Upwind Schemes for the Two-Dimensional Shallow Water Equations with Variable Depth Using Unstructured Meshes

Alfredo Bermúdez, Alain Dervieux, Jean-Antoine Desideri, Maria Elena Vázquez

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Upwind schemes for the two-dimensional shallow water equations with variable depth using unstructured meshes

A. Bermúdez, A. Dervieux, J.A. Désidéri & M.E. Vázquez

N° 2738

Décembre 1995

PROGRAMME 6

ISSN 0249-6399

INRIA
SOPHIA ANTIPOLIS

Upwind schemes for the two-dimensional shallow water equations with variable depth using unstructured meshes

A. Bermúdez*, A. Dervieux**, J.A. Désidéri** & M.E. Vázquez*

Programme 6 — Calcul scientifique, modélisation et logiciel numérique
Projet Sinus

Rapport de recherche n° 2738 — Décembre 1995 — 50 pages

Abstract: In this paper, certain well-known upwind schemes for hyperbolic equations are extended to solve the two-dimensional Saint-Venant (or shallow water) equations. We consider unstructured meshes and a new type of finite volume to obtain a suitable treatment of the boundary conditions. The source term involving the gradient of the depth is upwinded in a similar way as the flux terms. The resulting schemes are compared in terms of a *conservation property*. For the time discretization we consider both explicit and implicit schemes. Finally we present the numerical results for tidal flows in the Pontevedra ria, Galicia, Spain.

Key-words: Shallow water equations, variable depth, upwind schemes, unstructured meshes

(Résumé : tsvp)

*Departamento de Matematica Aplicada, 15706 Santiago de Compostela, Spain
**INRIA, 2004 Route des Lucioles, BP 93, F-06902 Sophia Antipolis cedex, France

Schémas décentrés en maillages non-structurés pour les équations de Saint-Venant bidimensionnelles avec profondeur variable

Résumé : Ce rapport présente l'extension de plusieurs méthodes de décomposition de flux à l'approximation des équations de Saint-Venant. Nous considérons des maillages non-structurés et un nouveau type de volumes finis bien adaptés au traitement des conditions aux bords. Le terme source contenant le gradient de la profondeur est décentré de manière consistante au traitement des flux; les schémas résultants sont comparés vis à vis d'une *propriété de conservation*. En ce qui concerne la discrétisation en temps, des schémas explicites et implicites sont considérés. On présente finalement des résultats numériques de calculs de marées dans la Ria de Pontevedra (Galice, Espagne).

Mots-clé : Equations de Saint-Venant, profondeur variable, schémas décentrés, maillages non structurés

Contents

1 Shallow water equations **5**

2 Discretization **8**

2.1 Finite volume of the edge-type 8

2.2 Flux discretization. 10

2.2.1 Q-schemes 11

2.2.2 Flux splitting techniques 12

2.2.3 Numerical results 14

2.3 Source term discretization 17

2.4 A conservation property 20

2.5 Discretization of the boundary conditions 22

2.6 Numerical Results 24

3 Implicit discretizations. **31**

3.1 Implicit discretizations of 1D problems 31

3.2 A simplified linearized implicit scheme for the 2D shallow water equations 35

3.3 Discretization of the boundary conditions 37

3.4 A conservation property 38

3.5 Numerical results 39

4 Implicit discretizations. **40**

4.1 Implicit discretizations of 1D problems 40

4.2 A simplified linearized implicit scheme for the 2D shallow water equations 44

4.3 Discretization of the boundary conditions 46

4.4 A conservation property 47

4.5 Numerical results 48

!r inriarap/intro.tex

1 Shallow water equations

The shallow water equations are frequently used as a mathematical model for water flow in coastal areas, lakes, estuaries, etc. Thus they are an important tool to simulate a variety of problems related to coastal engineering, environment, ecology, etc. (see Gambolati *et al.* [6]).

These equations can be obtained by integrating the incompressible Euler equations in depth and then taking into account the kinematic and kinetic boundary conditions on both the free and the bottom surfaces. For the sake of simplicity we consider neither wind stress nor Coriolis effect nor bottom friction. Then we get the following generalized conservation law (see for instance Stoker [12]):

$$\left\{ \begin{array}{rcl} \dfrac{\partial w}{\partial t}(x,y,t) + \dfrac{\partial F_1}{\partial x}(w(x,y,t)) + \dfrac{\partial F_2}{\partial y}(w(x,y,t)) & = & G(x,y,w(x,y,t)) \\ & & \\ (x,y) \in \Omega \subset I\!R^2 \qquad t \in [0,T] & & \end{array} \right. \tag{1}$$

$$w = \begin{pmatrix} h \\ hu_1 \\ hu_2 \end{pmatrix} = \begin{pmatrix} h \\ q_1 \\ q_2 \end{pmatrix}, \quad G(x,y,w) = \begin{pmatrix} 0 \\ gh\dfrac{\partial H}{\partial x}(x,y) \\ gh\dfrac{\partial H}{\partial y}(x,y) \end{pmatrix}$$

$$F_1(w) = \begin{pmatrix} q_1 \\ \dfrac{q_1^2}{h} + \dfrac{1}{2}gh^2 \\ \dfrac{q_1 q_2}{h} \end{pmatrix}, \quad F_2(w) = \begin{pmatrix} q_2 \\ \dfrac{q_1 q_2}{h} \\ \dfrac{q_2^2}{h} + \dfrac{1}{2}gh^2 \end{pmatrix}.$$

where $h(x,y,t)$ denotes the height of the fluid at point (x,y) at time t and $H(x,y)$ the depth of the same point but from a fixed reference level. The vector field (u_1,u_2) is the averaged horizontal velocity. Finally, the conservative variable q is given by $(q_1,q_2) = (hu_1, hu_2)$ and Ω denotes the projection of the domain occupied by the fluid onto the xy plane.

If we use a symbolic notation for the vectors $(F_i(w))_{i=1,2}$ in equation (1),

$$\mathcal{F}(w) = (F_1(w), F_2(w)), \tag{2}$$

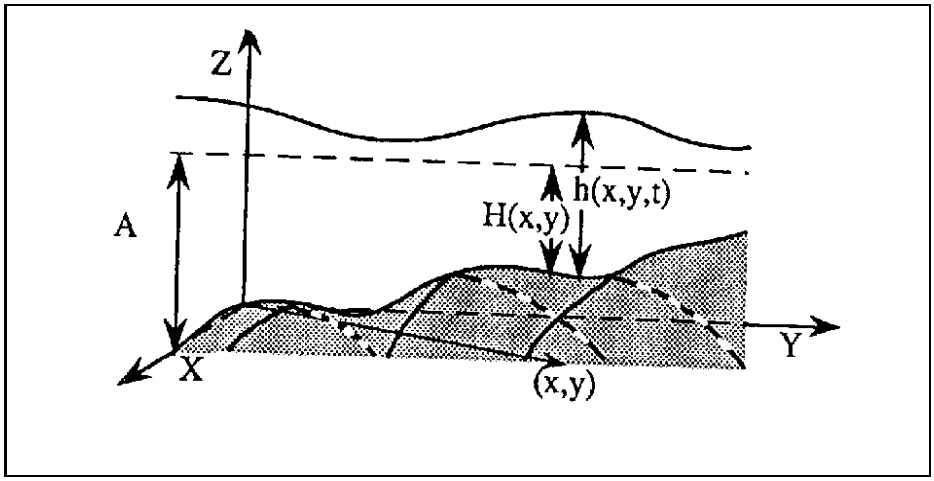

Figure 1: Shallow domain

then the conservative system becomes

$$\frac{\partial w}{\partial t} + \nabla \cdot \mathcal{F}(w) = G(x, y, w). \tag{3}$$

The boundary conditions consist of a slip condition for the "coast" boundary and the specification of the height (h) of the fluid on a "open sea" boundary.

Non-homogeneity of the flux

If h does not vanish, the system of partial differential equations is strictly hyperbolic. Indeed, for any (α, β) and any $w \in \Omega$, the matrix

$$\mathcal{A}(w, (\alpha, \beta)) = \alpha \mathcal{A}_1(w) + \beta \mathcal{A}_2(w),$$

in which $\mathcal{A}_i \, (i = 1, 2)$ are the Jacobian matrices of the two components of the flux F_i has 3 distinct real eigenvalues given by:

$$\lambda_1 = \alpha \frac{q_1}{h} + \beta \frac{q_2}{h}, \tag{4}$$

$$\lambda_2 = \alpha \frac{q_1}{h} + \beta \frac{q_2}{h} + ||(\alpha, \beta)|| \sqrt{gh}, \tag{5}$$

$$\lambda_3 = \alpha \frac{q_1}{h} + \beta \frac{q_2}{h} - ||(\alpha, \beta)|| \sqrt{gh}. \tag{6}$$

In what follows, upwind techniques combined with finite volume discretizations are applied to approximate the solution of (3). As is well known, many flux-vector and flux-difference splitting techniques have been developed to solve the Euler equations for aerodynamics. Some of them use the assumption of a homogeneous flux function.

Unfortunately the functions F_i given by (2) do not satisfy this property as it can be easily deduced from the expression of their respective Jacobian matrices

$$\mathcal{A}_1(w) = \begin{pmatrix} 0 & 1 & 0 \\ -\frac{q_1^2}{h^2} + gh & 2\frac{q_1}{h} & 0 \\ -\frac{q_1 q_2}{h^2} & \frac{q_2}{h} & \frac{q_1}{h} \end{pmatrix}, \quad \mathcal{A}_2(w) = \begin{pmatrix} 0 & 0 & 1 \\ -\frac{q_1 q_2}{h^2} & \frac{q_2}{h} & \frac{q_1}{h} \\ -\frac{q_2^2}{h^2} + gh & 0 & 2\frac{q_2}{h} \end{pmatrix}. \tag{7}$$

Thus in this paper we introduce two other matrices $\mathcal{A}_i^\star$ $(i = 1, 2)$ such that

$$F_i(W) = \mathcal{A}_i^\star(W)W \quad (i = 1, 2), \tag{8}$$

to apply the mentioned techniques as well as to get linearizations of the implicit schemes.

The expressions of these matrices for the shallow water equations are the following

$$\mathcal{A}_1^\star(w) = \begin{pmatrix} 0 & 1 & 0 \\ -\frac{q_1^2}{h^2} + \frac{1}{2}gh & 2\frac{q_1}{h} & 0 \\ -\frac{q_1 q_2}{h^2} & \frac{q_2}{h} & \frac{q_1}{h} \end{pmatrix}, \quad \mathcal{A}_2^\star(w) = \begin{pmatrix} 0 & 0 & 1 \\ -\frac{q_1 q_2}{h^2} & \frac{q_2}{h} & \frac{q_1}{h} \\ -\frac{q_2^2}{h^2} + \frac{1}{2}gh & 0 & 2\frac{q_2}{h} \end{pmatrix} \tag{9}$$

and the eigenvalues of matrix $\mathcal{A}^\star = \alpha\mathcal{A}_1^\star + \beta\mathcal{A}_2^\star$ are

$$\lambda_1^\star = \alpha\frac{q_1}{h} + \beta\frac{q_2}{h}, \tag{10}$$

$$\lambda_2^\star = \alpha\frac{q_1}{h} + \beta\frac{q_2}{h} + ||(\alpha, \beta)||\sqrt{\frac{gh}{2}}, \tag{11}$$

$$\lambda_3^\star = \alpha\frac{q_1}{h} + \beta\frac{q_2}{h} - ||(\alpha, \beta)||\sqrt{\frac{gh}{2}}. \tag{12}$$

2 Discretization

In this section we introduce an explicit upwind discretization by using finite volume techniques and Riemann solvers.

2.1 Finite volume of the edge-type

We introduce a set of degrees of freedom relying on mid-edges. We note that such degrees of freedom have been considered by Crouzeix and Raviart in the context of non-conforming P_1 finite-element [3].

We assume the computational domain Ω to be a polygonal bounded domain of $\mathbb{R}^2$. Let $\mathcal{T}_h$ be a standard finite element triangulation of Ω and h, as usually, the maximal length of the sides in $\mathcal{T}_h$.

The domains we deal with may have nonsmooth boundaries. The correct treatment of the boundary conditions has led us to introduce a new type of finite volumes, the centers of which are the midpoints of the edges of triangles N_i $(i = 1, \cdots, n_h)$. This is why we call them *finite volumes of the edge-type*. We define cell C_i as follows:

- Every triangle is subdivided in 6 subtriangles by means of its medians.
- The cell C_i is then defined to be the union of the resulting subtriangles having N_i as a vertex.

We also introduce the following definitions:

- $\mathcal{K}_i$ is the set of the indices of neighbouring nodes of N_i.
- The "edge" Γ_{ij} is the cell interface between the cells C_i and C_j (see Figure 2).
- If node i belongs to the boundary, then the "boundary edge" Γ_{i_F} is the side of the triangle to which node N_i belongs.
- η_{ij} is the outward normal vector to Γ_{ij} having the same length as Γ_{ij} and $\tilde{\eta}_{ij} = \dfrac{1}{||\ \eta_{ij}||}\ \eta_{ij}$. For the boundary edge Γ_{i_F}, we denote by η_{i_F} the corresponding outward normal vector.
- The *subcell* T_{ij} is defined to be the subtriangle $N_i \overset{\triangle}{I} P L$, (see Figure 2). Its area is

$$A_{T_{ij}} = \frac{\|\eta_{ij}\| d_{ij}}{2}, \tag{13}$$

where d_{ij} is its height.

- Finally, A_i denotes the area of the cell C_i and $\Gamma_i = \cup_{j \in \mathcal{K}_i} \Gamma_{ij}$ its boundary.

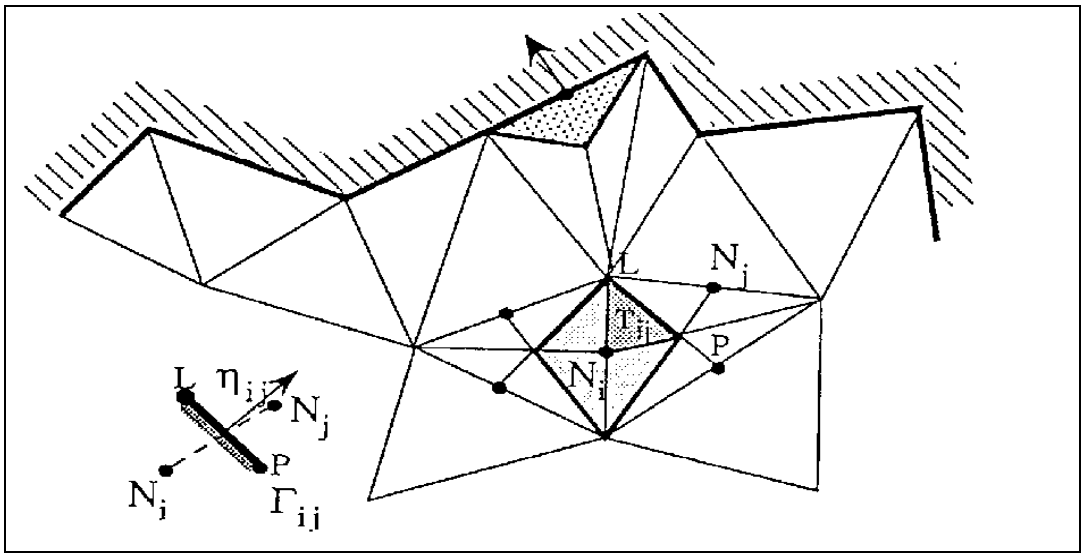

Figure 2: Finite volume of the edge-type

Remark that to define the "dual finite-volume mesh" we do not need to impose restrictions over the finite element triangulation.

The union of all these control volumes $\mathcal{C}_h$ is a partition of the domain Ω. Related to this partition, we consider the following discrete space

$$\mathcal{W}_h = \left\{ W_h : W_{h|C_i} = constant, \quad \forall i = 1, \cdots, n_h \right\}.$$

We start by considering the explicit Euler method for time discretization

$$\frac{W^{n+1}(x,y) - W^n(x,y)}{\Delta t} + \nabla \cdot \mathcal{F}(W^n(x,y)) = G(x,y,W^n(x,y)). \tag{14}$$

where W^n is an approximation of the exact solution $w(.,.,t_n)$. In Section 4 we will use an implicit time discretization.

The two-dimensional extension of the class of one-dimensional three-point upwind first-order accurate schemes is done as follows (see Dervieux and Desideri [4]):
First we integrate (14) over the cell C_i

$$\int\int_{C_i} \frac{W^{n+1}(x,y) - W^n(x,y)}{\Delta t} dx\, dy + \int\int_{C_i} \nabla \cdot \mathcal{F}(W^n(x,y)) dx\, dy = \int\int_{C_i} G(x,y,W^n(x,y)) dx\, dy, \tag{15}$$

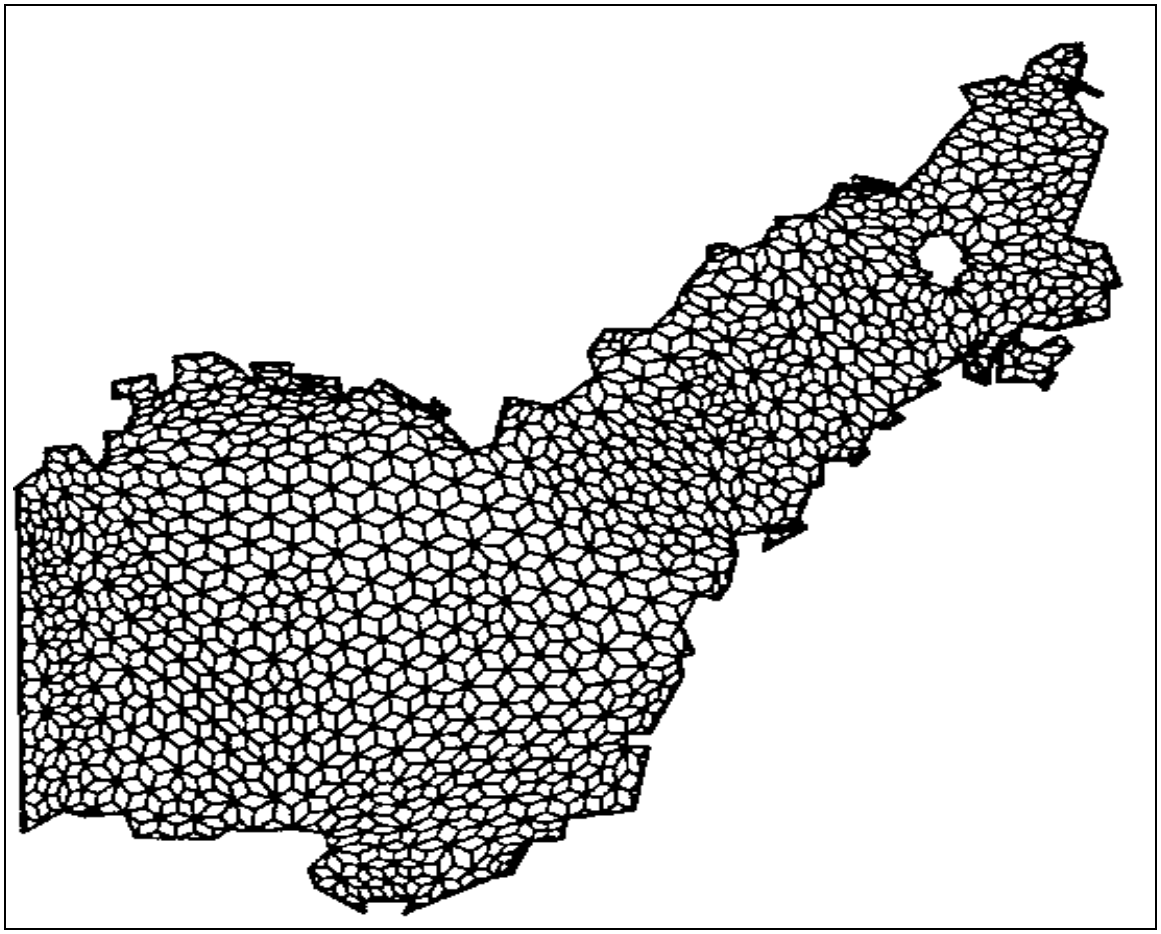

Figure 3: Finite volume of the edge-type mesh for the Pontevedra ria

and then we apply Gauss theorem to the flux term

$$A_i \frac{W_i^{n+1} - W_i^n}{\Delta t} + \int_{\Gamma_i} \mathcal{F}(W^n(x,y)) \cdot \tilde{\eta}_{ij} d\sigma = \int \int_{C_i} G(x,y,W^n(x,y)) dx\, dy, \qquad (16)$$

where W_i^n denotes the value of W^n at node N_i.

2.2 Flux discretization.

To give the approximation of the flux integral, we split the boundary Γ_i of the cell C_i into the cell interfaces Γ_{ij}, where $j \in \mathcal{K}_i$.

$$A_i \frac{W_i^{n+1} - W_i^n}{\Delta t} + \sum_{j \in \mathcal{K}_i} \int_{\Gamma_{ij}} \mathcal{F}(W^n) \cdot \tilde{\eta}_{ij}\, d\sigma = \int \int_{C_i} G(x,y,W^n) dx\, dy. \qquad (17)$$

Notice that if node N_i belongs to the boundary ($\Gamma_{i\mathrm{F}}$), the integral of the normal flux over $\Gamma_{i\mathrm{F}}$ must be added.

We now specify the integral over Γ_{ij}. Actually, the evaluation of this term corresponds to the one-dimensional calculation of the flux along the direction $\overline{N_iN_j}$. For upwinding, we introduce a *numerical flux function* ϕ

$$\int_{\Gamma_{ij}} \mathcal{F}(W)\cdot\tilde{\eta}_{ij}\; d\sigma \;\;\approx\;\; \|\eta_{ij}\|\phi\left(W_i^n, W_j^n, \tilde{\eta}_{ij}\right). \tag{18}$$

Then the approximate system is rewritten by

$$A_i\frac{W_i^{n+1}-W_i^n}{\Delta t}+\sum_{j\in\mathcal{K}_i}\|\eta_{ij}\|\phi\left(W_i^n, W_j^n, \tilde{\eta}_{ij}\right)=\int\int_{C_i} G(x,y,W^n)dx\,dy. \tag{19}$$

The expression of the numerical flux ϕ depends on the upwind scheme. In this paper we will consider the *Q-schemes* of Roe and van Leer and the *flux splitting techniques* of Steger-Warming and Vijayasundaram.

2.2.1 Q-schemes

These schemes are a family of upwind schemes (see van Leer [14]) corresponding to numerical fluxes of the form

$$\phi\left(W_i^n, W_j^n, \tilde{\eta}_{ij}\right)=\frac{Z\left(W_i^n,\tilde{\eta}_{ij}\right)+Z\left(W_j^n,\tilde{\eta}_{ij}\right)}{2}-\frac{1}{2}\left|Q\left(W_{\mathrm{Q}}^n\left(W_i{}^n, W_j^n\right),\tilde{\eta}_{ij}\right)\right|\left(W_j^n-W_i^n\right),$$

where Q is the Jacobian matrix of the function $Z(W,\eta)=F(W)\cdot\eta$ and W_{Q} is defined by

$$W_{\mathrm{Q}}\left(U,V\right)=\begin{cases}\dfrac{U+V}{2} & (Q\text{-scheme of van Leer})\\ \widetilde{W}(U,V) & (\text{Roe scheme})\end{cases} \tag{20}$$

and $\widetilde{W}(U,V)$ denotes the *Roe average* of U and V given by the equation

$$Z\left(V,\eta\right)-Z\left(U,\eta\right)=\mathcal{A}\left(\widetilde{W}(U,V),\eta\right)\left(V-U\right). \tag{21}$$

In [7] Glaister gives $\widetilde{W}$ for the one-dimensional shallow water equations. In this case it is easy to see that this average has the expression

$$\widetilde{W}\left(W_i,W_j\right)=\begin{pmatrix}\widetilde{h}\\ \widetilde{q_1}\\ \widetilde{q_2}\end{pmatrix}=\begin{pmatrix}\sqrt{h_ih_j}\\ \theta\dfrac{q_i^1}{h_i}+(1-\theta)\dfrac{q_j^1}{h_j}\\ \theta\dfrac{q_i^2}{h_i}+(1-\theta)\dfrac{q_j^2}{h_j}\end{pmatrix}, \tag{22}$$

where θ is

$$\theta = \frac{\sqrt{h_i}}{\sqrt{h_j} + \sqrt{h_i}}. \tag{23}$$

2.2.2 Flux splitting techniques

These techniques are also generalizations of the three-point upstream differencing scheme to the homogeneous problems of nonlinear conservation laws (see [10]). The physical flux is split in a *forward flux* F_i^+ and a *backward flux* F_i^-, that is

$$F_i(W) = F_i^+(W) + F_i^-(W) \quad (i = 1, 2), \tag{24}$$

or, in terms of the *normal flux* Z,

$$Z(W, \eta) = Z^+(W, \eta) + Z^-(W, \eta), \tag{25}$$

and then the numerical flux is defined using the above splitting by

$$\phi(U, V, \eta) = \phi^+(U, V, \eta) + \phi^-(U, V, \eta), \tag{26}$$

with

$$\phi^+(U, V, \eta) = \mathcal{B}^+(U, V, \eta)U, \tag{27}$$

$$\phi^-(U, V, \eta) = \mathcal{B}^-(U, V, \eta)V, \tag{28}$$

and $\mathcal{B}^+$ and $\mathcal{B}^-$ are two matrices such that:

i) $\begin{cases} Z^+(W, \eta) = \mathcal{B}^+(W, W, \eta)W, \\ Z^-(W, \eta) = \mathcal{B}^-(W, W, \eta)W. \end{cases}$

ii) $\mathcal{B}^+$ (resp. $\mathcal{B}^-$) only has real positive (resp. negative) eigenvalues.

Note that conditions i) imply consistency with the physical flux.

We consider in particular the flux-splitting schemes of Steger-Warming and Vijayasundaram.

Steger and Warming [11] introduced the notion of flux vector splitting for the equations of gas dynamics. They took advantage of the fact that in gas dynamics Z is a homogeneous function of W of degree one, that is

$$Z(W, \eta) = \mathcal{A}(W, \eta)W, \tag{29}$$

where $\mathcal{A}$ is the Jacobian matrix of the flux Z.

Using the diagonalization of $\mathcal{A}$, we have

$$Z(W,\eta) = \mathcal{A}^+(W,\eta)W + \mathcal{A}^-(W,\eta)W. \tag{30}$$

Then, the flux-splitting of Steger and Warming corresponds to the following choice of matrices $\mathcal{B}^+$ and $\mathcal{B}^-$

$$\mathcal{B}^+(U,V,\eta) = \mathcal{A}^+(U,\eta), \tag{31}$$

$$\mathcal{B}^-(U,V,\eta) = \mathcal{A}^-(V,\eta) \tag{32}$$

yielding the following numerical flux

$$\phi(U,V,\eta) = \mathcal{A}^+(U,\eta)U + \mathcal{A}^-(V,\eta)V. \tag{33}$$

Another similar flux splitting technique has been proposed by Vijayasundaram [16] by taking

$$\mathcal{B}^+(U,V,\eta) = \mathcal{A}^+\left(\frac{U+V}{2},\eta\right), \tag{34}$$

$$\mathcal{B}^-(U,V,\eta) = \mathcal{A}^-\left(\frac{U+V}{2},\eta\right). \tag{35}$$

As we mentioned in Section 1.1 the flux of the shallow water equations is not homogeneous; thus in order to define matrices $\mathcal{B}^\pm$ for the flux splitting of Steger-Warming and Vijayasundaram we consider instead the matrices $\mathcal{A}_i^\star$ $(i = 1, 2)$ defined in Section 1.1, to obtain $\mathcal{A}^\star$ by

$$\mathcal{A}^\star = \mathcal{A}_1^\star\eta_1 + \mathcal{A}_2^\star\eta_2, \tag{36}$$

and then we define the matrices $\mathcal{B}^\pm$ by replacing the Jacobian matrix $\mathcal{A}$ by $\mathcal{A}^\star$.

Remark 1

Notice that for all of the numerical fluxes presented above

$$\|\eta_{ij}\|\phi\left(W_i^n, W_j^n, \tilde{\eta}_{ij}\right) = \phi\left(W_i^n, W_j^n, \eta_{ij}\right), \tag{37}$$

and the approximation scheme (41) can be rewritten as

$$\frac{W_i^{n+1} - W_i^n}{\Delta t} + \frac{1}{A_i}\sum_{j\in\mathcal{K}_i}\phi\left(W_i^n, W_j^n, \eta_{ij}\right) + \text{boundary terms}$$

$$= \frac{1}{A_i}\int\int_{C_i} G(x,y,W^n)dx\,dy. \tag{38}$$

$\diamond$

2.2.3 Numerical results

In order to compare the behaviour of the above numerical schemes in situations of constant depth ($G \equiv 0$, i.e. no source terms), we have applied them to a problem reported by R. J. Fennema and M. H. Chaudhry [5], P. Glaister [8] and F. Alcrudo and P. García-Navarro [1].

This problem corresponds to a partial breach. A dam is assumed to fail instantaneously or their sluice gates are assumed to be opened instantly.

This test is similar to the shock tube problem for the Euler equations.

The computational domain comprises a 200-m-long and 200-m-wide channel. The nonsymmetrical breach or sluice gates are 75-m wide and the structure of the dam is 10-m thick in the direction of the flow (see Figure 4).

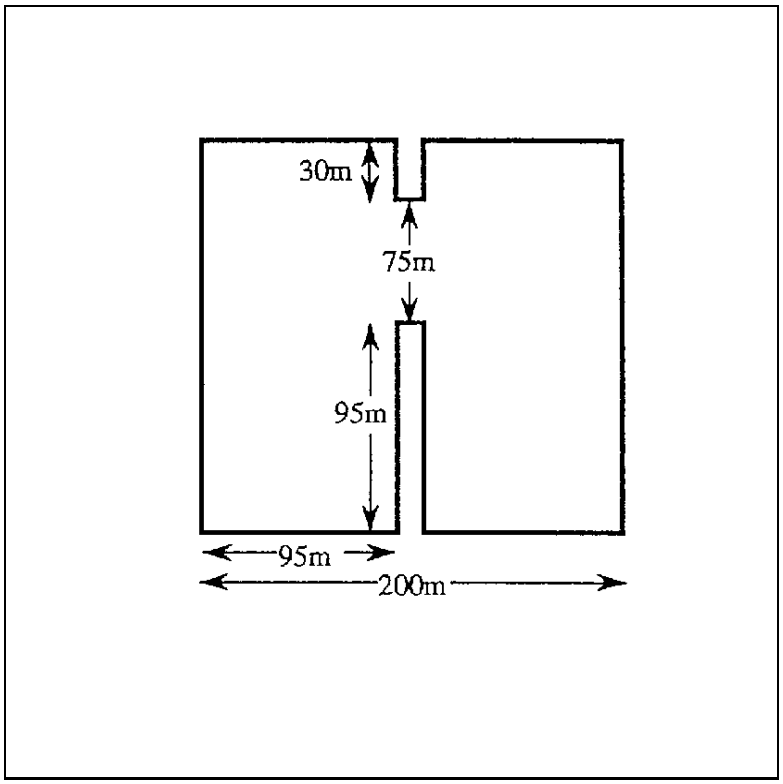

Figure 4: Definition sketch for partial dam breach

As initial conditions two levels of water are considered $h_1 = 10m.$ and $h_0 = 5m.$. Function $H(x, y)$ is taken to be constant and equal to h_0.

Numerical results with the Q-scheme of van Leer for time $t = 7.2$ are shown in Figures 5–7. First a 3D view of the water surface elevation is presented and for the same time the corresponding map of the level lines for h is shown in Figure 6. Finally the velocity field is plotted in Figure 7.

The obtained results are similar to those presented by the different authors mentioned. It is important to remark that in our case the schemes under study are only first-order accurate.

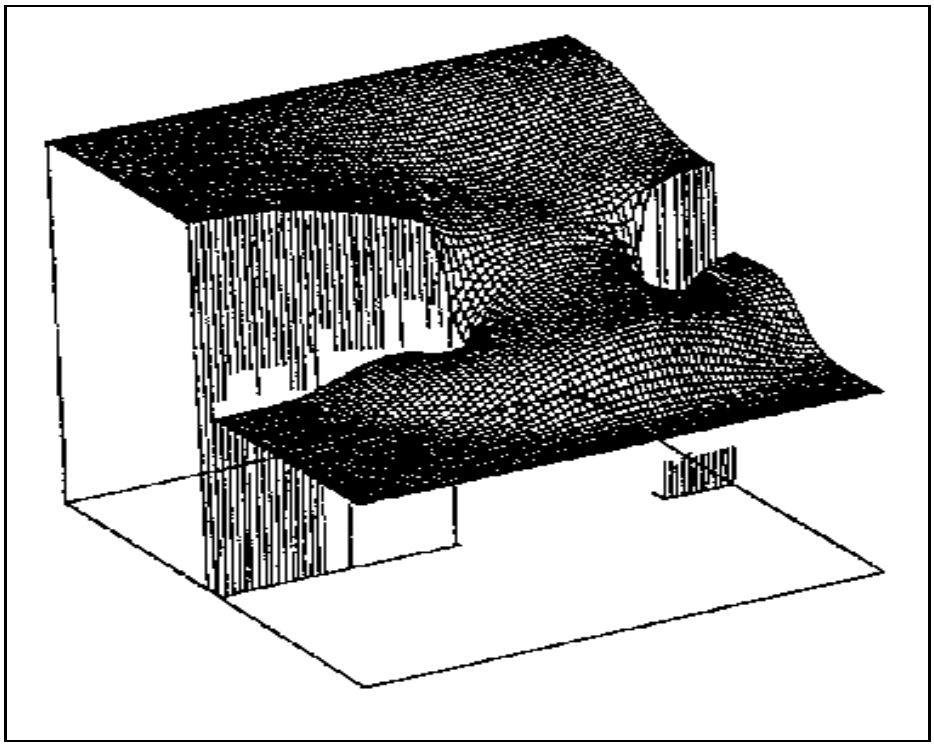

Figure 5: Water surface

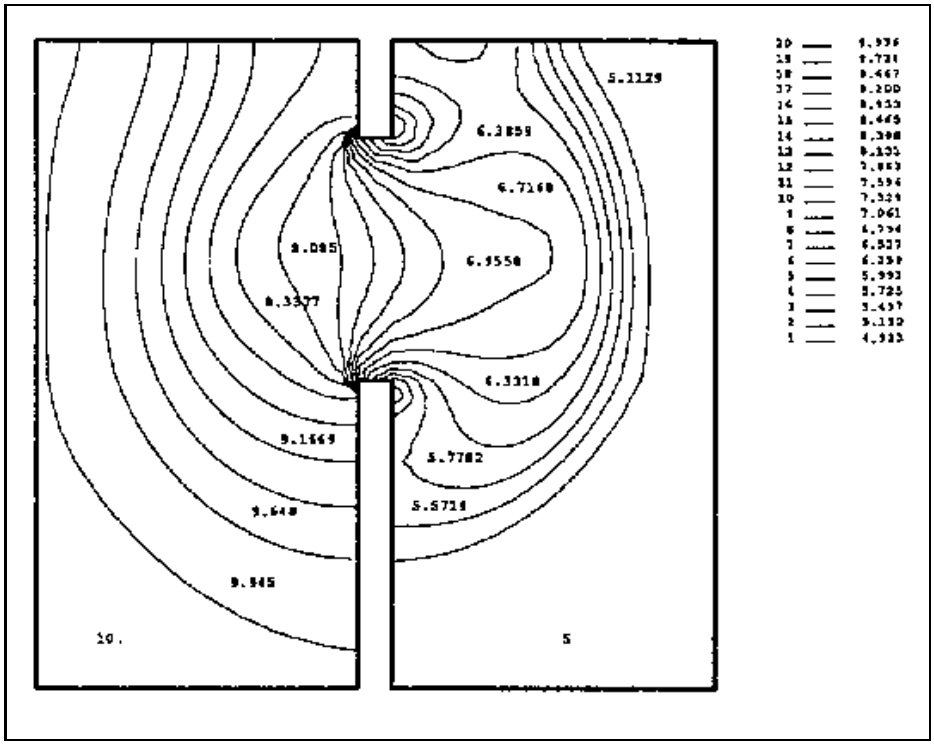

Figure 6: Level lines h

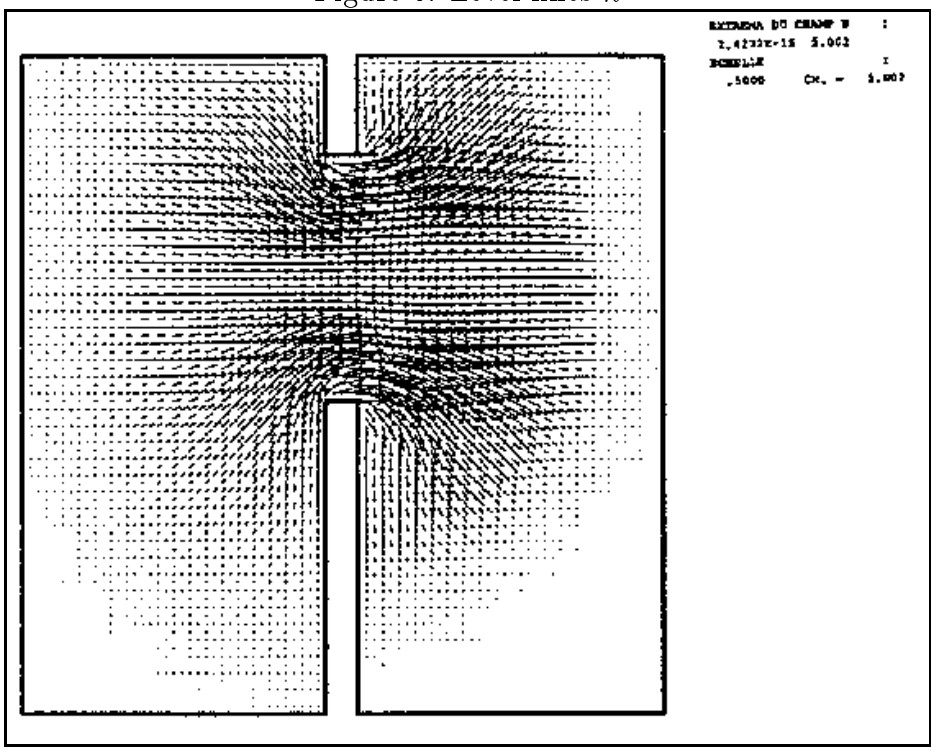

Figure 7: Velocity u

2.3 Source term discretization

In a previous paper (see A. Bermúdez and M.E. Vázquez [2]) the interest of using an upwind discretization of the source term was justified in the case of one space dimension as a means to avoid the propagation of spurious waves. In some sense, it was shown that the discretization of the source term should mimic that of the flux. In this paper, we extend these ideas to two-dimensional problems.

In that follows we only propose extensions of the two Q-schemes considered due to the mentioned study of the one-dimensional shallow water equations in relation with a conservation property [2].

First, to apply to the source term a treatment similar to that of the flux, the integral of the source term is separated into sums which involve all of the neighbouring nodes of N_i.

The idea is to decompose cell C_i into subcells T_{ij} and to compute at the same time the integral of the flux over Γ_{ij} and the integral of the source term over the corresponding subcell.

More precisely, the integral of the source term over the cell C_i is written in the form

$$\frac{1}{A_i}\int\int_{C_i} G(x,y,W^n)dx\,dy = \frac{1}{A_i}\sum_{j\in\mathcal{K}_i}\int\int_{T_{ij}} G(x,y,W^n)dx\,dy, \tag{39}$$

and then the value of the integral of G in each subcell T_{ij} is approximated by a function $\widehat{G}$ (a mean value of G in the nodes N_i and N_j) to obtain an approximation of the left hand side of (39), taking all of the neighbouring nodes of N_i into account.

As in [2] the second step is to upwind this approximation. To such end, the function $\widehat{G}$ is replaced by a *numerical source* ψ. Thus we obtain the following discretization

$$\frac{1}{A_i}\int\int_{C_i} G(x,y,W^n)dx\,dy \simeq \frac{1}{A_i}\sum_{j\in\mathcal{K}_i} A_{T_{ij}}\psi(N_i,N_j,W_i^n,W_j^n,\tilde{\eta}_{ij}). \tag{40}$$

Let us remark that in (40), ψ depends on the normal vector. This fact is equivalent to define the *left* numerical source function in the one dimensional case, because now we also have one direction given by the outward normal vector.

Then the discrete scheme takes the form

$$A_i \frac{W_i^{n+1} - W_i^n}{\Delta t} + \sum_{j \in \mathcal{K}_i} \|\eta_{ij}\| \phi\left(W_i^n, W_j^n, \tilde{\eta}_{ij}\right) + \text{ boundary terms} =$$

$$\sum_{j \in \mathcal{K}_i} A_{T_{ij}} \psi(N_i, N_j, W_i^n, W_j^n, \tilde{\eta}_{ij}), \tag{41}$$

In order to construct the upwind numerical source, we follow the ideas given in [2] in the one-dimensional case: the projection of the centred source term, $\widehat{G}$, onto the eigenvectors of Q associated with positive (resp. negative) eigenvalues is added to the state W_j (resp. W_i).

In other words, for each edge Γ_{ij} the contribution of the source term to node iN_i is defined as the projection of the centred source onto the eigenvectors of the Jacobian matrix of the flux corresponding to negative eigenvalues. Let us remark that η_{ij} is the "outward" normal vector, so the numerical source function is just the part of the source that "flows into" the finite volume. In that way, it seems natural to define the *2D numerical source function* in each subcell T_{ij}, by the following expression

$$\psi\left(N_i, N_j, {W_i}^n, W_j^n, \tilde{\eta}_{ij}\right) = \left(I - |Q| Q^{-1}\right) \widehat{G}\left(N_i, N_j, {W_i}^n, W_j^n, \tilde{\eta}_{ij}\right).$$

To simplify the notation, the dependence of the matrices Q and Q^{-1} on $(W_{\mathrm{Q}}^n, \tilde{\eta}_{ij})$ has been omitted. The state $W_{\mathrm{Q}}^n = W_{\mathrm{Q}}^n\left({W_i}^n, W_j^n\right)$ is defined in (20).

Remark 2

Contrary to what happens for the flux function (see (37)), the numerical source term is independent of the norm of η and

$$\psi\left(N_i, N_j, {W_i}^n, W_j^n, \tilde{\eta}_{ij}\right) = \psi\left(N_i, N_j, {W_i}^n, W_j^n, \eta_{ij}\right), \tag{42}$$

Indeed, for the diagonal matrix Λ given by the eigenvalues of Q we have

$$\Lambda\left(W_{\mathrm{Q}}^n\left(W_i^n, W_j^n\right), \tilde{\eta}_{ij}\right) = \|\eta_{ij}\| \Lambda\left(W_{\mathrm{Q}}^n\left(W_i^n, W_j^n\right), \eta_{ij}\right), \tag{43}$$

and then

$$\left|\Lambda\left(W_{\mathrm{Q}}^n\left(W_i^n, W_j^n\right), \eta_{ij}\right)\right| \Lambda^{-1}\left(W_{\mathrm{Q}}^n\left(W_i^n, W_j^n\right), \eta_{ij}\right) =$$

$$\left|\Lambda\left(W_{\mathrm{Q}}^n\left(W_i^n, W_j^n\right), \tilde{\eta}_{ij}\right)\right| \Lambda^{-1}\left(W_{\mathrm{Q}}^n\left(W_i^n, W_j^n\right), \tilde{\eta}_{ij}\right).$$

$\diamond$

In order to obtain the expression of the numerical source function ψ, it only remains to define $\widehat{G}$. The election of this function is related with the verification of a *conservation property*, which is introduced immediately afterwards.

In the one-dimensional case, $\widehat{G}$ is defined as an approximation of G in each subcell by taking the average of values of G at the two neighbouring nodes; similarly, for the two-dimensional case, we define $\widehat{G}$ in each subcell T_{ij} as a centred approximation of G using the two states W_i and W_j.

If we consider the analytical expression of G to obtain $\widehat{G}$, two difficulties appear:

- The first difficulty is related with the announced *Property C*. As will be proved, to use the analytical expression of the gradient of the depth prevents the verification of this property (see Section 2.4).
- In the applications, the water depth is obtained from the bathymetric charts, by discrete values of H. In this case the analytical H is not known, and it is necessary to approximate its gradient.

Thus we suppose that the values of the depth at the nodes N_i and N_j are known; the gradient is then discretized using the directional derivative in the direction $\overline{N_iN_j}$.

More precisely, the two elections of the function $\widehat{G}$ that we will consider are

$$\widehat{G}\left(N_i, N_j, W_i^n, W_j^n, \tilde{\eta}_{ij}\right) = \begin{pmatrix} 0 \\ g\left(\dfrac{h_i+h_j}{2}\right)\left(\dfrac{H_j-H_i}{d_{ij}}\right)\tilde{\eta}_{ij}^1 \\ g\left(\dfrac{h_i+h_j}{2}\right)\left(\dfrac{H_j-H_i}{d_{ij}}\right)\tilde{\eta}_{ij}^2 \end{pmatrix}, \tag{44}$$

for the extension of the Q-scheme of van Leer and

$$\widehat{G}\left(N_i, N_j, W_i^n, W_j^n, \tilde{\eta}_{ij}\right) = \begin{pmatrix} 0 \\ g\sqrt{h_ih_j}\left(\dfrac{H_j-H_i}{d_{ij}}\right)\tilde{\eta}_{ij}^1 \\ g\sqrt{h_ih_j}\left(\dfrac{H_j-H_i}{d_{ij}}\right)\tilde{\eta}_{ij}^2 \end{pmatrix}, \tag{45}$$

for the extension of the scheme of Roe. In both cases H_i denotes the value of depth at the node N_i.

2.4 A conservation property

In the one-dimensional case it was proved that a requirement for the numerical scheme to have a satisfactory behaviour when the bottom is not flat is that it approximates properly the trivial stationary solution $q(x,t) = 0$, $h(x,t) = H(x)$ when it is an exact solution to the continuous problem.

Similarly, we notice that a stationary solution for the two-dimensional shallow water equations is given by

$$\left\{ \begin{array}{rcl} h & \equiv & H, \\ q & \equiv & (0,0), \end{array} \right. \tag{46}$$

in which case the nonvanishing terms are

$$\left\{ \begin{array}{rcl} \dfrac{\partial}{\partial x}\left(\dfrac{1}{2}gh^2\right) & = & gh\dfrac{\partial H}{\partial x}, \\ \dfrac{\partial}{\partial y}\left(\dfrac{1}{2}gh^2\right) & = & gh\dfrac{\partial H}{\partial y}. \end{array} \right. \tag{47}$$

Therefore when the bottom is not flat, a good test problem for a numerical scheme for the shallow water equations should be (47).

The equations in (47) indicate that an appropriate scheme for the source should mimic the discretization of the flux terms in the left hand side of (47).

The importance of this requirement leads us to introduce the following definitions.

Definition 1

Exact $\mathcal{C}$-property

A scheme is said to satisfy the exact $\mathcal{C}$-property if it is exactly compatible with the stationary solution (46).

Definition 2

Approximate $\mathcal{C}$-property to the order p

A scheme is said to satisfy the approximate $\mathcal{C}$-property to the order p, if, when computing the stationary solution (46), it is formally $O(h^p)$-accurate.

In what follows, it is proved that the extensions of the Q-schemes verify the Property $\mathcal{C}$ in a exact or approximate way depending on the choice of the function $\widehat{G}$.

Proposition 1

i) The proposed extensions of the Q-schemes of Roe and van Leer to the shallow water equations verify the exact $\mathcal{C}$-property, if we consider the choice of $\widehat{G}$ given by (44).

ii) If we consider the function $\widehat{G}$ given by (45), then the extension of the Roe scheme verifies the approximate $\mathcal{C}$-property to the order 2.

Proof:

We assume that the initial conditions are $q \equiv 0$, $h \equiv H$. Thus the numerical flux and the flux term are given by

$$\phi\left(W_i^n, W_j^n, \eta_{ij}\right) = \frac{1}{2}\begin{pmatrix} 0 \\ \frac{1}{2}\left(h_i^2 + h_j^2\right)\eta_{ij_1} \\ \frac{1}{2}\left(h_i^2 + h_j^2\right)\eta_{ij_2} \end{pmatrix} - \frac{1}{2}\begin{pmatrix} c_{ij}\|\eta_{ij}\|\left(h_j - h_i\right) \\ 0 \\ 0 \end{pmatrix}, \tag{48}$$

$$\sum_{j\in\mathcal{K}_i}\phi\left(W_i^n, W_j^n, \eta_{ij}\right) = \begin{pmatrix} \displaystyle\sum_{j\in\mathcal{K}_i} -\frac{1}{2}c_{ij}\left(h_j - h_i\right)\|\eta_{ij}\| \\ \frac{1}{4}g\left[h_i^2\displaystyle\sum_{j\in\mathcal{K}_i}\eta_{ij_1} + \sum_{j\in\mathcal{K}_i} h_j^2\eta_{ij_1}\right] \\ \frac{1}{4}g\left[h_i^2\displaystyle\sum_{j\in\mathcal{K}_i}\eta_{ij_2} + \sum_{j\in\mathcal{K}_i} h_j^2\eta_{ij_2}\right] \end{pmatrix}, \tag{49}$$

where $c_{ij} = \sqrt{g\left(\dfrac{h_i + h_j}{2}\right)}$.

The numerical source function is given by

$$\psi\left(N_i, N_j, W_i^n, W_j^n, \tilde{\eta}_{ij}\right) = \begin{pmatrix} -c_{ij}\dfrac{\left(H_j - H_i\right)}{dij} \\ c_{ij}^2\dfrac{\left(H_j - H_i\right)}{dij}\tilde{\eta}_{ij_1} \\ c_{ij}^2\dfrac{\left(H_j - H_i\right)}{dij}\tilde{\eta}_{ij_2} \end{pmatrix}, \tag{50}$$

Then the contribution of the source term is

$$\sum_{j\in\mathcal{K}_i}\frac{\|\eta_{ij}\|d_{ij}}{2}\psi\left(N_i, N_j, W_i^n, W_j^n, \eta_{ij}\right) =$$

$$\begin{pmatrix} -\frac{1}{2}\sum_{j\in\mathcal{K}_i} c_{ij}\|\eta_{ij}\|\,(H_j - H_i) \\ \frac{1}{2}\sum_{j\in\mathcal{K}_i} g\left(\frac{h_i+h_j}{2}\right)(H_j - H_i)\,\eta_{ij1} \\ \frac{1}{2}\sum_{j\in\mathcal{K}_i} g\left(\frac{h_i+h_j}{2}\right)(H_j - H_i)\,\eta_{ij2} \end{pmatrix} \quad \begin{array}{c} = \\ \uparrow \\ H\equiv h \end{array} \quad \begin{pmatrix} \sum_{j\in\mathcal{K}_i} -\frac{1}{2}c_{ij}\,(h_j - h_i)\,\|\eta_{ij}\| \\ \frac{1}{4}g\left[-h_i^2\sum_{j\in\mathcal{K}_i}\eta_{ij1} + \sum_{j\in\mathcal{K}_i} h_j^2\eta_{ij1}\right] \\ \frac{1}{4}g\left[-h_i^2\sum_{j\in\mathcal{K}_i}\eta_{ij2} + \sum_{j\in\mathcal{K}_i} h_j^2\eta_{ij2}\right] \end{pmatrix}.$$

In order to prove the exact $\mathcal{C}$-property the following equality must be verified

$$\sum_{j\in\mathcal{K}_i}\phi\left(W_i^n, W_j^n, \eta_{ij}\right) = \sum_{j\in\mathcal{K}_i}\frac{\|\eta_{ij}\|d_{ij}}{2}\psi\left(N_i, N_j, \frac{W_i^n + W_j^n}{2}, \eta_{ij}\right), \tag{51}$$

This equality is trivial for the first components. To establish it for the other two, it suffices to observe that the sum of the coordinates of the vertices of a closed polygon is null, that is

$$\sum_{j\in\mathcal{K}_i}\eta_{ij} = 0, \tag{52}$$

thus part *i)* is proved.

In the second part of the proposition, the function $\widehat{G}$ is given by (45) and the proof is obtained as a consequence of the equality

$$\sqrt{h_i^n h_j^n} = \frac{h_i^n + h_j^n}{2} + O(h^2). \tag{53}$$

□

2.5 Discretization of the boundary conditions

The difficulty of the treatment of the boundary conditions comes from the fact that for a boundary node some *neighbouring nodes* which are necessary to define the numerical flux do not exist.

For this reason, the value of the solution in the *supposed neighbouring node* is taken to be the same as the value of the solution in the boundary node which, for the consistency of the numerical flux, is equivalent to taking the numerical flux equal to the physical flux.

More precisely, according to the type of boundary conditions, the procedure will be the following:

1. *Open sea boundary*

   The value of the height of the water is imposed, more precisely:

   - To update the variable q, the numerical flux is taken equal to the physical flux.
   - Then the value of h is imposed

$$h_i^{n+1} = \varphi\left(t_{n+1}\right) + H(x_i, y_i) \tag{54}$$

2. *Coast boundary*

   In this case a slip condition is considered: $q \cdot \eta = 0$. To impose this condition *strongly*, we subtract the normal components, i.e.

$$q_i^{n+1} = q_i^n + (\Delta q_i - < \Delta q_i, \eta_{i\mathrm{F}} > \eta_{i\mathrm{F}}) \tag{55}$$

   where Δq_i are the last two components of the vector ΔW_i^n given by

$$\Delta W_i^n = -\sum_{j \in \mathcal{K}_i} \phi\left(W_i^n, W_j^n, \eta_{ij}\right) - Z\left(W_i^n, \eta_{i\mathrm{F}}\right) + \sum_{j \in \mathcal{K}_i} \psi\left(N_i, N_j, W_i^n, W_j^n, \tilde{\eta}_{ij}\right).$$

2.6 Numerical Results

Propagation of a tidal wave in a ria

We present a numerical test for the computation of tidal currents in the Pontevedra ria, Galicia, Spain. The corresponding bathymetry is given in Figure 8. This domain has evidently a nonsmooth boundary

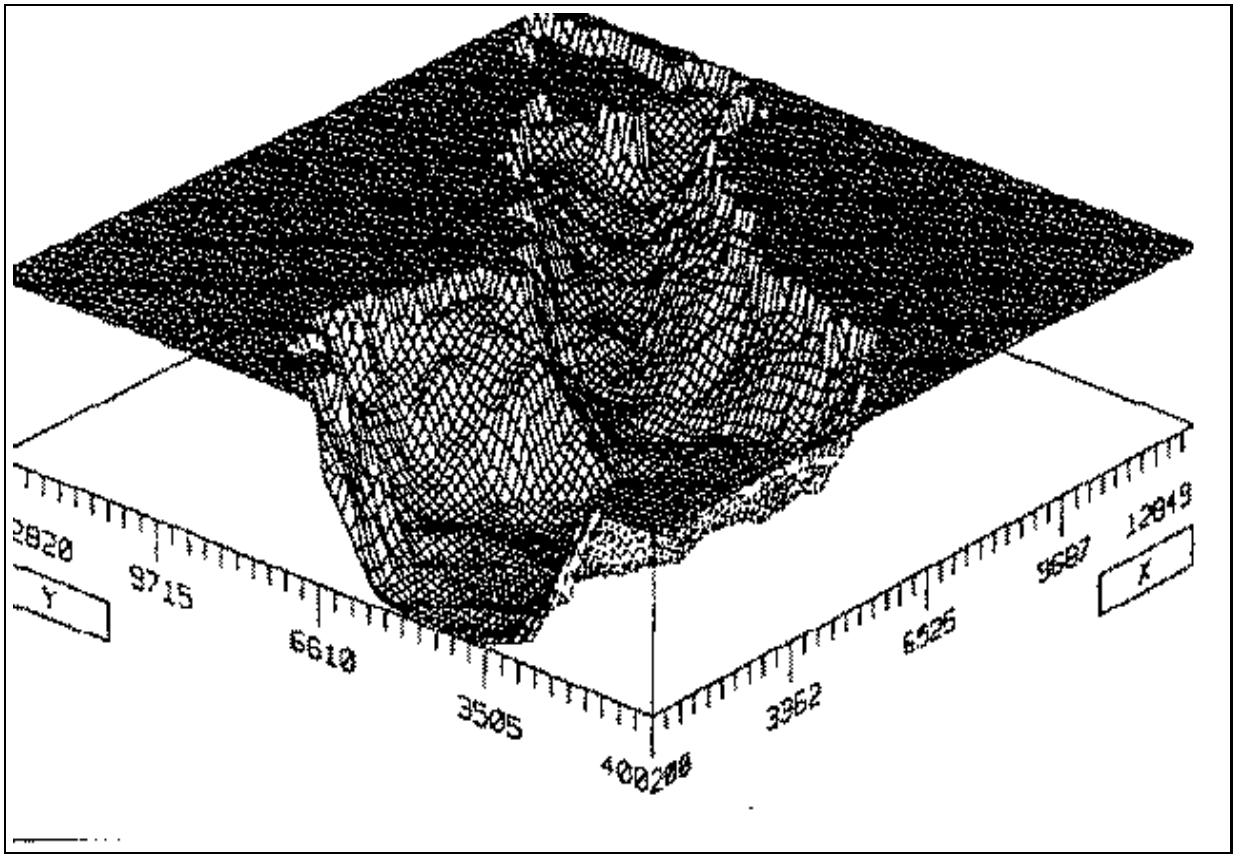

Figure 8: $z(x,y) = -H(x,y)$ for the Pontevedra ria.

We consider the following initial and boundary conditions

$$\underline{\text{Initial Conditions}} \quad \begin{cases} h_0(x,y) &= H(x,y) \\ q_0(x,y,0) &= 0 \end{cases}$$

$$\underline{\text{Boundary Conditions}} \quad \begin{cases} h(x,y,t) = H(x,y) + \varphi(t) & \text{if } (x,y) \in \Gamma_1 \\ q \cdot \eta \equiv 0 & \text{if } (x,y) \in \Gamma_2 \end{cases} \tag{56}$$

where

$$\varphi(t) = 4\left(1 + sin\left(\pi\left(\frac{4t}{86400} + \frac{1}{2}\right)\right)\right). \tag{57}$$

We start with the numerical results obtained with the type of finite volume more extensively applied in the papers devoted to the Euler equations for unstructured meshes, that we refer to as finite volume of the "vertex-type". As is well known, the nodes are the vertices of the standard finite element triangulation, and the problem appears in the definition of the normal vector at boundary nodes. This difficulty prevents the correct treatment of the boundary condition $q \cdot \eta = 0$.

More precisely, for the time $t = 10800$s the integral of the flux through the coast boundary Γ_2 is $197.65 m^3 s^{-1}$ instead of 0. The corresponding velocity field is given in Figure 10.

To prevent the numerical viscosity of the mentioned Q-schemes from vanishing when some of the eigenvalues of the Jacobian matrix of the flux is zero, we apply the Harten regularization (see Harten [9] and [15] for details). To illustrate the need of this regularization for this type of problems, the numerical results are given in Figure 9 "without" regularization whereas the other ones are obtained with the mentioned regularization.

In the velocity field of Figure 9, the oscillations on the remarkable zones are produced by the existence of zero eigenvalues.

Finally, Figures 11–14 show the numerical results of four points in time of the first tidal cycle. As can be noticed, the results confirm the good qualitative behavior of the proposed method.

In particular, the difference on $h(x, y, t) - H(x, y)$ between two arbitrary points of the ria is small, which is natural for a domain of small length. With respect to the velocities, the results seems to be acceptably accurate except in the area near the "open sea boundary", due to the "boundary layer" which is generated when imposing the boundary condition over the water height.

In the following table the values of the typical velocity obtained at an interior point P for the four mentioned points in time are shown.

Let us remark that for times $t = 21600$ and $t = 43200$, the function $h - H$ is smooth because the velocities are small, contrary to the cases when times $t = 10800$ and $t = 32400$.

Cycle step	time (s)	$\varphi(t)$ (m)	Velocity $(m\,s^{-1})$
"half"-rising tide	10800	4	0.1485
high water	21600	8	0.3119E-02
"half"-ebb tide	32400	4	0.1461
low water	43200	0	0.4059E-02

Table 1: Velocities for the first tidal cycle at point P.

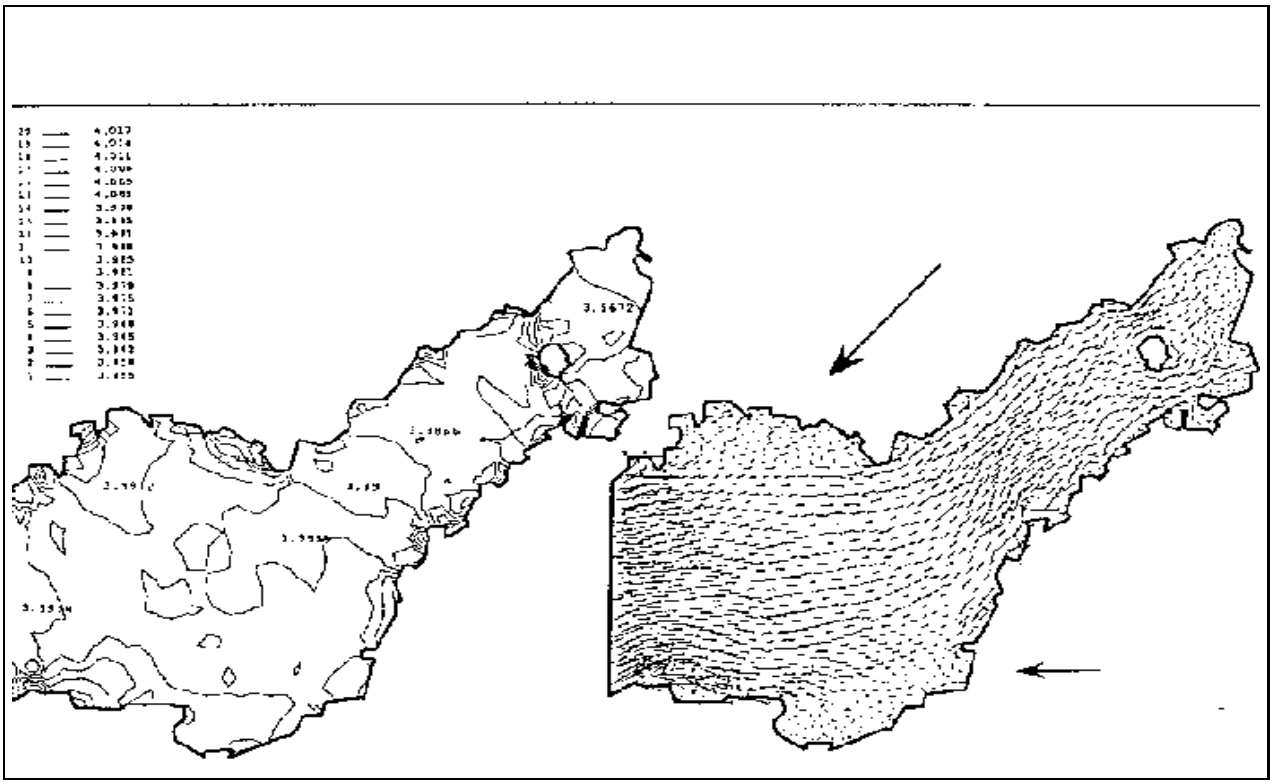

Figure 9: Extension of the Q-scheme of van Leer "without" Harten regularization.

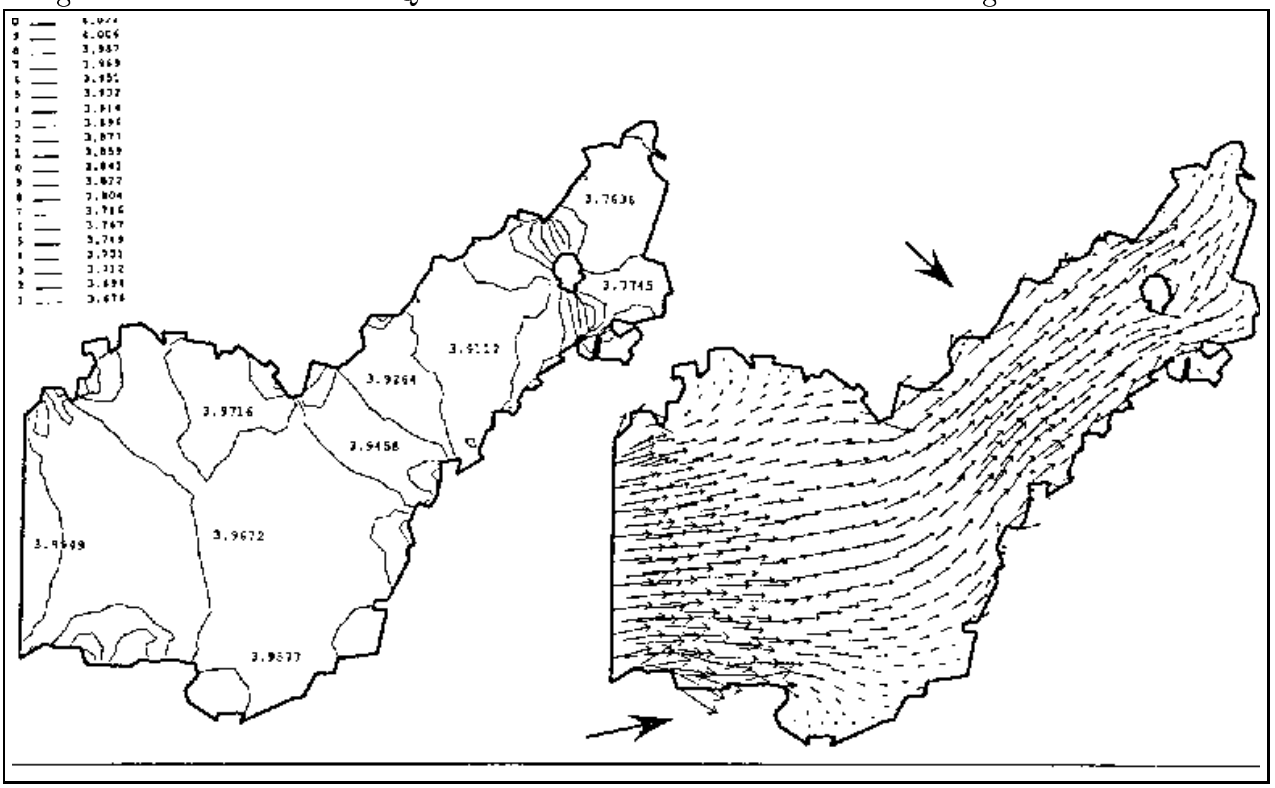

Figure 10: Extension of the Q-scheme of van Leer finite volume of the vertex-type

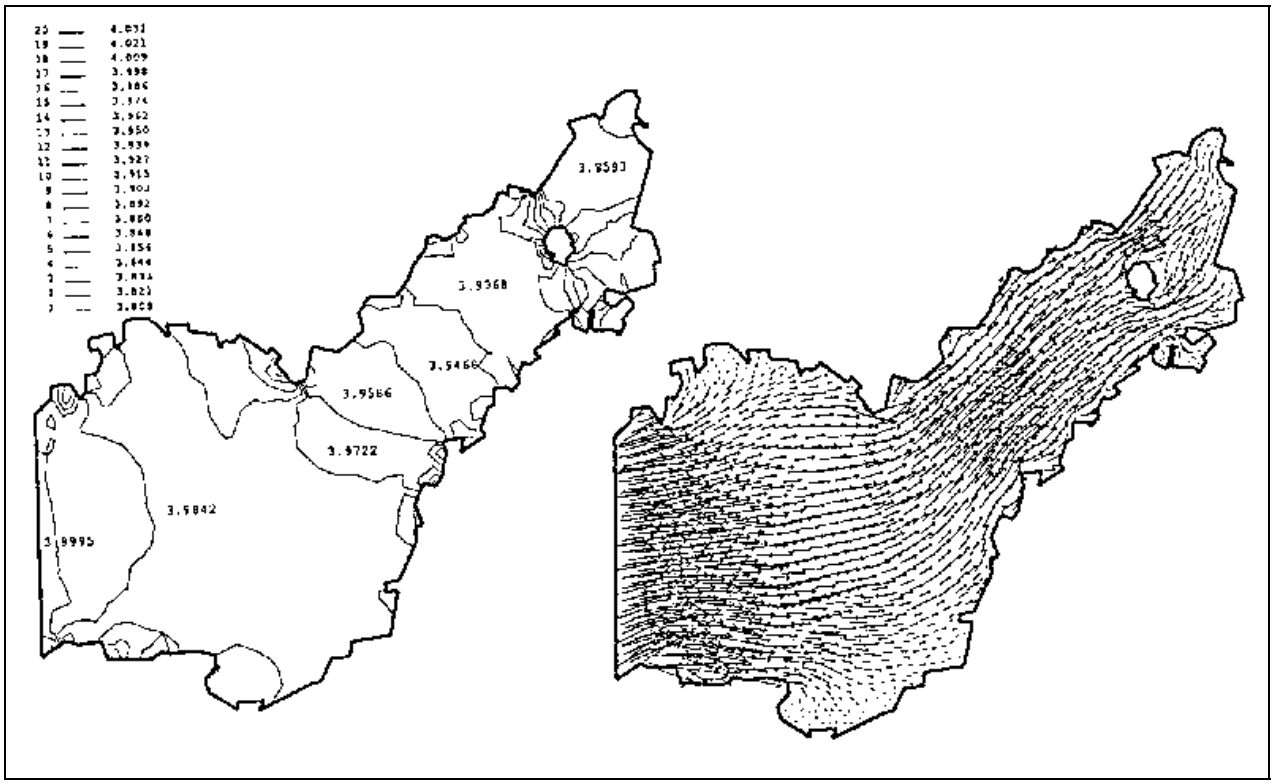

Figure 11: Extension of the Q-scheme of van Leer; $t = 10800$ (finite volume edge-type)

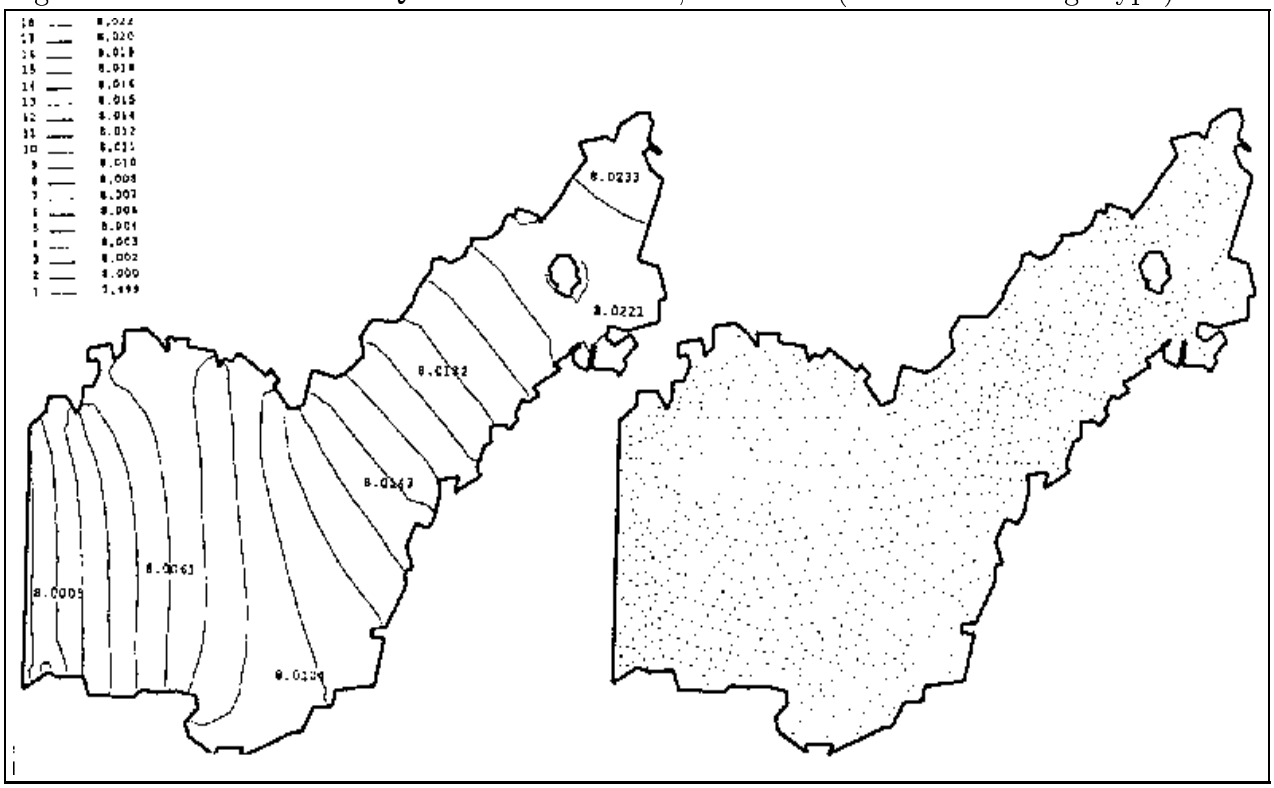

Figure 12: Extension of the Q-scheme of van Leer; $t = 21600$ (finite volume edge-type)

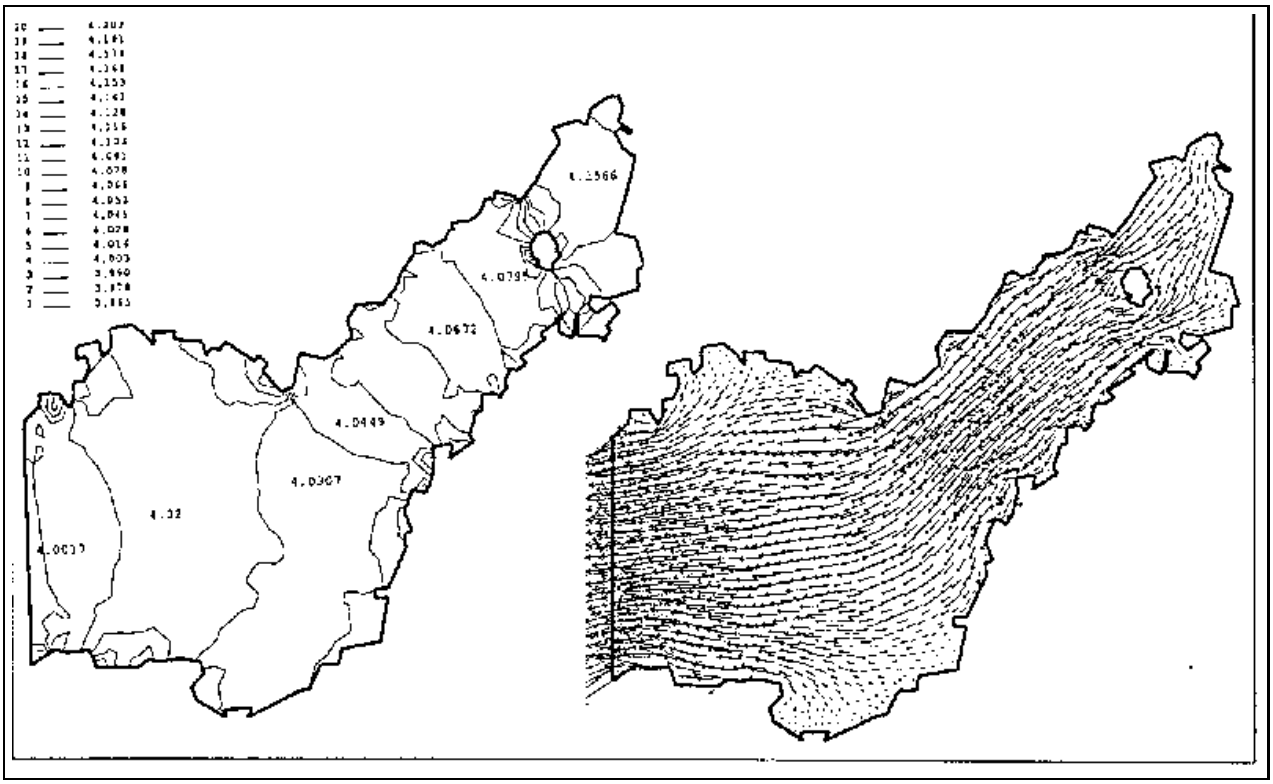

Figure 13: Extension of the Q-scheme of van Leer; $t = 32400$ (finite volume edge-type)

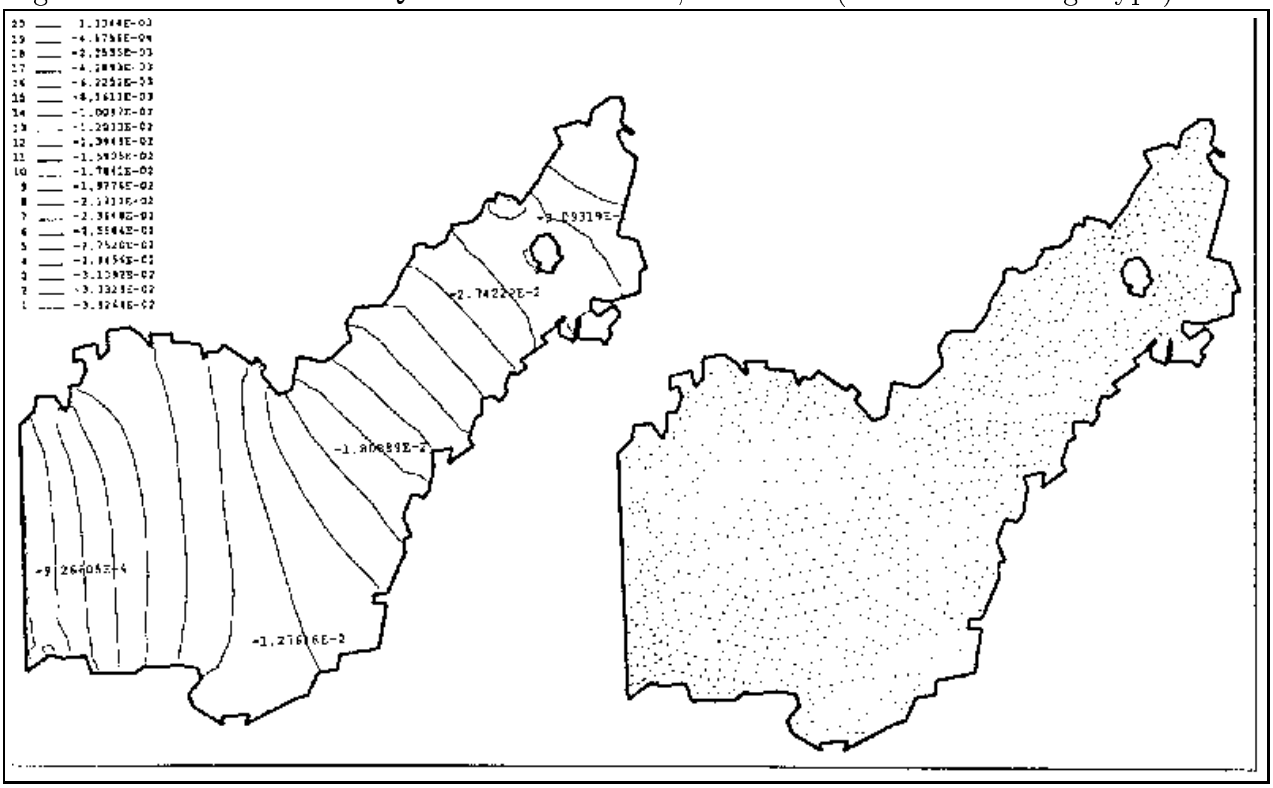

Figure 14: Extension of the Q-scheme of van Leer; $t = 43200$ (finite volume edge-type)

3 Implicit discretizations.

3.1 Implicit discretizations of 1D problems

In the previous section, the time discretization was explicit, and this implied a severe restriction on time step.

Indeed, for a given mesh, according to the Courant-Friedrichs-Lewy condition, the time step is determined by the largest eigenvalue of the Jacobian matrix of the flux. For the tidal current computation, this eigenvalue is of the same order of magnitude as the velocity of propagations of waves, i.e. $c = \sqrt{gh}$ which, in our previous example, is about $20m/s$, i.e. two orders of magnitude higher than the velocity of particles u $\left(u = \frac{q}{h}\right)$ which is the relevant unknown in the this kind of applications.

The main aim of this section is to combine the upwind schemes presented above with an implicit discretization in time to obtain unconditional linear stability.

For the sake of simplicity, the presentation of the resulting schemes is done first in one dimension. The *simplified linearized implicit scheme* we apply is similar to that introduced by B. Stoufflet in [13] for systems of conservation laws without source terms and in the case where the flux is a homogeneous function of degree one.

The problem to be solved in this section is a hyperbolic nonlinear system of the form

$$\frac{\partial w}{\partial t}(x,t) + \frac{\partial F}{\partial x}\left(w(x,t)\right) = G(x, w(x,t)). \tag{58}$$

The vector form of an implicit scheme to solve (111) using upwind schemes for flux and source terms is

$$\frac{W^{n+1} - W^n}{\Delta t} + \mathcal{H}\left(W^{n+1}\right) = 0, \tag{59}$$

where

$$\mathcal{H}(W) = \mathcal{H}_{\text{F}}(W) + \mathcal{H}_{\text{G}}(W) \tag{60}$$

$$\left(\mathcal{H}_{\text{F}}\left(W^{n+1}\right)\right)_j = \frac{\phi\left(W_j^{n+1}, W_{j+1}^{n+1}\right) - \phi\left(W_{j-1}^{n+1}, W_j^{n+1}\right)}{A_j} \tag{61}$$

$$\begin{aligned}\left(\mathcal{H}_{\text{G}}\left(W^{n+1}\right)\right)_j = & -\frac{1}{A_j}\left\{A_{T_{j\text{L}}}\psi_{\text{L}}\left(x_{j-1}, x_j, W_{j-1}^{n+1}, W_j^{n+1}\right)\right. \\ & \left. + A_{T_{j\text{R}}}\psi_{\text{R}}\left(x_j, x_{j+1}, W_j^{n+1}, W_{j+1}^{n+1}\right)\right\}.\end{aligned} \tag{62}$$

The difficulty with this scheme is the high required computing power. Hence it is quite reasonable to replace $\mathcal{H}$ by a linealization

$$\mathcal{H}\left(W^{n+1}\right) \simeq \mathcal{H}\left(W^n\right) + \mathcal{H}'\left(W^n\right)\left(W^{n+1} - W^n\right). \tag{63}$$

The resulting scheme is called *linearized implicit scheme* (see Dervieux and Desideri [4]). It can be rewritten in the following δ-form:

$$\left(\frac{I}{\Delta t} + \mathcal{H}'\left(W^n\right)\right)\delta W^{n+1} = -\mathcal{H}\left(W^n\right). \tag{64}$$

where $\delta W^{n+1} = W^{n+1} - W^n$. Nevertheless, this scheme also presents difficulties. For instance in the case of the Q-schemes, due to the presence of the absolute value of matrix $\mathcal{A}$ in the expression of the flux, operator $\mathcal{H}'_{\mathrm{F}}$ is not defined if the eigenvalues of this matrix change their sign. Moreover, in case that such operator exists, its computational cost may be high.

In such situations it is necessary to replace $\mathcal{H}'$ with another approximate linear operator, denoted $\mathcal{P}^n$, so that (117) takes the form

$$\left(\frac{I}{\Delta t} + \mathcal{P}^n\left(W^n\right)\right)\delta W^{n+1} = -\mathcal{H}\left(W^n\right). \tag{65}$$

These types of schemes are known as *simplified linearized implicit schemes* (see [13]).

Next, a procedure to realize one such construction is detailed.

Construction of the operator $\mathcal{P}^n$

In order to obtain the expression of $\mathcal{P}^n = \mathcal{P}^n_{\mathrm{F}} + \mathcal{P}^n_{\mathrm{G}}$, the corresponding part of the flux ($\mathcal{P}^n_{\mathrm{F}}$) and the source term ($\mathcal{P}^n_{\mathrm{G}}$) are introduced in parallel.

1. First we suppose that the flux and the source numerical functions verify the following property:
   There exist $\mathcal{R}_L$, $\mathcal{R}_R$ and $\mathcal{S}_L$ y $\mathcal{S}_R$ such that:

$$\phi\left(U,V\right) = \mathcal{R}_{\mathrm{L}}\left(U,V\right)U + \mathcal{R}_{\mathrm{R}}\left(U,V\right)V. \tag{66}$$

$$\psi_{\mathrm{L}}\left(x,y,U,V\right) = \mathcal{S}_{\mathrm{L}}\left(x,y,U,V\right)\left(U+V\right), \tag{67}$$

$$\psi_{\mathrm{R}}\left(x,y,U,V\right) = \mathcal{S}_{\mathrm{R}}\left(x,y,U,V\right)\left(U+V\right). \tag{68}$$

2. If the previous property is verified, then operators $\mathcal{H}_{\text{F}}$ and $\mathcal{H}_{\text{G}}$ can be written in the form

$$\begin{aligned}\left(\mathcal{H}_{\text{F}}\left(W^{n+1}\right)\right)_j &= \frac{1}{A_j}\left[\mathcal{R}_{\text{L}}\left(W_j^{n+1},W_{j+1}^{n+1}\right)W_j^{n+1}+\mathcal{R}_{\text{R}}\left(W_j^{n+1},W_{j+1}^{n+1}\right)W_{j+1}^{n+1}\right]\\ &- \frac{1}{A_j}\left[\mathcal{R}_{\text{L}}\left(W_{j-1}^{n+1},W_j^{n+1}\right)W_{j-1}^{n+1}+\mathcal{R}_{\text{R}}\left(W_{j-1}^{n+1},W_j^{n+1}\right)W_j^{n+1}\right]. \quad (69)\end{aligned}$$

$$\begin{aligned}\left(\mathcal{H}_{\text{G}}\left(W^{n+1}\right)\right)_j &= -\frac{A_{T_{j\text{R}}}}{A_j}\left[\mathcal{S}_{\text{R}}\left(x_j,x_{j+1},W_j^{n+1},W_{j+1}^{n+1}\right)\left(W_j^{n+1}+W_{j+1}^{n+1}\right)\right]\\ &+\frac{A_{T_{j\text{L}}}}{A_j}\left[\mathcal{S}_{\text{L}}\left(x_{j-1},x_j,W_{j-1}^{n+1},W_j^{n+1}\right)\left(W_{j-1}^{n+1}+W_j^{n+1}\right)\right]. \quad (70)\end{aligned}$$

3. Then we change (122) and (123) by evaluating matrix $\mathcal{R}_{\text{L}}$, $\mathcal{R}_{\text{R}}$, $\mathcal{S}_{\text{L}}$ and $\mathcal{S}_{\text{R}}$ at time t_n instead of t_{n+1}. The resulting approximations are

$$\begin{aligned}\left(\mathcal{H}_{\text{F}}\left(W^{n+1}\right)\right)_j &\simeq \frac{1}{A_j}\left[\mathcal{R}_{\text{L}}\left(W_j^{n},W_{j+1}^{n}\right)W_j^{n+1}+\mathcal{R}_{\text{R}}\left(W_j^{n},W_{j+1}^{n}\right)W_{j+1}^{n+1}\right]\\ &+ \frac{1}{A_j}\left[-\mathcal{R}_{\text{L}}\left(W_{j-1}^{n},W_j^{n}\right)W_{j-1}^{n+1}-\mathcal{R}_{\text{R}}\left(W_{j-1}^{n},W_j^{n}\right)W_j^{n+1}\right]. \quad (71)\end{aligned}$$

$$\begin{aligned}\left(\mathcal{H}_{\text{G}}\left(W^{n+1}\right)\right)_j &\simeq \frac{T_{j\text{R}}}{A_j}\left[-\mathcal{S}_{\text{R}}\left(x_j,x_{j+1},W_j^{n},W_{j+1}^{n}\right)\left(W_j^{n+1}+W_{j+1}^{n+1}\right)\right]\\ &+ \frac{T_{j\text{L}}}{A_j}\left[\mathcal{S}_{\text{L}}\left(x_{j-1},x_j,W_{j-1}^{n},W_j^{n}\right)\left(W_{j-1}^{n+1}+W_j^{n+1}\right)\right] \quad (72)\end{aligned}$$

4. Finally, if we introduce $\mathcal{H}_{\text{F}}\left(W^n\right)$ and $\mathcal{H}_{\text{G}}\left(W^n\right)$ in (124) and (125) respectively, the following expressions $\mathcal{H}_{\text{F}}\left(W^{n+1}\right)$ and $\mathcal{H}_{\text{G}}\left(W^{n+1}\right)$ are obtained,

$$\mathcal{H}_{\text{F}}(W^{n+1}) \simeq \mathcal{H}_{\text{F}}(W^n)+\mathcal{P}_{\text{F}}^n(W^n)\delta W^{n+1} \quad (73)$$

$$\mathcal{H}_{\text{G}}(W^{n+1}) \simeq \mathcal{H}_{\text{G}}(W^n)+\mathcal{P}_{\text{G}}^n(W^n)\delta W^{n+1}, \quad (74)$$

where

$$\left(\mathcal{P}_{\text{F}}^n\left(W^n\right)\right)_{jj}=\frac{1}{A_j}\left[\mathcal{R}_{\text{L}}\left(W_j^n,W_{j+1}^n\right)-\mathcal{R}_{\text{R}}\left(W_{j-1}^n,W_j^n\right)\right] \quad (75)$$

$$\left(\mathcal{P}_{\text{F}}^n\left(W^n\right)\right)_{j\,j+1}=\frac{1}{A_j}\mathcal{R}_{\text{R}}\left(W_j^n,W_{j+1}^n\right), \quad (76)$$

$$\left(\mathcal{P}_{\text{F}}^n\left(W^n\right)\right)_{j-1\,j}=-\frac{1}{A_j}\mathcal{R}_{\text{L}}\left(W_{j-1}^n,W_j^n\right). \quad (77)$$

$$\left(\mathcal{P}_{\mathrm{G}}^{n}\left(W^{n}\right)\right)_{jj} = \frac{T_{j\mathrm{R}}}{A_j}\mathcal{S}_{\mathrm{R}}\left(x_j, x_{j+1}, W_j^n, W_{j+1}^n\right) + \frac{T_{j\mathrm{L}}}{A_j}\mathcal{S}_{\mathrm{L}}\left(x_{j-1}, x_j, W_{j-1}^n, W_j^n\right), \quad (78)$$

$$\left(\mathcal{P}_{\mathrm{G}}^{n}\left(W^{n}\right)\right)_{j\,j+1} = \frac{T_{j\mathrm{R}}}{A_j}\mathcal{S}_{\mathrm{R}}\left(x_j, x_{j+1}, W_j^n, W_{j+1}^n\right), \quad (79)$$

$$\left(\mathcal{P}_{\mathrm{G}}^{n}\left(W^{n}\right)\right)_{j-1\,j} = \frac{T_{j\mathrm{L}}}{A_j}\mathcal{S}_{\mathrm{L}}\left(x_{j-1}, x_j, W_{j-1}^n, W_j^n\right). \quad (80)$$

Remark 3

When the numerical flux verifies (119), the consistency property reduces to

$$F(W) = \left[\mathcal{R}_{\mathrm{L}}\left(W, W\right) + \mathcal{R}_{\mathrm{R}}\left(W, W\right)\right] W. \quad (81)$$

If the flux is a homogeneous function of degree one, (134) is satisfied in particular when

$$A(W) = \mathcal{R}_{\mathrm{L}}\left(W, W\right) + \mathcal{R}_{\mathrm{R}}\left(W, W\right) \quad (82)$$

where A is the Jacobian matrix of F. In fact, this is the hypothesis that B. Stoufflet considers in [13] when introducing matrices $\mathcal{R}_{\mathrm{L}}$ and $\mathcal{R}_{\mathrm{R}}$ to factorize the flux. ◇

Operator $\mathcal{P}^n$ for the Q-scheme of van Leer

The numerical flux of the Q-scheme of van Leer for the shallow water equations is a particular case for which property (119) holds.

Recall the expression of the numerical flux

$$\phi\left(U, V\right) = \frac{F(U) + F(V)}{2} - \frac{1}{2}\left|A\left(\frac{U+V}{2}\right)\right|\left(V - U\right). \quad (83)$$

As mentioned in the first section, the flux is not a homogeneous function. Nevertheless there exists a matrix $A^\star$ such that $F(W) = A^\star(W)W$ (see [15] for details). Then matrices $\mathcal{R}_{\mathrm{L}}$ and $\mathcal{R}_{\mathrm{R}}$ can be defined in the following way:

$$\mathcal{R}_{\mathrm{L}}\left(U, V\right) = \frac{1}{2}\left(A^\star(U) + \left|A\left(\frac{U+V}{2}\right)\right|\right), \quad (84)$$

$$\mathcal{R}_{\mathrm{R}}\left(U, V\right) = \frac{1}{2}\left(A^\star(V) - \left|A\left(\frac{U+V}{2}\right)\right|\right). \quad (85)$$

Next, we prove that the numerical source functions defined in [2] for the extension of the Q-scheme of van Leer for the shallow water equations verify (120) and (121).

First, recall the expressions of the mentioned numerical source:

$$\psi_{\mathrm{L}}(x,y,U,V) = \left\{ I + \left| A\left(\frac{U+V}{2}\right)\right| A^{-1}\left(\frac{U+V}{2}\right)\right\} \widehat{G}(x,y,U,V) \tag{86}$$

$$\psi_{\mathrm{R}}(x,y,U,V) = \left\{ I - \left| A\left(\frac{U+V}{2}\right)\right| A^{-1}\left(\frac{U+V}{2}\right)\right\} \widehat{G}(x,y,U,V). \tag{87}$$

then

$$\mathcal{S}_{\mathrm{L}}(x,y,U,V) = \frac{1}{2}\left[I + \left| A\left(\frac{U+V}{2}\right)\right| A^{-1}\left(\frac{U+V}{2}\right)\right] \mathcal{M}(x,y) \tag{88}$$

$$\mathcal{S}_{\mathrm{R}}(x,y,U,V) = \frac{1}{2}\left[I - \left| A\left(\frac{U+V}{2}\right)\right| A^{-1}\left(\frac{U+V}{2}\right)\right] \mathcal{M}(x,y). \tag{89}$$

where

$$\mathcal{M}(x,y) = \begin{pmatrix} 0 & 0 \\ g\dfrac{H(y)-H(x)}{y-x}. & 0 \end{pmatrix}. \tag{90}$$

3.2 A simplified linearized implicit scheme for the 2D shallow water equations

In this section we introduce a simplified linearized implicit scheme to the two-dimensional shallow water equations. As in the explicit case, the main aim of this section is to solve the system of conservation laws given by (3). The corresponding implicit scheme is

$$\frac{W^{n+1}-W^n}{\Delta t_n} + \mathcal{H}(W^{n+1}) = 0, \tag{91}$$

where, in this case

$$\mathcal{H}(W) = \mathcal{H}_{\mathrm{z}}(W) + \mathcal{H}_{\mathrm{G}}(W) \tag{92}$$

$$\left(\mathcal{H}_{\mathrm{z}}\left(W^{n+1}\right)\right)_i = \frac{1}{A_i}\sum_{j\in\mathcal{K}_i} \phi\left(W_i^{n+1}, W_j^{n+1}, \eta_{ij}\right) \tag{93}$$

$$\tag{94}$$

$$\left(\mathcal{H}_{\mathrm{G}}(W^{n+1})\right)_i = -\frac{1}{A_i}\sum_{j\in\mathcal{K}_i} A_{T_{ij}}\psi\left(N_i, N_j, W_i^{n+1}, W_j^{n+1}, \widetilde{\eta}_{ij}\right). \tag{95}$$

The procedure to obtain the operator $\mathcal{P}^n$ can be deduced from to the one-dimensional case for each pair of neighbouring nodes. Thus it is necessary to obtain matrices $\mathcal{R}_{\rm L}$, $\mathcal{R}_{\rm R}$ and $\mathcal{S}$ such that

$$\phi\left(U,V,\eta\right) = \mathcal{R}_{\rm L}\left(U,V,\eta\right)U + \mathcal{R}_{\rm R}\left(U,V,\eta\right)V. \tag{96}$$

$$\psi\left(N_1,N_2,U,V,\tilde{\eta}\right) = \mathcal{S}\left(N_1,N_2,U,V,\tilde{\eta}\right)\left(U+V\right) \tag{97}$$

Let us remark that requiring only one matrix $\mathcal{S}$, instead of $\mathcal{S}_{\rm L}$ and $\mathcal{S}_{\rm R}$ for the one-dimensional problems, is analogous to defining only one numerical source function. It is related with the dependence of ψ and $\mathcal{S}$ on the normal vector $\tilde{\eta}$.

For the flux of the Q-scheme of van Leer, matrices $\mathcal{R}_{\rm L}$ and $\mathcal{R}_{\rm R}$ are given by

$$\mathcal{R}_{\rm L}\left(U,V,\eta\right) = \frac{1}{2}\left\{\mathcal{A}^\star(U,\eta) + \left|\mathcal{A}\left(\frac{U+V}{2},\eta\right)\right|\right\} \tag{98}$$

$$\mathcal{R}_{\rm R}\left(U,V,\eta\right) = \frac{1}{2}\left\{\mathcal{A}^\star(V,\eta) - \left|\mathcal{A}\left(\frac{U+V}{2},\eta\right)\right|\right\}, \tag{99}$$

where, matrix $\mathcal{A}^\star$ is given in Section 2.
Then

$$\left(\mathcal{P}^n_{\rm z}\right)_{ii} = \frac{1}{A_i}\sum_{j\in\mathcal{K}_i}\mathcal{R}_{\rm L}\left(W^n_i,W^n_j,\eta_{ij}\right), \quad \left(\mathcal{P}^n_{\rm z}\right)_{ij} = \frac{1}{A_i}\mathcal{R}_{\rm R}\left(W^n_i,W^n_j,\eta_{ij}\right),\ j\in\mathcal{K}_i. \tag{100}$$

Finally, matrix $\mathcal{S}$ is given by

$$\mathcal{S}\left(N_1,N_2,U,V,\tilde{\eta}\right) = \left[I - |Q(U,V,\tilde{\eta})|Q^{-1}(U,V,\tilde{\eta})\right]\mathcal{M}(N_1,N_2,\tilde{\eta})$$

where

$$\mathcal{M}\left(N_i,N_j,\tilde{\eta}_{ij}\right) = \begin{pmatrix} 0 & 0 & 0 \\ g\left(\dfrac{H_j-H_i}{2d_{ij}}\right)\tilde{\eta}_{ij1} & 0 & 0 \\ g\left(\dfrac{H_j-H_i}{2d_{ij}}\right)\tilde{\eta}_{ij2} & 0 & 0 \end{pmatrix}. \tag{101}$$

The linear operator $\mathcal{P}^n_{\rm G}$ is thus:

$$\left(\mathcal{P}^n_{\rm G}\right)_{ii} = \frac{1}{A_i}\sum_{j\in\mathcal{K}_i}A_{{\rm T}_{ij}}\mathcal{S}\left(N_i,N_j,W^n_i,W^n_j,\widetilde{\eta}_{ij}\right), \tag{102}$$

$$\left(\mathcal{P}^n_{\rm G}\right)_{ij} = \frac{1}{A_i}A_{{\rm T}_{ij}}\mathcal{S}\left(N_i,N_j,W^n_i,W^n_j,\widetilde{\eta}_{ij}\right),\ j\in\mathcal{K}_i. \tag{103}$$

3.3 Discretization of the boundary conditions

In this section, we describe the treatment of the two types of boundary conditions for the implicit schemes. In the first stage, the diagonal blocks of operator $\mathcal{P}_{\mathrm{z}}^{n}$ are modified, and in the second one the value of a conservative variable is imposed.

1. *Open sea boundary*

   To modify operator $\mathcal{P}_{\mathrm{z}}^{n}$ we propose the following development:

   - Split the variable W as the sum of the projections on the first and the last two components respectively, that is,

   $$W = \widetilde{W} + \widetilde{\widetilde{W}} \tag{104}$$

   where

   $$\widetilde{W} = \begin{pmatrix} h \\ 0 \\ 0 \end{pmatrix}, \quad \widetilde{\widetilde{W}} = \begin{pmatrix} 0 \\ q_1 \\ q_2 \end{pmatrix}, \tag{105}$$

   the flux is split analogously:

   $$Z\left(W,\eta\right) = \widetilde{Z}\left(W,\eta\right) + \widetilde{\widetilde{Z}}\left(W,\eta\right), \tag{106}$$

   where

   $$\widetilde{Z}\left(W,\eta\right) = \mathcal{A}^{\star}(W,\eta)\widetilde{W}, \quad \widetilde{\widetilde{Z}} = \mathcal{A}^{\star}(W,\eta)\widetilde{\widetilde{W}}. \tag{107}$$

   - To modify $\mathcal{P}_{\mathrm{z}}^{n}$ we only take $\widetilde{\widetilde{Z}}$ into account, since for the boundary nodes belonging to this boundary, the solution of the system only affects q_1 and q_2.

   Since the following identity holds

   $$\widetilde{\widetilde{Z}}\left(W,\eta\right) = \begin{pmatrix} 0 & \eta_1 & \eta_2 \\ 0 & 2\eta_1\dfrac{q_1}{h} + \eta_2\dfrac{q_2}{h} & \eta_2\dfrac{q_1}{h} \\ 0 & \eta_1\dfrac{q_2}{h} & \eta_1\dfrac{q_1}{h} + 2\eta_2\dfrac{q_2}{h} \end{pmatrix} W, \tag{108}$$

the update of $\mathcal{P}_{\mathrm{z}}^n$ is equivalent to adding the matrix which appears in (162) to the diagonal block of the corresponding boundary node.

On this boundary (Γ_1), as in the explicit formulations, the value of the height of the fluid is imposed.

2. *Coast boundary*

The procedure is analogous to the previous one: to compute $\mathcal{P}_{\mathrm{z}}^n$ only $\widetilde{Z}$ is taken into account, due to the fact that for this type of boundary condition the solution of the system only affects variable h.

>From the identity

$$\widetilde{Z}\,(W,\eta) = \begin{pmatrix} 0 & 0 & 0 \\ \eta_1\left(-\dfrac{q_1^2}{h^2}+\dfrac{1}{2}gh\right)+\eta_2\left(-\dfrac{q_1q_2}{h^2}\right) & 0 & 0 \\ \eta_1\left(-\dfrac{q_1q_2}{h^2}\right)+\eta_2\left(-\dfrac{q_2^2}{h^2}+\dfrac{1}{2}gh\right) & 0 & 0 \end{pmatrix} W, \tag{109}$$

if the condition of null flux is introduced in (162), one obtains

$$\widetilde{Z}\,(W,\eta) = \begin{pmatrix} 0 & 0 & 0 \\ \eta_1\dfrac{1}{2}gh & 0 & 0 \\ \eta_2\dfrac{1}{2}gh & 0 & 0 \end{pmatrix} W \tag{110}$$

since $q_1\eta_1 + q_2\eta_2 = 0$.

Then updating $\mathcal{P}_{\mathrm{z}}^n$ it is equivalent to adding the matrix in (163) to the diagonal block of the corresponding boundary node.

Finally, over the boundary (Γ_2) the slip condition is imposed strongly, in a way similar to the explicit case.

3.4 A conservation property

Recall that the problem considered in the definition of the $\mathcal{C}$-*Property* is stationary. Thus, for this property to be satisfied by the numerical scheme, it is necessary that

$\delta W^{n+1} = 0$. Now then, since the *exact* $\mathcal{C}$*-Property* is satisfied by the explicit scheme, if follows that $\mathcal{H}(W^n)$ is null.

The aim is to solve a homogeneous system, and we are interested on its trivial solution. The unicity of this solution is guaranteed, at least for Δt sufficiently small, since in that case the matrix is diagonally dominant.

3.5 Numerical results

The problem of propagation of the tidal wave in the Pontevedra ria is also used for the implicit schemes.

The numerical results are very similar to those obtained with the implicit ones. However, in the preset methods the Courant number μ was set equal to 150, whereas the explicit scheme could only be operated stably with $\mu = 0.9$.

The computations were carried on an IBM RISC/6000 and the CPU time necessary to compute a tidal cycle was 126390^{sec} with the explicit scheme, and only 6592^{sec} for the implicit one. Thus, a substantial gain in efficiency was realized by the implicit formulation.

4 Implicit discretizations.

4.1 Implicit discretizations of 1D problems

In the previous section, the time discretization was explicit, and this implied a severe restriction on time step.

Indeed, for a given mesh, according to the Courant-Friedrichs-Lewy condition, the time step is determined by the largest eigenvalue of the Jacobian matrix of the flux. For the tidal current computation, this eigenvalue is of the same order of magnitude as the velocity of propagations of waves, i.e. $c = \sqrt{gh}$ which, in our previous example, is about $20m/s$, i.e. two orders of magnitude higher than the velocity of particles u $\left(u = \frac{q}{h}\right)$ which is the relevant unknown in the this kind of applications.

The main aim of this section is to combine the upwind schemes presented above with an implicit discretization in time to obtain unconditional linear stability.

For the sake of simplicity, the presentation of the resulting schemes is done first in one dimension. The *simplified linearized implicit scheme* we apply is similar to that introduced by B. Stoufflet in [13] for systems of conservation laws without source terms and in the case where the flux is a homogeneous function of degree one.

The problem to be solved in this section is a hyperbolic nonlinear system of the form

$$\frac{\partial w}{\partial t}(x,t) + \frac{\partial F}{\partial x}\left(w(x,t)\right) = G(x, w(x,t)). \tag{111}$$

The vector form of an implicit scheme to solve (111) using upwind schemes for flux and source terms is

$$\frac{W^{n+1} - W^n}{\Delta t} + \mathcal{H}\left(W^{n+1}\right) = 0, \tag{112}$$

where

$$\mathcal{H}(W) = \mathcal{H}_{\text{F}}(W) + \mathcal{H}_{\text{G}}(W) \tag{113}$$

$$\left(\mathcal{H}_{\text{F}}\left(W^{n+1}\right)\right)_j = \frac{\phi\left(W_j^{n+1}, W_{j+1}^{n+1}\right) - \phi\left(W_{j-1}^{n+1}, W_j^{n+1}\right)}{A_j} \tag{114}$$

$$\begin{aligned}\left(\mathcal{H}_{\text{G}}\left(W^{n+1}\right)\right)_j = & -\frac{1}{A_j}\left\{A_{T_{j\text{L}}}\psi_{\text{L}}\left(x_{j-1}, x_j, W_{j-1}^{n+1}, W_j^{n+1}\right)\right. \\ & \left. + A_{T_{j\text{R}}}\psi_{\text{R}}\left(x_j, x_{j+1}, W_j^{n+1}, W_{j+1}^{n+1}\right)\right\}.\end{aligned} \tag{115}$$

The difficulty with this scheme is the high required computing power. Hence it is quite reasonable to replace $\mathcal{H}$ by a linealization

$$\mathcal{H}\left(W^{n+1}\right) \simeq \mathcal{H}\left(W^{n}\right) + \mathcal{H}'\left(W^{n}\right)\left(W^{n+1} - W^{n}\right). \tag{116}$$

The resulting scheme is called *linearized implicit scheme* (see Dervieux and Desideri [4]). It can be rewritten in the following δ-form:

$$\left(\frac{I}{\Delta t} + \mathcal{H}'\left(W^{n}\right)\right)\delta W^{n+1} = -\mathcal{H}\left(W^{n}\right). \tag{117}$$

where $\delta W^{n+1} = W^{n+1} - W^{n}$. Nevertheless, this scheme also presents difficulties. For instance in the case of the Q-schemes, due to the presence of the absolute value of matrix $\mathcal{A}$ in the expression of the flux, operator $\mathcal{H}'_{\text{F}}$ is not defined if the eigenvalues of this matrix change their sign. Moreover, in case that such operator exists, its computational cost may be high.

In such situations it is necessary to replace $\mathcal{H}'$ with another approximate linear operator, denoted $\mathcal{P}^n$, so that (117) takes the form

$$\left(\frac{I}{\Delta t} + \mathcal{P}^{n}\left(W^{n}\right)\right)\delta W^{n+1} = -\mathcal{H}\left(W^{n}\right). \tag{118}$$

These types of schemes are known as *simplified linearized implicit schemes* (see [13]).

Next, a procedure to realize one such construction is detailed.

Construction of the operator $\mathcal{P}^n$

In order to obtain the expression of $\mathcal{P}^n = \mathcal{P}^n_{\text{F}} + \mathcal{P}^n_{\text{G}}$, the corresponding part of the flux ($\mathcal{P}^n_{\text{F}}$) and the source term ($\mathcal{P}^n_{\text{G}}$) are introduced in parallel.

1. First we suppose that the flux and the source numerical functions verify the following property:
   There exist $\mathcal{R}_L$, $\mathcal{R}_R$ and $\mathcal{S}_L$ y $\mathcal{S}_R$ such that:

$$\phi\left(U,V\right) = \mathcal{R}_{\text{L}}\left(U,V\right)U + \mathcal{R}_{\text{R}}\left(U,V\right)V. \tag{119}$$

$$\psi_{\text{L}}\left(x,y,U,V\right) = \mathcal{S}_{\text{L}}\left(x,y,U,V\right)\left(U+V\right), \tag{120}$$

$$\psi_{\text{R}}\left(x,y,U,V\right) = \mathcal{S}_{\text{R}}\left(x,y,U,V\right)\left(U+V\right). \tag{121}$$

2. If the previous property is verified, then operators $\mathcal{H}_{\mathrm{F}}$ and $\mathcal{H}_{\mathrm{G}}$ can be written in the form

$$\begin{aligned}\left(\mathcal{H}_{\mathrm{F}}\left(W^{n+1}\right)\right)_j &= \frac{1}{A_j}\left[\mathcal{R}_{\mathrm{L}}\left(W_j^{n+1},W_{j+1}^{n+1}\right)W_j^{n+1}+\mathcal{R}_{\mathrm{R}}\left(W_j^{n+1},W_{j+1}^{n+1}\right)W_{j+1}^{n+1}\right]\\ &- \frac{1}{A_j}\left[\mathcal{R}_{\mathrm{L}}\left(W_{j-1}^{n+1},W_j^{n+1}\right)W_{j-1}^{n+1}+\mathcal{R}_{\mathrm{R}}\left(W_{j-1}^{n+1},W_j^{n+1}\right)W_j^{n+1}\right] \quad (122)\end{aligned}$$

$$\begin{aligned}\left(\mathcal{H}_{\mathrm{G}}\left(W^{n+1}\right)\right)_j &= -\frac{A_{T_{j\mathrm{R}}}}{A_j}\left[\mathcal{S}_{\mathrm{R}}\left(x_j,x_{j+1},W_j^{n+1},W_{j+1}^{n+1}\right)\left(W_j^{n+1}+W_{j+1}^{n+1}\right)\right]\\ &+\frac{A_{T_{j\mathrm{L}}}}{A_j}\left[\mathcal{S}_{\mathrm{L}}\left(x_{j-1},x_j,W_{j-1}^{n+1},W_j^{n+1}\right)\left(W_{j-1}^{n+1}+W_j^{n+1}\right)\right]. \quad (123)\end{aligned}$$

3. Then we change (122) and (123) by evaluating matrix $\mathcal{R}_{\mathrm{L}}$, $\mathcal{R}_{\mathrm{R}}$, $\mathcal{S}_{\mathrm{L}}$ and $\mathcal{S}_{\mathrm{R}}$ at time t_n instead of t_{n+1}. The resulting approximations are

$$\begin{aligned}\left(\mathcal{H}_{\mathrm{F}}\left(W^{n+1}\right)\right)_j &\simeq \frac{1}{A_j}\left[\mathcal{R}_{\mathrm{L}}\left(W_j^{n},W_{j+1}^{n}\right)W_j^{n+1}+\mathcal{R}_{\mathrm{R}}\left(W_j^{n},W_{j+1}^{n}\right)W_{j+1}^{n+1}\right]\\ &+ \frac{1}{A_j}\left[-\mathcal{R}_{\mathrm{L}}\left(W_{j-1}^{n},W_j^{n}\right)W_{j-1}^{n+1}-\mathcal{R}_{\mathrm{R}}\left(W_{j-1}^{n},W_j^{n}\right)W_j^{n+1}\right]. \quad (124)\end{aligned}$$

$$\begin{aligned}\left(\mathcal{H}_{\mathrm{G}}\left(W^{n+1}\right)\right)_j &\simeq \frac{T_{j\mathrm{R}}}{A_j}\left[-\mathcal{S}_{\mathrm{R}}\left(x_j,x_{j+1},W_j^{n},W_{j+1}^{n}\right)\left(W_j^{n+1}+W_{j+1}^{n+1}\right)\right]\\ &+ \frac{T_{j\mathrm{L}}}{A_j}\left[\mathcal{S}_{\mathrm{L}}\left(x_{j-1},x_j,W_{j-1}^{n},W_j^{n}\right)\left(W_{j-1}^{n+1}+W_j^{n+1}\right)\right] \quad (125)\end{aligned}$$

4. Finally, if we introduce $\mathcal{H}_{\mathrm{F}}\left(W^n\right)$ and $\mathcal{H}_{\mathrm{G}}\left(W^n\right)$ in (124) and (125) respectively, the following expressions $\mathcal{H}_{\mathrm{F}}\left(W^{n+1}\right)$ and $\mathcal{H}_{\mathrm{G}}\left(W^{n+1}\right)$ are obtained,

$$\mathcal{H}_{\mathrm{F}}(W^{n+1}) \simeq \mathcal{H}_{\mathrm{F}}(W^n)+\mathcal{P}_{\mathrm{F}}^n(W^n)\delta W^{n+1} \quad (126)$$

$$\mathcal{H}_{\mathrm{G}}(W^{n+1}) \simeq \mathcal{H}_{\mathrm{G}}(W^n)+\mathcal{P}_{\mathrm{G}}^n(W^n)\delta W^{n+1}, \quad (127)$$

where

$$\left(\mathcal{P}_{\mathrm{F}}^n\left(W^n\right)\right)_{jj} = \frac{1}{A_j}\left[\mathcal{R}_{\mathrm{L}}\left(W_j^n,W_{j+1}^n\right)-\mathcal{R}_{\mathrm{R}}\left(W_{j-1}^n,W_j^n\right)\right] \quad (128)$$

$$\left(\mathcal{P}_{\mathrm{F}}^n\left(W^n\right)\right)_{j\,j+1} = \frac{1}{A_j}\mathcal{R}_{\mathrm{R}}\left(W_j^n,W_{j+1}^n\right), \quad (129)$$

$$\left(\mathcal{P}_{\mathrm{F}}^n\left(W^n\right)\right)_{j-1\,j} = -\frac{1}{A_j}\mathcal{R}_{\mathrm{L}}\left(W_{j-1}^n,W_j^n\right). \quad (130)$$

$$\left(\mathcal{P}_{\mathrm{G}}^{n}\left(W^{n}\right)\right)_{jj}=\frac{T_{j\mathrm{R}}}{A_{j}}\mathcal{S}_{\mathrm{R}}\left(x_{j},x_{j+1},W_{j}^{n},W_{j+1}^{n}\right)+\frac{T_{j\mathrm{L}}}{A_{j}}\mathcal{S}_{\mathrm{L}}\left(x_{j-1},x_{j},W_{j-1}^{n},W_{j}^{n}\right), \tag{131}$$

$$\left(\mathcal{P}_{\mathrm{G}}^{n}\left(W^{n}\right)\right)_{j\,j+1}=\frac{T_{j\mathrm{R}}}{A_{j}}\mathcal{S}_{\mathrm{R}}\left(x_{j},x_{j+1},W_{j}^{n},W_{j+1}^{n}\right), \tag{132}$$

$$\left(\mathcal{P}_{\mathrm{G}}^{n}\left(W^{n}\right)\right)_{j-1\,j}=\frac{T_{j\mathrm{L}}}{A_{j}}\mathcal{S}_{\mathrm{L}}\left(x_{j-1},x_{j},W_{j-1}^{n},W_{j}^{n}\right). \tag{133}$$

Remark 4

When the numerical flux verifies (119), the consistency property reduces to

$$F(W)=\left[\mathcal{R}_{\mathrm{L}}\left(W,W\right)+\mathcal{R}_{\mathrm{R}}\left(W,W\right)\right]W. \tag{134}$$

If the flux is a homogeneous function of degree one, (134) is satisfied in particular when

$$A(W)=\mathcal{R}_{\mathrm{L}}\left(W,W\right)+\mathcal{R}_{\mathrm{R}}\left(W,W\right) \tag{135}$$

where A is the Jacobian matrix of F. In fact, this is the hypothesis that B. Stoufflet considers in [13] when introducing matrices $\mathcal{R}_{\mathrm{L}}$ and $\mathcal{R}_{\mathrm{R}}$ to factorize the flux. ◇

Operator $\mathcal{P}^n$ for the Q-scheme of van Leer

The numerical flux of the Q-scheme of van Leer for the shallow water equations is a particular case for which property (119) holds.

Recall the expression of the numerical flux

$$\phi\left(U,V\right)=\frac{F(U)+F(V)}{2}-\frac{1}{2}\left|A\left(\frac{U+V}{2}\right)\right|\left(V-U\right). \tag{136}$$

As mentioned in the first section, the flux is not a homogeneous function. Nevertheless there exists a matrix $A^{\star}$ such that $F(W)=A^{\star}(W)W$ (see [15] for details). Then matrices $\mathcal{R}_{\mathrm{L}}$ and $\mathcal{R}_{\mathrm{R}}$ can be defined in the following way:

$$\mathcal{R}_{\mathrm{L}}\left(U,V\right)=\frac{1}{2}\left(A^{\star}(U)+\left|A\left(\frac{U+V}{2}\right)\right|\right), \tag{137}$$

$$\mathcal{R}_{\mathrm{R}}\left(U,V\right)=\frac{1}{2}\left(A^{\star}(V)-\left|A\left(\frac{U+V}{2}\right)\right|\right). \tag{138}$$

Next, we prove that the numerical source functions defined in [2] for the extension of the Q-scheme of van Leer for the shallow water equations verify (120) and (121).

First, recall the expressions of the mentioned numerical source:

$$\psi_{\mathrm{L}}(x,y,U,V) = \left\{ I + \left| A\left(\frac{U+V}{2}\right)\right| A^{-1}\left(\frac{U+V}{2}\right)\right\} \widehat{G}(x,y,U,V) \tag{139}$$

$$\psi_{\mathrm{R}}(x,y,U,V) = \left\{ I - \left| A\left(\frac{U+V}{2}\right)\right| A^{-1}\left(\frac{U+V}{2}\right)\right\} \widehat{G}(x,y,U,V)\,. \tag{140}$$

then

$$\mathcal{S}_{\mathrm{L}}(x,y,U,V) = \frac{1}{2}\left[I + \left| A\left(\frac{U+V}{2}\right)\right| A^{-1}\left(\frac{U+V}{2}\right)\right] \mathcal{M}(x,y) \tag{141}$$

$$\mathcal{S}_{\mathrm{R}}(x,y,U,V) = \frac{1}{2}\left[I - \left| A\left(\frac{U+V}{2}\right)\right| A^{-1}\left(\frac{U+V}{2}\right)\right] \mathcal{M}(x,y)\,. \tag{142}$$

where

$$\mathcal{M}(x,y) = \begin{pmatrix} 0 & 0 \\ g\dfrac{H(y)-H(x)}{y-x}. & 0 \end{pmatrix}. \tag{143}$$

4.2 A simplified linearized implicit scheme for the 2D shallow water equations

In this section we introduce a simplified linearized implicit scheme to the two-dimensional shallow water equations. As in the explicit case, the main aim of this section is to solve the system of conservation laws given by (3). The corresponding implicit scheme is

$$\frac{W^{n+1}-W^n}{\Delta t_n} + \mathcal{H}(W^{n+1}) = 0, \tag{144}$$

where, in this case

$$\mathcal{H}(W) = \mathcal{H}_{\mathrm{z}}(W) + \mathcal{H}_{\mathrm{G}}(W) \tag{145}$$

$$\left(\mathcal{H}_{\mathrm{z}}\left(W^{n+1}\right)\right)_i = \frac{1}{A_i}\sum_{j\in\mathcal{K}_i} \phi\left(W_i^{n+1}, W_j^{n+1}, \eta_{ij}\right) \tag{146}$$

$$\tag{147}$$

$$\left(\mathcal{H}_{\mathrm{G}}(W^{n+1})\right)_i = -\frac{1}{A_i}\sum_{j\in\mathcal{K}_i} A_{T_{ij}}\psi\left(N_i, N_j, W_i^{n+1}, W_j^{n+1}, \widetilde{\eta}_{ij}\right). \tag{148}$$

The procedure to obtain the operator $\mathcal{P}^n$ can be deduced from to the one-dimensional case for each pair of neighbouring nodes. Thus it is necessary to obtain matrices $\mathcal{R}_{\rm L}$, $\mathcal{R}_{\rm R}$ and $\mathcal{S}$ such that

$$\phi\left(U,V,\eta\right)=\mathcal{R}_{\rm L}\left(U,V,\eta\right)U+\mathcal{R}_{\rm R}\left(U,V,\eta\right)V. \tag{149}$$

$$\psi\left(N_1,N_2,U,V,\tilde{\eta}\right)=\mathcal{S}\left(N_1,N_2,U,V,\tilde{\eta}\right)\left(U+V\right) \tag{150}$$

Let us remark that requiring only one matrix $\mathcal{S}$, instead of $\mathcal{S}_{\rm L}$ and $\mathcal{S}_{\rm R}$ for the one-dimensional problems, is analogous to defining only one numerical source function. It is related with the dependence of ψ and $\mathcal{S}$ on the normal vector $\tilde{\eta}$.

For the flux of the Q-scheme of van Leer, matrices $\mathcal{R}_{\rm L}$ and $\mathcal{R}_{\rm R}$ are given by

$$\mathcal{R}_{\rm L}\left(U,V,\eta\right) = \frac{1}{2}\left\{\mathcal{A}^{\star}(U,\eta)+\left|\mathcal{A}\left(\frac{U+V}{2},\eta\right)\right|\right\} \tag{151}$$

$$\mathcal{R}_{\rm R}\left(U,V,\eta\right) = \frac{1}{2}\left\{\mathcal{A}^{\star}(V,\eta)-\left|\mathcal{A}\left(\frac{U+V}{2},\eta\right)\right|\right\}, \tag{152}$$

where, matrix $\mathcal{A}^{\star}$ is given in Section 2.
Then

$$\left(\mathcal{P}_{\rm z}^n\right)_{ii}=\frac{1}{A_i}\sum_{j\in\mathcal{K}_i}\mathcal{R}_{\rm L}\left(W_i^n,W_j^n,\eta_{ij}\right),\quad \left(\mathcal{P}_{\rm z}^n\right)_{ij}=\frac{1}{A_i}\mathcal{R}_{\rm R}\left(W_i^n,W_j^n,\eta_{ij}\right),\ j\in\mathcal{K}_i. \tag{153}$$

Finally, matrix $\mathcal{S}$ is given by

$$\mathcal{S}\left(N_1,N_2,U,V,\tilde{\eta}\right)=\left[I-|Q(U,V,\tilde{\eta})|Q^{-1}(U,V,\tilde{\eta})\right]\mathcal{M}(N_1,N_2,\tilde{\eta})$$

where

$$\mathcal{M}\left(N_i,N_j,\tilde{\eta}_{ij}\right)=\begin{pmatrix} 0 & 0 & 0\\ g\left(\dfrac{H_j-H_i}{2d_{ij}}\right)\tilde{\eta}_{ij1} & 0 & 0\\ g\left(\dfrac{H_j-H_i}{2d_{ij}}\right)\tilde{\eta}_{ij2} & 0 & 0\end{pmatrix}. \tag{154}$$

The linear operator $\mathcal{P}_{\rm G}^n$ is thus:

$$\left(\mathcal{P}_{\rm G}^n\right)_{ii} = \frac{1}{A_i}\sum_{j\in\mathcal{K}_i}A_{{\rm T}_{ij}}\mathcal{S}\left(N_i,N_j,W_i^n,W_j^n,\widetilde{\eta}_{ij}\right), \tag{155}$$

$$\left(\mathcal{P}_{\rm G}^n\right)_{ij} = \frac{1}{A_i}A_{{\rm T}_{ij}}\mathcal{S}\left(N_i,N_j,W_i^n,W_j^n,\widetilde{\eta}_{ij}\right),\ j\in\mathcal{K}_i. \tag{156}$$

4.3 Discretization of the boundary conditions

In this section, we describe the treatment of the two types of boundary conditions for the implicit schemes. In the first stage, the diagonal blocks of operator $\mathcal{P}_{\mathrm{z}}^n$ are modified, and in the second one the value of a conservative variable is imposed.

1. *Open sea boundary*

   To modify operator $\mathcal{P}_{\mathrm{z}}^n$ we propose the following development:

   - Split the variable W as the sum of the projections on the first and the last two components respectively, that is,

   $$W = \widetilde{W} + \widetilde{\widetilde{W}} \tag{157}$$

   where

   $$\widetilde{W} = \begin{pmatrix} h \\ 0 \\ 0 \end{pmatrix}, \quad \widetilde{\widetilde{W}} = \begin{pmatrix} 0 \\ q_1 \\ q_2 \end{pmatrix}, \tag{158}$$

   the flux is split analogously:

   $$Z(W,\eta) = \widetilde{Z}(W,\eta) + \widetilde{\widetilde{Z}}(W,\eta), \tag{159}$$

   where

   $$\widetilde{Z}(W,\eta) = \mathcal{A}^\star(W,\eta)\widetilde{W}, \quad \widetilde{\widetilde{Z}} = \mathcal{A}^\star(W,\eta)\widetilde{\widetilde{W}}. \tag{160}$$

   - To modify $\mathcal{P}_{\mathrm{z}}^n$ we only take $\widetilde{\widetilde{Z}}$ into account, since for the boundary nodes belonging to this boundary, the solution of the system only affects q_1 and q_2.

   Since the following identity holds

   $$\widetilde{\widetilde{Z}}(W,\eta) = \begin{pmatrix} 0 & \eta_1 & \eta_2 \\ 0 & 2\eta_1\dfrac{q_1}{h} + \eta_2\dfrac{q_2}{h} & \eta_2\dfrac{q_1}{h} \\ 0 & \eta_1\dfrac{q_2}{h} & \eta_1\dfrac{q_1}{h} + 2\eta_2\dfrac{q_2}{h} \end{pmatrix} W, \tag{161}$$

the update of $\mathcal{P}_z^n$ is equivalent to adding the matrix which appears in (162) to the diagonal block of the corresponding boundary node.

On this boundary (Γ_1), as in the explicit formulations, the value of the height of the fluid is imposed.

2. *Coast boundary*

The procedure is analogous to the previous one: to compute $\mathcal{P}_z^n$ only $\widetilde{Z}$ is taken into account, due to the fact that for this type of boundary condition the solution of the system only affects variable h.

>From the identity

$$\widetilde{Z}(W,\eta) = \begin{pmatrix} 0 & 0 & 0 \\ \eta_1\left(-\frac{q_1^2}{h^2}+\frac{1}{2}gh\right)+\eta_2\left(-\frac{q_1q_2}{h^2}\right) & 0 & 0 \\ \eta_1\left(-\frac{q_1q_2}{h^2}\right)+\eta_2\left(-\frac{q_2^2}{h^2}+\frac{1}{2}gh\right) & 0 & 0 \end{pmatrix} W, \tag{162}$$

if the condition of null flux is introduced in (162), one obtains

$$\widetilde{Z}(W,\eta) = \begin{pmatrix} 0 & 0 & 0 \\ \eta_1\frac{1}{2}gh & 0 & 0 \\ \eta_2\frac{1}{2}gh & 0 & 0 \end{pmatrix} W \tag{163}$$

since $q_1\eta_1 + q_2\eta_2 = 0$.

Then updating $\mathcal{P}_z^n$ it is equivalent to adding the matrix in (163) to the diagonal block of the corresponding boundary node.

Finally, over the boundary (Γ_2) the slip condition is imposed strongly, in a way similar to the explicit case.

4.4 A conservation property

Recall that the problem considered in the definition of the *C-Property* is stationary. Thus, for this property to be satisfied by the numerical scheme, it is necessary that

$\delta W^{n+1} = 0$. Now then, since the *exact* $\mathcal{C}$*-Property* is satisfied by the explicit scheme, if follows that $\mathcal{H}(W^n)$ is null.

The aim is to solve a homogeneous system, and we are interested on its trivial solution. The unicity of this solution is guaranteed, at least for Δt sufficiently small, since in that case the matrix is diagonally dominant.

4.5 Numerical results

The problem of propagation of the tidal wave in the Pontevedra ria is also used for the implicit schemes.

The numerical results are very similar to those obtained with the implicit ones. However, in the preset methods the Courant number μ was set equal to 150, whereas the explicit scheme could only be operated stably with $\mu = 0.9$.

The computations were carried on an IBM RISC/6000 and the CPU time necessary to compute a tidal cycle was 126390^{sec} with the explicit scheme, and only 6592^{sec} for the implicit one. Thus, a substantial gain in efficiency was realized by the implicit formulation.

References

[1] F. Alcrudo and P. García-Navarro. A high-resolution Godunov-type scheme in finite volumes for the 2d shallow-water equations. *International Journal for Numerical Methods in Fluids*, 16:489–505, 1993.

[2] A. Bermúdez and M.E. Vázquez. Upwind methods for hyperbolic conservation laws with source terms. *Computers and Fluids*, 23:1049–1071, 1994.

[3] M. Crouzeix and P.A. Raviart. Conforming and non-conforming finite-element methods for solving the stationary stokes equations. *Rairo*, R-3:33–75, 1973.

[4] A. Dervieux and J. A. Desideri. Compressible flow solvers using unstructured grids. Rapports de Recherche 1732, INRIA, 1992.

[5] R. J. Fennema and M. H. Chaudhry. Explicit methods for 2-D transient free-surface flows. *J. Hyd. Eng.*, 116-No.8:1013–1034, 1990.

[6] G. Gambolati, A. Rinaldo, C. A. Brebbia, W. G. Gray, and G. F.Pinder. *Computational Methods in Surface Hydrology.* Computational Mechanics Publications, Southampton, 1990.

[7] P. Glaister. Approximate Riemann solutions of the shallow water equations. *J. Hydraulic Research*, 26:293–305, 1988.

[8] P. Glaister. An efficient numerical scheme for the two-dimensional shallow water equations using arithmetic averaging. *Computers Math. Applic.*, 27:97–117, 1994.

[9] A. Harten. On a class of high resolution total-variation-stable finite-difference schemes. *SIAM J. Numer. Anal.*, 21(1):1–23, February 1984.

[10] A. Harten, P. Lax, and A. van Leer. On upstream differencing and Godunov-type schemes for hyperbolic conservation laws. *SIAM Rev.*, 25:35–61, 1983.

[11] J. Steger and R.F. Warming. Flux vector splitting of the inviscid gasdynamic equations with application to finite-difference methods. *J. Comput. Phys.*, 40, 1981.

[12] J.J. Stoker. *Water Waves.* Interscience, New York, 1957.

[13] B. Stoufflet. *Résolution numérique des équations d'Euler des fluides parfaits compressibles par des schémas implicites en éléments finis.* PhD thesis, Université de Paris VI, 1984.

[14] B. van Leer. Progress in multi-dimensional upwind differencing. ICASE Report 92/43, NASA Langley Research Center, Hampton,VA, 1984. Also to appear Proc. of the 13th Int. Conf. on Numerical Methods in Fluid Dynamics held in Rome, July 6-10,1992.

[15] M. E. Vázquez. *Estudio de esquemas descentrados para su aplicación a las leyes de conservación hiperbólicas con términos fuente.* PhD thesis, Departamento de Matemática Aplicada. Universidad de Santiago de Compostela. Spain, 1994.

[16] G. Vijayasundaram. Transonic flow simulations using an upstream centered scheme of Godunov in finite elements. *J. Comput. Phys.*, 63:416–433, 1986.

Unité de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unité de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unité de recherche INRIA Rhône-Alpes, 46 avenue Félix Viallet, 38031 GRENOBLE Cedex 1
Unité de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unité de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex

Éditeur

INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)

ISSN 0249-6399