x_0, y_0) + \mathbf{r}(x, y)\mathbf{L}_{u,i} = u_{i,0} + \mathbf{r}(x, y)\mathbf{L}_{u,i} \tag{28}$$

where $\mathbf{r}$ is the position vector from the centroid, and $\mathbf{L}$ is the cell slope. The same applies to $\boldsymbol{\tau}$. Different forms to define the cell slope will be described later. Figure 4 shows the position vectors of the middle points of edge k.

As u has a constant slope in this case, the first integral in (7) can be evaluated as follows:

$$\frac{\partial}{\partial t}\int_{\Omega_i} u \, \mathrm{d}\Omega \approx \frac{1}{\Delta t}\left(u_i^{t+\Delta t} - \frac{1}{\mathrm{NE}}\sum_{k=1}^{\mathrm{NE}} (u_{i,0} + \mathbf{r}_{i,e_k}\mathbf{L}_{u,i})\right) = \frac{u_i^{n+1} - u_i^n}{\Delta t} \tag{29}$$

with $u_i^n = u_{i,0}$ and $\mathbf{r}_{i,e_k}$ the position vector of vertex e_k.

For the second integral in (7), $\delta\mathbf{f}$ is evaluated at the mid-edge (I, J) so that, taking advantage of the linear cell distribution:

$$\sum_{k=1}^{\mathrm{NE}} \int_{e_k}^{e_{k+1}} (\boldsymbol{\lambda}\mathbf{n})_k (u_j(x, y) - u_i(x, y)) \, \mathrm{d}l = \sum_{k=1}^{\mathrm{NE}} \widetilde{\lambda}_{JI,k} (u_J - u_I)_k l_k \tag{30}$$

with a linearized advection velocity $\widetilde{\lambda}_{JI,k} = (\widetilde{\boldsymbol{\lambda}\mathbf{n}})_{JI,k}$ defined as

$$\widetilde{\lambda}_{JI,k} = \frac{(\mathbf{f}_J - \mathbf{f}_I)_k}{(u_J - u_I)_k}\mathbf{n}_k \tag{31}$$

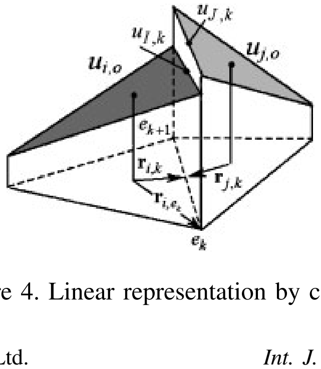

Figure 4. Linear representation by cells.

and

$$\begin{aligned} u_{I,k} &= u_{i,0} + \mathbf{r}_{i,k}\mathbf{L}_{u,i}, \quad u_{J,k} = u_{j,0} + \mathbf{r}_{j,k}\mathbf{L}_{u,j} \\ \mathbf{r}_{i,k} &= \tfrac{1}{2}(\mathbf{r}_{i,e_k} + \mathbf{r}_{i,e_{k+1}}), \quad \mathbf{r}_{j,k} = \tfrac{1}{2}(\mathbf{r}_{j,e_k} + \mathbf{r}_{j,e_{k+1}}) \end{aligned} \tag{32}$$

where $\mathbf{r}_{j,k}$ is the position vector of the mid-point edge from the centroid of cell j. In this case, due to the cell variation, the term $\mathbf{f}$ in the third integral in (7) does not vanish:

$$\sum_{k=1}^{\mathrm{NE}} \int_{e_k}^{e_{k+1}} \mathbf{f}_i \mathbf{n}_k \, \mathrm{d}l = \sum_{k=1}^{\mathrm{NE}} \mathbf{f}_{I,k}\mathbf{n}_k l_k \tag{33}$$

with $\mathbf{f}_{I,k} = (\widetilde{\boldsymbol{\lambda}\mathbf{n}})u_{I,k}$ in the case of a first-order homogeneous flux function. For a more general case, (33) can be rewritten using property (6) as

$$\sum_{k=1}^{\mathrm{NE}} \mathbf{f}_{I,k}\mathbf{n}_k l_k - \mathbf{f}_{i,0}\sum_{k=1}^{\mathrm{NE}} \mathbf{n}_k l_k = \sum_{k=1}^{\mathrm{NE}} (\mathbf{f}_{I,k} - \mathbf{f}_{i,0})\mathbf{n}_k l_k = \sum_{k=1}^{\mathrm{NE}} (u_{I,k} - u_{i,0})\widetilde{\lambda}_{Ii,k} l_k \tag{34}$$

Now $\delta\boldsymbol{\tau}(x, y)$ in (7) is approximated as

$$\sum_{k=1}^{\mathrm{NE}} \int_{e_k}^{e_{k+1}} \delta\boldsymbol{\tau}\mathbf{n}_k \, \mathrm{d}l = \sum_{k=1}^{\mathrm{NE}} (\delta\boldsymbol{\tau})_{JI,k}\mathbf{n}_k l_k \tag{35}$$

with $\delta\boldsymbol{\tau}_{JI,k} = \boldsymbol{\tau}_{J,k} - \boldsymbol{\tau}_{I,k}$. The term $\boldsymbol{\tau}$ in the third integral does not vanish either

$$\sum_{k=1}^{\mathrm{NE}} \int_{e_k}^{e_{k+1}} \boldsymbol{\tau}_i \mathbf{n}_k \, \mathrm{d}l = \sum_{k=1}^{\mathrm{NE}} \boldsymbol{\tau}_{I,k}\mathbf{n}_k l_k \tag{36}$$

but can also be rewritten using (6) as

$$\sum_{k=1}^{\mathrm{NE}} \boldsymbol{\tau}_{I,k}\mathbf{n}_k l_k - \boldsymbol{\tau}_{i,0}\sum_{k=1}^{\mathrm{NE}} \mathbf{n}_k l_k = \sum_{k=1}^{\mathrm{NE}} (\boldsymbol{\tau}_{I,k} - \boldsymbol{\tau}_{i,0})\mathbf{n}_k l_k = \sum_{k=1}^{\mathrm{NE}} \delta\boldsymbol{\tau}_{Ii,k}\mathbf{n}_k l_k \tag{37}$$

leading to the following updating scheme:

$$u_i^{n+1} = u_i^n - \sum_{k=1}^{\mathrm{NE}} (\delta\mathbf{f}\mathbf{n}_k - \delta\boldsymbol{\tau}\mathbf{n}_k)^-_{JI,k} \frac{l_k}{A_i}\Delta t - \sum_{k=1}^{\mathrm{NE}} (\delta\mathbf{f}\mathbf{n}_k - \delta\boldsymbol{\tau}\mathbf{n}_k)_{Ii,k} \frac{l_k}{A_i}\Delta t \tag{38}$$

that defines the second-order upwind finite volume method for scalar equations with source terms and unified discretization.

The updating formula in (38) can also be expressed in a compact form as

$$u_i^{n+1} = u_i^n - \sum_{k=1}^{\mathrm{NE}} v^*_{JI,k}\delta u_{JI,k} - \sum_{k=1}^{\mathrm{NE}} v^*_{Ii,k}\delta u_{Ii,k} \tag{39}$$

This is a second-order formulation that reduces to (15) as a simpler case and where

$$v^*_{JI,k} = \frac{\lambda^*_{JI,k}}{(A_i/l_k)}\Delta t, \quad v^*_{Ii,k} = \frac{\lambda^*_{Ii,k}}{(A_i/l_k)}\Delta t \tag{40}$$

DOI: 10.1002/fld

with the definitions

$$\lambda^*_{JI,k} = \widetilde{\lambda}^{-}_{JI,k}\theta_{JI,k}, \quad \lambda^*_{Ii,k} = \widetilde{\lambda}_{Ii,k}\theta_{Ii,k} \tag{41}$$

that contains information on the relative importance of source terms and flux differences and separates at the same time the upwind contributions through the cell edges from the central contributions due to non-uniform cell representation.

2.2.1. Influence of θ_k on the stability condition. To ensure monotonicity over u, as expressed in (18), when (39) is used both $v^*_{JI,k}$ and $v^*_{Ii,k}$ must be bounded

$$-1 \leqslant v^*_{JI,k} \leqslant 0, \quad -1 \leqslant v^*_{Ii,k} \leqslant 0 \tag{42}$$

and, therefore, the following conditions over the θ coefficients expressing the influence of the source terms, appear

$$\theta_{JI,k} \geqslant 0, \quad \theta_{Ii,k} \geqslant 0 \tag{43}$$

As the cell representation function of u changes in second order, the criterion over the time step for stability is affected and must be revisited. The following condition is necessary to limit the new contributions:

$$|\delta u_{Ii,k} + \delta u_{JI,k}| \leqslant |\delta u_k| \tag{44}$$

with $\delta u_{JI,k} = (u_{J,k} - u_{I,k})$ and $\delta u_{Ii,k} = \mathbf{r}_{i,k}\mathbf{L}_{u,i}$.

For that purpose, the values of the function u at the cell edge, as computed from (32), have to be bounded as follows:

$$u^{\min}_k \leqslant u_{J,k} \leqslant u^{\max}_k, \quad u^{\min}_k \leqslant u_{I,k} \leqslant u^{\max}_k \tag{45}$$

this is controlled by the extra restrictions placed upon the u values by the slope limiters (see Appendix A) and leads, at the same time, to

$$\begin{aligned} |\delta u_{JI,k}| &= |u_{J,k} - u_{I,k}| \leqslant |\delta u_k| \\ |\delta u_{Ii,k}| &= |\mathbf{r}_{i,k}\mathbf{L}_{u,i}| \leqslant |\delta u_k| \end{aligned} \tag{46}$$

The cell contributions in (39) can be expressed as

$$\sum_{k=1}^{\mathrm{NE}} (\lambda^*_{JI,k}\delta u_{JI,k} + \lambda^*_{Ii,k}\delta u_{Ii,k}) l_k \Delta t \tag{47}$$

If $(\delta u_{JI,k}\delta u_{Ii,k}) \geqslant 0$ the modulus of these contributions can be limited by

$$\left| \sum_{k=1}^{\mathrm{NE}} \lambda^*_{k,\sup}(\delta u_{JI,k} + \delta u_{Ii,k}) l_k \right| \Delta t \tag{48}$$

with $\lambda^*_{k,\sup} = \max\{|\lambda^*_{JI,k}|, |\lambda^*_{Ii,k}|, |\lambda^*_k|\}$. Considering (44), (48) in turn is bounded by

$$\left| \sum_{k=1}^{\mathrm{NE}} \lambda^*_{k,\sup}(\delta u_k) l_k \right| \Delta t \tag{49}$$

DOI: 10.1002/fld

If $(\delta u_{JI,k}\delta u_{Ii,k})\leqslant 0$ (47) can be directly limited by (49), and now

$$\left|\sum_{k=1}^{\mathrm{NE}} \lambda^*_{k,\sup}(\delta u_k) l_k\right| \Delta t \leqslant 3 \max_k \{\lambda^*_{k,\sup} l_k\} |\delta u_0| \Delta t = 3\lambda^*_{k,\max} l_k |\delta u_0| \Delta t \tag{50}$$

where $\delta u_0 = \max_k\{\delta u_k\}$. At this point, the cell contributions are limited as in (22) leading to the new stability condition for second-order approach in space that expressed per edges as in (23) is

$$\begin{gathered}\Delta t = \mathrm{CFL}\,\Delta t_{\max}, \quad \mathrm{CFL}\leqslant\tfrac{1}{3}\\ \Delta t_{\max} = \min\{\Delta t_k\}_{k=1,N\mathrm{edge}}, \quad \Delta t_k = \frac{A_{\min,k}}{\lambda^*_{k,\max} l_k}, \quad \theta_{JI,k}\geqslant 0, \quad \theta_{Ii,k}\geqslant 0\end{gathered} \tag{51}$$

This condition guarantees strict positivity and is more restrictive than the stability criterion for the first-order explicit scheme. In the particular case $\theta_{JI,k} = 1$ and $\theta_{Ii,k} = 1$, the homogeneous stability condition is obtained.

Considering steady state as before, the unified discretization requires

$$(\delta \mathbf{fn}_k - \delta \boldsymbol{\tau}\mathbf{n}_k)^-_{JI,k} = 0, \quad (\delta \mathbf{fn}_k - \delta \boldsymbol{\tau}\mathbf{n}_k)_{Ii,k} = 0 \tag{52}$$

for all k, which is only feasible if second order is reduced to first order. This can be more clearly seen using the notation of (27) in the second-order formulation (39)

$$u_i^{n+1} = u_i^n - \sum_{k=1}^{\mathrm{NE}} \widetilde{\widetilde{\lambda}}^-_{JI,k} \delta d_{JI,k} \frac{l_k}{A_i}\Delta t - \sum_{k=1}^{\mathrm{NE}} \widetilde{\widetilde{\lambda}}_{Ii,k} \delta d_{Ii,k} \frac{l_k}{A_i}\Delta t \tag{53}$$

with

$$\delta d_{JI,k} = \delta u_{JI,k}\theta_{JI,k}, \quad \delta d_{Ii,k} = \delta u_{Ii,k}\theta_{Ii,k} \tag{54}$$

It is not difficult to see that the stability condition formulated as in (51) requires a limitation over the variable d as follows:

$$|\delta d_{Ii,k} + \delta d_{JI,k}| \leqslant |\delta d_k| \tag{55}$$

In the second-order representation, a unique interpolation function must be defined for d in each cell in order to preserve the conservative character of the numerical scheme:

$$d_i(x, y) = d_{i,0} + \mathbf{r}(x, y)\mathbf{L}_{d,i} \tag{56}$$

The construction of (56) requires expressing a modified source term $\hat{\tau}$ also as a piecewise linear function:

$$\hat{\tau}_i(x, y) = \hat{\tau}_{i,0} + \mathbf{r}(x, y)\mathbf{L}_{\hat{\tau},i}, \quad \hat{\tau}_{i,0} = \left(\frac{\tau}{\widetilde{\widetilde{\lambda}}}\right)_{i,k} \tag{57}$$

so that $d_{i,0} = u_{i,0} - \hat{\tau}_{i,0}$. For the sake of simplicity, in this work the source terms will be expressed assuming first-order approximation, setting $\mathbf{L}_{\hat{\tau},i} = \mathbf{0}$, so in each cell $\mathbf{L}_{u,i} = \mathbf{L}_{d,i}$.

With this modification the quantities $\delta d_{JI,k} = (d_{J,k} - d_{I,k})$ and $\delta d_{Ii,k} = \mathbf{r}_{i,k}\mathbf{L}_{d,i}$ can be correctly defined and the function d at the cell edges can be bounded to satisfy (55) as follows:

$$d_k^{\min} \leqslant d_{J,k} \leqslant d_k^{\max}, \quad d_k^{\min} \leqslant d_{I,k} \leqslant d_k^{\max} \tag{58}$$

DOI: 10.1002/fld

where $d_k^{\max}=\max\{d_{j,0}^n,d_{i,0}^n\}_k$ and $d_k^{\min}=\min\{d_{j,0}^n,d_{i,0}^n\}_k$, and it is possible to achieve equilibrium in steady-state cases by requiring

$$\delta d_{JI,k}=0,\quad \delta d_{Ii,k}=0 \tag{59}$$

for all k, over the single variable d. This is equivalent to (52) but much simpler and efficient and also implies that the second-order scheme reduces to first order automatically ($\mathbf{L}_d=\mathbf{0}$).

The new formulation in terms of the d variable is useful and efficient when dealing with the second-order scheme in presence of source terms. However, it must be stressed that the behaviour of the conserved variable u must be always controlled so that, for all k, the following is never allowed:

$$|\delta u_{JI,k}|>|\delta u_k|,\quad u_{I,k}=d_{I,k}+\hat{\tau}_{I,k},\quad u_{J,k}=d_{J,k}+\hat{\tau}_{J,k} \tag{60}$$

If (60) happens, the numerical scheme must be reduced to first order, otherwise condition (58) as applied to variable u would be violated. At the same time, it may happen that $\theta_k<0$. In this case the source term dominates over the numerical flux. Considering that condition (60) must be avoided and that the definition of γ becomes of relevance in these cases, the numerical scheme must be also reduced to first order when $0\leqslant\gamma<1$.

2.3. *Explicit second order in time and space scheme for the scalar equation*

The second order in space approach in (39) or (53) may produce stable but oscillatory solutions even in simple homogeneous cases [14]. The numerical solution can be improved by extending the numerical scheme to second order also in time. For that reason the MUSCL-Hancock, or MHM, approach was proposed [15, 16]. This method can be formulated as based on two steps. In the first step the solution must be reconstructed using the $\mathbf{L}_i$ gradient vectors (as in (28) or (32)) and then intermediate values are re-calculated at a half time step at cell edges as

$$u_{I,k}^{n+1/2}=u_{I,k}^{n}-\sum_{k=1}^{\mathrm{NE}}(\delta\mathbf{fn}_k-\delta\boldsymbol{\tau}\mathbf{n}_k)_{Ii,k}^{n}\frac{l_k}{A_i}\frac{\Delta t}{2} \tag{61}$$

which is equivalent to redefining the interpolation function in each cell. The intermediate values (61) must ensure monotonicity and positivity as previously defined and must be reduced to first order when $\theta_k<0$, as stated in the preceding section. The updated variable is constructed as

$$u_i^{n+1}=u_i^n-\sum_{k=1}^{\mathrm{NE}}(\delta\mathbf{fn}_k-\delta\boldsymbol{\tau}\mathbf{n}_k)_{JI,k}^{n+1/2,-}\frac{l_k}{A_i}\Delta t-\sum_{k=1}^{\mathrm{NE}}(\delta\mathbf{fn}_k-\delta\boldsymbol{\tau}\mathbf{n}_k)_{Ii,k}^{n+1/2}\frac{l_k}{A_i}\Delta t \tag{62}$$

where $\delta\mathbf{f}_{JI,k}^{n+1/2}=\mathbf{f}_{J,k}^{n+1/2}-\mathbf{f}_{I,k}^{n+1/2}$ and $\delta\mathbf{f}_{Ii,k}^{n+1/2}=\mathbf{f}_{I,k}^{n+1/2}-\mathbf{f}_{i,k}^{n}$, with an upwind part (first term) and a central part (second term) that reduces again the time step to (51).

2.4. *Application to the linear scalar equation*

This first test case, taken from Batten *et al.* [17], is run using uniform triangular elements generated by dividing square elements along the top-left to bottom-right diagonal Figure 5(a) that will be referred as type M1 and by dividing square elements along the top-left to bottom-right diagonal and the bottom-left to top-right (Figure 5(b)) that will be referred as type M2. The square domain

DOI: 10.1002/fld

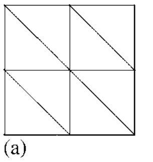

Figure 5. M1 (left) and M2 (right) triangular grids.

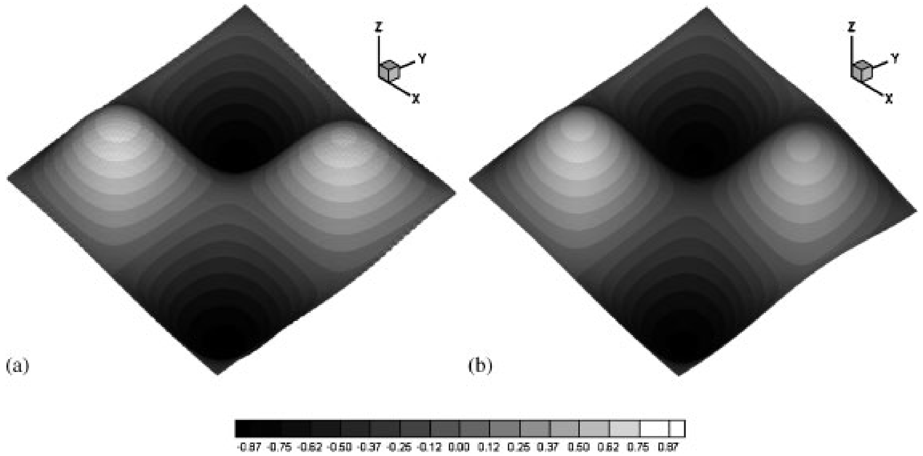

Figure 6. Numerical results in first order using M1 (a) and M2 (b) with $l = 0.0125$.

Table I. Order of convergence in M1.

	$l = 0.025$	$l = 0.0125$	$l = 0.00625$	$l = 0.003125$
$\log(L_1)/\log(l)$, first order	0.66	0.69	0.72	0.75
$\log(L_1)/\log(l)$, MLG	1.26	1.33	1.39	1.45
$\log(L_1)/\log(l)$, MLG-Wierse	1.28	1.34	1.39	1.41

is defined by $0 \leqslant x \leqslant 1$ and $0 \leqslant y \leqslant 1$. A constant diagonal velocity $\boldsymbol{\lambda} = (1, 1)^{\mathrm{T}}$ is used to advect an initial condition defined by the double sin function,

$$u = \sin(2\pi x)\sin(2\pi y) \tag{63}$$

and a periodic boundary condition is imposed to ensure that the initial solution and the solution after every second are equal, as the periodic wave returns to its initial position. A constant value of $\mathrm{CFL} = \frac{1}{3}$ is used in all cases. Figure 6(a) and (b) shows the result for first-order approach in M1 and M2, respectively, with $l = 0.0125$, showing that M2 meshes provide less accurate results.

The order of accuracy was estimated from the rate of convergence of L_1 error for the sine function when the cell edge l varies from 0.025 to 0.003125 on both M1 and M2 meshes. Table I shows the results when using M1 meshes and the second order in space and time approach. As defined in

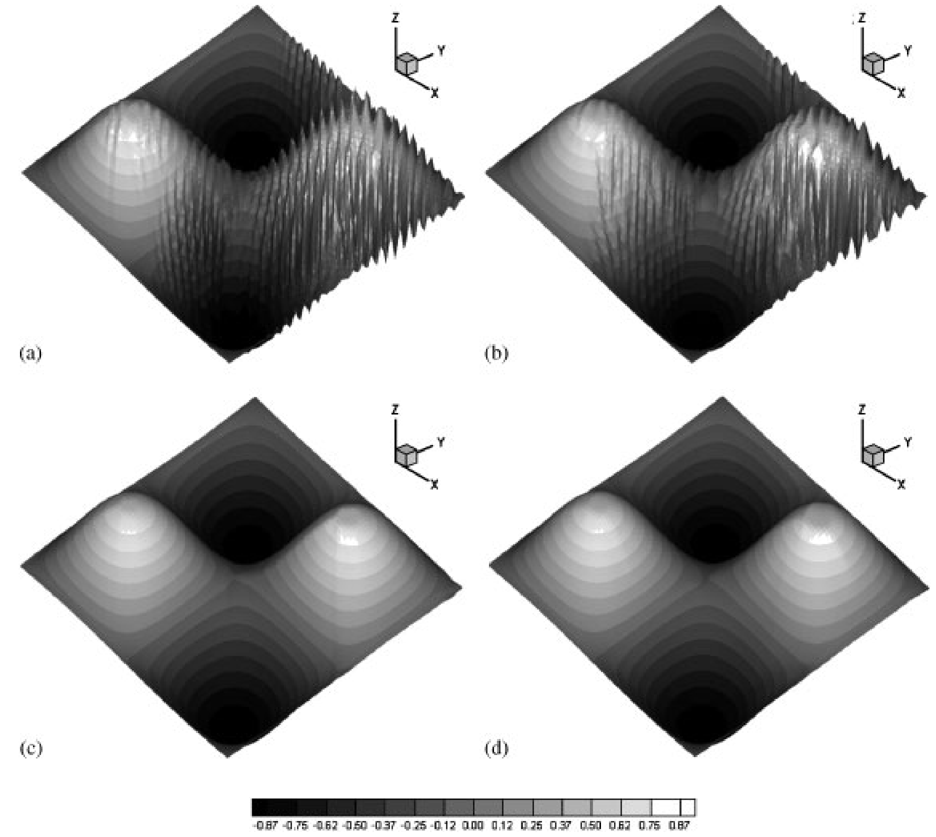

Figure 7. Second-order schemes on M1 mesh with $l = 0.0125$: (a) MLG interpolation technique for second order in space; (b) MLG-Wierse interpolation techniques for second order in space; (c) MLG interpolation technique for second order in space and time; and (d) MLG-Wierse interpolation technique for second order in space and time.

Appendix A, both MLG and MLG-Wierse techniques provide similar order of accuracy. Figure 7(a) and (b) displays the results when using the MLG and the MLG-Wierse interpolation techniques, respectively, for second order in space approach showing how oscillations in the solution appear, as expected. When second order in time and space is used the oscillations are eliminated for both MLG and MLG-Wierse interpolation techniques as Figure 7(c) and (d) displays, respectively.

Table II shows the results when using M2 type mesh comparing the first order and the second order in space and time approaches. Both MLG and MLG-Wierse techniques provide similar order of accuracy, being the MLG-Wierse slightly bigger, which decreases as l decreases.

Figure 8(a) and (b) displays the results when using the MLG and the MLG-Wierse interpolation techniques, respectively, for second order in space approach showing how oscillations in the solution appear. When second order in time and space is applied the MLG-Wierse technique proves to eliminate these oscillations efficiently as Figure 8(d) displays, but when using the MLG limiter spurious oscillations remain as Figure 8(c) displays.

DOI: 10.1002/fld

Table II. Order of convergence in M2.

	$l=0.025$	$l=0.0125$	$l=0.00625$	$l=0.003125$
$\log(L_1)/\log(l)$, first order	0.40	0.78	0.80	0.71
$\log(L_1)/\log(l)$, MLG	1.01	0.99	0.93	0.85
$\log(L_1)/\log(l)$, MLG-Wierse	1.04	1.03	1.03	0.90

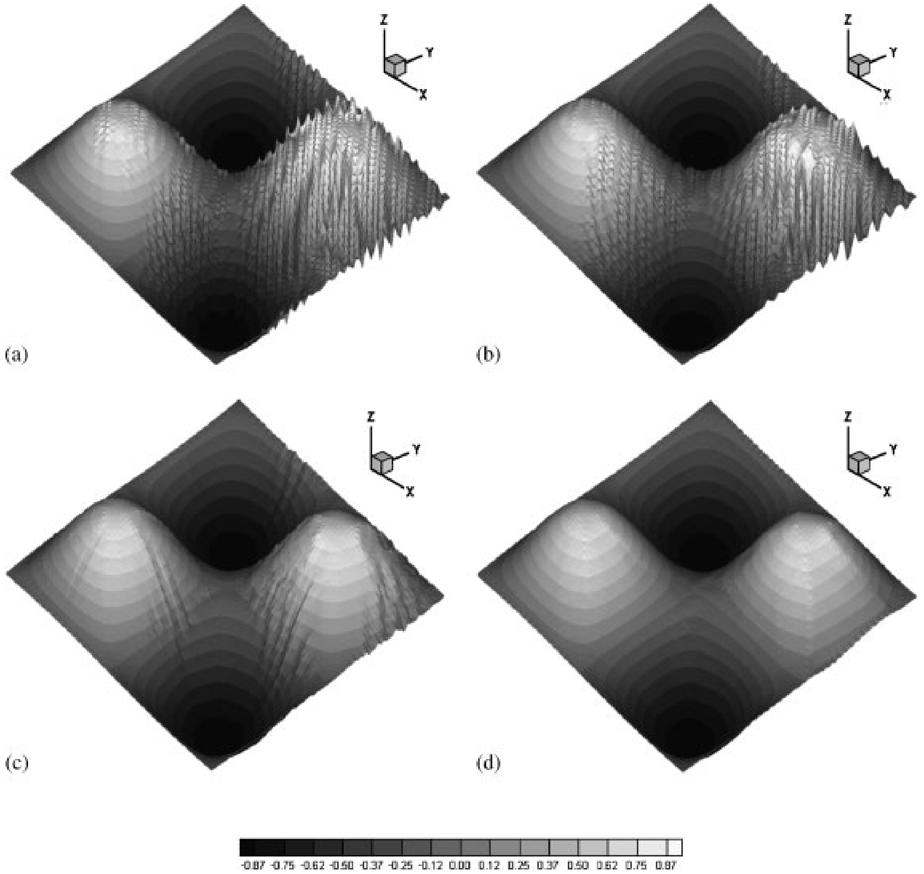

Figure 8. M2 mesh with $l=0.0125$: (a) MLG interpolation technique for second order in space; (b) MLG-Wierse interpolation techniques for second order in space; (c) MLG interpolation technique for second order in space and time; and (d) MLG-Wierse interpolation technique for second order in space and time.

2.5. Application to the Burgers' equation with source terms

In this test case the inviscid two-dimensional Burgers' equation has been modified by adding a source term function

$$\frac{\partial u}{\partial t}+\frac{\partial}{\partial x}\left(\frac{1}{2}u^2\right)+\frac{\partial}{\partial y}\left(\frac{1}{2}u^2\right)=-u\left(\frac{\partial z}{\partial x}+\frac{\partial z}{\partial y}\right) \tag{64}$$

DOI: 10.1002/fld

where $z=z(x,y)$. The function z in our test is defined by

$$z(x,y)=\sin^2\left((x-y)\frac{\pi}{4}\right) \tag{65}$$

and the boundary conditions are given by

$$u(x,y,t)=\cos^2\left((x-y)\frac{\pi}{4}\right) \tag{66}$$

so that in steady state the solution is

$$z(x,y)+u(x,y)=d=0 \tag{67}$$

The simulation is performed in the squared domain M2 with $l=0.0125$ defined in Section 1.4, from the initial condition

$$u(x,y,t=0)=1 \tag{68}$$

From the numerical point of view, the linearized local advection velocity is

$$\widetilde{\lambda}_k=\widetilde{\boldsymbol{\lambda}}_k\mathbf{n}_k=(\widetilde{u},\widetilde{u})_k\mathbf{n}_k=\widetilde{u}_k(n_x+n_y)_k \tag{69}$$

with $\widetilde{u}_k=\frac{1}{2}(u_{i,0}+u_{j,0})_k$. Even though an analytical $\boldsymbol{\tau}$ cannot be identified, in order to ensure a correct discrete balance in steady state, the source terms are discretized as follows:

$$\delta\widetilde{\tau}_k=\delta\boldsymbol{\tau}\mathbf{n}_k=-\widetilde{u}_k\delta z(n_x+n_y) \tag{70}$$

so that $\widetilde{\lambda}_k^*$ and $\hat{\tau}_i$ are in this case

$$\widetilde{\lambda}_k^*=\widetilde{\lambda}_k^-\left(1+\frac{\delta z}{\delta u}\right)_k,\quad \hat{\tau}_{i,0}=z_{i,0} \tag{71}$$

and variable d can be constructed as

$$d_{i,0}=u_{i,0}+z_{i,0} \tag{72}$$

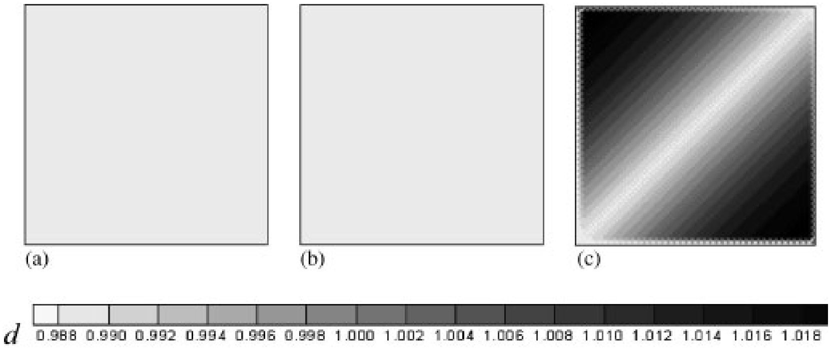

Figure 9. Contour plot of the numerical solution: (a) for first order; (b) second order interpolating over d; and (c) second order interpolating over u.

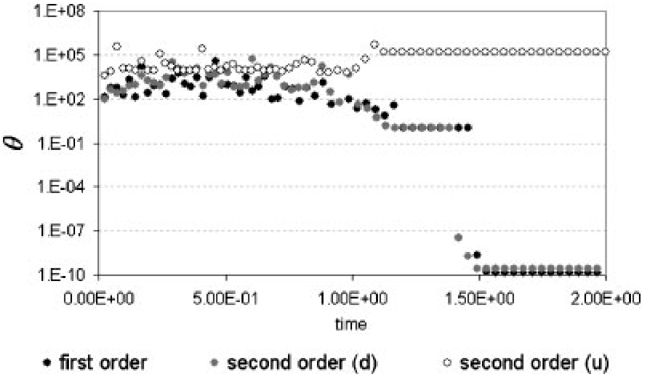

Figure 10. Burger's equation test case. Time evolution of the maximum value of θ using first order, second order over d and second order over u.

Figure 9(a) shows a contour plot of the steady-state solution using first-order approximation and Figure 9(b) shows the steady-state solution when using second-order approximation (MLG algorithm) over the variable d at $t = 3$. In both cases the L_1 error is nil. When second order is applied and the interpolated variable is u, numerical perturbations appear in the solution and the steady-state solution cannot be reached, Figure 9(c).

Figure 10 shows the evolution of the maximum value of θ_k in each time step for first-order approximation, second-order approximation over d and second-order approximation over u, showing that no convergence can be reached in the latter case.

3. SYSTEMS OF CONSERVATION LAWS WITH SOURCE TERMS

The numerical methods are extended in this chapter to solve hyperbolic non-linear systems of equations with source terms, of the form:

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial \mathbf{F}(\mathbf{U})}{\partial x} + \frac{\partial \mathbf{G}(\mathbf{U})}{\partial y} = \mathbf{S}(\mathbf{U}, x, y) \tag{73}$$

It will be first assumed that the source term $\mathbf{S}$ can be expressed as

$$\mathbf{S}(\mathbf{U}) = \frac{\partial \mathbf{S}_1}{\partial x} + \frac{\partial \mathbf{S}_2}{\partial y} \tag{74}$$

so that, calling $\mathbf{E} = (\mathbf{F}, \mathbf{G})^{\mathrm{T}}$ and $\mathbf{T} = (\mathbf{S}_1, \mathbf{S}_2)^{\mathrm{T}}$, (73) becomes

$$\frac{\partial \mathbf{U}}{\partial t} + \nabla(\mathbf{E}(\mathbf{U}) - \mathbf{T}(\mathbf{U})) = 0 \tag{75}$$

The mathematical properties of the hyperbolic system of equations include the existence of a Jacobian matrix, $\mathbf{J_n}$, of the normal flux $(\mathbf{E} \cdot \mathbf{n})$ defined as

$$\mathbf{J_n} = \frac{\partial(\mathbf{E} \cdot \mathbf{n})}{\partial \mathbf{U}} = \frac{\partial(\mathbf{F})}{\partial \mathbf{U}} n_x + \frac{\partial(\mathbf{G})}{\partial \mathbf{U}} n_y \tag{76}$$

DOI: 10.1002/fld

From its eigenvectors, two matrices $\mathbf{P}$ and $\mathbf{P}^{-1}$ can be constructed with the property that they diagonalize the Jacobian $\mathbf{J_n}$,

$$\mathbf{J_n} = \mathbf{P\Lambda P}^{-1} \tag{77}$$

where $\mathbf{\Lambda}$ is a diagonal matrix with eigenvalues in the main diagonal.

The equivalent to (5) for the system is

$$\frac{\partial}{\partial t}\int_{\Omega_i} \mathbf{U}(x,y)\,\mathrm{d}\Omega + \sum_{k=1}^{\mathrm{NE}} \int_{e_k}^{e_{k+1}} (\mathbf{E}_j - \mathbf{T}_j)\mathbf{n}_k\,\mathrm{d}l = 0 \tag{78}$$

where j subindex labels the surrounding cells to cell i, as in the scalar case. Adding and subtracting both the contour integrals of $\mathbf{E}_i(x,y)$ and $\mathbf{T}_i(x,y)$ on the left- and right-hand side:

$$\frac{\partial}{\partial t}\int_{\Omega_i} \mathbf{U}(x,y)\,\mathrm{d}\Omega + \sum_{k=1}^{\mathrm{NE}} \int_{e_k}^{e_{k+1}} (\delta\mathbf{E} - \delta\mathbf{T})_k\mathbf{n}_k\,\mathrm{d}l + \sum_{k=1}^{\mathrm{NE}} \int_{e_k}^{e_{k+1}} (\mathbf{E}_i - \mathbf{T}_i)\mathbf{n}_k\,\mathrm{d}l = 0 \tag{79}$$

3.1. Explicit first-order upwind scheme

In first order the vector quantities $\mathbf{U}$, $\mathbf{E}$ and $\mathbf{T}$ are uniform per cell. In particular, the first integral in (79) can be approximated by

$$\frac{\partial}{\partial t}\int_{\Omega} \mathbf{U}(x,y)\,\mathrm{d}\Omega = \frac{\partial \mathbf{U}_i}{\partial t}A_i \cong \frac{\mathbf{U}_i^{n+1} - \mathbf{U}_i^n}{\Delta t}A_i \tag{80}$$

where $\mathbf{U}_i^n = \mathbf{U}_{i,0}$. In the second integral of (79)

$$\sum_{k=1}^{\mathrm{NE}} \int_{e_k}^{e_{k+1}} \delta\mathbf{E}(x,y)_k\mathbf{n}_k\,\mathrm{d}l = \sum_{k=1}^{\mathrm{NE}} (\delta\mathbf{E}_k\mathbf{n}_k l_k)^n \tag{81}$$

with $\delta\mathbf{E}_k = \mathbf{E}(\mathbf{U}_{j,0}) - \mathbf{E}(\mathbf{U}_{i,0}) = \mathbf{E}_j - \mathbf{E}_i$. Due to the non-linear character of the flux $\mathbf{E}$, the definition of an approximated flux Jacobian, $\widetilde{\mathbf{J}}_{\mathbf{n},k}$ [13] allows for a local linearization and is exploited here.

Matrices $\widetilde{\mathbf{P}}^{-1}$and $\widetilde{\mathbf{P}}$ can be built so that they diagonalize the approximate Jacobian matrix $\widetilde{\mathbf{J}}_{\mathbf{n},k}$

$$\widetilde{\mathbf{J}}_{\mathbf{n},k} = (\widetilde{\mathbf{P}}\widetilde{\mathbf{\Lambda}}\widetilde{\mathbf{P}}^{-1})_k, \quad \widetilde{\mathbf{P}}_k = [\widetilde{\mathbf{e}}^1, \ldots, \widetilde{\mathbf{e}}^{N\lambda}]_k \tag{82}$$

and $\widetilde{\mathbf{\Lambda}}$ is the diagonal eigenvalues matrix. From the approximate Jacobian [13]

$$\widetilde{\mathbf{J}}_{\mathbf{n},k}\widetilde{\mathbf{e}}_k^m = (\widetilde{\lambda}\widetilde{\mathbf{e}})_k^m, \quad m = 1, \ldots, N\lambda \tag{83}$$

where $N\lambda$ is the number of eigenvalues, $\widetilde{\lambda}^m$. The problem is reduced to a one-dimensional Riemann problem projected onto the direction $\mathbf{n}$ at each cell edge [11]. Following a flux difference procedure, the difference in vector $\mathbf{U}$ across the grid edge is projected onto the matrix eigenvectors basis

$$\delta\mathbf{U}_k = \mathbf{U}_{j,0} - \mathbf{U}_{i,0} = \sum_{m=1}^{N\lambda} (\alpha\widetilde{\mathbf{e}})_k^m \tag{84}$$

with $\alpha_k^m = \alpha^m(\mathbf{U}_{j,0}, \mathbf{U}_{i,0})$. The contributions in (81) are written as

$$\sum_{k=1}^{\mathrm{NE}} \widetilde{\mathbf{J}}_{\mathbf{n},k}\delta\mathbf{U}_k l_k = \sum_{k=1}^{\mathrm{NE}} \sum_{m=1}^{N\lambda} (\widetilde{\lambda}^m \alpha^m \widetilde{\mathbf{e}}^m)_k l_k \tag{85}$$

DOI: 10.1002/fld

Since all the variables defined at the cell are uniform, the term $\delta\mathbf{T}$ in the second integral in (79) is approximated by

$$\sum_{k=1}^{\mathrm{NE}} \delta\mathbf{T}_k \mathbf{n}_k l_k \tag{86}$$

with $\delta\mathbf{T}_k = \mathbf{T}_{j,0} - \mathbf{T}_{i,0}$. The normal source difference $(\delta\mathbf{Tn})_k$ can also be expressed in function of the eigenvalues and eigenvectors of $\widetilde{\mathbf{J}}_{\mathbf{n},k}$, using the approximate matrix $\widetilde{\mathbf{P}}_k$ [13] in order to reach a unified formulation:

$$\delta\mathbf{T}_k \mathbf{n}_k = \sum_{m=1}^{N\lambda} (\beta\widetilde{\mathbf{e}})_k^m l_k, \quad \boldsymbol{\beta}_k = \widetilde{\mathbf{P}}^{-1}(\delta\mathbf{Tn})_k \tag{87}$$

with $\boldsymbol{\beta}_k = [\beta^1, \ldots, \beta^m]_k^{\mathrm{T}}$. In order to discriminate the sense of advection linked to the sign of the different eigenvalues, two matrices $\widetilde{\boldsymbol{\Lambda}}^{\pm}$ are defined:

$$\widetilde{\boldsymbol{\Lambda}}^{\pm} = (\widetilde{\boldsymbol{\Lambda}} \pm |\widetilde{\boldsymbol{\Lambda}}|)/2 \tag{88}$$

The flux difference across each edge k is split into contributions directed in the sense of the normal vector (out going or positive waves) and contributions directed against the sense of the normal vector (in going or negative waves). Note that this is always relative to the chosen normal direction and that, as the normal vector is defined pointing outward to a cell, the contributions exchange their character as we move from one cell to the neighbour cell:

$$\delta(\mathbf{E}\cdot\mathbf{n})_k = \widetilde{\mathbf{P}}_k \widetilde{\boldsymbol{\Lambda}}^{-} \widetilde{\mathbf{P}}_k^{-1} \delta\mathbf{U}_k + \widetilde{\mathbf{P}}_k \widetilde{\boldsymbol{\Lambda}}^{+} \widetilde{\mathbf{P}}_k^{-1} \delta\mathbf{U}_k \tag{89}$$

For the updating algorithm, as defined for a given cell i, only the in-going contributions generated at the edges are of interest, as in the scalar case. The contour integral of the numerical normal flux is equivalent to the sum of the in-going waves:

$$\sum_{k=1}^{\mathrm{NE}} \widetilde{\mathbf{J}}_{\mathbf{n},k} \delta\mathbf{U}_k l_k \cong \sum_{k=1}^{\mathrm{NE}} \sum_{m=1}^{N\lambda} (\widetilde{\lambda}^{-m} \alpha^m \widetilde{\mathbf{e}}^m)_k l_k \tag{90}$$

where $\lambda^{-} = \frac{1}{2}(\lambda - |\lambda|)$.

In order to enforce equilibrium in steady-state cases the normal source difference $(\delta\mathbf{Tn})_k$ can also be split in two kinds of waves:

$$\delta\mathbf{T}_k \mathbf{n}_k = \underbrace{(\delta\mathbf{Tn})_k^{-}}_{\text{in going}} + \underbrace{(\delta\mathbf{Tn})_k^{+}}_{\text{out going}} \tag{91}$$

with the same philosophy as before and where

$$\delta\mathbf{T}_k^{-} \mathbf{n}_k = \sum_{m=1}^{N\lambda} (\beta\widetilde{\mathbf{e}})_k^{m-} l_k \tag{92}$$

with $\beta^{m-} = \frac{1}{2}(1 - \mathrm{sign}(\widetilde{\lambda}^m))\beta^m$. The third integral on the left-hand side of (79), assuming piecewise constant values per cell, vanishes:

$$\sum_{k=1}^{\mathrm{NE}} \int_{e_k}^{e_{k+1}} (\mathbf{E}_{i,0} - \mathbf{T}_{i,0})\mathbf{n}_k \,\mathrm{d}l = (\mathbf{E}_{i,0} - \mathbf{T}_{i,0}) \sum_{k=1}^{\mathrm{NE}} \mathbf{n}_k l_k = 0 \tag{93}$$

DOI: 10.1002/fld

and the first-order upwind scheme gets the form:

$$\mathbf{U}_i^{n+1} = \mathbf{U}_i^n - \sum_{k=1}^{\mathrm{NE}} \sum_{m=1}^{N\lambda} ((\widetilde{\lambda}^{-}\alpha - \beta^{-})\widetilde{\mathbf{e}})_k^m l_k \frac{\Delta t}{A_i} \tag{94}$$

This is a compact form with the focus on the waves generated at the cell edges, made of both the normal flux difference and the normal source term, and governed by the sign of the eigenvalues of the normal flux Jacobian. This formulation is closely related to the best-known numerical flux formulation for finite volume schemes but is more convenient for our purposes in the present work.

As was done in the scalar case, the numerical scheme in (94) can be rewritten as

$$\mathbf{U}_i^{n+1} = \mathbf{U}_i^n - \sum_{k=1}^{\mathrm{NE}} \sum_{m=1}^{N\lambda} (v^* \delta \mathbf{U})_k^m, \quad v_k^{m,*} = \frac{\lambda_k^{m,*}}{(A_i/l_k)} \Delta t \tag{95}$$

with

$$\widetilde{\lambda}_k^{m,*} = \widetilde{\lambda}_k^{m,-} \theta_k^m, \quad \theta_k^m = 1 - \left(\frac{\beta^-}{\alpha \widetilde{\lambda}^-} \right)_k^m \tag{96}$$

where the ratio $\beta/(\alpha\widetilde{\lambda})$ expresses the influence of the source terms over that of the flux differences.

In absence of source terms, the numerical scheme (94) or (95) is stable provided that

$$-1 \leqslant v_{i,k}^m \leqslant 0, \quad v_k^m = \frac{\widetilde{\lambda}_k^m}{(A_i/l_k)} \Delta t, \quad m = 1, \ldots, N\lambda \tag{97}$$

and the following condition over the conserved variables applies:

$$U_{s,k}^{\min} \leqslant U_{s,i}^{n+1} \leqslant U_{s,k}^{\max}, \quad \mathbf{U} = (U_1, \ldots, U_s, \ldots, U_{nc})^{\mathrm{T}} \tag{98}$$

where $U_{s,k}^{\max} = \max\{U_{s,i,0}, U_{s,j,0}\}_k$ and $U_{s,k}^{\min} = \min\{U_{s,i,0}, U_{s,j,0}\}_k$.

Monotonicity in the conserved variables in presence of source terms requires that for all m

$$-1 \leqslant v_{i,k}^{m,*} \leqslant 0 \tag{99}$$

which means that

$$\lambda_{i,k}^{m,*} \leqslant 0, \quad \theta_{i,k}^m \geqslant 0 \tag{100}$$

Under these assumptions the numerical stability when $\theta_{i,k}^m \geqslant 0$ for all m, is provided by the intersection of the stability regions defined for each celerity $\widetilde{\lambda}_k^{m,*}$

$$\begin{aligned} &\Delta t = \mathrm{CFL}\, \Delta t_{\max}, \quad \mathrm{CFL} \leqslant 1 \\ &\Delta t_{\max} = \min\{\Delta t_k\}_{k=1,N\mathrm{edge}}, \quad \Delta t_k = \frac{A_{\min,k}}{\max_m \{|\lambda_k^{m,*}|\} l_k}, \quad \theta_{i,k}^m \geqslant 0, \quad m = 1, \ldots, N\lambda \end{aligned} \tag{101}$$

In the particular case of $\theta_{i,k}^m = 1$ for all m, (101) expresses the stability condition without source terms (CFL condition).

DOI: 10.1002/fld

Equilibrium in steady-state cases is ensured if the discretization of the source term has been constructed enforcing

$$(\widetilde{\lambda}^{-}\alpha - \beta^{-})_k^m = 0, \quad m = 1, \ldots, N\lambda \tag{102}$$

which is equivalent to $\lambda_k^{m,*} = 0$, leading to an unconditionally stable scheme in this particular case.

To extend the formulation in (28) from scalar to systems, Equation (95) can also be expressed

$$\mathbf{U}_i^{n+1} = \mathbf{U}_i^n - \sum_{k=1}^{\mathrm{NE}} \sum_{m=1}^{N\lambda} (\widetilde{\lambda}^{-}\delta\mathbf{D})_k^m l_k \frac{\Delta t}{A_i}, \quad \delta\mathbf{D}_k^m = \theta_k^m \delta\mathbf{U}_k^m \tag{103}$$

The stability region for (103) with source terms is enlarged if $0 \leqslant \theta_k^m \leqslant 1$ for all m, as in this particular case $|\delta\mathbf{D}_k^m| \leqslant |\delta\mathbf{U}_k^m|$. Equilibrium is achieved when all $\delta\mathbf{D}_k^m = 0$, condition that can be automatically derived from (103).

On the other hand, as seen in the scalar case, when $\theta_k^m < 0$ the source terms dominate over the flux differences. If (101) is still desired, the definition of γ has to be based on the specific necessities of the physical problem. A direct extension of (26) is not possible, as it is not feasible to define a γ_k^m coefficient if $\theta_k^m < 0$ for all m, since there is no correspondence between the m-waves and the s-variables. In the special case trying to preserve the sign over the solution in the s component, expressed as

$$\begin{aligned} U_{s,i}^{n+1} &\leqslant 0, \quad U_{s,i}^n, U_{s,j=1,2,3}^n \leqslant 0 \\ U_{s,i}^{n+1} &\geqslant 0, \quad U_{s,i}^n, U_{s,j=1,2,3}^n \geqslant 0 \end{aligned} \tag{104}$$

the time step in the stability region must be computed following:

$$\Delta t_k = \gamma \frac{A_{\min,k}}{\max_m\{|\lambda_k^{*,m}|\} l_k}, \quad \gamma = \frac{\min_{i,j}\{|U_{s,i}|, |U_{s,j}|, |\delta U_s|\}}{|\delta U_s|} \tag{105}$$

where $0 \leqslant \gamma \leqslant 1$. If the conserved variable is gradually varied, the coefficient γ is 1 and the time step in (105) reduces to (101). In the particular case $\gamma = 0$, no flux information can cross the edge.

3.2. Second order in space upwind scheme for systems

The spatial accuracy of the scheme can be increased by using piecewise linear instead of piecewise constant representations of the different conserved variables at the cells. As the linear reconstruction is conservative, the first integral in (83) can still be approximated by

$$\frac{\partial}{\partial t} \int_{\Omega} \mathbf{U}(x, y)\, \mathrm{d}\Omega = \frac{\partial \mathbf{U}_i}{\partial t} A_i \cong \frac{\mathbf{U}_i^{n+1} - \mathbf{U}_i^n}{\Delta t} A_i \tag{106}$$

where $\mathbf{U}_i^n = \mathbf{U}_{i,0}$, the value at the cell centroid. The flux difference contour integral can be expressed in terms of the jump of the variable at the cell edge:

$$\sum_{k=1}^{\mathrm{NE}} \int_{e_k}^{e_{k+1}} \delta\mathbf{E}(x, y)_k \mathbf{n}_k \, \mathrm{d}l = \sum_{k=1}^{\mathrm{NE}} \delta\mathbf{E}_{JI,k} \mathbf{n}_k l_k \tag{107}$$

DOI: 10.1002/fld

where $\delta\mathbf{E}_{JI,k} = \mathbf{E}(\mathbf{U}_{J,k}) - \mathbf{E}(\mathbf{U}_{I,k}) = \mathbf{E}_{J,k} - \mathbf{E}_{I,k}$, with the same interpretation as in the scalar case (33). Following (91):

$$\sum_{k=1}^{\mathrm{NE}} \delta\mathbf{E}(x, y)_{JI,k}\mathbf{n}_k l_k \cong \sum_{k=1}^{\mathrm{NE}} \sum_{m=1}^{N\lambda} (\widetilde{\lambda}^{-m} \alpha^m \widetilde{\mathbf{e}}^m)_{JI,k} l_k \tag{108}$$

In the third integral on the left-hand side of (83):

$$\sum_{k=1}^{\mathrm{NE}} \int_{e_k}^{e_{k+1}} \mathbf{E}_i(x, y)\mathbf{n}_k \,\mathrm{d}l = \sum_{k=1}^{\mathrm{NE}} \mathbf{E}_{I,k}\mathbf{n}_k l_k \tag{109}$$

and does not vanish in general. In order to get a final compact form for the scheme as a sum of waves directed along the eigenvectors directions, a flux decomposition is required in (109). It may happen that the property

$$\mathbf{E}\mathbf{n}_k = \widetilde{\mathbf{J}}_{\mathbf{n},k}\mathbf{U}_k \tag{110}$$

does not hold in general. Hence, the following transformation based on (7) is proposed:

$$\sum_{k=1}^{\mathrm{NE}} \mathbf{E}_{I,k}\mathbf{n}_k l_k - \mathbf{E}_{i,0}\sum_{k=1}^{\mathrm{NE}} \mathbf{n}_k l_k + \mathbf{E}_{i,0}\sum_{k=1}^{\mathrm{NE}} \mathbf{n}_k l_k = \sum_{k=1}^{\mathrm{NE}} (\mathbf{E}_{I,k} - \mathbf{E}_{i,0})\mathbf{n}_k l_k = \sum_{k=1}^{\mathrm{NE}} \delta\mathbf{E}_{Ii,k}\mathbf{n}_k l_k \tag{111}$$

so that an eigenvalue decomposition is possible:

$$\sum_{k=1}^{\mathrm{NE}} \delta\mathbf{E}_{Ii,k}\mathbf{n}_k l_k \cong \sum_{k=1}^{\mathrm{NE}} \sum_{m=1}^{N\lambda} (\widetilde{\lambda}^m \alpha^m \widetilde{\mathbf{e}}^m)_{Ii,k} l_k \tag{112}$$

The normal source difference in the second term of (83) is projected onto the eigenvectors of $\widetilde{\mathbf{J}}_{JI,k}$

$$\sum_{k=1}^{\mathrm{NE}} \delta\mathbf{T}_{JI,k}\mathbf{n}_k l_k = \sum_{k=1}^{\mathrm{NE}} \sum_{m=1}^{N\lambda} (\beta^- \widetilde{\mathbf{e}})^m_{JI,k} l_k \tag{113}$$

with $\delta\mathbf{T}_{JI,k} = \mathbf{T}_{J,k} - \mathbf{T}_{I,k}$, and the source term in the last integral in (83) is approximated by

$$\sum_{k=1}^{\mathrm{NE}} \int_{e_k}^{e_{k+1}} \mathbf{T}_i(x, y)\mathbf{n}_k \,\mathrm{d}l = \sum_{k=1}^{\mathrm{NE}} \mathbf{T}_{I,k}\mathbf{n}_k l_k \tag{114}$$

that using (7), can be written as

$$\sum_{k=1}^{\mathrm{NE}} \mathbf{T}_{I,k}\mathbf{n}_k l_k - \mathbf{T}_{i,0}\sum_{k=1}^{\mathrm{NE}} \mathbf{n}_k l_k + \mathbf{T}_{i,0}\sum_{k=1}^{\mathrm{NE}} \mathbf{n}_k l_k = \sum_{k=1}^{\mathrm{NE}} (\mathbf{T}_{I,k} - \mathbf{T}_{i,0})\mathbf{n}_k l_k = \sum_{k=1}^{\mathrm{NE}} \delta\mathbf{T}_{Ii,k}\mathbf{n}_k l_k \tag{115}$$

which, projecting again onto the eigenvectors of $\widetilde{\mathbf{J}}_{\mathbf{n},Ii,k}$, yields:

$$\sum_{k=1}^{\mathrm{NE}} \sum_{m=1}^{N\lambda} (\beta \widetilde{\mathbf{e}})^m_{Ii,k} l_k \tag{116}$$

The updating second-order scheme for the system with source terms is

$$\mathbf{U}_i^{n+1} = \mathbf{U}_i^n - \sum_{k=1}^{\mathrm{NE}} \sum_{m=1}^{N\lambda} ((\widetilde{\lambda}^- \alpha - \beta^-)\widetilde{\mathbf{e}})^m_{JI,k} l_k \frac{\Delta t}{A_i} - \sum_{k=1}^{\mathrm{NE}} \sum_{m=1}^{N\lambda} ((\widetilde{\lambda}\alpha - \beta)\widetilde{\mathbf{e}})^m_{Ii,k} l_k \frac{\Delta t}{A_i} \tag{117}$$

As in the case of the first-order upwind scheme, an effort has been made to formulate the cell-updating algorithm as a sum of waves generated at the cell edges by the joint contribution of normal flux differences and normal source terms. In the second-order method (117) these waves are of two kinds, those generated by the jump across the edge and those generated by the jump between the edge and the cell centre values. This is different from the most usual formulation based on numerical fluxes and, as before, will be better suited to our further analysis.

At this point, (117) can be expressed as

$$\mathbf{U}_i^{n+1} = \mathbf{U}_i^n - \sum_{k=1}^{\mathrm{NE}} \sum_{m=1}^{N\lambda} (v^* \delta \mathbf{U})^m_{JI,k} - \sum_{k=1}^{\mathrm{NE}} \sum_{m=1}^{N\lambda} (v^* \delta \mathbf{U})^m_{Ii,k} \tag{118}$$

where

$$v^m_{JI,k} = \frac{\lambda^{m,*}_{JI,k}}{(A_i/l_k)} \Delta t, \quad v^m_{Ii,k} = \frac{\lambda^{m,*}_{Ii,k}}{(A_i/l_k)} \Delta t \tag{119}$$

with

$$\widetilde{\lambda}^*_{JI,k} = \widetilde{\lambda}^-_{JI,k} \theta_{JI,k}, \quad \theta_{JI,k} = 1 - \left(\frac{\beta^-}{\alpha \widetilde{\lambda}^-} \right)_{JI,k}, \quad \widetilde{\lambda}^*_{Ii,k} = \widetilde{\lambda}^-_{Ii,k} \theta_{Ii,k}, \quad \theta_{Ii,k} = 1 - \left(\frac{\beta^-}{\alpha \widetilde{\lambda}^-} \right)_{Ii,k} \tag{120}$$

To ensure the monotonicity of the conserved variables in (118), both $v^{m,*}_{JI,k}$ and $v^{m,*}_{Ii,k}$ must be limited as

$$-1 \leqslant v^{m,*}_{JI,k} \leqslant 0, \quad -1 \leqslant v^{m,*}_{Ii,k} \leqslant 0 \tag{121}$$

and therefore

$$\theta_{JI,k} \geqslant 0, \quad \theta_{Ii,k} \geqslant 0 \tag{122}$$

The numerical instabilities are avoided if for the m components of $\delta \mathbf{U}_k$

$$|\delta \mathbf{U}^m_{Ii,k} + \delta \mathbf{U}^m_{JI,k}| \leqslant |\delta \mathbf{U}^m_k| \tag{123}$$

and as each $\delta \mathbf{U}^m_k$ is a combination of the different U_s components, (123) is fulfilled if

$$|\delta U_{s,Ii,k} + \delta U_{s,JI,k}| \leqslant |\delta U_{s,k}|, \quad s = 1, \ldots, Nc \tag{124}$$

For that reason the values of the function U_s at the middle-edge point, computed as in (28), have to be bounded as follows:

$$U^{\min}_{s,k} \leqslant U_{s,J,k} \leqslant U^{\max}_{s,k}, \quad U^{\min}_{s,k} \leqslant U_{s,I,k} \leqslant U^{\max}_{s,k} \tag{125}$$

Under these assumptions the stability region is defined under

$$\begin{gathered} \Delta t = \mathrm{CFL}\, \Delta t_{\max}, \quad \mathrm{CFL} \leqslant \tfrac{1}{3} \\ \Delta t_{\max} = \min\{\Delta t_k\}_{k=1,N\mathrm{edge}}, \quad \Delta t_k = \frac{A_{\min,k}}{\max_m \{\lambda^{m,*}_{k,\max}\} l_k}, \quad m = 1, \ldots, N\lambda, \quad \theta^m_{JI,k} \geqslant 0, \quad \theta^m_{Ii,k} \geqslant 0 \end{gathered} \tag{126}$$

with $\lambda^{m,*}_{k,\max} = \max\{|\lambda^{m,*}_{JI,k}|, |\lambda^{m,*}_{Ii,k}|, |\lambda^{m,*}_k|\}$.

DOI: 10.1002/fld

Now, for equilibrium in steady state it is necessary that

$$(\widetilde{\lambda}^{-}\alpha-\beta^{-})^{m}_{JI,k}=0, \quad (\widetilde{\lambda}\alpha-\beta)^{m}_{Ii,k}=0, \quad m=1,\ldots,N\lambda \tag{127}$$

for all k, which is only feasible if the local discretization follows the conditions stated for first order. As in the scalar case, condition (127) can only be reached if a careful interpolation procedure is made. If (118) is expressed as

$$\mathbf{U}_i^{n+1}=\mathbf{U}_i^n-\sum_{k=1}^{\mathrm{NE}}\sum_{m=1}^{N\lambda}(\widetilde{\lambda}^{-}\delta\mathbf{D})^{m}_{JI,k}l_k\frac{\Delta t}{A_i}-\sum_{k=1}^{\mathrm{NE}}\sum_{m=1}^{N\lambda}(\widetilde{\lambda}^{-}\delta\mathbf{D})^{m}_{Ii,k}l_k\frac{\Delta t}{A_i} \tag{128}$$

with

$$\delta\mathbf{D}^{m}_{JI,k}=\theta^{m}_{JI,k}\delta\mathbf{U}^{m}_{JI,k}, \quad \delta\mathbf{D}^{m}_{Ii,k}=\theta^{m}_{Ii,k}\delta\mathbf{U}^{m}_{Ii,k} \tag{129}$$

condition (126) is ensured and enlarges the stability region of the homogeneous case if

$$0\leqslant\theta^{m}_{JI,k}\leqslant 1, \quad 0\leqslant\theta^{m}_{Ii,k}\leqslant 1 \tag{130}$$

for all m.

In the case where $\theta^{m}_{JI,k}\geqslant 0$, assuming $\theta_{JI,k}\geqslant 0$ and $\theta_{Ii,k}\geqslant 0$, the numerical instabilities are avoided if

$$|\delta\mathbf{D}^{m}_{Ii,k}+\delta\mathbf{D}^{m}_{JI,k}|\leqslant|\delta\mathbf{D}^{m}_{k}| \tag{131}$$

which is equivalent to

$$|\delta D^{m}_{s,Ii,k}+\delta D^{m}_{s,JI,k}|\leqslant|\delta D^{m}_{s,k}|, \quad s=1,\ldots,Nc \tag{132}$$

and the values at the cell edge have to be bounded as follows:

$$D^{\min}_{s,k}\leqslant D_{s,J,k}\leqslant D^{\max}_{s,k}, \quad D^{\min}_{s,k}\leqslant D_{s,I,k}\leqslant D^{\max}_{s,k} \tag{133}$$

where $D^{\min}_{s,k}=\min\{D_{s,j,0}, D_{s,i,0}\}_k$, $D^{\max}_{s,k}=\max\{D_{s,j,0}, D_{s,i,0}\}_k$. This is only feasible if $\delta\mathbf{D}^{m}_{JI,k}$ and $\delta\mathbf{D}^{m}_{Ii,k}$ can be written as

$$\delta\mathbf{D}^{m}_{JI,k}=\mathbf{D}^{m}_{J,k}-\mathbf{D}^{m}_{I,k}, \quad \delta\mathbf{D}^{m}_{Ii,k}=\mathbf{D}^{m}_{I,k}-\mathbf{D}^{m}_{i,k} \tag{134}$$

with the appropriate relations among the conserved variables. Now in equilibrium

$$\delta\mathbf{D}^{m}_{JI,k}=0, \quad \delta\mathbf{D}^{m}_{Ii,k}=0 \tag{135}$$

and the numerical scheme reduces automatically to first-order approximation, leading to $\theta^{m}_{JI,k}=0$ and $\theta^{m}_{Ii,k}=0$, which also provides an unconditional scheme in that case. It is important to remark that the extension of the scalar case to systems requires the identification of independent $\mathbf{D}^{m}$ functions in order to ensure (135) in steady state.

As in the scalar case, the behaviour of the conserved variables must be always controlled in order to avoid situations where

$$|\delta U_{s,JI,k}|>|\delta U_{s,k}| \tag{136}$$

In that case it is necessary to reduce the numerical scheme to first-order approach. In those cases where $\theta^{m}_{k}<0$ the analysis performed for the second order in space approach assuming $\theta^{m}_{JI,k}\geqslant 0$ and

DOI: 10.1002/fld

$\theta^m_{Ii,k}\geqslant 0$ is not valid. The requirements to preserve the sign in the solution in the s component (98) in combination with (136) are ensured if the numerical scheme is reduced to first order when $0\leqslant\gamma<1$.

3.3. *Explicit second-order in time and space scheme for systems*

The MUSCL-Hancock scheme extended to systems of equations can also be formulated based on two steps. In the first step the solution must be reconstructed using the $\mathbf{L}_i$ gradient vectors and then intermediate values are re-calculated at a half time step at cell edges as

$$\mathbf{U}^{n+1/2}_{I,k}=\mathbf{U}^n_{I,k}-\sum_{k=1}^{\mathrm{NE}}(\delta\mathbf{En}_k-\delta\mathbf{Tn}_k)^n_{Ii,k}\frac{l_k}{A_i}\frac{\Delta t}{2} \tag{137}$$

which is equivalent to redefining the interpolation function in each cell. The intermediate values (137) must ensure monotonicity and positivity as previously defined and must be reduced to first order when $\theta^m_k<0$, as stated in the preceding section. The updated variable is constructed as

$$\mathbf{U}^{n+1}_i=\mathbf{U}^n_i-\sum_{k=1}^{\mathrm{NE}}(\delta\mathbf{En}_k-\delta\mathbf{Tn}_k)^{n+1/2,-}_{JI,k}\frac{l_k}{A_i}\Delta t-\sum_{k=1}^{\mathrm{NE}}(\delta\mathbf{En}_k-\delta\mathbf{Tn}_k)^{n+1/2}_{Ii,k}\frac{l_k}{A_i}\Delta t \tag{138}$$

where $\delta\mathbf{E}^{n+1/2}_{JI,k}=\mathbf{E}^{n+1/2}_{J,k}-\mathbf{E}^{n+1/2}_{I,k}$ and $\delta\mathbf{E}^{n+1/2}_{Ii,k}=\mathbf{E}^{n+1/2}_{I,k}-\mathbf{E}^n_{i,k}$, with an upwind part (first term) and a central part (second term).

4. APPLICATION TO THE SHALLOW WATER EQUATIONS

4.1. *Two-dimensional mathematical model of the shallow water equations*

In this work the schemes are used to solve the following system of equations:

$$\begin{aligned}&\mathbf{U}=(h,q_x,q_y)^{\mathrm{T}}\\&\mathbf{F}=\left(q_x,\frac{q_x^2}{h}+\frac{gh^2}{2},\frac{q_xq_y}{h}\right)^{\mathrm{T}},\quad \mathbf{G}=\left(q_y,\frac{q_xq_y}{h},\frac{q_y^2}{h}+\frac{gh^2}{2}\right)^{\mathrm{T}}\\&\mathbf{S}=(0,gh(S_{0x}-S_{fx}),gh(S_{0y}-S_{fy}))^{\mathrm{T}}\end{aligned} \tag{139}$$

where h is the water depth, g is the acceleration of the gravity, $q_x=uh$, $q_y=vh$ the unit discharge components, with (u,v) the depth-averaged components of the velocity vector $\mathbf{u}$ along the x and y, coordinates, respectively.

4.2. *Application of the explicit upwind numerical scheme*

As stated in Section 3, the mathematical properties of the hyperbolic system of equations include the existence of a Jacobian matrix. In the case of (139) it is convenient to work with the normal flux Jacobian matrix $\mathbf{J_n}$. Upwind schemes were first developed for the Euler equations [16]. In those equations the numerical flux is first-order homogeneous and (110) holds. This is not the case for the Saint-Venant equations [18], and an approximate normal flux Jacobian matrix $\widetilde{\mathbf{J}}_{\mathbf{n},k}$ has to be defined. The eigenvalues of $\widetilde{\mathbf{J}}_{\mathbf{n},k}$ are used. For more details on the normal flux Jacobian and its properties see [19–21].

The vectors $\mathbf{S}_1$ and $\mathbf{S}_2$ in (74) for the discretization of the source term are defined as

$$\mathbf{S}_1 = (0, gh(-z+H), 0, 0)^{\mathrm{T}}, \quad \mathbf{S}_2 = (0, 0, gh(-z+H), 0)^{\mathrm{T}} \tag{140}$$

where H is the total energy head:

$$H = h + z + \frac{|\mathbf{u}|^2}{2g} \tag{141}$$

The normal source difference is

$$\delta\mathbf{T}\mathbf{n} = \delta(\mathbf{S}_1, \mathbf{S}_2)^{\mathrm{T}}\mathbf{n} = \begin{pmatrix} 0 \\ gh(-\delta z + \delta H)n_x \\ gh(-\delta z + \delta H)n_y \\ 0 \end{pmatrix} \tag{142}$$

and the spatial bed and energy level variations acting as source terms are:

$$-\nabla z = -\left(\frac{\partial z}{\partial x}, \frac{\partial z}{\partial y}\right) = (S_{0x}, S_{0y}), \quad \nabla H = \left(\frac{\partial H}{\partial x}, \frac{\partial H}{\partial y}\right) = (-S_{fx}, -S_{fy}) \tag{143}$$

Following (87) the coefficients β^m are defined as

$$\beta^1 = -\frac{\tilde{c}}{2}(\delta z + d_{\mathbf{n}} S_f), \quad \beta^1 = -\beta^3, \quad \beta^2 = \beta^4 = 0 \tag{144}$$

having used the diagonalization matrices

$$\widetilde{\mathbf{P}} = \begin{pmatrix} 1 & 0 & 1 \\ \tilde{u} + \tilde{c}n_x & -\tilde{c}n_y & \tilde{u} - \tilde{c}n_x \\ \tilde{v} + \tilde{c}n_y & \tilde{c}n_x & \tilde{v} - \tilde{c}n_y \end{pmatrix}, \quad \widetilde{\mathbf{P}}^{-1} = \frac{1}{2\tilde{c}} \begin{pmatrix} -\tilde{\mathbf{u}}\cdot\mathbf{n} + c & n_x & n_y \\ 2(\tilde{u}n_y - \tilde{v}n_x) & -2n_y & 2n_x \\ \tilde{\mathbf{u}}\cdot\mathbf{n} + c & -n_x & -n_y \end{pmatrix} \tag{145}$$

where $\tilde{c} = \sqrt{g\tilde{h}}$ and $\tilde{h} = \frac{1}{2}(h_i + h_j)$.

The difference δH_k is computed as

$$\delta H_k = -d_{\mathbf{n}} S_{f,k} = -d_{\mathbf{n}} \frac{n^2 \tilde{\mathbf{u}}\mathbf{n}|\tilde{\mathbf{u}}|}{\tilde{h}^{4/3}} \tag{146}$$

where $d_{\mathbf{n}}$ is the distance between cell centroids projected onto the $\mathbf{n}$ direction. The empirical Manning friction formula has been applied [22, 23] and cannot be transformed into a linear space variation over the cell. For that reason, when second order is imposed H is considered uniform in the cell and terms like $\delta H_{Ii,k}$ are nil. Also, for the sake of simplicity, the interpolation vector $\mathbf{L}_z$ will be set nil in those cases.

4.3. Conservation properties and equilibrium at steady state

The unified discretization of the source terms is successfully constructed when it ensures an exact balance in first-order approximation [24]. For that reason, the discretization of the fluxes and source terms proposed in Section 4.1 is here analysed. Steady state in the first-order scheme is expressed as

$$\widetilde{\mathbf{J}}_{\mathbf{n},k}\delta\mathbf{U}_k = \delta\widetilde{\mathbf{T}}_k\mathbf{n}_k \tag{147}$$

According to the form of the matrices involved, the first equation in (147) gives

$$\delta(hu)_k n_x + \delta(hv)_k n_y = \delta(\mathbf{q}_{j,0} - \mathbf{q}_{i,0})\mathbf{n}_k = \delta\mathbf{q}_k\mathbf{n}_k = 0 \tag{148}$$

which means that the normal discharge is constant at the edge. The second line in (147) yields

$$\delta h(\widetilde{c}^2 n_x - \widetilde{\mathbf{u}}\cdot\mathbf{n}\widetilde{u}) + \delta q_x(\widetilde{u}n_x + \widetilde{\mathbf{u}}\cdot\mathbf{n}) + \delta q_y(\widetilde{u}n_y) = g\widetilde{h}(-\delta z_k + \delta H_k)n_x \tag{149}$$

or

$$\delta h(\widetilde{c}^2 n_x - \widetilde{\mathbf{u}}\cdot\mathbf{n}\widetilde{u}) + \delta q_x(\widetilde{\mathbf{u}}\cdot\mathbf{n}) = g\widetilde{h}(-\delta z_k + \delta H_k)n_x \tag{150}$$

and the third line can be easily reduced to

$$\delta h(\widetilde{c}^2 n_y - \widetilde{\mathbf{u}}\cdot\mathbf{n}\widetilde{v}) + \delta q_y(\widetilde{\mathbf{u}}\cdot\mathbf{n}) = g\widetilde{h}(-\delta z_k + \delta H_k)n_y \tag{151}$$

If (150) and (151) are multiplied by n_x and n_y, respectively, and next combined, the expression for the water depth profile for uniform flows appears

$$\delta h(1 - Fr_k^2) + \delta z_k = \delta H_k = -d_{\mathbf{n}}S_{f,k} \tag{152}$$

where Fr_k is the Roe average Froude number in the normal edge direction, $Fr_k = (\widetilde{\mathbf{u}}\mathbf{n})_k/\widetilde{c}_k \geqslant 1$. Equation (152) can be derived directly enforcing

$$\delta\mathbf{D}_k^m = 0 \tag{153}$$

or $(\widetilde{\lambda}^{-}\alpha - \beta^{-})_k^m = 0$ for $m = 1$ and 3. Therefore, the discretization in (153) ensures equilibrium in steady-state cases correctly for first-order approximation.

When second order is used, equilibrium in steady state can only be satisfied if (135) holds ($\delta\mathbf{D}_{JI,k}^m = 0$ and $\delta\mathbf{D}_{Ii,k}^m = 0$). That results in the following conditions over the interpolated edge variables:

$$\begin{gathered}
\delta h_{JI,k}(1 - Fr_{JI,k}^2) + \delta z_{JI,k} + \delta H_{JI,k} = 0 \\
\delta h_{Ii,k}(1 - Fr_{Ii,k}^2) + \delta z_{Ii,k} = 0
\end{gathered} \tag{154}$$

But the construction of interpolating planes for each conserved variable can only satisfy (135) and (154) in the case of still water where

$$\begin{gathered}
\delta h_{JI,k} + \delta z_{JI,k} = 0 \\
\delta h_{Ii,k} + \delta z_{Ii,k} = 0
\end{gathered} \tag{155}$$

condition that can only be satisfied if the interpolated variable is $\xi = h + z$,

$$\xi(x, y) = \xi_{i,0} + \mathbf{r}\mathbf{L}_i \tag{156}$$

Following (136) it is necessary to reduce the numerical scheme to first order when

$$|\delta h_{JI,k}| > |\delta h_k| \tag{157}$$

where $h_{I,k} = \xi_{I,k} - z_{I,k}$ and $h_{J,k} = \xi_{J,k} - z_{J,k}$.

When no source terms are present and second order is required it is also impossible to satisfy (154) and a perfect equilibrium or an exact steady-state solution can never be reached unless in trivial cases.

4.4. Interpretation of the θ coefficient

According to (96) and previous section in steady state:

$$\theta_k^1 = \theta_k^3 = \theta = 1 + \frac{S_f d_{\mathbf{n}} + \delta z}{(1 - Fr^2)\delta h} \tag{158}$$

or, calling $\delta z' = -S_f d_{\mathbf{n}} - \delta z$ and $\delta h' = (1 - Fr^2)\delta h$,

$$\theta = 1 - \frac{\delta z'}{\delta h'} \tag{159}$$

This compact formulation is useful to analyse the time step limits following Figure 3.

4.4.1. Bed slope term. When the bed slope term dominates over the friction term and convergence to equilibrium is analysed, the monotonicity over the solution can be expressed in terms of water level surface instead of the water depth h,

$$\xi_k^{\min} \leqslant \xi_i^{n+1} \leqslant \xi_k^{\max} \tag{160}$$

This is only feasible considering gradually varied water surface elevation, $|\delta\xi_k| < \xi_{i,0}, \xi_{j,0}$ and that $|\delta\xi_k| \leqslant |\delta h_k|$, so the stability region is defined by $0 \leqslant \theta_k^m \leqslant 1$.

Instabilities arise in the presence of adverse slope in wetting processes where both $\delta h'$ and $\delta z'$ become negative, or in recession surface processes where both become positive, leading to values of $\theta_k^m < 0$. Preservation of positivity over the water depth must be applied in these cases. The stability region requires computing the time step as in (105) where

$$\begin{aligned} &0 \leqslant \gamma < 1, \quad \theta_k^m < 0 \\ &\gamma = \frac{\min\{h_{i,0}, h_{j,0}, |\delta h_k|\}}{|\delta h_k|} \end{aligned} \tag{161}$$

In the special case $\gamma = 0$ the cell edge acts as a solid wall for any value of time step. As in both drying/wetting and wetting/drying interfaces our desire is mainly focused on conserving water volume, the following condition must be applied to the future solution at the cells sharing edge k:

$$(\mathbf{u}_i \mathbf{n}_{i,k})^{n+1} = (\mathbf{u}_j \mathbf{n}_{i,k})^{n+1} = 0 \tag{162}$$

DOI: 10.1002/fld

In the cases where $0<\gamma<1$, the reduction of the magnitude of the time step can be avoided by means of a conservative strategy based on the redistribution of updating fluxes [8], involving the local time step $\Delta t_{h,i}$, that replaces the γ coefficient in (170). That technique proved successful for first-order approach and can be directly applied to the second-order scheme, as it must be reduced to first order in the cases where $0\leqslant\gamma<1$. It is worth remarking that, when dealing with second order, in those cases where strong variations in the discrete water level surface are present, $|\delta\xi_k|>|\delta h_k|$, the numerical scheme must be reduced to first order, otherwise negative values of water depth can be obtained in the interpolation process.

4.4.2. Bed friction term. Near wetting/drying fronts, characterized by small values of water depth, the bed friction term may dominate over the bed slope terms. Under this hypothesis, and assuming a negative gradient in the water depth value in the direction towards the shoreline as in wetting/drying fronts, $\delta h'<0$ and $\delta z'\cong -S_f d_{\mathbf{n}}<0$, the region $\theta<0$ is met. Again, the preservation of positivity over the solution is required for the water depth as in the previous case. In the case $\gamma=0$ the cell edge acts as a solid wall for any value of time step. When $0<\gamma<1$ numerical instabilities are avoided by requiring that friction alone is not able to change the sign of the discharge, so the following conditions are enforced over the unit discharge function hu:

$$\begin{aligned} (hu)_i^{n+1} &\geqslant 0, \quad (hu)_i^n, (hu)_{j=1,2,3}^n \geqslant 0 \\ (hu)_i^{n+1} &\leqslant 0, \quad (hu)_i^n, (hu)_{j=1,2,3}^n \leqslant 0 \end{aligned} \tag{163}$$

and similarly over hv. These conditions must be included to determine the maximum allowable time step. Let us assume without loss of generality the one-dimensional case, where the updated value can be expressed as

$$(hu)_i^{n+1} = (hu)_i^n + (g\widetilde{h}\delta H)_k^n \frac{l_k}{A_i}\Delta t = (hu)_i^n - \left(g\widetilde{h}\frac{n^2\widetilde{u}|\widetilde{u}|}{\widetilde{h}^{4/3}} d_{\mathbf{n}} l\right)_k^n \frac{\Delta t}{A_i} \tag{164}$$

that can be rewritten as

$$(hu)_i^{n+1} = (hu)_i^n \left[1 - \frac{(\widetilde{h}\widetilde{u})_k^n}{(hu)_i^n}\left(\frac{gn^2|\widetilde{u}|}{\widetilde{h}^{4/3}} d_{\mathbf{n}} l\right)_k^n \frac{\Delta t}{A_i}\right] \tag{165}$$

The second term on right hand of (165) must be positive to ensure (163). Hence, in general, the time step Δt_k, taking into account also condition (126), is limited by

$$\Delta t_k = \min\left\{\left(\frac{n^2|\widetilde{\mathbf{u}}|}{\widetilde{h}^{4/3}}\frac{d_{\mathbf{n}} l}{A_{\min,k}} g\right)^{-1}, \frac{A_{\min,k}}{\max_m\{\widetilde{\lambda}_k^{m,*}|\}l}\right\}_k \tag{166}$$

It is remarkable that, considering that $\delta z' \propto d_{\mathbf{n}}$ the stability region $0\leqslant\theta\leqslant 1$ can be recovered by decreasing the size of the cells, as pointed out in Burguete *et al.* [25]. This option is not always affordable due to the high computational cost associated. In the case of a pointwise explicit

DOI: 10.1002/fld

discretization the following limit must be imposed:

$$\Delta t \leqslant \min\left\{\left(\frac{n^2|\mathbf{u}_i|}{h_i^{4/3}}g\right)^{-1}\right\}_{i=1,N\text{cell}} \tag{167}$$

to prevent instabilities. All the limitations related to the friction term can be overcome using a pointwise implicit discretization [20], although an exact equilibrium among fluxes and friction source term cannot be achieved in steady state.

5. NUMERICAL RESULTS

5.1. Two-dimensional steady state with variable bed slope and friction

A two-dimensional steady flow test case with analytical solution involving friction is used to study the behaviour of the numerical schemes in presence of source terms when second order is enforced. The flow discharge is constant in the entire domain and equal to

$$q_x(x,y) = q_y(x,y) = q_x = q_y \tag{168}$$

and the steady-state water depth and bed slope analytical functions are

$$h(x,y) = a + q_x x + q_y y, \quad z(x,y) = -\frac{1}{2g}\frac{(q_x^2+q_y^2)+2gh^3}{h^2} + \frac{3}{7}\frac{|q_x|n^2\sqrt{2}}{h^{7/3}} \tag{169}$$

The performance of the schemes are tested using $q_x = q_y = 0.1$, $a = 0.5$ and Manning friction coefficient $n = 0.03$ in a squared domain $10\times 10\,\text{m}$ discretized using a Delaunay mesh with 2064 cells. Figure 11(a) displays the computational mesh, (b) a contour map of the exact bed elevation in meters and (c) a three-dimensional view of the exact water level surface elevation in meters.

Before analysing the conditions for flow in movement, the special case of still water, $u = v = 0$ as bed level function (169), is considered, assuming a constant value of water level surface ξ equal to zero. Figure 12 shows the solution for the water surface elevation after one time step using second-order approximation and interpolating the water depth using the MLG technique. The solutions only becomes equal to the solution for first-order approach if second order is enforced over the water level surface, as the numerical scheme reduces automatically to first order.

In the non-trivial case, the steady water level is computed starting from initial condition of still water, with $\xi = 0$. Figure 13(a) shows the exact water level surface, the plot of the values as computed with first-order scheme is shown in Figure 13(b), those from second order over h, hu, hv using MLG in Figure 13(c), the result from second order over ξ, hu, hv using MLG in Figure 13(d) and those from second order over ξ, hu, hv using MLG-Wierse in Figure 13(e). The solution has been computed for 300 s. No remarkable differences appear between the solutions computed using first or second order interpolating over ξ. Figure 14 shows the evolution of the L_1 error in time and how the most accurate solution is provided using first order. When applying a second-order approach a perfect equilibrium can never be attained and a constant error cannot be determined. The MLG-Wierse provides the closest result to

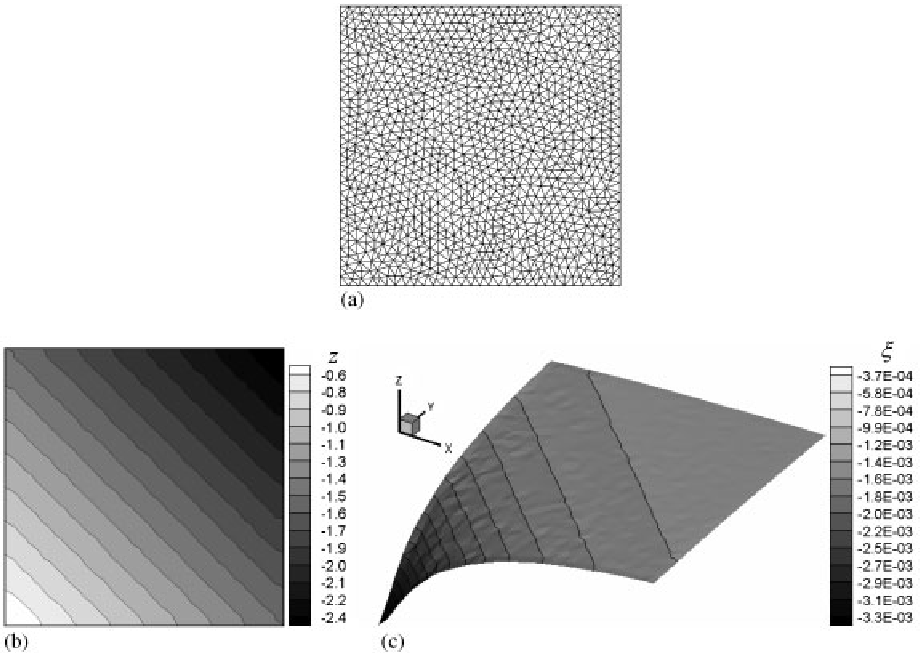

Figure 11. (a) Computational mesh; (b) contour map of the bed elevation in meters; and (c) three dimensional-view of the water level surface elevation in meters.

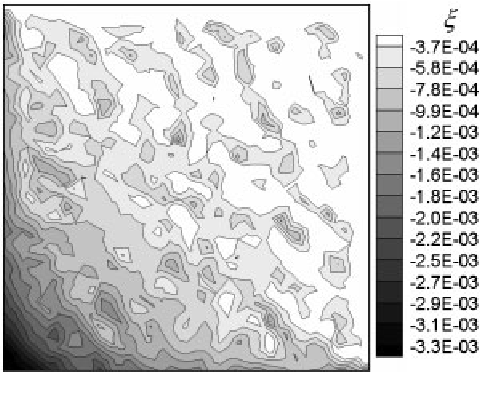

Figure 12. Water level surface with second order over h after one time step using the MLG technique in the case of still water.

first-order approach with small oscillations in the L_1 error function compared with the MLG technique.

In this case the bed level source term dominates over the friction term that has almost no influence over the solution. To show the importance of the correct discretization of the source

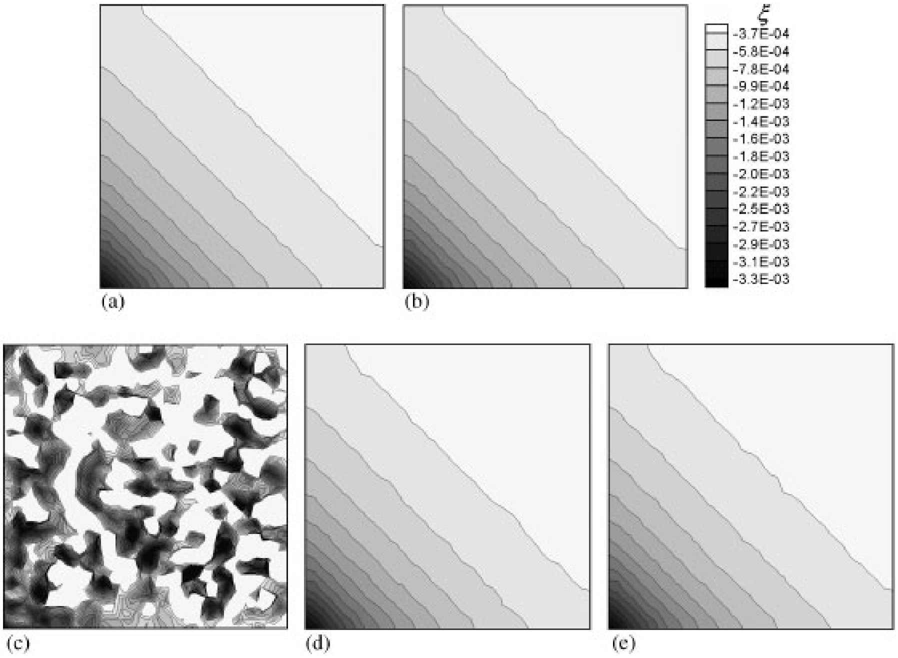

Figure 13. Exact water level surface (a) first-order solution (b) second order over h, hu, hv using MLG (c) second order over ξ, hu, hv using MLG (d) second order over ξ, hu, hv using MLG-Wierse (e) after 300 s.

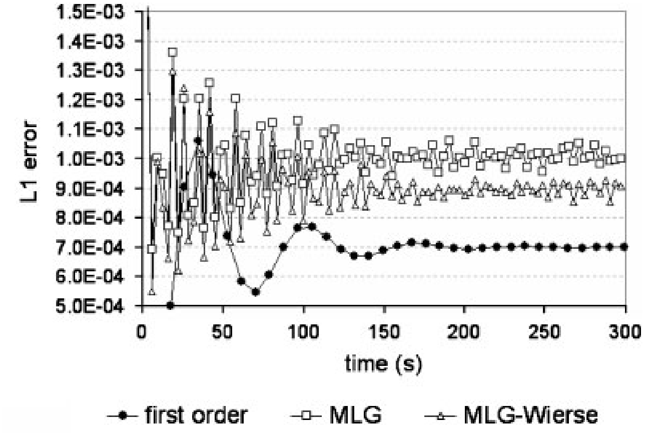

Figure 14. L_1 error in time for first- and second-order approaches.

terms an extremely high value of roughness coefficient is now used, $n = 0.3$ keeping unchanged the other values. Figure 15 shows a three-dimensional view of the new water level surface. The steady-state water level is computed starting from the initial condition of still water. Figure 16

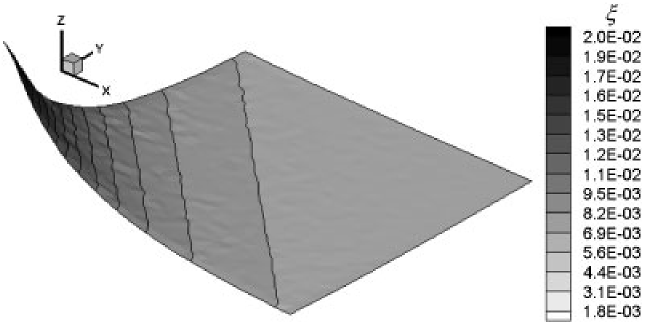

Figure 15. Three-dimensional view of the water level surface (m).

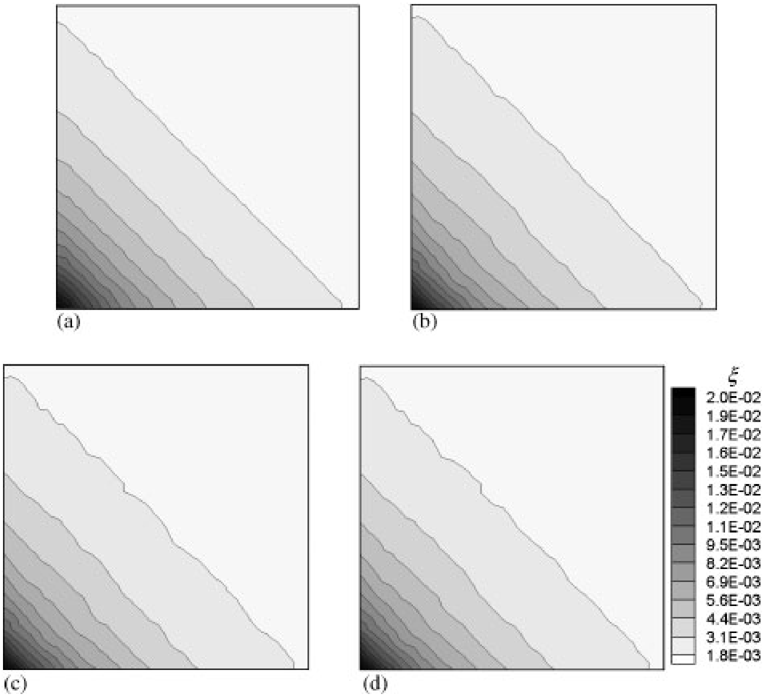

Figure 16. Exact water level surface (a) with first order (b), second order over ξ, hu, hv using MLG (c) over ξ, hu, hv using MLG-Wierse (d) after 200 s.

shows the exact water level surface (a) with first order (b) second order over ξ, hu, hv using MLG (c) and second order over ξ, hu, hv using MLG-Wierse (d) after 200 s. Again no remarkable differences appear among the solutions. Figure 17 shows the evolution of the L_1 error in time and how the most accurate solution is provided using first order. The same conclusions regarding equilibrium are achieved. In this case no differences between the results for the MLG and the MLG-Wierse techniques can be noted.

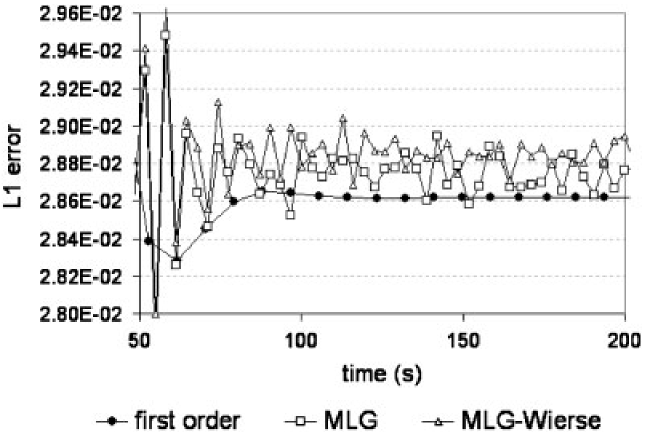

Figure 17. L_1 error in time for first- and second-order approaches.

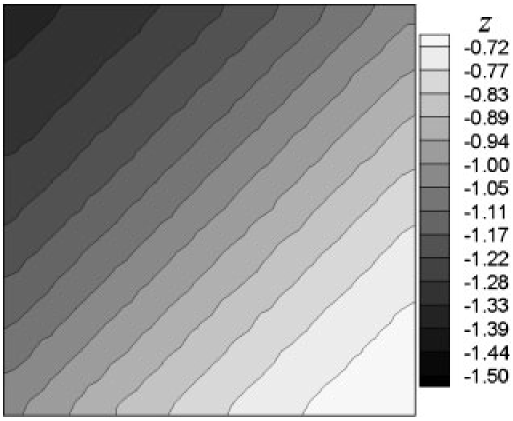

Figure 18. Bottom level in meters.

5.2. *Two-dimensional frictionless steady state with sinusoidal bed slope*

This test case is again a frictionless steady flow over variable bed characterized by a uniform horizontal surface level $\xi(x, y, t) = 0$. The corresponding bed analytical function is

$$z(x, y) = -h_0 + \frac{q_0}{a}\sin(a(x-y)) \tag{170}$$

with the unit discharges varying in space:

$$q_x(x, y) = q_y(x, y) = q_0\cos(a(x-y)) \tag{171}$$

The schemes are tested using $q_0 = 0.05$, $a = 2\pi/(\sqrt{2}\,30)$ and $h_0 = 3q_0/a$. Figure 18 shows the contour plot of the bottom elevation.

DOI: 10.1002/fld

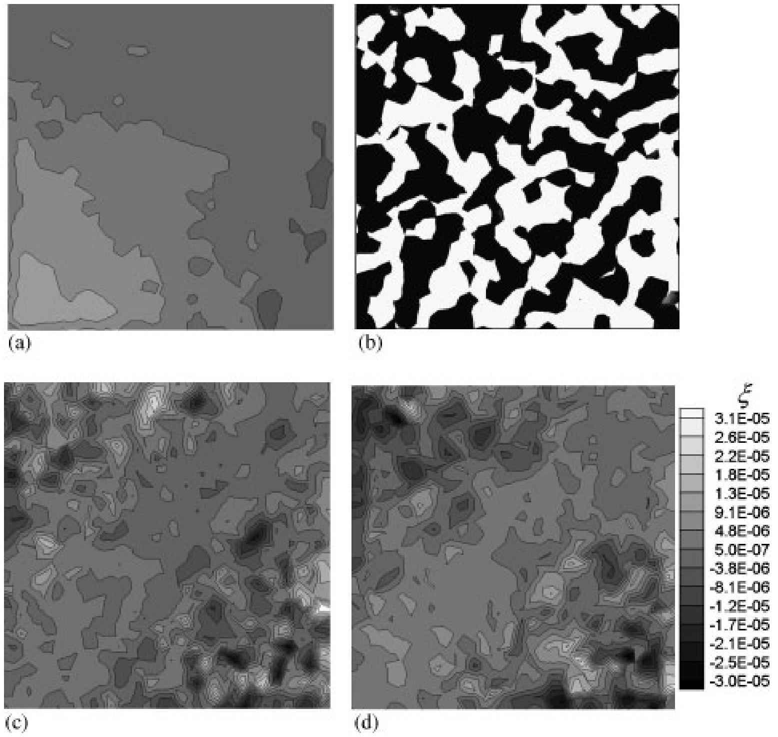

Figure 19. Water level surface with first order (a), second order over h, hu, hv using MLG (b) and over ξ, hu, hv using MLG (c) over ξ, hu, hv using MLG-Wierse (d) after 300 s.

The boundary conditions imposed at the upstream sides (south and west sides) are the x and y unit discharges and, at the downstream sides (north and east sides) the water depth, as before. The same squared domain of the previous test case is discretized using the mesh detailed in Section 5.2. When second order is used in this test case, the same conclusion about the necessity of computing the water depth h from the second order extrapolated surface level ξ is derived. Figure 19(a) shows the solution for first-order approach and Figure 19(b) the distortions in the water surface level when second order is imposed using the MLG limiter function over h. When the limiter acts over ξ, accurate results are obtained as Figure 19(c) and (d) shows for the MLG and the MLG-Wierse techniques, respectively. Figure 20 shows the evolution of the L_1 error in time. Again, the most accurate solution is provided using first order and the MLG-Wierse provides the closest result to first-order approach.

5.3. *Frictionless steady-state hydraulic jump with flat bed*

This test case is used to check the behaviour of the solutions in the presence of a discontinuous flow. A supercritical uniform flow, over flat and frictionless bed, is deflected by a solid wall at an angle σ generating an oblique hydraulic jump as shown in Figure 21 (right).

There is an exact relationship between the water depths upstream the shock, h_1, and downstream of it, h_2, the Froude number of the incoming flow normal to the jump, Fr_1, and the angle formed

DOI: 10.1002/fld

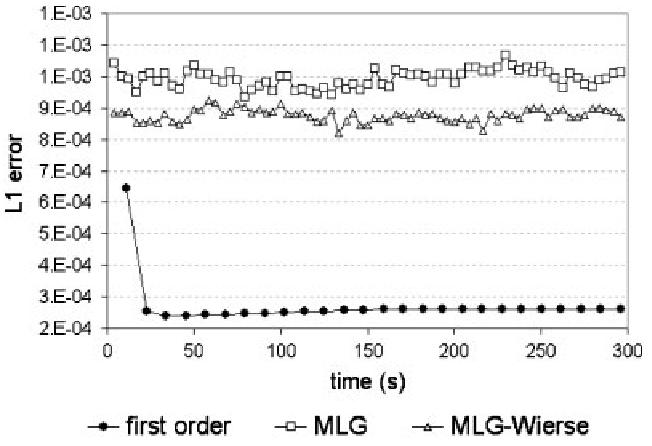

Figure 20. L_1 error in time for first- and second-order approaches.

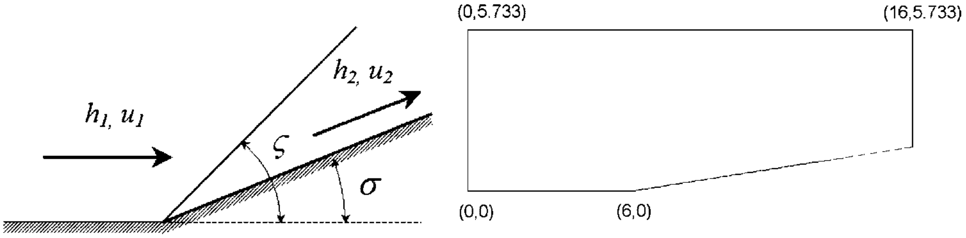

Figure 21. Oblique hydraulic jump.

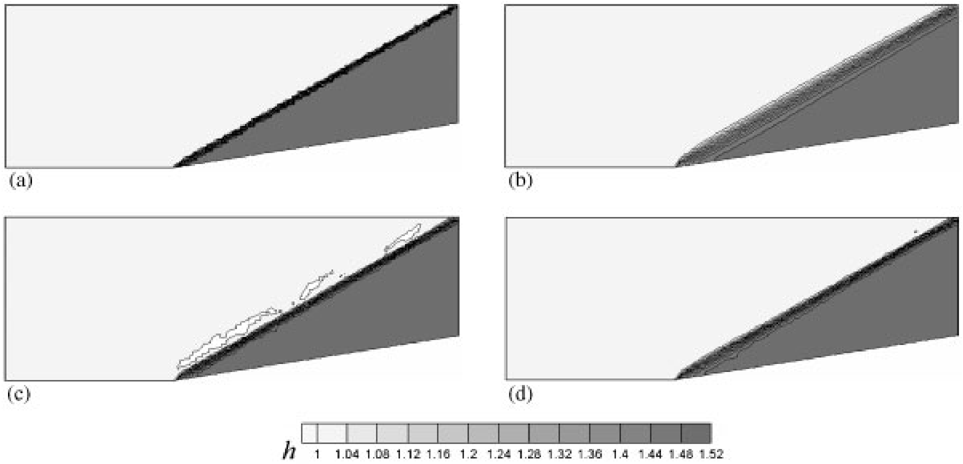

Figure 22. Exact water depth (a) with first order (b), second order over h, hu, hv using MLG (c) over h, hu, hv using MLG-Wierse (d) after 10 s.

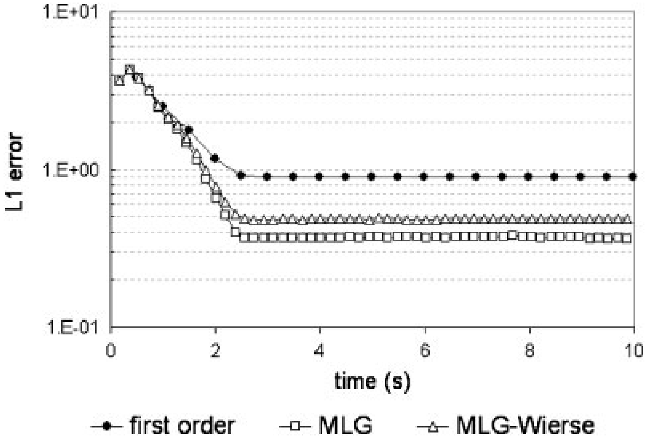

Figure 23. $\log(L_1)$ error in time for first- and second-order approaches.

Table III. L_1 error in h for different mesh refinement.

	$n=4052$	$n=9069$	$n=14\,474$
L_1, first order	1.586	1.130	0.898
L_1, MLG	0.698	0.482	0.369
L_1, MLG-Wierse	0.960	0.663	0.486

by the jump, ς

$$\frac{h_2}{h_1}=\frac{1}{2}\left(\sqrt{1+8Fr_1^2\sin^2\varsigma}-1\right) \tag{172}$$

and also an exact relationship that links the deflection angle σ and the jump angle ς:

$$\tan\sigma=\tan\varsigma\frac{\sqrt{1+8Fr_1^2\sin^2\varsigma}-3}{2\tan^2\varsigma-1+\sqrt{1+8Fr_1^2\sin^2\varsigma}} \tag{173}$$

The computational domain, Figure 21 (left), is represented by three different meshes divided in 4052, 9069 and 14 474 cells. The steady flow numerical solution is obtained from a constant value of water depth h_1 and velocity v_1, after simulating 10 s, enough to converge to steady condition where a hydraulic jump is produced. In this case, the upstream and downstream values are, respectively, $Fr_1=2.74$, $h_1=1$ and $Fr_2=2.74$, $h_2=1$. The deflexion angle is $\sigma=8.95°$ and the jump angle is $\varsigma=30°$. Figure 22(a) displays the exact solution and (b) the solution obtained for first-order approach in the most refined mesh. Second order using the MLG technique provides the sharpest jump (Figure 22(c)) compared with the results for the MLG-Wierse technique (Figure 22(d)), but the excessive antidiffusive effect of the MLG technique results in the presence of undershoots

DOI: 10.1002/fld

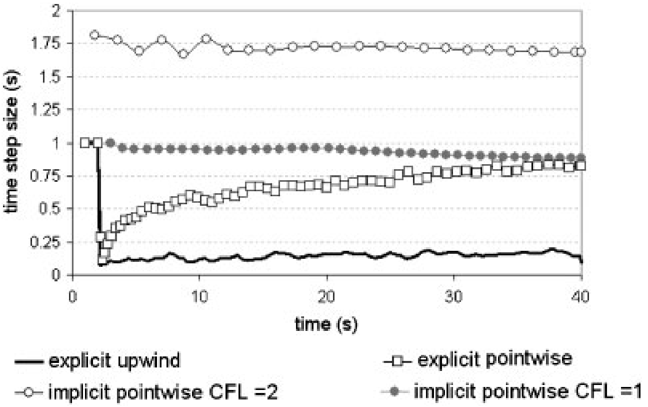

Figure 24. Time step size for the different discretizations of the source term.

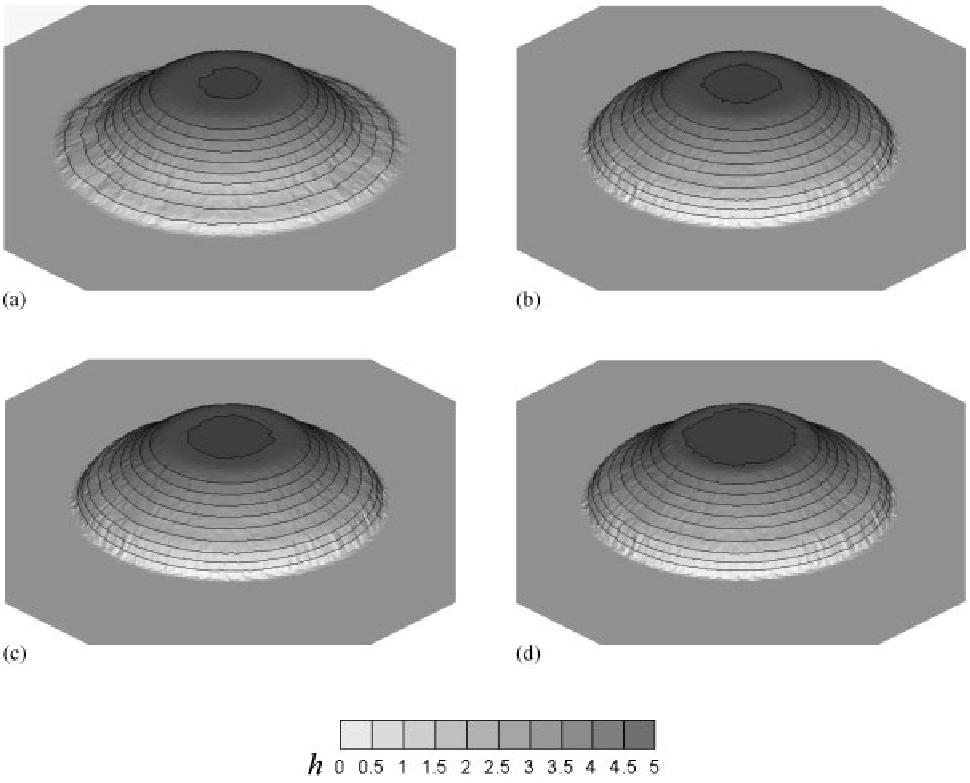

Figure 25. Three-dimensional contour plot of the water depth using the upwind explicit discretization (a), the pointwise discretization (b), the pointwise implicit with CFL = 1 (c) and the pointwise implicit with CFL = 2 (d) at time $t = 40$ s.

DOI: 10.1002/fld

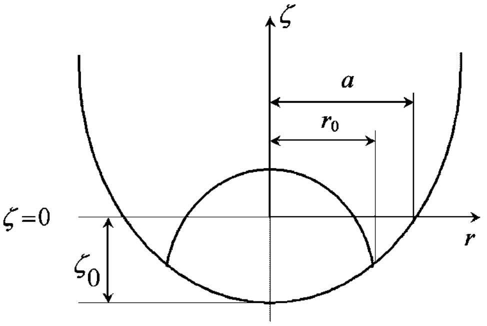

Figure 26. Initial free surface and water depth profile for the parabolic basin test.

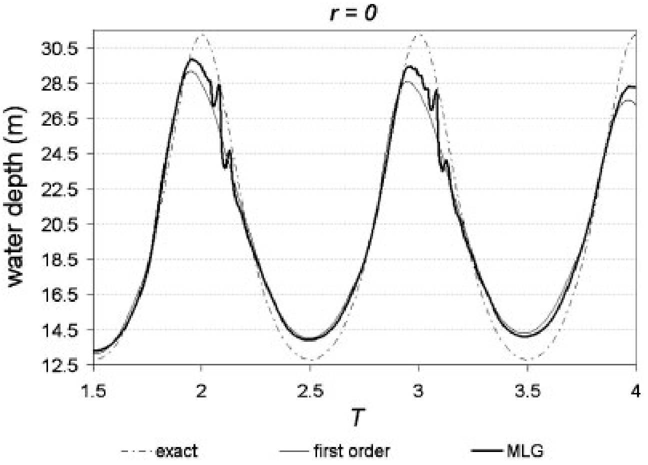

Figure 27. Time evolution of the water depth at the central point of the basin. The exact solution is compared with the results computed with the first-order scheme and the spatial second-order scheme using the MLG limiter.

in the supercritical region. To reveal the difference among first- and second-order approaches, Figure 23 displays the logarithm of the L_1 error in the mesh divided in 14 474 cells. In this case the second-order solution varies less in time than in the previous cases, but also a perfect equilibrium can never be attained. Table III shows the L_1 error obtained for the different meshes.

5.4. Circular dam break with friction

A circular dam break with friction is next presented in a flat squared domain 2000 × 2000 m divided in 32 672 triangular cells, generated by means of a Delaunay mesh solver. The coordinate origin is located at the centre of the domain, and the initial water depth elevation is given by

$$h(t=0)=\begin{cases}0.01 & \text{if } r>800\\ 5 & \text{if } r\leqslant 800\end{cases} \tag{174}$$

DOI: 10.1002/fld

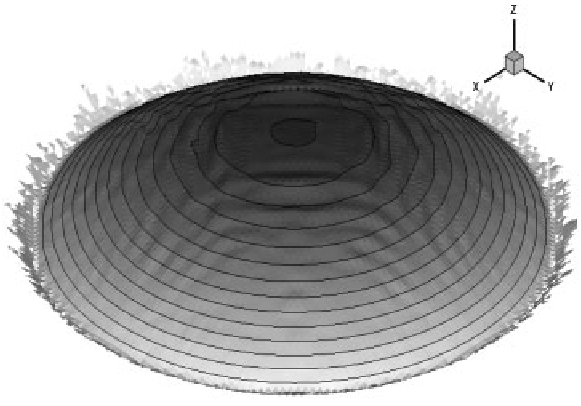

Figure 28. Three-dimensional plot of the water surface elevation at $t = T$ as computed with the second order in space scheme.

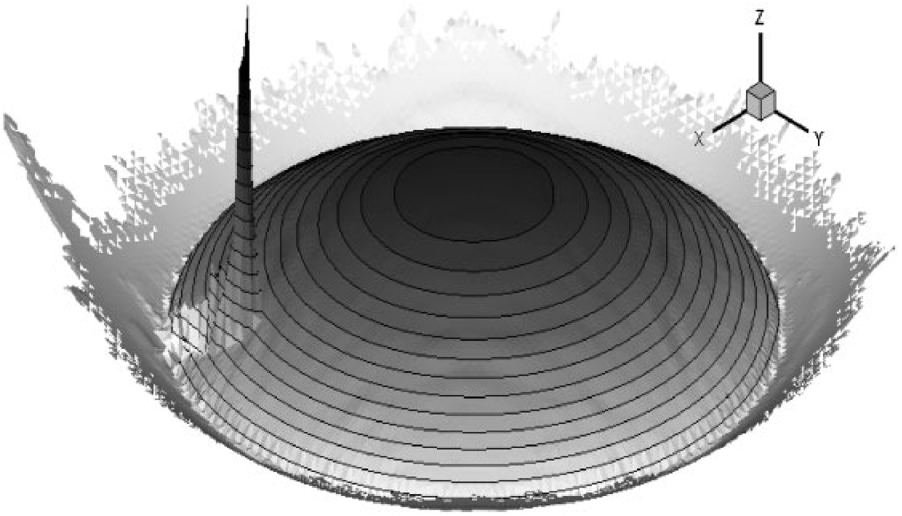

Figure 29. Water surface elevation at $t = 0.95T$ as computed with the second-order scheme in space and time. Destructive oscillations start to appear if the scheme is not reduced to first order when $|\delta h_{JI,k}|>|\delta h_k|$.

with h in meters and where r, in meters, is the radial distance from the centre of the domain considering a Manning's roughness parameter equal to $n = 0.04$.

Figure 24 displays the time step size for three different discretization techniques of the source term. It is remarkable that for explicit discretization, upwind and pointwise, the limits in the time step size in (166) and (167) must be imposed, respectively, otherwise the computation blows up at approximately $t = 5$ s. The implicit pointwise discretization does not involve any restrictions over the time step size, so it is suitable to be combined with the upwind scheme extended to values of CFL greater than one [9]. Figure 25 displays a three-dimensional view of the water depth

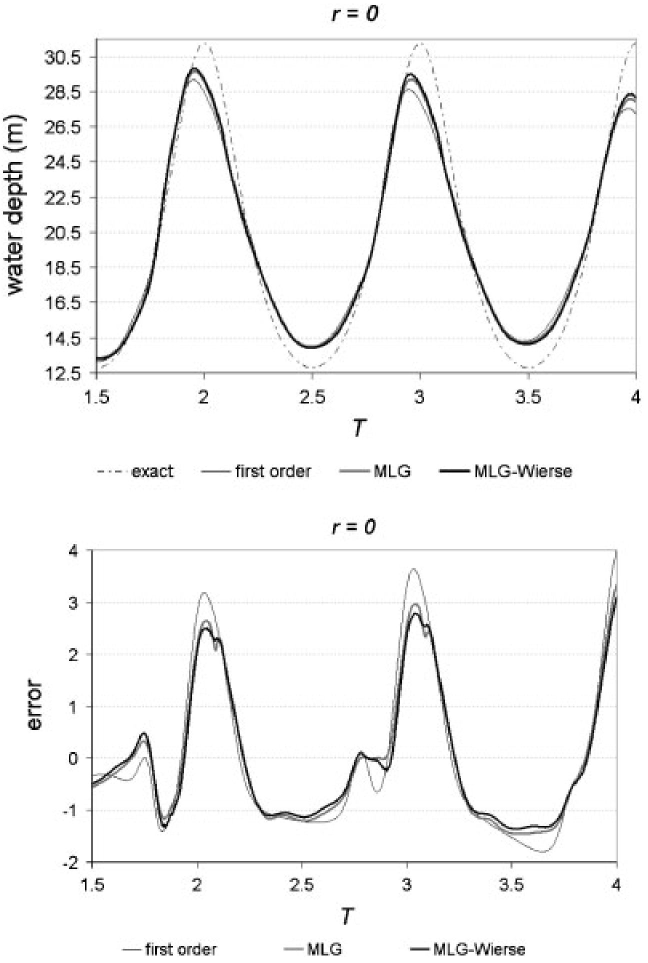

Figure 30. Water depth at the central point r = 0, exact and computed with first order, MLG, MLG-Wierse.

when using the upwind explicit discretization (a), the pointwise discretization (b), the pointwise implicit with $\mathrm{CFL}=1$ (c) and the pointwise implicit with $\mathrm{CFL}=2$ (d) at time $t=40\,\mathrm{s}$. The less diffusive solution is provided using the pointwise implicit discretization with $\mathrm{CFL}=2$, requiring the minimum computational effort.

5.5. Long wave resonance in a circular parabolic frictionless basin

The analytical solution of a long wave resonating in a circular parabolic basin was presented by Thacker [26] for the shallow water equations, where the free surface displacement is given by

$$\zeta(r,t)=\zeta_0\left(\frac{(1-A^2)^{1/2}}{1-A\cos\omega t}-1-\frac{r^2}{a^2}\left\{\frac{1-A^2}{(1-A\cos\omega t)^2}-1\right\}\right) \tag{175}$$

DOI: 10.1002/fld

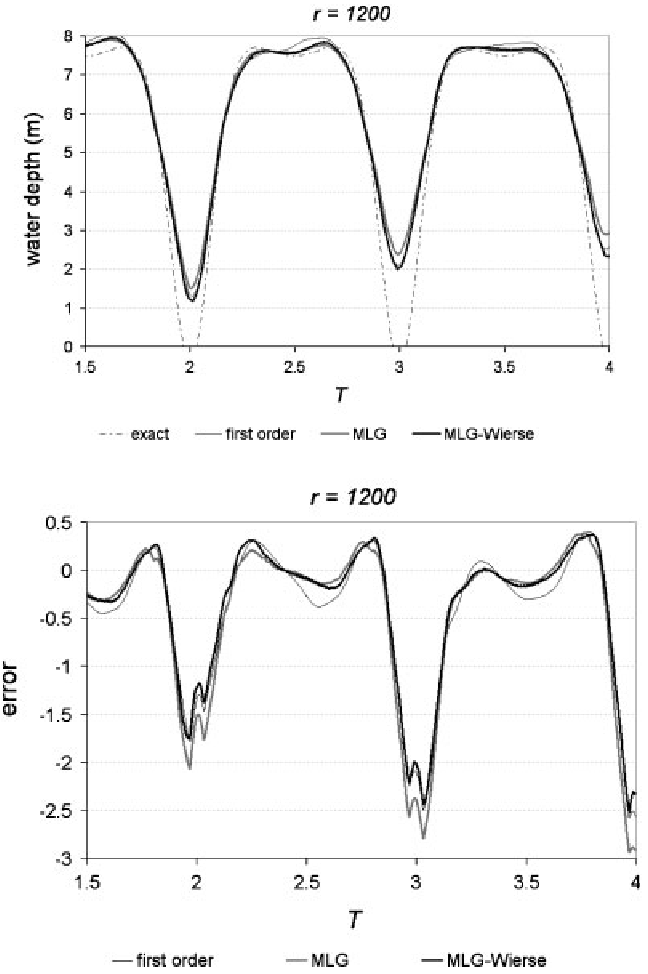

Figure 31. Water depth at the central point r = 1200, exact and computed with first order, MLG, MLG-Wierse.

and the basin shape is given as

$$z(r,t) = -\zeta_0\left(1-\frac{r^2}{a^2}\right) \tag{176}$$

with $A=(a^4-r_0^4)(a^4+r_0^4)$ and $\omega = a^{-1}\sqrt{8g\zeta_0}$, where ζ_0 is the centre point water depth, r is the distance from the centre point, a is the radial distance from the centre point to the zero elevation on the shoreline and r_0 is the distance from the centre point to the point where the water depth is initially nil. Those values are represented in Figure 26. No bottom friction is considered in this case. The numerical values used for this test are $\zeta_0 = 20.0\,\text{m}$, $r_0 = 1200\,\text{m}$, $a = 1500\,\text{m}$. The domain is divided in triangular cells with $l = 25\,\text{m}$ generated using the discretization shown in Figure 5(b). This test case illustrates the wetting/drying fronts and the generation of dry regions from wet areas.

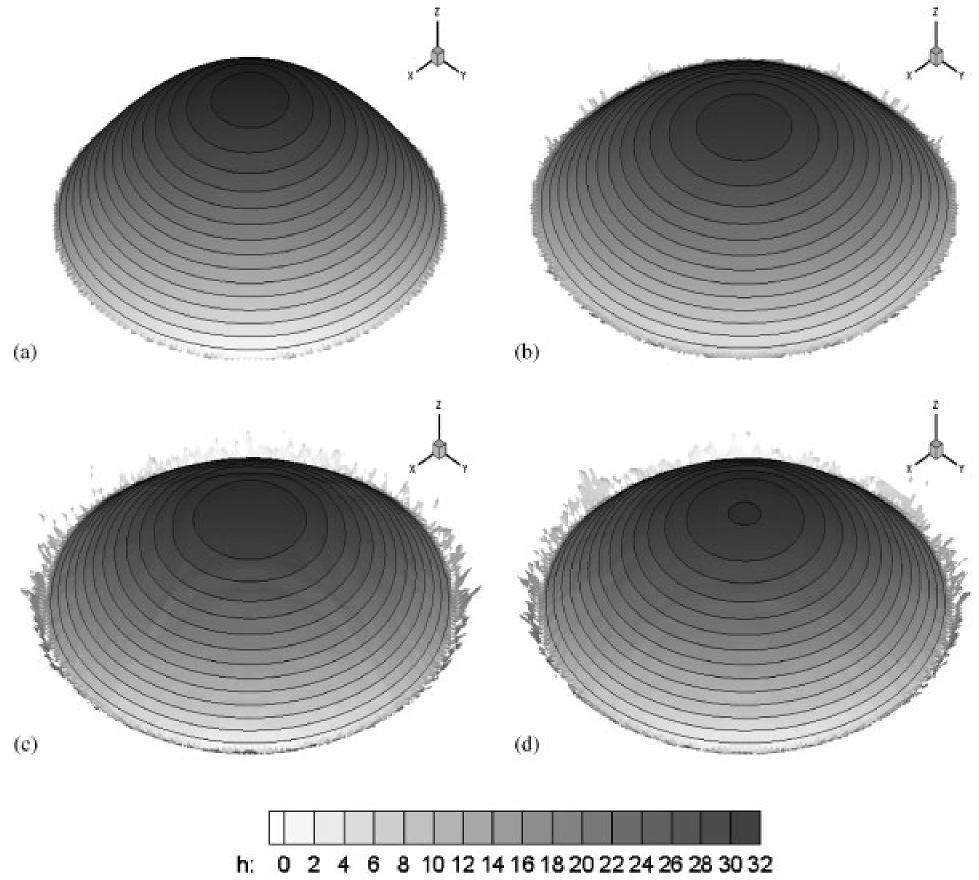

Figure 32. Three-dimensional view for the exact water level surface (a) at $4T$, using first-order approximation (b), second order with MLG (c) and second order with MLG-Wierse at $4T$ (d).

In particular, the advance of the wetting/drying front is produced in the first half period, during the wave expansion, while during the wave contraction both wetting/drying fronts and drying process are present. The numerical treatment at the wet/dry fronts follows the methodology described in Murillo *et al.* [8]. A strategy of conservative redistribution of the cell-updating information has been applied in this kind of problems to avoid the required extreme reduction of the time step [8]. Figure 27 shows the time evolution of the water depth at the central point of the basin where the exact solution is compared with the results computed with the first-order scheme and the spatial second-order scheme using the MLG limiter in order to display the stable but oscillatory solution provided by the latter. Figure 28 is a three-dimensional plot of the water surface elevation at $t=T$ as computed with the second order in space scheme to emphasize the bad quality of the solutions provided by this approach in some cases. Figure 29 is a plot of the water surface elevation at $t=0.95T$ as computed with the second-order scheme in space and time. Even though this scheme is able to cure the oscillations of the second order in space method, destructive oscillations start to appear if the scheme is not reduced to first order when $|\delta h_{JI,k}|>|\delta h_k|$.

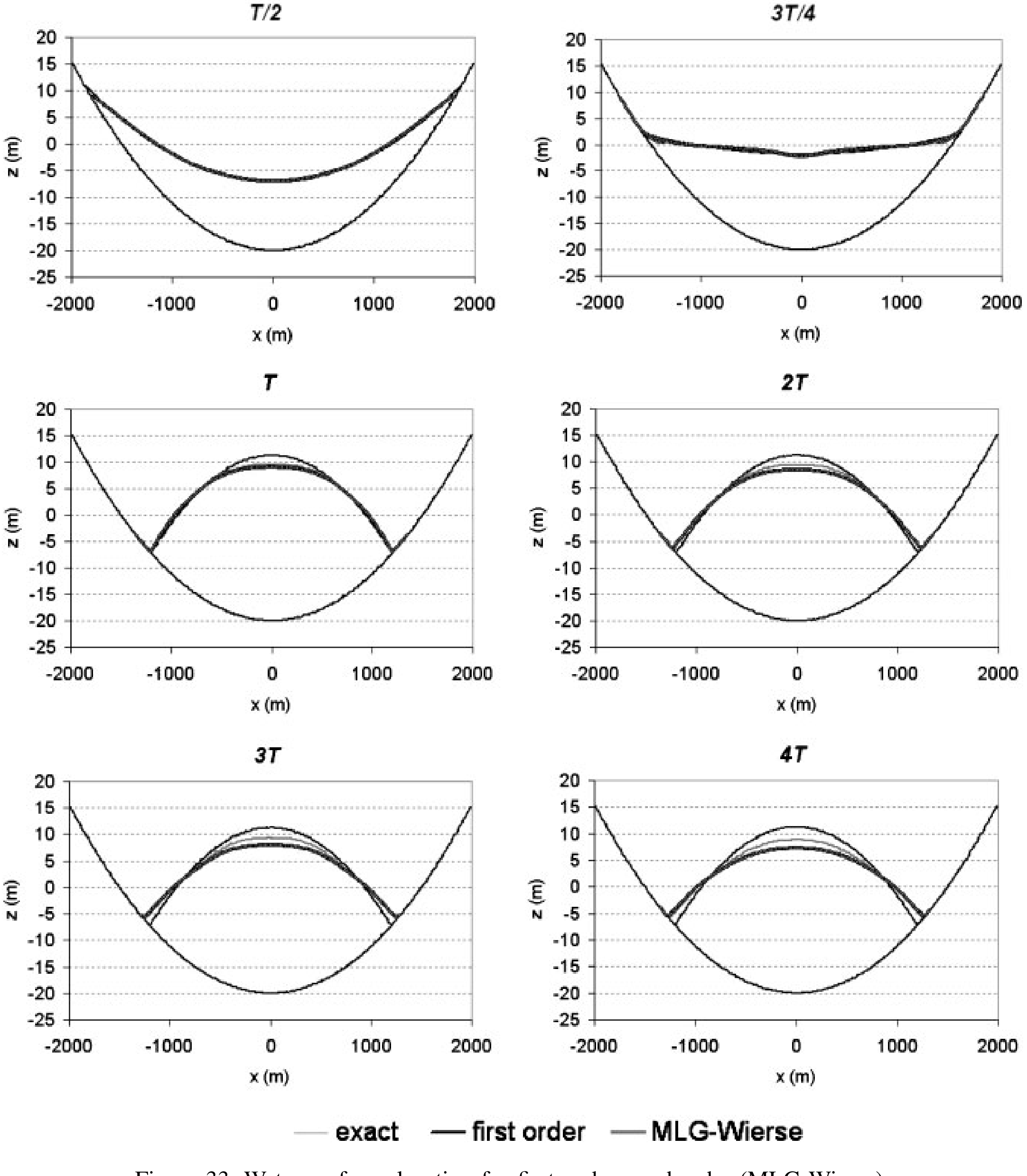

Figure 33. Water surface elevation for first and second order (MLG-Wierse).

Figure 30 (top) displays the water depth at the central point ($r = 0$) for first- and second-order approximations using the MLG and the MLG-Wierse techniques while Figure 30 (bottom) shows the error computed as the difference between the exact and the computed solutions, showing how the MLG-Wierse technique provides the most accurate results. In Figure 31 (top) the water depth at r_0 for first- and second-order approximations using the MLG and the MLG-Wierse techniques is presented, whereas Figure 31 (bottom) shows the error as defined in the previous figure, showing how the MLG-Wierse technique provides the most accurate results. In this case the MLG technique is remarkably less accurate than first-order approach when the water depth recovers the initial position.

Figure 32(a) shows a three-dimensional view for the exact water level surface at $4T$, using first-order approximation (b), second order with MLG (c) and second order with MLG-Wierse showing how in all cases the numerical model conserves perfectly the symmetrical shape of the initial water distribution. The differences between first- and second-order approximation (MLG-Wierse) can be observed for the times $T, 2T, 3T$ and $4T$ in Figure 33.

6. CONCLUSIONS

A general formulation of finite volume upwind schemes on triangular grids has been presented in order to study the properties of first order, spatial second-order and spatial and temporal second-order approximations. The main focus of the work has been the relative performance of these schemes as applied to conservation laws with source terms and the influence of the latter on the numerical stability conditions.

Spatial second-order accuracy has been built using MLG and MLG-Wierse limited cell gradient methods (MUSCL). The second has proved the most efficient in all cases but especially in homogeneous problems such as the linear advection and the oblique hydraulic jump test cases. Second order in time and space accuracy has been implemented by means of a two-step MUSCL-Hancock scheme. In cases of transient flow, the MUSCL-Hancock scheme is not only more accurate, but able to eliminate oscillatory solutions that otherwise appear. This has also been shown in the linear advection test case.

The presence of source terms in the equations, both in scalar case and for systems of equations, has been taken into account in order to derive a systematic series of restrictions over the time step compatible with numerical stability that are reduced to the CFL condition in the homogeneous case. This analysis helps to understand the form in which the source terms get involved in the time step restrictions and identifies the cases in which the time step must be further reduced due to the source terms.

The careful formulation of the second-order spatial representation in presence of source terms has led to a twofold conclusion: first, that an exact balance at steady state can only be achieved if the method reduces automatically to first order and, second, that linear reconstruction has to be made over a secondary variable that combines the conserved variables and the source terms. This has been shown in the inviscid Burgers' equation with source term test case.

The same can be concluded when solving the shallow water system of equations with source terms. In this particular case, it has been shown that the only form to achieve a perfect discrete balance in the still water steady state when using a second-order scheme is to replace water depth by water surface level in the interpolated set of variables. In general, it cannot be concluded that second-order approximations lead to better quality steady-state solutions.

The restrictions in time step associated specifically to the friction source terms in the shallow water equations have been studied separately. It can be concluded that the unified upwind discretization of these terms, although accurate and conservative, is overly restrictive and can lead to excessively small time steps in transient calculations. For these cases, the option of

DOI: 10.1002/fld

a pointwise implicit discretization of the friction source terms has proved to be the most efficient strategy.

In complex cases of transient shallow water flow with source terms and wet/dry fronts there are several reasons that can force the reduction of the time step size in order to preserve numerical stability and monotonicity. A strategy of conservative redistribution of the cell-updating information has been applied in this kind of problems to avoid the required extreme reduction of the time step. The test case of long wave resonance in a parabolic basin has been useful to see that, even though second-order schemes are not able to produce a better quality steady-state solution, they are superior in transient calculations. At the same time, the proposed conservative redistribution strategies have been proved valid and useful.

APPENDIX A: CONSTRUCTION OF NON-OSCILLATORY PIECEWISE LINEAR L FUNCTIONS

In Batten *et al.* [17] a new algorithm, the maximum limited gradient (MLG) approach, was presented to define reconstruction second order in space functions. It was based on the limited central difference (LCD) approach and the compressive limiter of Durlofsky. The LCD approach generates a reconstruction function formulated with a matrix $\mathbf{L}^1$ computed at a cell i using the information stored in the three immediate neighbour cells, $\mathbf{L}^1 = \nabla(123)$. In order to enforce condition (45) $\mathbf{L}^1$ is limited by

$$\mathbf{L}^1 = \alpha\mathbf{L}^1, \quad 0\leqslant\alpha\leqslant 1 \tag{A1}$$

where the coefficient α is a damping factor maximized to avoid overshoots or undershoots at the edge mid-points, with the following rule:

$$\alpha = \begin{cases} (u_k^{\max} - u_{i,0})/\delta u_{JI,k} & \text{if } u_{I,k} > u_k^{\max} \\ (u_k^{\min} - u_{i,0})/\delta u_{JI,k} & \text{if } u_{I,k} < u_k^{\min} \\ 1 & \text{otherwise} \end{cases} \tag{A2}$$

for $k = 1, 2, 3$. In the Durlofsky technique [27] three different gradient planes, $\mathbf{L}^2 = \nabla(12i)$, $\mathbf{L}^3 = \nabla(13i)$, $\mathbf{L}^4 = \nabla(23i)$ participate. The compressive limiter suggested by Durlofsky *et al.* [27] selects the gradient matrix with maximum $|\mathbf{L}^s|$ that fulfils (45), otherwise first order is imposed. The MLG algorithm [17] is a combination of the two previous slope limiter functions, where every gradient operator $\mathbf{L}^s$ $(s = 1, 4)$ is limited according to (A2) and the $\mathbf{L}^s$ for which $|\mathbf{L}^s|$ is maximum is finally used.

This method is extended including an extra limiter function presented by Wierse [28] and will be referred to as the MLG-Wierse limiter. Despite being more restrictive, this algorithm proves to be the best option, as results in an optimum evaluation of the updating pointwise part. The algorithm is:

(i) Construction of the operators $\mathbf{L}^s$ $(s = 1, 4)$.

DOI: 10.1002/fld

(ii) For each $\mathbf{L}^s$, for $k=1,2,3$ redefine the value of $\delta d^s_{Ii,k} = r_{i,k}\mathbf{L}^s$ as:

$$\delta d^s_{Ii,k} = \begin{cases} \delta d^s_{Ii,k} & \text{if } \delta d^s_{Ii,k} \cdot \delta d^s_k > 0 \\ 0 & \text{otherwise} \end{cases} \tag{A3}$$

(a) If $\delta d^s_{Ii,k} = 0$ only for one k (denoted by k_1) and $\delta d^s_{Ii,k_2} \cdot \delta\, d^s_{Ii,k_3} < 0$

$$\begin{aligned} \delta d^s_{Ii,k_2} &= \text{sign}(\delta d^s_{Ii,k_2}) \cdot \text{minvalue} \\ \delta d^s_{Ii,k_3} &= \text{sign}(\delta d^s_{Ii,k_3}) \cdot \text{minvalue} \end{aligned} \tag{A4}$$

where minvalue $= \min(|\delta d^s_{Ii,k_2}|, |\delta d^s_{Ii,k_3}|)$, and reconstruct $\mathbf{L}^s$.

(b) If $\delta d^s_{Ii,k} = 0$ only for one k (denoted by k_1) and $\delta d^s_{Ii,k_2} \cdot \delta d^s_{Ii,k_3} \geqslant 0$ set $\mathbf{L}^s = \mathbf{0}$.

(c) If $\delta d^s_{Ii,k} = 0$ for two or more k's set $\mathbf{L}^s = \mathbf{0}$.

(iii) Limit each $\mathbf{L}^s$ as in (A1).

(iv) Selection of the $\mathbf{L}^s$ for which $|\mathbf{L}^s|$ is maximum.

REFERENCES

1. Kawahara M, Umetsu T. Finite element method for moving boundary problems in river flow. *International Journal for Numerical Methods in Fluids* 1986; **6**:365–386.
2. Zhao DH, Shen HW, Tabios GQ, Lai JS, Tan WY. Finite volume two-dimensional unsteady flow model for river basins. *Journal of Hydraulic Engineering* 1994; **120**(7):863–883.
3. Khan AA. Modeling flow over an initially dry bed. *Journal of Hydraulic Research* 2000; **38**(5):383–389.
4. Heniche M, Secretan Y, Boudreau P, Leclerc M. A two-dimensional finite element drying-wetting shallow water model for rivers and estuaries. *Advances in Water Research* 2000; **23**:359–372.
5. Bradford SF, Sanders BF. Finite-volume model for shallow water flooding of arbitrary topography. *Journal of Hydraulic Engineering* 2002; **128**(3):289–298.
6. Audusse E, Bristeau MO. A well-balanced positivity preserving 'second-order' scheme for shallow water flows on unstructured meshes. *Journal of Computational Physics* 2005; **206**(1):311–333.
7. Begnudelli L, Sanders F. Unstructured grid finite-volume algorithm for shallow-water flow and scalar transport with wetting and drying. *Journal of Hydraulic Engineering* 2006; **132**(4):371–384.
8. Murillo J, García-Navarro P, Burguete J, Brufau P. A conservative 2d model of inundation flow with solute transport over dry bed. *International Journal for Numerical Methods in Fluids*. Published online in www.interscience.wiley.com, 2006.
9. Murillo J, García-Navarro P, Brufau P, Burguete J. Extension of an explicit finite volume method to large time steps (CFL>1): application to shallow water flows. *International Journal for Numerical Methods in Fluids* 2005; **50**:63–102.
10. Courant R, Isaacson E, Rees M. On the solution of nonlinear hyperbolic differential equations by finite differences. *Communications on Pure and Applied Mathematics* 1952; **5**:243–255.
11. Godlewsky E, Raviart PA. *Numerical Approximation of Hyperbolic Systems of Conservation Laws*. Springer: Berlin, 1996.
12. Hubbard ME, García-Navarro P. Flux difference splitting and the balancing of source terms and flux gradients. *Journal of Computational Physics* 2000; **165**:89–125.
13. Roe PL. *A Basis for Upwind Differencing of the Two-Dimensional Unsteady Euler Equations*. Numerical methods in fluid dynamics, vol. II. Oxford University Press: Oxford, 1986.
14. Burguete J, García-Navarro P. Efficient construction of high-resolution TVD conservative schemes for equations with source terms: application to shallow water flows. *International Journal for Numerical Methods in Fluids* 2001; **37**:209–248.
15. Van Leer B. On the relation between the upwind differencing schemes of Godunov, Engquist-Osher and Roe. *SIAM Journal on Scientific and Statistical Computing* 1985; **5**(1):1–20.

DOI: 10.1002/fld

16. Toro EF. *Riemann Solvers and Numerical Methods for Fluid Dynamics*. Springer: Berlin, 1997.
17. Batten P, Lambert C, Causon DM. Positively conservative high-resolution convection schemes for unstructured elements. *International Journal for Numerical Methods in Engineering* 1996; **39**:1821–1838.
18. Vázquez-Cendón ME. Improved treatment of source terms in upwind schemes for the shallow water equations in channels with irregular geometry. *Journal of Computational Physics* 1999; **148**(2):497–526.
19. Alcrudo F, García-Navarro P. A high resolution Godunov-type scheme in finite volumes for the 2D shallow water equations. *International Journal for Numerical Methods in Fluids* 1993; **16**(6):489–505.
20. Brufau P, Vázquez-Cendón ME, García-Navarro P. A numerical model for the flooding and drying of irregular domains. *International Journal for Numerical Methods in Fluids* 2002; **39**:247–275.
21. Brufau P, García-Navarro P, Vázquez-Cendón ME. Zero mass error using unsteady wetting/drying conditions in shallow flows over dry irregular topography. *International Journal for Numerical Methods in Fluids* 2004; **45**:1047–1082.
22. Chow VT. *Open Channel Flow*. McGraw-Hill: New York, 1959.
23. Cunge JA, Holly FM, Vervey A. *Practical Aspects of Computational River Hydraulics*. Pitman: London, 1980.
24. Bermúdez A, Vázquez-Cendón ME. Upwind methods for hyperbolic conservation laws with source terms. *Computers and Fluids* 1998; **8**:1049–1071.
25. Burguete J, García-Navarro P, Murillo J. Analysis of the friction term in one-dimensional shallow water model: application to open channel and river flow. *Journal of Hydraulic Engineering* 2006 (under revision).
26. Thacker WC. Some exact solutions to the non linear shallow water equations. *Journal of Fluid Mechanics* 1981; **107**:499–508.
27. Durlofsky LJ, Engquist B, Osher S. Triangle based adapted stencils for the solution of hyperbolic conservation laws. *Journal of Computational Physics* 1992; **98**:64–73.
28. Wierse M. A new theoretically motivated higher order upwind scheme on unstructured grids of simplices. *Advances in Computational Mathematics* 1997; **7**:303–335.

DOI: 10.1002/fld