On well-balanced Finite Volume Methods for non-conservative non-homogeneous hyperbolic systems. [*]

Tomás Chacón Rebollo [†] Enrique D. Fernández-Nieto [‡]
Manuel J. Castro Díaz [§]& Carlos Parés

February 2, 2006

Abstract

In this work we introduce a general family of finite volume methods for non-homogeneous hyperbolic systems with non-conservative terms. We prove that all of them are "asymptotically well-balanced": They preserve all smooth stationary solutions in all the domain but a set whose measure tends to zero as Δx tends to zero. This theory is applied to solve the bilayer Shallow-Water equations with arbitrary cross-section. Finally, some numerical tests are presented for simplified but meaningful geometries, comparing the computed solution with approximated asymptotic analytical solutions.

Short title : Well-balanced Finite volume solvers.

Keywords : Well-balanced, Finite Volume Method, upwinding, shallow water, source terms, two-layer flows.

Subject Classifications : AMS (MOS) : 65N06, 76B15, 76M20, 76N99.

1 Introduction.

This work deals with finite volume solvers for non-homogeneous hyperbolic systems with non-conservative terms. Our goal is to design numerical schemes that calculate with high accuracy the stationary solutions of the system.

The accurate computation of stationary solutions of hyperbolic systems with source terms has been found in the past years as closely related to the accurate computation of transient solutions. Numerical schemes that do not solve steady solutions up to second order, at least, yield transient solutions that present large errors, unacceptable from the physical point of view. To avoid this difficulty, in [1] Bermúdez and Vázquez introduced the so-called "C-property" in the context of Shallow-Water equations. They ask the numerical schemes to exactly calculate at grid nodes the stationary solution corresponding to water at rest. They also present some numerical evidence of the large errors presented by schemes that do not verify this property. In that work, the scheme of Roe is generalized to non homogeneous hyperbolic systems, in such a way that it verifies the C-property.

This property has given fruitful extensions to more general situations: numerical models, hyperbolic equations, and extensions of the condition itself.

[*] This research was partially supported by Spanish Government Research Projects BFM2003-07530-C02-01 and BFM2003-07530-C02-02.

[†] Departamento de Ecuaciones Diferenciales y Análisis Numérico, Universidad de Sevilla. C/ Tarfia, s/n. 41080 Sevilla, Spain (chacon@us.es).

[‡] Departamento de Matemática Aplicada I, Universidad de Sevilla. 41012 Sevilla, Spain (edofer@us.es).

[§] Dpto. Análisis Matemático, Universidad de Málaga, Campus Teatinos s/n, Málaga, Spain (castro@anamat.cie.uma.es),(pares@anamat.cie.uma.es).

For instance, an extension to kinetic schemes was introduced by Perthame and Simeoni in [16]. This type of schemes is based upon the discretization of an equation for the density of particles. Also, Kurganov and Levy in [14] introduced the central upwind schemes for Shallow-Water equations, characterized by only including scalar numerical diffusion. No upwinding matrix is needed. Both schemes ensure the positiveness of the water height, and verify the C-property.

The relevance of this property as a consistency condition was pointed out in the work of Perthame and Simeoni [17]. They prove the convergence, in the sense of distributions, of the numerical solution computed by a finite volume scheme for a non-homogeneous equation under some hypothesis of consistency of the numerical scheme. This hypothesis is fulfilled if the numerical scheme preserves certain equilibrium states of the system.

An extension of the C-condition to a more general condition was introduced in Greenberg and Leroux [13]: the concept of "well-balanced"scheme, as a scheme that preserves all equilibria of the system at grid nodes. In this work, a numerical scheme for non-homogeneous scalar equations satisfying this condition is introduced.

An extension to a class of non-conservative hyperbolic system of equations is introduced in [2] by Castro, Macías and Parés by means of a generalized Roe's scheme. The systems considered include a non-conservative product with the structure $B(W)\dfrac{\partial W}{\partial x}$ where B is a matrix and W is the unknown. The numerical stability of this scheme is achieved by using upwinding matrices defined not only from the Jacobian of the flux, but also from the matrix B involved by the non-conservative terms. The scheme introduced is applied to the 1d two-layer shallow water system for channels with constant width. The scheme exactly computes water at rest. More recently, in [4] this Roe's scheme has been extended to the more general case of channels with irregular geometry and not necessarily rectangular cross-section. This extended numerical scheme was applied to the simulation of the Strait of Gibraltar.

The derivation of systematic techniques to build numerical schemes satisfying the C-property is also a relevant issue, as this is far from being a straightforward condition to obtain. In [5] Chacón, Domínguez and Fernández-Nieto introduced the idea of separately compensate the centered and decentered components of the numerical flux by centered and decentered discretizations of the source terms. In [6] the same authors introduced the idea of adding an additional source term to the parabolic equivalent equation. This term is specifically designed to compensate for the first-order error generated by the numerical diffusion term on steady solutions. Then, centered discretizations automatically provide second-order accurate approximations of all steady solutions.

In this work, we step on this process by considering a general non-homogeneous hyperbolic system of equations with non-conservative terms, which also depends upon the space variable. We derive a general class of numerical schemes that we prove to be "asymptotically well-balanced" in the following sense: they calculate, with at least order two, all stationary solutions of the system in all the domain but on a set whose measure tends to zero as Δx tends to zero. One of the free parameters of this class is the diffusion matrix, so we include both flux-splitting and flux-difference or Roe's methods as particular cases.

We also give general sufficient conditions under which the schemes exactly calculate a given stationary solution. These conditions may be read as meaning on one hand that the numerical flux of the scheme must compensate the centered part of the numerical source term, and on another hand that the numerical diffusion must compensate the decentered part of the numerical source term.

This paper is organized as follows: In Section 2, a family of numerical schemes for a general non-homogeneous hyperbolic systems of equations with non-conservative terms is introduced. In Section 3, we introduce the concept of "asymptotically well-balanced" scheme and we prove that the introduced numerical schemes satisfy this property. Section 5 is devoted to the construction of numerical diffusion matrices that ensure the stability of our schemes for linear systems, while ensuring the asymptotically well-balanced computation of all steady solutions. In Section 5, we apply the theory developed to the bilayer Shallow-Water equations with arbitrary cross-sections to construct a family of asymptotically well-balanced numerical schemes. The Roe-type scheme introduced in [4] is a particular case. In Section 6, we present some numerical tests for simplified geometries, where exact solutions can be determined. The purpose of these experiments is to compare the properties of the here introduced numerical schemes for the two-layer shallow water system. More precisely, we focus on the well-balance property and the quality of the numerical approximations for smooth transitions and shocks. We present our conclusions in Section 7. In particular we conclude that the use of pointwise adapted discretizations of the centered part of the numerical flux helps to construct accurate solvers for non-conservative equations.

2 Non-homogeneous non-conservative hyperbolic systems.

We shall consider the following system:

$$\begin{cases} \dfrac{\partial W(x,t)}{\partial t} + \dfrac{\partial}{\partial x}[F(x,W(x,t))] = G(x,W(x,t)) + B(x,W(x,t))\dfrac{\partial W(x,t)}{\partial x}, \\ \hfill x \in (0,L), \quad t \in (0,T); \\ W(x,0) = W_0(x) \hfill x \in (0,L); \end{cases} \tag{1}$$

where by $F : [0,L] \times \mathbf{R}^N \to \mathbf{R}^N$ we denote the physical flux function, by $G : [0,L] \times \mathbf{R}^N \to \mathbf{R}^N$ the source term, and $B : [0,L] \times \mathbf{R}^N \to \mathbf{R}^N \times \mathbf{R}^N$ is a regular matrix function of x and W. We will suppose that F, G and B are C^2 functions. By L we denote the length of the domain and by T the final time. We will complete (1) with appropriate boundary conditions, that we set in each problem.

System (1) includes the non-conservative product

$$B(x,W)\frac{\partial W}{\partial x},$$

which is not a distribution if W is discontinuous. Its presence requires an adaptation of the usual concept of weak solutions: Across a discontinuity of W, the non conservative product can be defined along paths connecting the two states, following DalMasso-LeFloch & Murat [8].

System (1) also includes a dependency upon the space variable of the flux F. Both characteristics appear in two layer Shallow Water equations in channels of variable breadth; as we shall see later on. However, these characteristics also appear in other flow models such as boiling flows and two-phase flows among other (See Fowler [11]).

We suppose that the system (1) is hyperbolic, in the sense that $\forall x \in [0,L]$ and $\forall W \in \mathbf{R}^d$ the matrix

$$M(x,W) = A(x,W) - B(x,W) \tag{2}$$

can be diagonalized and all its eigenvalues are real and different, where $A(x,W) = \dfrac{\partial F}{\partial W}(x,W)$ is the Jacobian matrix of the flux function F.

We consider a partition $\{x_i\}_{i=0}^{M+1}$ of the interval $[0,L]$, and a partition $\{t^n\}_{n=0}^{m+1}$ of $[0,T]$. By W_i^n we define an approximation to the average of the solution in the control volume, $(x_{i-1/2}, x_{i+1/2})$ at $t = t^n$.

In this work, for the space discretization, we develop finite volume schemes that can be written in viscous form. For the homogeneous conservative system, these schemes are second order approximations in space, of a parabolic equivalent system:

$$\frac{\partial W}{\partial t} + \frac{\partial}{\partial x}F(x,W) - \nu\frac{\partial}{\partial x}(\mathcal{D}(x,W)\frac{\partial}{\partial x}W) = 0,$$

where ν is equal to the half of the space step ($\nu = \Delta x/2$) and $\mathcal{D}(x,W)$ is the numerical diffusion matrix: The term

$$-\nu\frac{\partial}{\partial x}(\mathcal{D}(x,W)\frac{\partial W}{\partial x}), \tag{3}$$

represents the numerical diffusion introduced by the scheme, stemming from the upwinding of the convection terms. The diffusion matrix $\mathcal{D}(x,W)$ must take into account that the effective transport term in system (1) is $M(x,W)\frac{\partial W}{\partial x}$, where M is the matrix defined by (2).

If a centered discretization of the remaining terms in the non-homogeneous and non-conservative system (1) is performed, then the equivalent parabolic system is

$$\frac{\partial W}{\partial t}+\frac{\partial}{\partial x}F(x,W)-\nu\frac{\partial}{\partial x}(\mathcal{D}(x,W)\frac{\partial}{\partial x}W)=G(x,W)+B(x,W)\frac{\partial W}{\partial x}. \tag{4}$$

Notice that, if $\bar{W}$ is a stationary solution of (1), then it verifies

$$\frac{\partial}{\partial x}[F(x,\bar{W}(x))]=G(x,\bar{W}(x))+B(x,\bar{W}(x))\frac{\partial\bar{W}(x)}{\partial x}. \tag{5}$$

So, it is clear that $\bar{W}$ is not a stationary solution of (4). There is a first order error due to the numerical diffusion term (3) that is not compensated.

In order to obtain numerical schemes that calculate up to second order of accuracy all smooth stationary solutions, we propose to compensate the numerical diffusion term by an additional source term, that we denote by $C(x,W)$ (Correction term), that is, the numerical scheme must be a second order approximation of

$$\frac{\partial W}{\partial t}+\frac{\partial}{\partial x}\Big[F(x,W)\Big]-\nu\frac{\partial}{\partial x}\Big(\mathcal{D}(x,W)\frac{\partial W}{\partial x}\Big)=G(x,W)+B(x,W)\frac{\partial W}{\partial x}+C(x,W).$$

In order to define the term $C(x,W)$, we consider a smooth stationary solution $\bar{W}$ of (1). We want the identity

$$-\nu\frac{\partial}{\partial x}\Big(\mathcal{D}(x,\bar{W})\frac{\partial\bar{W}}{\partial x}\Big)=C(x,\bar{W}) \tag{6}$$

to hold. As $\bar{W}$ verifies (5), then

$$\frac{\partial F}{\partial x}(x,\bar{W}(x))+A(x,\bar{W}(x))\frac{\partial\bar{W}}{\partial x}(x)=\frac{\partial}{\partial x}[F(x,\bar{W}(x))]=G(x,\bar{W}(x))+B(x,\bar{W}(x))\frac{\partial\bar{W}}{\partial x}.$$

Thus,

$$M(x,\bar{W}(x))\frac{\partial\bar{W}}{\partial x}(x)=G(x,\bar{W}(x))-F_x(x,\bar{W}(x)), \tag{7}$$

where $F_x(x,W)=\dfrac{\partial F}{\partial x}(x,W)$. Moreover, if $M(x,\bar{W})$ is non-singular, we have

$$\frac{\partial\bar{W}(x)}{\partial x}=M^{-1}(x,\bar{W}(x))S(x,\bar{W}(x)),$$

where

$$S(x,W)=G(x,W)-F_x(x,W).$$

Therefore,

$$-\nu\frac{\partial}{\partial x}\Big(\mathcal{D}(x,\bar{W}(x))\frac{\partial\bar{W}}{\partial x}(x)\Big)=-\nu\frac{\partial}{\partial x}\Big(\mathcal{D}(x,\bar{W}(x))M^{-1}(x,\bar{W}(x))S(x,\bar{W}(x))\Big). \tag{8}$$

Comparing (8) with (6), we propose to define

$$C(x,W)=-\nu\frac{\partial}{\partial x}\Big(\mathcal{D}(x,W)M^{-1}(x,W)S(x,W)\Big);\quad\forall x\in[0,L],$$

$\forall W \in \mathbf{R}^N$ such that $M(x,W)$ is non-singular. The equivalent system for the hyperbolic system with source term and non-conservative terms is the following,

$$\frac{\partial W}{\partial t} + \frac{\partial}{\partial x} F(x,W) - \nu \frac{\partial}{\partial x}\left(\mathcal{D}(x,W)\frac{\partial W}{\partial x}\right) =$$

$$= G(x,W) + B(x,W)\frac{\partial W}{\partial x} - \nu \frac{\partial}{\partial x}\Big(\mathcal{D}(x,W)M^{-1}(x,W)S(x,W)\Big). \qquad (9)$$

Every stationary solution of (1) is also a stationary solution of this system at points where $M(x,W(x))$ is non-singular.

We conclude then that asymptotically well-balanced numerical schemes for the hyperbolic system (1) can be built by the use of approximations of second order in space of (9).

Notice that system (9) has a conservative structure when the original system (1) is conservative (i.e., $B=0$). Indeed, it can be re-written as

$$\frac{\partial W}{\partial t} + \frac{\partial}{\partial x}\left[F(x,W) - \nu\mathcal{D}(x,W)\left(\frac{\partial W}{\partial x} - M^{-1}S(x,W)\right)\right] = G(x,W) + B(x,W)\frac{\partial W}{\partial x}. \qquad (10)$$

Based upon this structure, we propose general three-points schemes as follows

$$\frac{W_i^{n+1} - W_i^n}{\Delta t} + \frac{\phi_R^S(x_i, x_{i+1}, W_i^n, W_{i+1}^n) - \phi_L^S(x_{i-1}, x_i, W_{i-1}^n, W_i^n)}{\Delta x} =$$

$$= G_C(x_{i-1}, x_i, x_{i+1}, W_{i-1}^n, W_i^n, W_{i+1}^n), \qquad (11)$$

where

$$\phi_R^S(x_i, x_{i+1}, W_i, W_{i+1}) = \phi^S(x_i, x_{i+1}, W_i, W_{i+1}) - \frac{1}{2}\mathcal{B}_R(x_i, x_{i+1}, W_i, W_{i+1})\,(W_{i+1} - W_i), \qquad (12)$$

$$\phi_L^S(x_i, x_{i+1}, W_i, W_{i+1}) = \phi^S(x_i, x_{i+1}, W_i, W_{i+1}) + \frac{1}{2}\mathcal{B}_L(x_i, x_{i+1}, W_i, W_{i+1})(W_{i+1} - W_i) \qquad (13)$$

and

$$\phi^S(x_i, x_{i+1}, W_i, W_{i+1}) = F_C(x_i, x_{i+1}, W_i, W_{i+1}) -$$

$$-\nu D(x_i, x_{i+1}, W_i, W_{i+1})\left(\frac{W_{i+1} - W_i}{\Delta x} - \widetilde{M^{-1}}(x_i, x_{i+1}, W_i, W_{i+1})S_D(x_i, x_{i+1}, W_i, W_{i+1})\right). \qquad (14)$$

Here, $\mathcal{B}_L$ and $\mathcal{B}_R$ are approximation of B at $x = x_{i+1/2}$, respectively. Moreover, G_C is an approximation of $G(x,W)$ in $x = x_i$. By F_C, D, $\widetilde{M^{-1}}$ and S_D we denote the approximations of F, $\mathcal{D}$, M^{-1} and S in $x = x_{i+1/2}$, respectively.

Let us suppose for instance that the singularities of M are concentrated on a continuous hypersurface $\gamma \subset \mathbf{R}^{n+1}$, where, say the j-th eigenvalue λ_j vanishes. In this case, we define $\widetilde{M^{-1}}(x,y,U,V)$ as

$$\widetilde{M^{-1}}(x,y,U,V) = X((x+y)/2, \widetilde{W})\widetilde{\Lambda^{-1}}(x,y,U,V)X^{-1}((x+y)/2, \widetilde{W}), \qquad (15)$$

where $\widetilde{W} = \widetilde{W}(x,y,U,V)$ is an intermediate state between U and V, by X we denote the matrix defined by the eigenvectors of $M((x+y)/2, \widetilde{W})$, and

$$\widetilde{\Lambda^{-1}}(x,y,U,V) = \mathrm{Diag}(\widetilde{\lambda_i^{-1}}(x,y,U,V), i = 1, \ldots, N)$$

with

$$\widetilde{\lambda_i^{-1}} = \lambda_i^{-1}, \quad \text{if } i \neq j,$$

and

$$\widetilde{\lambda_j^{-1}} = \begin{cases} \lambda_j^{-1} & \text{if } L[(x,U),(y,V)] \subset \mathbf{R}^{N+1} \setminus \gamma, \\ 0 & \text{otherwise;} \end{cases} \qquad (16)$$

where the λ_j are the eigenvalues of $M((x+y)/2, \widetilde{W})$ and $L[(x,U),(y,V)]$ is the closed segment in $\mathbf{R}^{N+1}$ that connects (x,U) and (y,V). It is straightforward to extend this definition to a more general case.

This definition of $\widetilde{M^{-1}}$ implies that, when an eigenvalue vanishes, no upwinding of the source term is performed at the direction defined by the corresponding eigenvector. For non-homogeneous scalar hyperbolic equations we may prove that this leads to stable schemes.

This definition of $\widetilde{M^{-1}}$ allows the schemes to be asymptotically well-balanced even for a class of stationary solutions for which matrix M is singular and, for convenient choices of the diffusion matrix D, for every stationary solutions. We shall prove this in the next section.

Remark 1 *In [7] it is proved that for homogeneous conservative systems like (10) both flux-difference and flux-splitting schemes fit into the general structure of fluxes given by (12) and (13) with convenient choices of the centered numerical flux F_C and the numerical diffusion matrix D. We point out that for non-conservative systems, it makes no sense to consider flux-splitting schemes, as these are intrinsically based upon the conservative formulation. Indeed, these schemes use a splitting of the flux into (we assume $B = 0$)*

$$F(W) = \mathcal{A}^*(W)\,W.$$

for some matrix $\mathcal{A}^*$. However, in general we may use flux-difference schemes, as these are based upon the use of the velocities and speeds of propagation of the total flux, given by the matrix M.

In addition, other kind of schemes fit into the general framework set by the general structure (11). We shall introduce some of them in Section 5.

3 Analysis of the “asymptotically well-balanced” property

The purpose of this section is to analyze the well-balanced property of the numerical schemes of the form (10)-(14). This analysis is performed by considering an arbitrary stationary solution and using Taylor expansions. As consequence, the results are valid for every stationary solution (and not only for particular cases as ’water at rest’ solutions for the shallow-water equations) but we shall only be able to analyze the well-balanced property in terms of orders of accuracy, and not exactly at grid nodes.

Due to the lack of regularity of M^{-1} (and also to the low regularity of D, usually only Lipschitz continuous) we cannot hope to balance steady solutions of system (1) up to second order of accuracy in the whole domain $(0,L)$. However, we shall prove that this property holds on a subset of $(0,L)$ except a subset whose measure tends to zero as Δx tends to zero for a large class of stationary solutions. We will say in this case that the numerical scheme is “asymptotically well-balanced”.

Based upon these observations, we introduce the following definitions:

Definition **1** *(Second order well-balanced schemes) We say that the numerical scheme (11)-(14) is well-balanced for a regular stationary solution $W(x)$ of the hyperbolic system (1) if the corresponding consistency error is second order in all nodes x_i inside the domain $(0,L)$.*

The following definition was introduced in [6]:

Definition **2** *(Asymptotically well-balanced schemes)*

We say that the scheme (11)-(14) is asymptotically well-balanced for a regular stationary solution $W(x)$ of the hyperbolic system (1) if there exists an increasing sequence of compacts $\{K_n\}_n$ of $[0,L]$ such that

1) *$\mu((0,L)\setminus \cup_n K_n) = 0$, where by μ we denote the Lebesgue measure in $\mathbf{R}$.*

2) *For all n there exists a $\delta_n > 0$ such that if $0 < \Delta x < \delta_n$, then the scheme balances system (1) up to second order in K_n.*

To perform our analysis, we will assume some hypothesis on the different elements defining the numerical scheme (11)-(14). Although these are somewhat cumbersome, this will allow to extend the analysis to a rather large class of schemes:

Hypothesis 1 *a) There exists a C^1 2-tensor $F_1 : [0,L]\times\mathbf{R}^N \to \mathcal{L}(\mathbf{R}^{n+1},\mathbf{R}^n)$ such that $\forall\, U,V \in \mathbf{R}^N$, $\forall\, x,y \in [0,L]$*

$$F_C(x,y,U,V) = F\left(\frac{x+y}{2},\frac{U+V}{2}\right) +$$
$$+F_1\left(\frac{x+y}{2},\frac{U+V}{2}\right)\left(\begin{array}{c} y-x \\ V-U \end{array}\right) + o(|x-y|^2+|U-V|^2).$$

b) There exists two C^1 3-tensors $B_{1,L}$ and $B_{1,R}$: $[0,L]\times\mathbf{R}^N \to \mathcal{L}(\mathbf{R}^{n+1},\mathcal{L}(\mathbf{R}^n,\mathbf{R}^n))$ such that $\forall\, U,V \in \mathbf{R}^N$, $\forall\, x,y \in [0,L]$

$$\mathcal{B}_{L/R}(x,y,U,V) = B\left(\frac{x+y}{2},\frac{U+V}{2}\right) +$$
$$+B_{1,L/R}\left(\frac{x+y}{2},\frac{U+V}{2}\right)\left(\begin{array}{c} y-x \\ V-U \end{array}\right) + o(|x-y|^2+|U-V|^2).$$

c) $\forall\, x,y,z,\in [0,L]$, $\forall\, U,V,W \in R^N$,

$$G_C(x,y,z,U,V,W) = G(y,V) + \mathcal{O}(|y-x|^2+|z-y|^2+|V-U|^2+|W-V|^2).$$

d) There exists a C^1 2-tensor $S_1 : [0,L]\times\mathbf{R}^N \to \mathcal{L}(\mathbf{R}^{n+1},\mathbf{R}^n)$ such that $\forall\, x,y \in [0,L]$, $\forall\, U,V,\in \mathbf{R}^N$,

$$S_D(x,y,U,V) = S\left(\frac{x+y}{2},\frac{U+V}{2}\right) +$$
$$+S_1\left(\frac{x+y}{2},\frac{U+V}{2}\right)\left(\begin{array}{c} y-x \\ V-U \end{array}\right) + \mathcal{O}(|x-y|^2+|V-U|^2).$$

e) There exists a continuous hypersurface $\gamma \subset \mathbf{R}^{N+1}$ such that M is non-singular and C^2 in $\mathbf{R}^{N+1}\setminus\gamma$. Moreover, there exists a 3-tensor $M_1 : \mathbf{R}^{N+1}\setminus\gamma \to \mathcal{L}(\mathbf{R}^{n+1},\mathcal{L}(\mathbf{R}^n,\mathbf{R}^n))$ which is C^1 in $\mathbf{R}^{N+1}\setminus\gamma$, that verifies

$$\widetilde{M^{-1}}(x,y,U,V) = M^{-1}\left(\frac{x+y}{2},\frac{U+V}{2}\right) +$$
$$+M_1\left(\frac{x+y}{2},\frac{U+V}{2}\right)\left(\begin{array}{c} y-x \\ V-U \end{array}\right) + \mathcal{O}(|x-y|^2+|V-U|^2).$$

for close enough points (x,U), (y,V) such that the closed segment $L[(x,U),(y,V)] \subset \mathbf{R}^{N+1}\setminus\gamma$.
f) There exists a continuous hypersurface $\widetilde{\gamma} \subset \mathbf{R}^{N+1}$ and a 3-tensor $D_1 : \mathbf{R}^{N+1}\setminus\widetilde{\gamma} \to \mathcal{L}(\mathbf{R}^{n+1},\mathcal{L}(\mathbf{R}^n,\mathbf{R}^n))$ which is C^1 in $\mathbf{R}^{N+1}\setminus\widetilde{\gamma}$ such that for all points (x,U), (y,V) with $L[(x,U),(y,V)] \subset \mathbf{R}^{N+1}\setminus\widetilde{\gamma}$,

$$D(x,y,U,V) = \mathcal{D}\left(\frac{x+y}{2},\frac{U+V}{2}\right) +$$
$$+D_1\left(\frac{x+y}{2},\frac{U+V}{2}\right)\left(\begin{array}{c} y-x \\ V-U \end{array}\right) + \mathcal{O}(|y-x|^2+|V-U|^2).$$

where, given two Banach spaces E, F, $\mathcal{L}(E,F)$ denotes the vector space of continuous linear maps.

Hypothesis *a*), *b*) and *c*) state that F_C, $(\mathcal{B}_L+\mathcal{B}_R)/2\Delta x$ and G_C respectively are second order approximations of F, $B\partial W/\partial x$ and G. Hypothesis *e*) formalizes the fact that M^{-1} presents singularities at points where one of its eigenvalues vanishes. Hypothesis *f*) takes into account the fact that usually matrix D will only present Lipschitz regularity when one of its eigenvalues vanishes. A typical choice is, for example, $D = |M|$.

To state our main result, we need another technical definition:

Definition **3** *Given a continuous function $W : [0, L] \to \mathbf{R}^N$ we call "sets of regular and singular points" of W with respect to the scheme (11)-(14) to be the subsets of $[0, L]$ defined respectively by*

$$\omega(W) = \{x \in (0, L) \text{ such that } W(x) \in (\mathbf{R}^N \setminus \gamma) \cap (\mathbf{R}^N \setminus \widetilde{\gamma})\}$$

$$\sigma(W) = (0, L) \setminus \omega(W)$$

Then we have

Theorem **1** *We consider $W : [0, L] \to \mathbf{R}^N$ a stationary solution of class C^2 of the system (1). Then,*
a) *If $M(x, W)$ does not have any singular point, the scheme (11)-(14) balances the system (1) for W up to the second order.*
b) *If the set of singular points of W, $\sigma(W)$, has zero measure, then the scheme (11)-(14) asymptotically balances the system (1) for W.*
c) *If $\mathcal{D}$ has the same eigenvectors as M and the eigenvalues of D vanish when the eigenvalues of M vanish, then the scheme (11)-(14) asymptotically balances the system (1) for W.*

The proof of this result is given in the Appendix, as it essentially uses Taylor expansions and finite-dimensional compactness arguments.

We also have the following result that yields sufficient conditions for the exact calculation of stationary solutions at grid points:

Theorem **2** *Let W be a stationary solution of (1). Assume that F_C, G_C, S_D, $\mathcal{B}_L$ and $\mathcal{B}_R$ evaluated on W verify*

$$\frac{F_C(x_i, x_{i+1}, W_i, W_{i+1}) - F_C(x_{i-1}, x_i, W_{i-1}, W_i)}{\Delta x} = G_C(x_{i-1}, x_i, x_{i+1}, W_{i-1}, W_i, W_{i+1}) +$$

$$+\frac{1}{2}\left(\mathcal{B}_L(x_{i-1}, x_i, W_{i-1}, W_i)\frac{W_i - W_{i-1}}{\Delta x} + \mathcal{B}_R(x_i, x_{i+1}, W_i, W_{i+1})\frac{W_{i+1} - W_i}{\Delta x}\right), \tag{17}$$

$$\frac{W_{i+1} - W_i}{\Delta x} = \widetilde{M^{-1}}(x_i, x_{i+1}, W_i, W_{i+1}) S_D(x_i, x_{i+1}, W_i, W_{i+1}), \tag{18}$$

then, the scheme defined by (11)-(14) balances exactly system (1) for the stationary solution at grid points. And this occurs independently of the choice of the upwind matrix $D(x, y, U, V)$.

To prove this Theorem it is enough to verify that under these hypothesis the numerical scheme is reduced to $W_i^{n+1} = W_i^n$.

These results mean that the general structure of our schemes allows to separately balance the numerical flux and the numerical diffusion terms, in a way largely independent of the discretization parameters appearing in each of them.

In fact, the role of the matrix D is mainly to give some stability to the scheme and not to contribute to the balancing of source terms. It remains, however, to construct matrices D that satisfy Hypothesis H1 f) of Theorem 1 in order to ensure the well-balance property for all stationary solutions. The construction of such matrices is discussed in Section 4

Remark 2 *The definition of $\widetilde{M^{-1}}$ given by (15) satisfies Hypothesis e) if the intermediate state $\widetilde{W}$ satisfies for (x, U) and (y, V) close enough*

$$\widetilde{W}(U, V) = \frac{U + V}{2} + \rho\left(\frac{U + V}{2}\right)(V - U) + \mathcal{O}(|V - U|^2), \tag{19}$$

where $\rho : \mathbf{R}^N \to \mathbf{R}$ is a C^1 function. Indeed, if $L[(x, U), (y, V)] \subset \mathbf{R}^{N+1} \setminus \gamma$, as $(\frac{x+y}{2}, \widetilde{W}(U, V)) \in \mathbf{R}^{N+1} \setminus \gamma$ and M^{-1} is C^2 in $\mathbf{R}^{N+1} \setminus \gamma$, we deduce

$$\widetilde{M^{-1}}(x, y, U, V) = M^{-1}\left(\frac{x + y}{2}, \widetilde{W}(U, V)\right) = M^{-1}\left(\frac{x + y}{2}, \frac{U + V}{2}\right) +$$

$$+\partial_W(M^{-1})\left(\frac{x+y}{2},\frac{U+V}{2}\right)\left(\widetilde{W}(U,V)-\frac{U+V}{2}\right)+\mathcal{O}\left(\left|\widetilde{W}(U,V)-\frac{U+V}{2}\right|^2\right)=$$

$$=M^{-1}\left(\frac{x+y}{2},\frac{U+V}{2}\right)+\partial_W(M^{-1})\left(\frac{x+y}{2},\frac{U+V}{2}\right)\rho\left(\frac{U+V}{2}\right)(V-U)+\mathcal{O}(|V-U|^2).$$

We thus obtain Hypothesis e).

Remark 3 *Hypothesis f) is also recovered if the diffusion matrix is defined similarly to $\widetilde{M^{-1}}$:*

$$D(x,y,U,V)=X((x+y)/2,\widetilde{W})\widetilde{\Lambda}_D(x,y,U,V)X^{-1}((x+y)/2,\widetilde{W}),$$

where X denotes the matrix composed by the eigenvectors of $\mathcal{D}$, and

$$\widetilde{\Lambda}_D(x,y,U,V)=Diag(\widetilde{d}_j(x,y,U,V),j=1,\ldots,N)$$

with

$$\widetilde{d}_j(x,y,U,V)=\begin{cases} d_j & \text{if } L[(x,U),(y,V)]\subset \mathbf{R}^N\setminus\gamma; \\ 0 & \text{if not;}\end{cases} \tag{20}$$

where d_j denote the eigenvalues of $\mathcal{D}((x+y)/2,\widetilde{W})$. Notice that this is an usual way to define the diffusion matrix.

Remark 4 *Property (19) is satisfied by intermediate states of usual solvers such as Van-Leer's or Roe's for Shallow Water equations.*

Remark 5 *The equalities (17) and (18) are discrete versions of the identities*

$$\frac{\partial}{\partial x}F(x,W(x))=G(x,W(x))+B(W(x))\frac{\partial W}{\partial x},$$

and

$$\frac{\partial W}{\partial x}(x)=M^{-1}(x,W(x))\cdot S(x,W(x)).$$

4 Some possible choices of matrix D.

In this section we discuss some possible choices of the diffusion matrix D with a double purpose: To ensure the asymptotically well-balanced property for all steady smooth solutions, and the stability of the numerical scheme at least for the linear homogeneous case.

To start with this last purpose, we consider the linear system,

$$\frac{\partial W}{\partial t}+A\frac{\partial W}{\partial x}=B\frac{\partial W}{\partial x}, \tag{21}$$

where A and B are constant matrices. We suppose that the system is hyperbolic, that is, the matrix $M=A-B$ has N real different eigenvalues $m_j\ \in\mathbf{R}$, $j=1,\cdots,m$.

We study under which conditions the family of schemes previously introduced is L^2 stable. To do this, we observe that these schemes can be rewritten in viscous form, so the L^2 stability condition is given from the following result (Cf. Godlewsky and Raviart [12])

Theorem **3** *Let us denote by $d_j\in\mathbf{R}$ $j=1,\ldots,N$ the eigenvalues of D. If D and M have the same eigenvectors and their eigenvalues verify*

$$\left(\frac{\Delta t}{\Delta x}m_j\right)^2\le\frac{\Delta t}{\Delta x}d_j\le 1,\quad j=1,\ldots,N, \tag{22}$$

then the scheme is L^2 stable.

We look for choices of D satisfying the hypothesis of the previous theorem (and therefore those of the statement c) of Theorem 1 and the hypothesis H1 f). Some possible choices of matrix D satisfying these conditions are the following:
If $M = X\Lambda X^{-1}$ with $\Lambda = \mathrm{diag}(m_j, j = 1, \ldots, N)$, then

$$D = X\Lambda_D X^{-1} \quad \text{with} \quad \Lambda_D = \mathrm{diag}(d_j, j = 1, \ldots, N)$$

and

- $d_j = |m_j|$,
- $d_j = \dfrac{\Delta t}{\Delta x} m_j^2$,
- $d_j = (1 - \varphi(r))|m_j| + \varphi(r)\dfrac{\Delta t}{\Delta x} m_j^2$, where $\varphi(r)$ is a flux-limiter (Cf. [12], [18]).

Moreover, (22) gives the corresponding CFL condition to set the time step in terms of the solution and the space step.

Remark 6 *We observe that for previous definitions, the singularities of matrix* $\widetilde{M^{-1}}$ *and* D *are located at same points, so the hipersurfaces* γ *and* $\widetilde{\gamma}$ *coincides. Moreover, as matrix* M *and* D *have the same eigenvectors we obtain*

$$\widetilde{M^{-1}}(x, y, U, V)D(x, y, U, V) = X((x + y)/2, \widetilde{W})\, \Lambda(x, y, U, V)\, X^{-1}((x + y)/2, \widetilde{W})$$

with $\Lambda = Diag(\widetilde{\lambda_j^{-1}}\, \widetilde{d}_j, j = 1, \ldots, N)$*, where by (16) and (20) we have*

$$\widetilde{\lambda_j^{-1}}\widetilde{d}_j(x, y, U, V) = \begin{cases} \lambda_j^{-1} d_j & \textit{if } L[(x, U), (y, V)] \subset \mathbf{R}^N \setminus \gamma; \\ 0 & \textit{if not;} \end{cases} \tag{23}$$

Finally, we also observe that in the last case we obtain a zero why $\widetilde{\lambda_j^{-1}} = 0$*, and* $\widetilde{d}_j = 0$*. In practice, in this case we define directly* $\widetilde{M^{-1}}D$ *by (23).*

5 Application to the two-layer shallow water system

In this section we apply the general discretization technique previously introduced to the numerical solution of the one dimensional two-layer shallow water system for channels with arbitrary cross-sections. We start by describing the model equations. Next, we propose a family of numerical schemes, by defining the different components of the general scheme (11)-(14), in such a way that hypothesis H1 a)-f) are satisfied. We also discuss some possible choices for the diffusion matrix D.

5.1 Equations

Let us consider the flow of a stratified fluid along an open channel. The channel is supposed to have a straight axis and to be symmetric with respect to a vertical plane passing through its axis. The channel shape of cross-sections is given by the function $\sigma(x, z)$ and the bottom is given by the function $z_b(x)$. The fluid is assumed to be composed of two shallow layers of immiscible fluids of constant density, ρ_1 and ρ_2. Moreover, we assume that the flow is one dimensional, i.e., at every layer the velocities $v_i(x, t)$, $i = 1, 2$ are uniform over the cross-section and the thickness only depends on the x coordinate and the time. In the following sections, index 1 makes reference to the upper layer and index 2 to the lower one (see Figure 1).

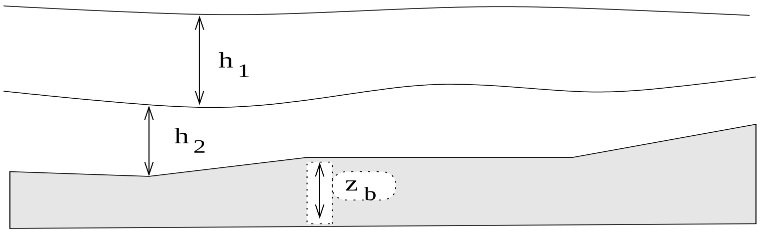


Figure 1: Sketch of a bilayer fluid

Let r be the ratio of densities $r = \rho_1/\rho_2$. Let $Q_i(x,t) = v_i(x,t)A_i(x,t)$, be the discharge of the i-th layer, being $A_i(x,t)$ the wetted cross-section of the i-th layer at position x and time t. The thickness of layer i is given by $h_i(x,t)$. Therefore, A_i, $\sigma(x,z)$, h_i and z_b are related through the equations:

$$A_1 = \int_{z_b+h_2}^{z_b+h_2+h_1} \sigma(x,z)dz, \quad A_2 = \int_{z_b}^{z_b+h_2} \sigma(x,z)dz.$$

In [3] a P.D.E. system modeling such a flow was deduced by imposing the mass and momentum conservations principles. Here, we consider the reformulation introduced in [4]:

$$\frac{\partial \mathrm{W}}{\partial t}(x,t) + \frac{\partial \mathrm{F}}{\partial x}(\varsigma(x,t), \mathrm{W}(x,t)) = G(x, \varsigma(x,t), \sigma^b(x), \mathrm{W}(x,t)) + \\ + B(\varsigma(x,t), \mathrm{W}(x,t))\frac{\partial \mathrm{W}}{\partial x} \quad \forall x \in (0,L), \;\; t \in (0,T), \tag{24}$$

where the unknowns are

$$\mathrm{W}(x,t) = \begin{bmatrix} W_1(x,t) \\ W_2(x,t) \end{bmatrix}, \quad W_j = \begin{bmatrix} A_j \\ Q_j \end{bmatrix}, \quad j = 1,2$$

and $\varsigma = \begin{bmatrix} \sigma_1 \\ \sigma_2 \end{bmatrix}$ where σ_1 and σ_2 are defined by

$$\sigma_1(x,t) = \sigma(x, z_b(x) + h_1(x,t) + h_2(x,t)),$$

$$\frac{1}{\sigma_2(x,t)} = \frac{1-r}{\sigma_3(x,t)} + \frac{r}{\sigma_1(x,t)},$$

being

$$\sigma_3(x,t) = \sigma(x, z_b(x) + h_2(x,t)).$$

We also define,

$$\sigma^b(x) = \sigma(x, z_b(x)).$$

The flux function, the coupling (non-conservative) terms and the source terms are given respectively by:

$$\mathrm{F}(\varsigma, \mathrm{W}) = \begin{bmatrix} F_1(\sigma_1, W_1) \\ F_2(\sigma_2, W_2) \end{bmatrix}, \quad F_j(\sigma_j, W_j) = \begin{bmatrix} Q_j \\ \dfrac{Q_j^2}{A_j} + \dfrac{g}{2\sigma_j}A_j^2 \end{bmatrix}, \quad j = 1,2,$$

$$B(\varsigma, \mathrm{W}) = \begin{bmatrix} 0 & B_1(\varsigma, W_1) \\ B_2(\varsigma, W_2) & 0 \end{bmatrix},$$

with

$$B_1(\varsigma, W_1) = \begin{bmatrix} 0 & 0 \\ -g\dfrac{1}{\sigma_1}A_1 & 0 \end{bmatrix}, \quad B_2(\varsigma, W_2) = \begin{bmatrix} 0 & 0 \\ -g\dfrac{r}{\sigma_1}A_2 & 0 \end{bmatrix}$$

and

$$G(x, \varsigma, \sigma^b, \mathrm{W}) = \mathrm{V}(\varsigma, \mathrm{W}) + \mathbf{S}(x, \varsigma, \sigma^b, \mathrm{W}),$$

being

$$\mathrm{V}(\varsigma, \mathrm{W}) = \begin{bmatrix} V_1(\sigma_1, W_1) \\ V_2(\sigma_2, W_2) \end{bmatrix}, \quad V_j(\sigma_j, W_j) = \begin{bmatrix} 0 \\ \dfrac{g}{2}\left(\dfrac{1}{\sigma_j}\right)_x A_j^2 \end{bmatrix}, \quad j = 1, 2.$$

We denote by $\mathbf{S}$ the source terms due to the variations of the geometry (depth and cross-section variations). If we define $\omega = z_b + h_1 + h_2$ and $\omega_r = z_b + h_2 + r\, h_1$, then $\mathbf{S}$ can be written

$$\mathbf{S} = \begin{pmatrix} 0 \\ g\dfrac{A_1}{\sigma_1}\dfrac{\partial}{\partial x}(A_1 + A_2 - \omega) \\ 0 \\ gA_2\dfrac{1}{\sigma_2}\dfrac{\partial A_2}{\partial x} + gA_2\dfrac{r}{\sigma_1}\dfrac{\partial A_1}{\partial x} - gA_2\dfrac{\partial \omega_r}{\partial x} \end{pmatrix}. \tag{25}$$

In order to easily describe the discretization of the terms V_j, these are rewritten as follows:

$$V_j(\sigma_j, W_j) = \begin{bmatrix} 0 \\ \dfrac{g}{2}\left(\dfrac{A_j^2}{\sigma_j}\right)_x - g\dfrac{A_j}{\sigma_j}(A_j)_x \end{bmatrix}, \quad j = 1, 2.$$

These terms appear due to the dependence of the flux F on the spatial variable x. In fact, $V = \partial_\varsigma F\ \partial_x \varsigma$. Therefore

$$G(x, \varsigma, \sigma^b, \mathrm{W}) - \mathrm{F}_\varsigma(\varsigma, \mathrm{W})\, \partial_x\varsigma(x) = \mathbf{S}(x, \varsigma, \sigma^b, \mathrm{W}).$$

It is easy to verify that this system can be written in matrix form as follows:

$$\frac{\partial W}{\partial t} + \mathcal{A}(\varsigma, \mathrm{W})\frac{\partial W}{\partial x} = \mathbf{S}(x, \varsigma, \sigma^b, \mathrm{W}), \tag{26}$$

where $\mathcal{A}$ is the 4×4 matrix whose expression is given by:

$$\mathcal{A}(\sigma, \mathrm{W}) = J(\varsigma, \mathrm{W}) - B(\varsigma, \mathrm{W}),$$

with:

$$J(\varsigma, \mathrm{W}) = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -\dfrac{Q_1^2}{A_1^2} + \dfrac{g}{\sigma_1}A_1 & 2\dfrac{Q_1}{A_1} & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -\dfrac{Q_2^2}{A_2^2} + \dfrac{g}{\sigma_2}A_2 & 2\dfrac{Q_2}{A_2} \end{bmatrix}.$$

The eigenvalues of $\mathcal{A}$ can be classified in two external and two internal eigenvalues. The external eigenvalues, $\lambda^{e\pm}$, are related to the propagation speed of barotropic perturbations and the internal ones $\lambda^{i\pm}$, to the propagation of baroclinic perturbations. In the case $r \cong 1$, first-order approximations of the eigenvalues were given in [20]. In most of the applications to geophysical flows, one has $\lambda^{e-} < 0$ and $\lambda^{e+} > 0$. Therefore, the transitions depend on the sign of the internal eigenvalues: the flow is said to be *subcritical* if the sign of the internal eigenvalues differs, *critical* if one of them takes the value zero, otherwise the flow is called *supercritical*. It can be deduced from the expression of the matrix $\mathcal{A}$ that critical, supercritical and subcritical sections can be characterized by $G^2 = 1$, $G^2 > 1$, $G^2 < 1$, respectively, where

$$G^2 = F_1^2 + F_2^2 - (1-r)\frac{\sigma_2}{\sigma_3}F_1^2F_2^2, \tag{27}$$

and

$$F_1^2 = \frac{v_1^2}{g'\left(\frac{\sigma_2}{\sigma_3}\right)\left(\frac{A_1}{\sigma_1}\right)}, \quad F_2^2 = \frac{v_2^2}{g'\left(\frac{A_2}{\sigma_3}\right)}. \tag{28}$$

In these definitions, $g' = g(1-r)$ is the reduced gravity, and G and F_i, $i = 1, 2$ are the appropriate definitions for this case of the *composite Froude number* and the *internal Froude numbers*, respectively.

The internal eigenvalues may become complex corresponding to the development of shear instabilities. Nevertheless, in the description of the numerical schemes only the case in which the matrix $\mathcal{A}$ has 4 different real eigenvalues is considered, i.e., the flow is supposed to be stable and the system strictly hyperbolic.

5.2 Discretization of the system.

In this subsection we introduce a one-parameter numerical scheme to solve system (24), with the structure (11)-(14). In order to obtain asymptotically well-balanced schemes, we must define the different components of the numerical scheme to fit the general Hypothesis H1 a)-f).

We will use the following notation:

$$\mathrm{W}_{i+\alpha} = (1-\alpha)\mathrm{W}_i + \alpha\mathrm{W}_{i+1}, \quad \alpha \in [0,1].$$

We also consider an analogous notation for ς:

$$\varsigma_{i+\alpha} = \left[\begin{array}{c} \sigma_{1,i+\alpha} \\ \sigma_{2,i+\alpha} \end{array}\right],$$

with

$$\sigma_{1,i+\alpha} = (1-\alpha)\sigma_{1,i} + \alpha\sigma_{1,i+1}, \quad \sigma_{3,i+\alpha} = (1-\alpha)\sigma_{3,i} + \alpha\sigma_{3,i+1}$$

and $\sigma_{2,i+\alpha}$ defined by

$$\frac{1}{\sigma_{2,i+\alpha}} = \frac{1-r}{\sigma_{3,i+\alpha}} + \frac{r}{\sigma_{1,i+\alpha}}.$$

With this notation, following [7] we define the centered approximation of the flux function by

$$F_C(x_i, x_{i+1}, W_i, W_{i+1}) = \frac{F(\varsigma_{i+\alpha}, \mathrm{W}_{i+\alpha}) + F(\varsigma_{i+1-\alpha}, \mathrm{W}_{i+1-\alpha})}{2}, \tag{29}$$

with $\alpha \in [0,1]$. This approximation satisfies hypothesis H1 *a*).

We propose to use this approximation of F_C due to our former experience with a similar one-parameter family of numerical schemes for solving the one-layer Shallow-Water system. In this case, a good choice of the parameter α allows to correctly approximate solutions including sonic rarefaction waves, without the use of entropy corrections. In particular, in [7] the value $\alpha = 1/8$ is proposed.

Matrix $\widetilde{M^{-1}}$ is defined by (15). In practice we set $\widetilde{\lambda_j^{-1}} = 0$ if the absolute value of λ_j is smaller than a certain ϵ nearly to zero. The de-centered source terms $\mathcal{B}_L$ and $\mathcal{B}_R$ are given by

$$\mathcal{B}_L(x_{i-1}, x_i, W_{i-1}, W_i) = \alpha\mathbf{B}\left(\varsigma_{i-\alpha/2}, \mathrm{W}_{i-\alpha/2}\right) + (1-\alpha)\mathbf{B}\left(\varsigma_{i-(1-\alpha)/2}, \mathrm{W}_{i-(1-\alpha)/2}\right) \tag{30}$$

and

$$\mathcal{B}_R(x_i, x_{i+1}, W_i, W_{i+1}) = \alpha\mathbf{B}\left(\varsigma_{i+\alpha/2}, \mathrm{W}_{i+\alpha/2}\right) + (1-\alpha)\mathbf{B}\left(\varsigma_{i+(1-\alpha)/2}, \mathrm{W}_{i+(1-\alpha)/2}\right). \tag{31}$$

To complete the definition of ϕ_L^S and ϕ_R^S we must define a discretization of S in $x = x_{i+1/2}$, $S_D(x_i, x_{i+1}, W_i, W_{i+1})$, satisfying H1 c). We propose the following one:

$$\mathcal{S}_D(x_i, x_{i+1}, W_i, W_{i+1}) = \frac{1}{\Delta x}\begin{pmatrix} 0 \\ \\ g\dfrac{A_{1,i+1/2}}{\sigma_{1,i+1/2}}(A_{1,i+1} + A_{2,i+1} - \omega_{i+1} - A_{1,i} - A_{2,i} + \omega_i) \\ \\ 0 \\ \\ g\dfrac{A_{2,i+1/2}}{\sigma_{2,i+1/2}}(A_{2,i+1} - A_{2,i}) + gr\dfrac{A_{2,i+1/2}}{\sigma_{1,i+1/2}}(A_{1,i+1} - A_{1,i}) - \\ -gA_{2,i+1/2}(\omega_{r,i+1} - \omega_{r,i}) \end{pmatrix}. \tag{32}$$

We define a centered discretization of G in $x = x_i$, G_C:

$$G_C = \mathbf{S}_i + V_i,$$

where $\mathbf{S}_i$ and V_i are centered approximations of $\mathbf{S}$ and V at $x = x_i$, respectively. They are constructed as follows:

$$\mathbf{S}_i = \frac{1}{2}\Big(\mathbf{S}_{i,L} + \mathbf{S}_{i,R}\Big), \tag{33}$$

where, if we denote by $[\cdot]_j$ the j-th component of the vector$[\cdot]$, then

$$[\mathcal{S}_{i,L}]_1 = 0, \tag{34}$$

$$\Delta x[\mathcal{S}_{i,L}]_2 = g\frac{A_{1,i-\alpha/2}}{\sigma_{1,i-\alpha/2}}(A_{1,i} + A_{2,i} - \omega_i - A_{1,i-\alpha} - A_{2,i-\alpha} + \omega_{i-\alpha}) +$$

$$+ g\frac{A_{1,i-(1-\alpha)/2}}{\sigma_{1,i-(1-\alpha)/2}}(A_{1,i} + A_{2,i} - \omega_i - A_{1,i-1+\alpha} - A_{2,i-1+\alpha} + \omega_{i-1+\alpha}), \tag{35}$$

$$[\mathcal{S}_{i,L}]_3 = 0 \tag{36}$$

and

$$\Delta x[\mathcal{S}_{i,L}]_4 = g\frac{A_{2,i-\alpha/2}}{\sigma_{2,i-\alpha 1/2}}(A_{2,i} - A_{2,i-\alpha}) + g\,r\frac{A_{2,i-\alpha/2}}{\sigma_{1,i-\alpha/2}}(A_{1,i} - A_{1,i-\alpha}) -$$

$$-gA_{2,i-\alpha/2}(\omega_{r,i} - \omega_{r,i-\alpha}) + g\frac{A_{2,i-(1-\alpha)/2}}{\sigma_{2,i-(1-\alpha)/2}}(A_{2,i} - A_{2,i-1+\alpha}) +$$

$$+g\,r\frac{A_{2,i-(1-\alpha)/2}}{\sigma_{1,i-(1-\alpha)/2}}(A_{1,i} - A_{1,i-1+\alpha}) - gA_{2,i-(1-\alpha)/2}(\omega_{r,i} - \omega_{r,i-1+\alpha}). \tag{37}$$

$\mathcal{S}_{i,R}$ is given by:

$$[\mathcal{S}_{i,R}]_1 = 0, \tag{38}$$

$$\Delta x[\mathcal{S}_{i,R}]_2 = g\frac{A_{1,i+\alpha/2}}{\sigma_{1,i+\alpha/2}}(A_{1,i+\alpha} + A_{2,i+\alpha} - \omega_{i+\alpha} - A_{1,i} - A_{2,i} + \omega_i) +$$

$$+g\frac{A_{1,i-(1-\alpha)/2}}{\sigma_{1,i-(1-\alpha)/2}}(A_{1,i+1-\alpha} + A_{2,i+1-\alpha} - \omega_{i+1-\alpha} - A_{1,i} - A_{2,i} + \omega_i), \tag{39}$$

$$[\mathcal{S}_{i,R}]_3 = 0, \tag{40}$$

and

$$\Delta x[\mathcal{S}_{i,R}]_4 = g\frac{A_{2,i+\alpha/2}}{\sigma_{2,i+\alpha/2}}(A_{2,i+\alpha} - A_{2,i}) + g\,r\frac{A_{2,i-\alpha/2}}{\sigma_{1,i-\alpha/2}}(A_{1,i+\alpha} - A_{1,i}) -$$

$$-gA_{2,i-\alpha/2}(\omega_{r,i+\alpha}-\omega_i)+g\frac{A_{2,i+(1-\alpha)/2}}{\sigma_{2,i+(1-\alpha)/2}}(A_{2,i+1-\alpha}-A_{2,i})+$$

$$g\,r\frac{A_{2,i+(1-\alpha)/2}}{\sigma_{1,i+(1-\alpha)/2}}(A_{1,i+1-\alpha}-A_{1,i})-gA_{2,i+(1-\alpha)/2}(\omega_{r,i+1-\alpha}-\omega_{r,i}). \tag{41}$$

The centered discretization of the term V at point $x=x_i$, V_i, is given by:

$$V_i=\begin{bmatrix}\mathcal{V}_1\\ \mathcal{V}_2\end{bmatrix} \tag{42}$$

being

$$[\mathcal{V}_j]_1=0, \quad \text{for} \quad j=1,2 \tag{43}$$

and

$$2\Delta x[\mathcal{V}_j]_2=g\frac{A_{j,i+\alpha}A_{j,i+\alpha}}{2\sigma_{j,i+\alpha}}-g\frac{A_{j,i-1+\alpha}A_{j,i-1+\alpha}}{2\sigma_{j,i-1+\alpha}}+$$

$$g\frac{A_{j,i+1-\alpha}A_{j,i+1-\alpha}}{2\sigma_{j,i+1-\alpha}}-g\frac{A_{j,i-\alpha}A_{j,i-\alpha}}{2\sigma_{j,i-\alpha}}-$$

$$-g\frac{A_{j,i+\alpha/2}}{\sigma_{j,i+\alpha/2}}(A_{j,i+\alpha}-A_{j,i})-g\frac{A_{j,i+(1-\alpha)/2}}{\sigma_{j,i+(1-\alpha)/2}}(A_{j,i+1-\alpha}-A_i)-$$

$$-g\frac{A_{j,i-\alpha/2}}{\sigma_{j,i-\alpha/2}}(A_i-A_{j,i-\alpha})-g\frac{A_{j,i-(1-\alpha)/2}}{\sigma_{j,i-(1-\alpha)/2}}(A_{j,i}-A_{j,i-1+\alpha}). \tag{44}$$

for $j=1,2$.

It remains to define a matrix D satisfying hypothesis H1 f). This may be done as indicated in Section 4. In all the numerical tests discussed in Section 6, we have chosen $D=|M|$ where M is the Roe-type matrix introduced in [4]. With this choice, the numerical scheme proposed in this latter article is the particular case of the previous one corresponding to $\alpha=0$.

5.3 Balance properties

As the different components defining the numerical scheme verify the hypothesis H1 a)-f), we can apply Theorem 1 and conclude the following result.

Theorem **4**

1) *For any of the possible choices of D presented in subsection 4, the numerical scheme (11)-(14), (29)-(44) asymptotically well-balances system (24) for all regular stationary solutions.*

2) *Independently of the choice of matrix D, the numerical scheme balances all regular stationary solutions $W(x)$ such that $M(x,W(x))$ is not singular $\forall x\in[0,L]$. If the set of points such that $M(x,W(x))$ is singular has zero measure, then the scheme asymptotically well-balances system (24) for $W(x)$.*

We can also apply Theorem 2 to the stationary solutions corresponding to water at rest:

$$\overline{W}=\begin{bmatrix}A_1\\0\\A_2\\0\end{bmatrix}, \quad \text{with} \quad A_1=\int_{\overline{h}_2}^{\overline{h}_2+\overline{h}_1}\sigma(x,z)dz, \quad A_2=\int_{z_b}^{\overline{h}_2}\sigma(x,z)dz, \tag{45}$$

with $\overline{h}_1$ constant and $\overline{h}_2=h_2+z_b(x)$ with $\overline{h}_2$ constant. In this solution the free surface of the fluid, given by $\overline{h}_1+\overline{h}_2+z_b$, is constant, as well as the interface given by $\overline{h}_2+z_b$.

We have the following result.

Theorem 5 *The numerical scheme defined by (29)-(44) exactly balances system (24) on the stationary solution (45) for any value of the parameter α and for any matrix D.*

To prove this theorem it is enough to verify conditions (17) and (18) on the stationary solution (45).

Remark 7 *As we previously mentioned, solving the one-layer Shallow-Water system with centered numerical fluxes F_C as (29) with a good choice of the parameter α ($\alpha = 1/8$) allows to correctly approximate sonic rarefaction waves, without the need of entropy-correction techniques. Also, in the case of stationary shocks the best choice is given by $\alpha = 0$. This suggests that a better method can be obtained if the parameter α is replaced by a function $\alpha(x)$ taking the value 0 at inter-cells where a stationary shock develops and 1/8 elsewhere.*

Taking in mind that transitions depend on the sign of the internal eigenvalues, a suitable approximation $\alpha_{j+1/2}$ to $\alpha(x_{j+1/2})$ can be given by

$$\begin{aligned} \alpha_{j+1/2} &= min(\alpha^-_{j+1/2}, \alpha^+_{j+1/2}); \quad with \\ \alpha^\pm_{j+1/2} &= \begin{cases} \bar{\alpha} & if\ sgn(\lambda^{i\pm}_j) = sgn(\lambda^{i\pm}_{j+1}) \\ \bar{\alpha}(1 - 0.5(1 + sgn(\lambda^{i\pm}_j - \lambda^{i\pm}_{j+1}))) & otherwise, \end{cases} \end{aligned} \tag{46}$$

where $\bar{\alpha}$ is a constant value ($\bar{\alpha} = 1/8$) and $\lambda^{i\pm}_j$ are the internal eigenvalues of the matrix $M(x_j, W_j)$.

In most applications, this numerically-defined function $\alpha(x)$ will be constant in all the domain but a set whose measure tends to zero (a bounded number of cells), so that the balance properties are preserved.

6 Numerical tests

In this section, some experiments illustrating the properties of the proposed family of numerical schemes are presented. The purposes of these experiments are mainly to validate the numerical schemes and to investigate the influence of the parameter α when the numerical viscosity vanishes. For this reason in all our tests we have chosen a unique numerical diffusion matrix D, given by the simplest choice in Section 4, i. e., $D = |M|$.

The first experiment has been designed to verify whether the still water solution is exactly computed in practice, even for very irregular geometries. The second one provides a numerical study of the order of the numerical scheme. This study confirms the theoretical results. In section 6.3 the influence of the parameter α in smooth transitions is investigated: we observe that any positive value yields good results. Finally, the last test has been designed to analyze the influence of the parameter α for a non smooth solution. In this case we see that the best results are obtained if the constant parameter α is replaced by the function α defined in Remark 6.

6.1 Water at rest in a channel with irregular geometry.

In this example, a numerical test is presented to verify how accurately the solution (45) is, in practice, computed for a channel with very irregular geometry. To do this, we have considered a 4m. long channel with rectangular cross-section and for which depth and breadth were randomly generated. The numerical scheme was applied for the values of the parameter $\alpha = 0$, 1/8, 1/4 and 3/8 (only the values for $\alpha = 1/8$ are shown, because all results are very similar), taking as initial condition a steady solution representing water at rest. The CFL parameter is set to 0.95 and $\Delta x = 0.05$.

Figures 2(a) and 2(b) show the bottom topography and the channel breadth, respectively. Figures 2(c) and 2(d) depict the free surface and the water interface at the initial time and after 15 seconds of simulation. Finally, figures 2(e) and 2(f) show the value of the discharge at each layer at the initial state and after 15 seconds of simulation, respectively. As predicted in Theorem 5, for any value of the parameter, the water elevation at the free surface and the interface as well as the discharges at each layer are accurately computed: only round-off errors are present and they do not grow with time.

Table 1: Test case 6.2 solved with $\alpha = 1/8$. h_1 and h_2.

N. cells	L^1 error h_1	L^1 order h_1	L^1 error h_2	L^1 order h_2
25	5.20×10^{-3}	–	5.50×10^{-3}	–
50	1.66×10^{-3}	1.65	1.87×10^{-3}	1.56
100	4.20×10^{-4}	1.98	4.83×10^{-4}	1.95
200	1.04×10^{-4}	2.02	1.15×10^{-4}	2.07
400	2.38×10^{-5}	2.12	2.72×10^{-5}	2.08
800	5.67×10^{-6}	2.07	6.48×10^{-6}	2.07

6.2 Well-balancing test

In this section, we test the order of the numerical scheme when it is applied to approach a stationary solution of the two-layer shallow water system with constant breadth function. The regular stationary solutions of this system satisfy:

$$\begin{cases} Q_1 = \text{constant}, \\ \dfrac{v_1^2}{2} - \dfrac{v_2^2}{2} + g' h_1 = \text{constant}, \\ Q_2 = \text{constant}, \\ \dfrac{v_1^2}{2} + g(h_1 + h_2 + z_b) = \text{constant}, \end{cases} \tag{47}$$

where $g' = (1 - r)g$ is the reduced gravity.

In this numerical test, we consider $\sigma(x, z) = 1.0$, $r = 0.98$ and the depth function

$$z_b(x) = 0.5e^{-x^2} - 2.0, \quad x \in [-3, 3].$$

In order to obtain a particular stationary solution, we impose first its values at $x = -3$: $h_1(-3, 0) = 0.5$, $h_2(-3, 0) = z_b(-3) + 1.5$, $Q_1(-3, 0) = 0.15$ and $Q_2(-3, 0) = -0.15$. From these values, we compute the constants in (47). Once the constants are known, the values of the solution at any point are obtained by solving (47).

In the numerical experiment, this stationary solution is taken as initial condition. As the stationary solution chosen does not correspond to water at rest, the numerical scheme should be second order accurate.

We have applied the numerical scheme for different values of the parameter $\alpha = 0$, $1/8$, $1/4$ and $3/8$ (only the values for $\alpha = 1/8$ are shown, because all results are very similar). The CFL coefficient has been taken as 0.9. The results obtained for $\alpha = 1/8$ are shown in Tables 1 and 2. As expected, the order obtained is 2.

6.3 Rectangular channel with a bump

The validation of numerical schemes for solving the two-layer shallow water system with transcritical solutions is not a simple task, as exact solutions with smooth transitions cannot be easily obtained for this system. Nevertheless, Farmer and Armi in [9] introduced, in the context of the study of exchange flows through channels, an asymptotic technique to obtain steady solutions for channels with simplified geometries. In this particular case, they consider a channel of uniform width with a sill on the floor, separating two broad and deep reservoirs.

Let us first briefly recall the main aspects of the theory developed by these authors. In their works, smooth steady state solutions are studied. The equations of the model are obtained by replacing the

Table 2: Test case 6.2 solved with $\alpha = 1/8$. Q_1 and Q_2.

N. cells	L^1 error q_1	L^1 order q_1	L^1 error q_2	L^1 order q_2
25	8.42×10^{-4}	–	6.31×10^{-4}	–
50	2.30×10^{-4}	1.87	1.84×10^{-4}	1.78
100	5.96×10^{-5}	1.95	4.89×10^{-5}	1.91
200	1.38×10^{-5}	2.11	1.23×10^{-5}	1.99
400	3.11×10^{-6}	2.14	2.97×10^{-6}	2.05
800	6.91×10^{-7}	2.17	6.79×10^{-7}	2.13

fourth equation in (47) by the *rigid lid* assumption:

$$h_1 + h_2 + z_b = \text{constant}. \tag{48}$$

The equations are first written in non-dimensional form using some dimensionless variables Q_i', h_i', z_b' that are such that $h_1' + h_2' + z_b' = 1$ at the section of minimal depth. Then, all the variables are expressed in terms of the internal Froude numbers, $F_i = v_i/\sqrt{(g'h_i)}$, $i = 1, 2$. Using the approximation $r = 1$ the following 2×2 nonilinear system is obtained from the second equation of (47) and (48):

$$F_1^{-\frac{2}{3}} + \frac{1}{2}F_1^{\frac{4}{3}} - \frac{1}{2}q_r^{-\frac{2}{3}}F_2^{\frac{4}{3}} = \Delta H'(Q_1')^{-\frac{2}{3}}, \tag{49}$$

$$F_1^{-\frac{2}{3}} + q_r^{-\frac{2}{3}}F_2^{-\frac{2}{3}} = \left[\frac{Q_1'}{(1-z_b')^{\frac{3}{2}}}\right]^{-\frac{2}{3}}, \tag{50}$$

where $\Delta H'$ is a constant corresponding to the dimensionless energy difference between the two layers and $q_r = Q_1/Q_2$. Here, only the case $|q_r| = 1$ will be considered.

The solutions of the equations can be identified with the family of curves in the (F_1^2, F_2^2)-plane defined by (49). These curves haven been plotted in detail in [9].

If the reservoirs are sufficiently deep, except in the neighborhood of the sill, the speed of the thicker layer is negligible away from the influence of the sill. In this case, the Froude numbers F_1 and F_2 have to be near of zero far from the sill at the reservoir corresponding to water of density ρ_1 and ρ_2, respectively. A close inspection of the family of curves (49) shows that there is only a family of curves for which these two requirements are satisfied. Among the curves of this family, there is one for which the value of the exchanged flow Q_1' is maximal. This curve corresponds to $\Delta H' = 1.5$ and is called the *maximal two-way exchange*: it represents the steady-state limit that occurs when the fluids are free to move in opposite directions across the sill. If the reservoir corresponding to water of density ρ_1 is placed to the left, this solution is critical at the section of minimal depth, subcritical to the right of this section and supercritical to the left. If the condition $F_2 = 0$ is imposed at the exit section of the channel (that is, if the depth of the reservoir to the left is assumed to be infinity) the flow is critical again at this section.

In the experiment, we consider a rectangular channel of 1 meter width and 6 meters long. Its bottom topography is defined by the function $z_b(x) = \exp(-x^2) - 2$, where $x \in [-3, 3]$. The ratio between fluxes is set to 1 and the ratio between densities is set to $r = \rho_1/\rho_2 = 0.98$. In order to reproduce the maximal two-way exchange with the numerical model, we solve the classical *lock exchange problem*. This problem consists in taking as initial state the two fluids separated by a vertical "artificial barrier" that, in this case, is located at $x = 2$ (see figure 4(a)). This barrier is dropped out at time $t = 0$ and the fluids are let to evolve until a stationary state is reached (see figure 4(b)). The CFL parameter is set to 0.9 and $\Delta x = 1/15$.

The boundary conditions in this test have been designed to simulate the presence of the infinite reservoirs at both ends of the channel and the fact that $q_r = 1$: the two layers must be allowed to go out of the channel as freely as possible, with the only constraint $Q_1 = -Q_2$. To do this, we have modified the usual numerical treatment based on the introduction of two ghost cells and the duplication of states, which is used very often in practice to simulate open boundaries. The idea is as follows: let us suppose that, at the n-th time iteration, the approximation obtained at the first cell is W_1^n. If the corresponding value of Q_2, $Q_{2,1}^n$, is equal to zero, the lower layer has not yet reached the left boundary. In this case, the state is duplicated at the ghost cell:

$$\mathrm{W}_0^n = \mathrm{W}_1^n,$$

and the first cell is treated as an internal one during the calculation of the approximations at time t^{n+1}. If, instead, $Q_{2,1}^n$ is not equal to zero, then the state at the ghost cell is modified so that the relation $Q_1 = -Q_2$ holds. More precisely, we define:

$$\mathrm{W}_0^n = \mathrm{W}_1^n - \frac{\Delta Q}{|Q_{1,1}^n| + |Q_{2,1}^n|} \begin{bmatrix} 0 \\ |Q_{1,1}^n| \\ 0 \\ |Q_{2,1}^n| \end{bmatrix}$$

being

$$\Delta Q = Q_{1,1}^n + Q_{2,1}^n.$$

Again, the first cell is treated as an internal one. A similar treatment is performed at the right boundary.

During the computations, complex eigenvalues appear. Following [10], when the appearance of complex eigenvalues is detected, a linear friction term is added to the momentum equations of both layers:

$$-CQ_k, \quad k = 1, 2,$$

where the local friction coefficient C is great enough to make the numerical scheme linearly stable. These coefficients are function of the imaginary part of the eigenvalues. This extra amount of friction is intended to simulate the loss of momentum induced by the instabilities in real flows.

Figures 4(b) and 4(c) show the interface at the steady state reached for the different values of the parameter $\alpha = 0$, $1/8$, $1/4$, and $3/8$. These steady states are compared with the interface corresponding to the maximal two-way exchange obtained in [9] (in the sequel, this solution will be called the F&A solution). To do this, once the value Q_1' has been calculated, the 2×2 nonlinear system (49)-(50) is solved numerically for every value of z_b' corresponding to the intercells of the mesh (the graphical study of the solutions performed in [9] gives the necessary hints about the adequate choice of the initial guess for the numerical algorithms).

Notice that all the numerical solutions differ slightly from the F&A solution at the right part of the channel (see 4(b)). These differences are due mainly to the rigid lid hypothesis. The main differences among the numerical solutions are located near the critical section, as expected. The parameter $\bar{\alpha}$ has practically no influence in regular zones, but it is important in transitions (regular or not). So, we focus only in a neighborhood of $x = 0$ ($x \in [-0.4, 0.4]$). Figure 4(c) shows the computed interfaces and the F&A interface in this zone. Notice that, if the parameter α is set to 0, the numerical scheme corresponds to a Roe-type scheme and it produces a non-entropy solution, as expected. Table 6.3 shows the L^∞ and L^1 errors corresponding to the comparison of the numerical interfaces with the analytical one. Observe that the values $\alpha = 1/8$, $\alpha = 1/4$, $\alpha = 3/8$ produce similar results.

The discharges computed by the model are $Q_1 = -Q_2 = 9.38\,10^{-2}$ m^3/s for $\alpha = 1/8$, $1/4$ and $3/8$, and $Q_1 = -Q_2 = 9.33\,10^{-2}$ m^3/s for $\alpha = 0$, while the F&A model gives $Q_1 = -Q_2 = 9.213\,10^{-2}$ m^3/s.

6.4 Non smooth solution

The purpose of this numerical test is to compare the behavior of the numerical schemes in the neighborhood of a discontinuity. We consider the two-layer shallow water system for channels with rectangular

α	L^∞-error	L^1-error
0	0.0331	0.0074
1/8	0.0085	0.0045
1/4	0.0085	0.0045
3/8	0.0085	0.0046

Table 3: L^∞ and L^1 errors (F&A interface against numerical computed ones) ($x = [-0.4, 0.4]$)

cross-section with constant breadth ($\sigma = 1$) and a flat bottom topography. In this case, the system reduces to:

$$\frac{\partial W}{\partial t} + \mathcal{A}(\mathrm{W}) \frac{\partial W}{\partial x} = 0, \tag{51}$$

where $\mathcal{A}$ is the 4×4 matrix whose expression is given by:

$$\mathcal{A}(\mathrm{W}) = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -\dfrac{Q_1^2}{A_1^2} + gA_1 & 2\dfrac{Q_1}{A_1} & gA_1 & 0 \\ 0 & 0 & 0 & 1 \\ rgA_2^2 & 0 & -\dfrac{Q_2^2}{A_2^2} + gA_2 & 2\dfrac{Q_2}{A_2} \end{bmatrix}.$$

The difficulty in defining a discontinuous solution of this system comes from the fact that the non-conservative products do not make sense within the distribution theory for such a solution. Nevertheless, there are several mathematical theories allowing to give a sense to these products in a weaker manner. In particular we consider here the definition of non-conservative product as Borel measures introduced by Volpert in [21]. This definition can be considered as a particular case of the concept of non-conservative product developed in [8]. In this latter work, the definition of the non-conservative product depends on the choice of a family of viscous paths. Volpert's definition is equivalent to the choice of the family of segments.

According to this definition, across a discontinuity with speed ξ a weak solution of the system must satisfy the generalized Rankine-Hugoniot condition

$$\int_0^1 \mathcal{A}\left(\mathrm{W}^- + s(\mathrm{W}^+ - \mathrm{W}^-)\right) \cdot (\mathrm{W}^+ - \mathrm{W}^-)\, ds = \xi(\mathrm{W}^+ - \mathrm{W}^-), \tag{52}$$

where W^-, W^+ are the left and right limits of the solution at the discontinuity.

In [15], a Roe-type numerical scheme for the two-layer shallow water system was studied which is the particular case of the here introduced family that corresponds to the choice $\alpha = 0$. In this reference it was shown that this numerical scheme is consistent with the notion of weak solutions associated to Volpert's definition of nonconservative product in the following sense: let us suppose that the approximations W_i^n and W_{i+1}^n obtained by the numerical scheme in two neighbor cells can be linked by a shock of speed ξ satisfying the jump condition (52). Then, ξ is an eigenvalue of the intermediate matrix of the scheme, $\mathcal{A}_{i+1/2}$, and $\mathrm{W}_{i+1}^n - \mathrm{W}_i^n$ is an associated eigenvector. As a consequence, the solution of the linear approximate Riemann solver used by the numerical scheme to advance in time is equal to that of the exact Riemann problem. In particular, this implies that the numerical scheme solves exactly stationary shocks. In order to numerically verify this property, in the cited reference the authors proposed a test consisting in solving the system with values $g = 10$ and $r = 0.02$, and the initial condition:

$$W(x,t) = \begin{cases} W_L \text{ if } x < 0 \\ W_R \text{ if } x > 0, \end{cases}$$

with

$$W_L = \begin{bmatrix} 1 \\ \sqrt{0.1} \\ 1 \\ \sqrt{20} \end{bmatrix}, \quad W_R = \begin{bmatrix} 0.3961560 \\ \sqrt{0.1} \\ 1.5820186 \\ \sqrt{20} \end{bmatrix}.$$

These states satisfy the generalized jump condition (52) with an error of about 10^{-5}.

We have applied to this test problem the numerical scheme with values of the parameter $\alpha = 0$, 1/8, 1/4, and 3/8. If the numerical schemes were consistent with the jump condition (52), they should obtain a stationary discontinuity placed at $x = 0$ with values of the unknowns close to the initial conditions. Figure 5 depicts the steady state solutions (water depths and discharges) for the different values of the parameter $\bar{\alpha}$ obtained with $\Delta x = 0.01$ and CFL= 0.90 for a rectangular channel of 1m long. The errors are shown in tables 2 and 3. Notice that in all the cases the numerical scheme behaves as expected: a stationary discontinuity placed at $x = 0$ and close to the initial condition is captured.

h_1		
α	L^∞-error	L^1-error
0	0.0111	$2.24{\cdot}10^{-4}$
1/8	0.0963	0.0017
1/4	0.4693	0.0059
3/8	0.4817	0.0061

h_2		
α	L^∞-error	L^1-error
0	0.0114	$1.14{\cdot}10^{-4}$
1/8	0.0972	0.0015
1/4	0.4434	0.0055
3/8	0.4574	0.0056

Table 4: L^∞ and L^1 errors (water depths)

Q_1		
α	L^∞-error	L^1-error
0	0.0114	$4.33{\cdot}10^{-4}$
1/8	0.0995	0.0014
1/4	0.1273	0.0018
3/8	0.1406	0.0020

Q_2		
α	L^∞-error	L^1-error
0	0.0127	$1.28{\cdot}10^{-4}$
1/8	0.1467	0.0021
1/4	0.2098	0.0031
3/8	0.2487	0.0039

Table 5: L^∞ and L^1 errors (discharges)

From Tables 2 and 3 and Figure 5 it is clear that the smallest errors correspond to $\alpha = 0$. For the other values an over-shooting near the stationary shock occurs (see figure 5), becoming larger for increasing values of $\bar{\alpha}$.

We have applied finally to the test problem the numerical scheme with the parameter α replaced by the function $\alpha(x)$ defined in Remark 7 with $\bar{\alpha} = 1/8$. Figure 6 depicts the steady state solution (water depths and discharges) obtained with again $\Delta x = 0.01$ and CFL= 0.90. It coincides with that obtained with $\alpha = 0$. In this case, once the stationary solution is reached the corresponding function $\alpha(x)$ takes the value 0 only at the inter-cell where the discontinuity stands and 1/8 at the others: figure 7 shows the values of $\alpha(x)$ close to the discontinuity.

7 Conclusion

We have addressed in this paper the accurate numerical solution of non-conservative non-homogeneous hyperbolic systems in one space dimension by finite volume solvers. We have introduced a general class of schemes that we have proved to be "asymptotically well-balanced" in the sense that they calculate, with at least order two, all stationary solutions of the system in all the domain but on a set whose measure tends to zero as Δx tends to zero. One of the free parameters of this class is the diffusion matrix, so we include both flux-splitting (for conservative equations) and flux-difference solvers as particular cases.

We also give general sufficient conditions under which the schemes exactly calculate a given stationary solution. These conditions may be read as meaning on one hand that the numerical flux of the scheme must compensate the centered part of the numerical source term, and on another hand that the numerical diffusion must compensate the decentered part of the numerical source term.

In the case of two-layer Shallow Water flows, we finally have introduced a specific solver to deal with the numerical difficulties associated to non-conservative products combined with nonlinear convection and source terms effects. On one hand, this solver is able to exactly compute water at rest. On another hand, it yields accurate computation of subcritical to supercritical transitions without need of entropy corrections, and also of non smooth solutions due to the presence of the non-conservative products. Our main technical innovation is an adaptive discretization of the centered part of the numerical flux which, in our opinion, may be extended to construct accurate solvers for more general non-conservative non-homogeneous equations.

Appendix

Proof of Theorem 1:

We consider first a regular stationary solution W of (1) and a compact set, $K \subset \omega(W)$.

Let T be the continuous operator appearing in system (1),

$$T(W) = \frac{\partial}{\partial x}[F(x,W)] - G(x,W) - B(x,W)\frac{\partial W}{\partial x}$$

and T_h the discrete operator appearing in the numerical scheme (11)-(14),

$$T_h(W) = \frac{\phi_R^S - \phi_L^S}{\Delta x} - G_C$$

(where we omit the arguments for brevity). We must prove that

$$T_h(W)(x_i) = T(W)(x_i) + \mathcal{O}((\Delta x)^2), \quad \forall x_i \in K.$$

The proof is based on identity (8):

$$\widetilde{T}(W) = \frac{\partial}{\partial x}\left[F(x,W) - \nu\mathcal{D}(x,W)(\frac{\partial W}{\partial x} - M^{-1}S(x,W))\right] - G(x,W) - B(x,W)\frac{\partial W}{\partial x}.$$

It is enough to prove that every term appearing in T_h:

$$\alpha_i = \frac{F_C(x_i,x_{i+1},W_i,W_{i+1}) - F_C(x_{i-1},x_i,W_{i-1},W_i)}{\Delta x},$$

$$\beta_i = \nu\frac{D(x_i,x_{i+1},W_i,W_{i+1})(W_{i+1}-W_i) - D(x_{i-1},x_i,W_{i-1},W_i)(W_i - W_{i-1})}{(\Delta x)^2} \tag{53}$$

$$\eta_i = \nu\frac{1}{\Delta x}\Big[D(x_i,x_{i+1},W_i,W_{i+1})\widetilde{M^{-1}}(x_i,x_{i+1},W_i,W_{i+1})S_D(x_i,x_{i+1},W_i,W_{i+1}) - \\ -D(x_{i-1},x_i,W_{i-1},W_i)\widetilde{M^{-1}}(x_{i-1},x_i,W_{i-1},W_i)S_D(x_{i-1},x_i,W_{i-1},W_i)\Big] \tag{54}$$

$$\delta_i = G_C(x_{i-1},x_i,x_{i+1},W_{i-1},W_i,W_{i+1}),$$

$$\lambda_i = \frac{1}{2\Delta x}\Big[\mathcal{B}_R(x_i, x_{i+1}, W_i, W_{i+1})\,(W_{i+1} - W_i) - \mathcal{B}_L(x_{i-1}, x_i, W_{i-1}, W_i)\,(W_i - W_{i-1})\Big]$$

balances up to the second order the corresponding term in $T(W)$. We prove the case of the more complex term η_i. The proof is similar for the other terms.

We define

$$R(x,U) = \mathcal{D}(x,U)M^{-1}(x,U)S(x,U),$$

$$R_C(x,y,U,V) = D(x,y,U,V)\widetilde{M^{-1}}(x,y,U,V)S_D(x,y,U,V),$$

and we want to prove that

$$\Delta R_C(x_i) = R_C(x_i, x_{i+1}, W_i, W_{i+1}) - R_C(x_{i-1}, x_i, W_{i-1}, W_i) =$$

$$= \Delta x \frac{\partial}{\partial x}[R(x, W(x))]_{|x=x_i} + \mathcal{O}((\Delta x)^2). \tag{55}$$

As K is a compact set and $\mathbf{R}^{N+1} \setminus \gamma$ is an open set, there exists $\delta_K > 0$, such that if

$$|U - W(x)| + \Delta x < \delta_K,$$

then $(x + \Delta x, U) \subset \mathbf{R}^{N+1} \setminus \gamma$, for all $x \in K$.

On another hand, if $W_{i+\theta} = (1-\theta)W_i + \theta W_i$, then using the Mean Value Theorem we prove

$$|W_{i+\theta} - W(x_{i+\theta})| \le \frac{\Delta x}{2}\left\|\frac{\partial W}{\partial x}\right\|_{L^\infty(K)}; \quad \forall\, x_i \in K.$$

So, if

$$\Delta x < \delta_K/(1 + \|\frac{\partial W}{\partial x}\|/2),$$

we obtain that $(x_{i+\theta}, W_{i+\theta}) \subset \mathbf{R}^{N+1} \setminus \gamma$, for $\theta \in [0,1]$. Then if Δx verifies the previous inequality, we obtain that $L[(x_i, W_i), (x_{i+1}, W_{i+1})] \subset \mathbf{R}^{N+1} \setminus \gamma$ for all $x_i \in K$. Then, we can apply the hypothesis H1 c), d) and f) and conclude that

$$R_C(x_i, x_{i+1}, W_i, W_{i+1}) = R\left(x_{i+1/2}, \frac{W_i + W_{i+1}}{2}\right) +$$

$$+R_1\left(x_{i+1/2}, \frac{W_i + W_{i+1}}{2}\right)\begin{pmatrix} x_{i+1} - x_i \\ W_{i+1} - W_i \end{pmatrix} + \mathcal{O}(|x_{i+1} - x_i|^2 + |W_{i+1} - W_i|^2);$$

for some C^1 matrix function R_1.

As W is of class C^2, we have that

$$R_1\left(x_{i+1/2}, \frac{W_i + W_{i+1}}{2}\right)\begin{pmatrix} x_{i+1} - x_i \\ W_{i+1} - W_i \end{pmatrix} - R_1\left(x_{i-1/2}, \frac{W_{i-1} + W_i}{2}\right)\begin{pmatrix} x_i - x_{i-1} \\ W_i - W_{i-1} \end{pmatrix} =$$

$$= \Delta x\Bigg[R_1(x_{i+1/2}, W(x_{i+1/2}))\begin{pmatrix} 1 \\ \dfrac{\partial W}{\partial x}(x_{i+1/2}) \end{pmatrix} -$$

$$-R_1(x_{i-1/2}, W(x_{i-1/2}))\begin{pmatrix} 1 \\ \dfrac{\partial W}{\partial x}(x_{i-1/2}) \end{pmatrix}\Bigg] + \mathcal{O}((\Delta x)^2) =$$

$$= (\Delta x)^2\frac{\partial}{\partial x}\left[R_1(x, W(x))\begin{pmatrix} 1 \\ \dfrac{\partial W}{\partial x} \end{pmatrix}\right]\Bigg|_{x=x_i} + \mathcal{O}((\Delta x)^2) = \mathcal{O}((\Delta x)^2). \tag{56}$$

Moreover,

$$R\left(x_{i+1/2}, \frac{W_i + W_{i+1}}{2}\right) - R\left(x_{i-1/2}, \frac{W_{i-1} + W_i}{2}\right) = \Delta x \frac{\partial}{\partial x}\Big[R(x, W(x))\Big]\Bigg|_{x=x_i} + \mathcal{O}((\Delta x)^2). \quad (57)$$

From (56) and (57) we conclude that (55) is verified.

We can now prove the statements *a*) and *b*) of Theorem 1:
a) If W has no singular points, then $\omega(W) = (0, L)$ and it is enough to use the preceding argument for some compact set $K \subset \omega(W)$ containing all the internal nodes of the discretization.
b) If $\text{meas}(\sigma(W)) = 0$, as $\omega(W)$ is an open set, we may find an increasing sequence of compact sets $\{K_n\}_{n\geq 1}$ such that $\cup_{n\geq 1} K_n = \omega(W)$. Then, $\text{meas}((0, L) \setminus \cup_{n\geq 1} K_n) = 0$. It is enough to use the preceding argument for each K_n.

To prove statement *c*) let us remark that, by the previous argument, the scheme is asymptotically well-balanced in $\omega(W)$. Thus, it only remains to prove this property in $\sigma(W)$ when this set has non-zero measure.

Let us denote by $\text{int}(\sigma(W))$ the interior of $\sigma(W)$. As

$$\text{meas}(\sigma(W) - \text{int}(\sigma(W))) = 0,$$

it is enough to prove the property in $\text{int}(\sigma(W))$. And, as we shall see this occurs because in $\text{int}(\sigma(W))$ $\mathcal{D}$ and M^{-1} coincide with smooth functions.

For sake of simplicity, we shall assume $N = 2$ and that the first eigenvalue of M vanishes in $\sigma(W)$.

Let us define the operator

$$\widetilde{T}(W) = \frac{\partial}{\partial x}\left[F(x, W) - \mathcal{H}(x, W)\left(\frac{\partial W}{\partial x} - \mathcal{C}(x, W)S(x, W)\right)\right] -$$

$$-G(x, W) - B(x, W)\frac{\partial W}{\partial x};$$

where

$$\mathcal{H}(x, W) = X(x, W)\begin{pmatrix} 0 & 0 \\ 0 & d_2(x, W) \end{pmatrix} X^{-1}(x, W);$$

$$\mathcal{C}(x, W) = X(x, W)\begin{pmatrix} 0 & 0 \\ 0 & \lambda_2^{-1}(x, W) \end{pmatrix} X^{-1}(x, W).$$

By d_2 and λ_2 we denote the second eigenvalue of $\mathcal{D}$ and M, respectively.

As W is a steady solution of (1),

$$M(x, W)\frac{\partial W}{\partial x}(x) = S(x, W).$$

Then,

$$\mathcal{C}(x, W)S(x, W) = \mathcal{C}(x, W)M(x, W)\frac{\partial W}{\partial x}(x).$$

But

$$\mathcal{C}(x, W)M(x, W)\frac{\partial W}{\partial x}(x) = X(x, W)\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} X^{-1}(x, W)\frac{\partial W}{\partial x}(x) =$$

$$= X(x, W)\left[\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 10 & 0 \\ & 0 \end{pmatrix}\right] X^{-1}(x, W)\frac{\partial W}{\partial x}(x) =$$

$$= \frac{\partial W}{\partial x} - \left[X^{-1}(x, W)\frac{\partial W}{\partial x}(x)\right]_1 X_1(x, W),$$

where by $[\,\cdot\,]_1$ we denote the first component of a vector and by X_1 the first eigenvector of M.

Then, as X_1 is also the first eigenvector of $\mathcal{H}$,

$$\mathcal{H}(x,W)\left[\frac{\partial W}{\partial x}(x)-\mathcal{C}(x,W)S(x,W)\right]=\lambda_{\mathcal{H},1}(x,W)\left[X^{-1}(x,W)\frac{\partial W}{\partial x}(x)\right]_1 X_1(x,W)=0,$$

as by construction, the first eigenvalue $\lambda_{\mathcal{H},1}$ of $\mathcal{H}$ is zero.

We deduce that

$$\widetilde{T}(W)=T(W).$$

It is then enough to prove that for any compact set $K\subset \operatorname{int}(\sigma(W))$ there exists $\delta_K>0$ such that if $0<\Delta x<\delta_K$, then

$$T_h(W)(x_i)=\widetilde{T}(W)(x_i)+\mathcal{O}((\Delta x)^2),\quad \forall\, x_i\in K. \tag{58}$$

Les us assume

$$\delta_K\leq \operatorname{dist}(K,\partial\sigma(W)).$$

If $x_i\in K$, then x_{i-1}, $x_{i+1}\in \operatorname{int}(\sigma(W))$ and we may ensure that $L[(x_i,W_i),(x_{i+1},W_{i+1})]$ cuts $\gamma=\widetilde{\gamma}$. Then, using our definition of $\widetilde{M^{-1}}$ and D, we have

$$\widetilde{M^{-1}}(x_i,x_{i+1},W_i,W_{i+1})=$$

$$=X(x_{i+1/2},\widetilde{W}_{i+1/2})\begin{pmatrix}0 & 0\\ 0 & \lambda_2^{-1}(x_{i+1/2},\widetilde{W}_{i+1/2})\end{pmatrix}X^{-1}(x_{i+1/2},\widetilde{W}_{i+1/2})=\mathcal{C}(x_{i+1/2},\widetilde{W}_{i+1/2}),$$

where we denote $\widetilde{W}_{i+1/2}=\widetilde{W}(x_i,x_{i+1},W_i,W_{i+1})$. Also,

$$D(x_i,x_{i+1},W_i,W_{i+1})=$$

$$=X(x_{i+1/2},\widetilde{W}_{i+1/2})\begin{pmatrix}0 & 0\\ 0 & d_2(x_{i+1/2},\widetilde{W}_{i+1/2})\end{pmatrix}X^{-1}(x_{i+1/2},\widetilde{W}_{i+1/2})=\mathcal{H}(x_{i+1/2},\widetilde{W}_{i+1/2}).$$

Now, T_h can be written as in the previous case by replacing D by $\mathcal{H}$ and $\widetilde{M^{-1}}$ by $\mathcal{C}$ in the definitions of β_i, η_i and the proof is similar.

□

References

[1] A. Bermúdez, M. E. Vázquez Cendón, *Upwind Methods for Hyperbolic Conservation Laws with Source Terms.* Computers & Fluids 23(8), pp. 1049-1071 (1994).

[2] M. J. Castro, J. Macías, C. Parés, *A Q-scheme for a class of systems of coupled conservation laws with source term. Application to a two-layer 1-d shallow water system.* M2AN 35(1), pp. 107-127 (2001).

[3] M. J. Castro, J. Macías, C. Parés, J. A. Rubal, M. E. Vázquez, *Two-layer numerical model for solving exchange flows through channels with irregular geometry.* Proceedings of “ECCOMAS 2001”, Swansea (2001).

[4] M. J. Castro, J. A. García, J. M. González, J. Macías, C. Parés, M. E. Vázquez, *Numerical simulation of two-layer Shallow Water flows in channels with irregular geometry.* Journal of Comp. Phys. 195, pp. 202-235 (2004).

[5] T. Chacón Rebollo, A. Domínguez Delgado, E. D. Fernández Nieto, *A family of stable numerical solvers for Shallow Water equations with source terms.* Comput. Methods Appl. Mech. Eng. 192, pp. 203-225 (2003).

[6] T. Chacón Rebollo, A. Domínguez Delgado, E. D. Fernández Nieto, *An entropy correction-free solver for non-homogeneous shallow-water equations.* M2AN 37(3), pp. 755-772 (2003).

[7] T. Chacón Rebollo, A. Domínguez Delgado, E. D. Fernández Nieto, *Asymptotically balanced schemes for non-homogeneous hyperbolic systems. Application to Shallow Water Equations.* C.R. Acad. Sci. Paris, Sr. I 338, pp. 85-90 (2004).

[8] G. Dal Maso, P. G. LeFloch, F. Murat, *Definition and weak stability of non-conservative products.* J. Math. Pures Appl 74, pp. 483-548 (1995).

[9] D. Farmer and L. Armi, *Maximal two-layer exchange over a sill and through a combination of a sill and contraction with barotropic flow.* J. Fluid Mech. 164, pp. 53-76 (1986).

[10] E. D. Fernández Nieto,*Aproximación numérica de leyes de conservación hiperbólicas no homogéneas. Aplicación a las ecuaciones de Aguas Someras.* Ph.D.Thesis Universidad de Sevilla, (2003).

[11] A.C. Fowler, *Mathematical Model in the Applied Sciences.* Cambridge (1997).

[12] E. Godlewski - P. A. Raviart, *Numerical Approximation of Hyperbolic Systems of Conservation Laws.* Springer - Verlag (1996).

[13] J.M. Greenberg, A. Y. Leroux, *A Well-Balanced scheme for the numerical processing of source terms in hyperbolic equations..* SIAM J. Numer. Anal. 33(1), pp. 1-16 (1996).

[14] A. Kurganov, D. Levy, *Central-upwind schemes for the Saint-Venant system.* M2AN 36(3), pp. 397-425 (2002).

[15] C. Parés, M. J. Castro,*On the well-balanced property of Roe's method for non-conservative hyperbolic systems. Applications to shallow-water systems.* M2AN 38(5), pp. 821-852, (2004).

[16] B. Perthame, C. Simeoni, *A kinetic scheme for the Saint-Venant system with a source term.* Calcolo 38(4), 201-231 (2001).

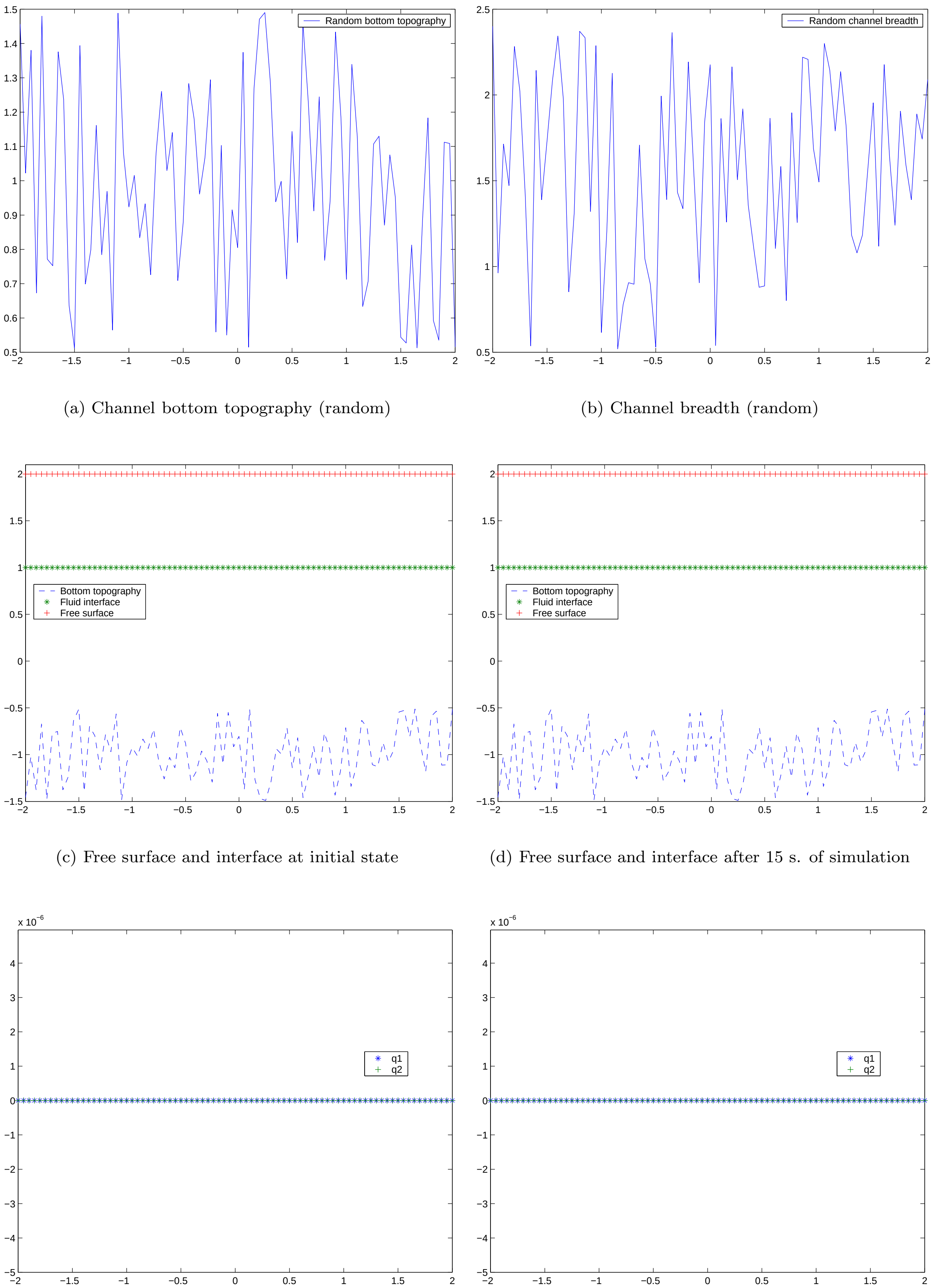
[17] B. Perthame, C. Simeoni, *Convergence of the Upwind Interface Source method for hyperbolic conservation laws.* Proceeding of Hyp 2002, Thou and E. Tadmor editor, Springer (2003).

[18] Randall J. LeVeque *Numerical Methods for conservation Laws.* Birkhauser Verlag, Zurich (1990).

[19] P.L. Roe, *Upwind differencing schemes for hyperbolic conservation laws with source terms. Nonlinear Hyperbolic Problems, C. Carraso, P.-A. Raviart and D. Serre, eds.,* Springer-Verlag, Lecture Notes in Matematics 1270, pp. 41-51 (1986).

[20] J.B. Schijf and J.C. Schonfeld. *Theoretical considerations on the motion of salt and fresh water.* In *Proc. of the Minn. Int. Hydraulics Conv.*, 321–333. Joint meeting IAHR and Hyd. Div. ASCE., Sept. 1953.

[21] A. I. Volpert, *The space BV and quasilinear equations*, Math. USSR Sbornik 73(115), pp. 225-267 (1967).



(a) Channel bottom topography (random)

(b) Channel breadth (random)

(c) Free surface and interface at initial state

(d) Free surface and interface after 15 s. of simulation

(e) Discharge at each layer at initial state

(f) Discharge at each layer after 15 s. of simulation

Figure 2: Water at rest in an irregular channel ($\alpha = 1/8$).

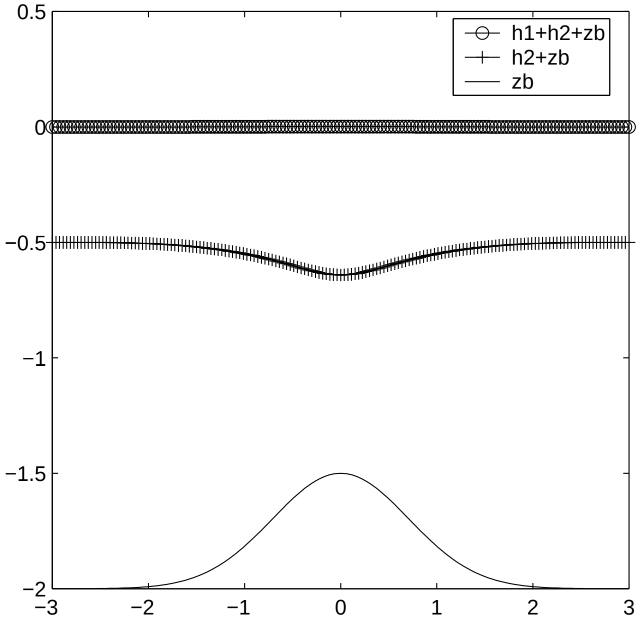
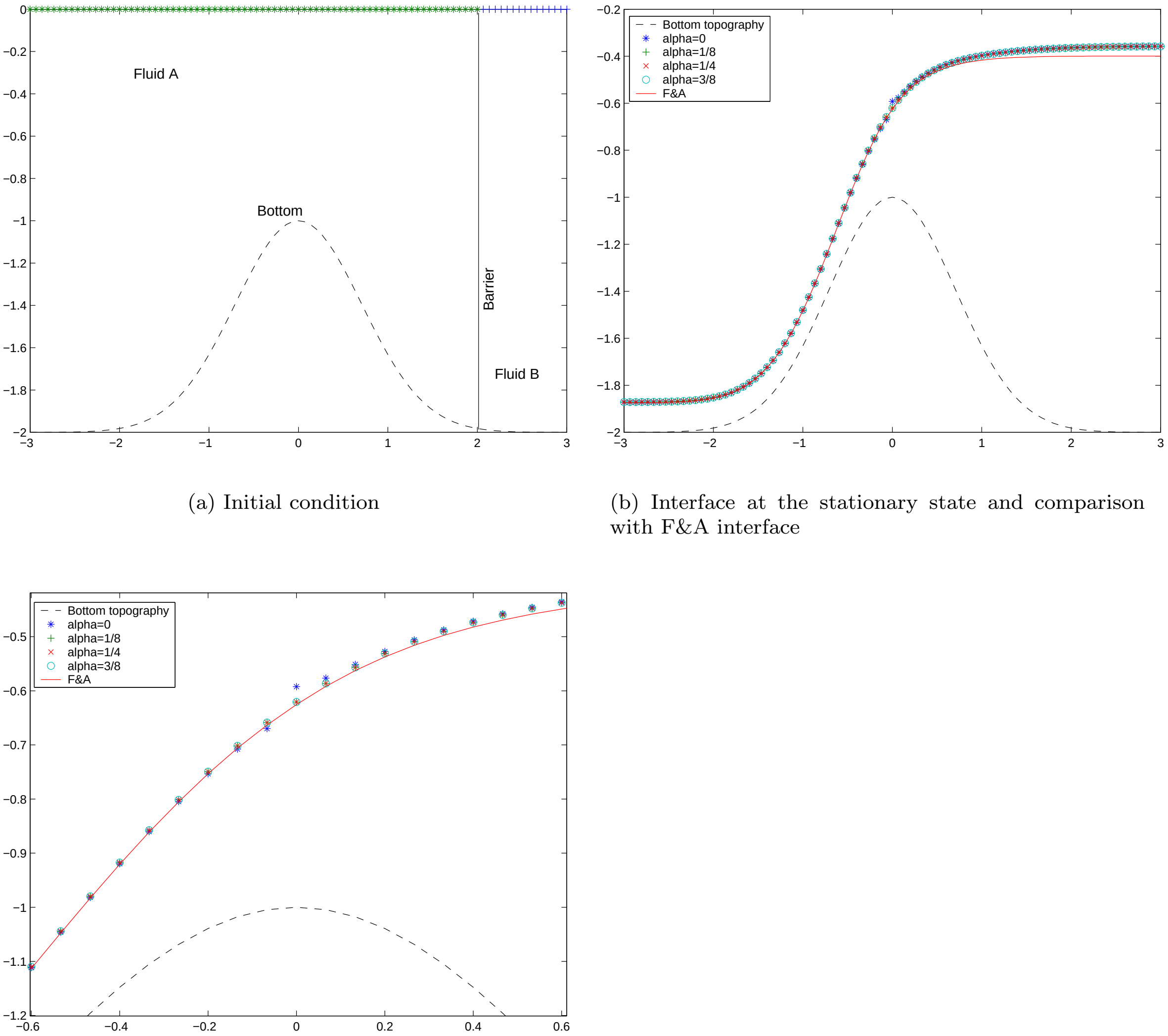


Figure 3: Stationary solution in test case 6.2. Elevations $\eta_1 = h_1 + h_2 + z_b(x)$, $\eta_2 = h_2 + z_b(x)$ and bottom topography $z_b(x)$.

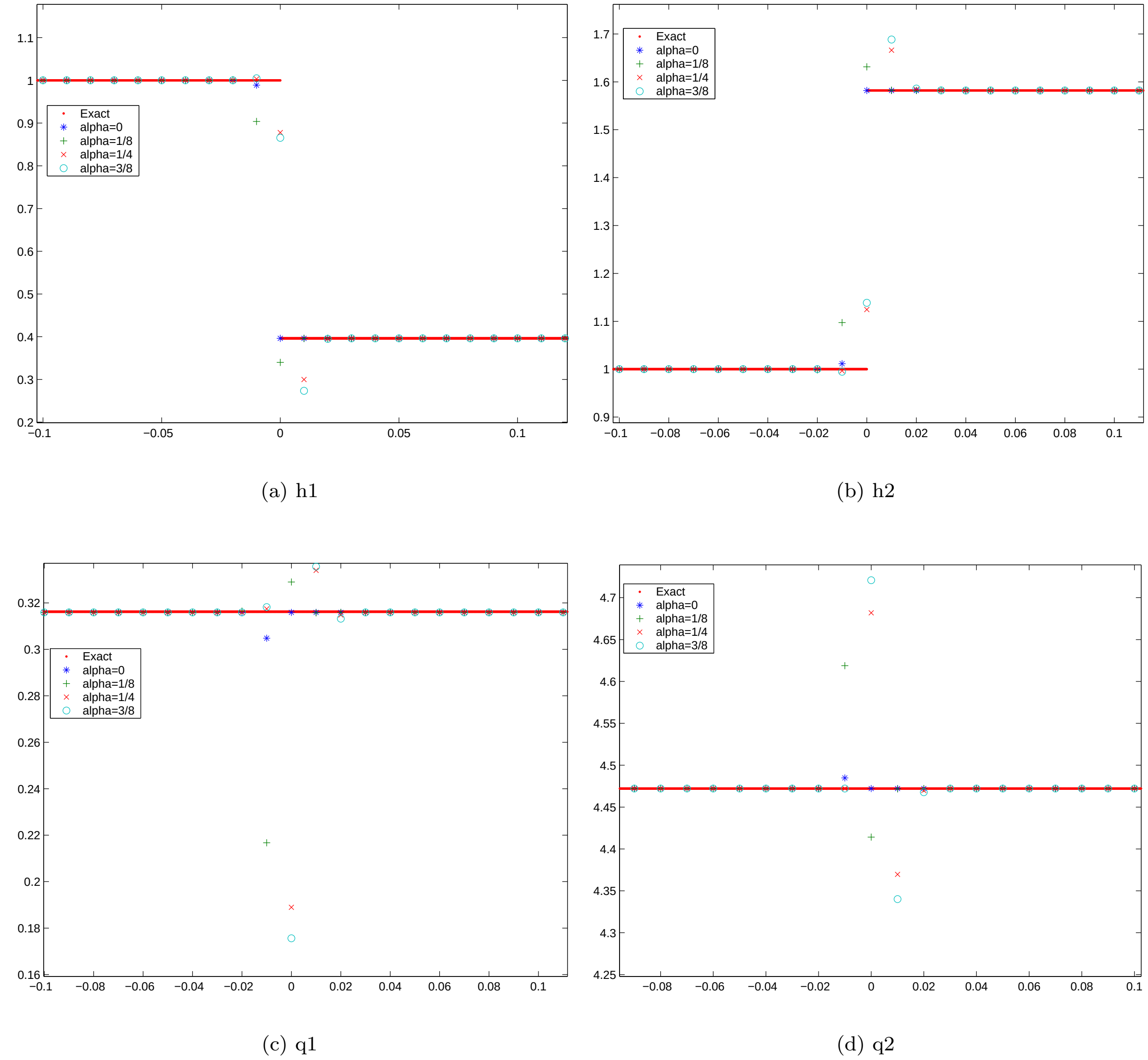


(a) Initial condition

(b) Interface at the stationary state and comparison with F&A interface

(c) Interface at the stationary state and comparison with F&A interface(zoom)

Figure 4: Maximal exchange flow through a rectangular channel with a single bump. Initial condition and comparison with F&A stationary interface ($\bar{\alpha} = 0,\ 1/8,\ 1/4,\ 3/8$).



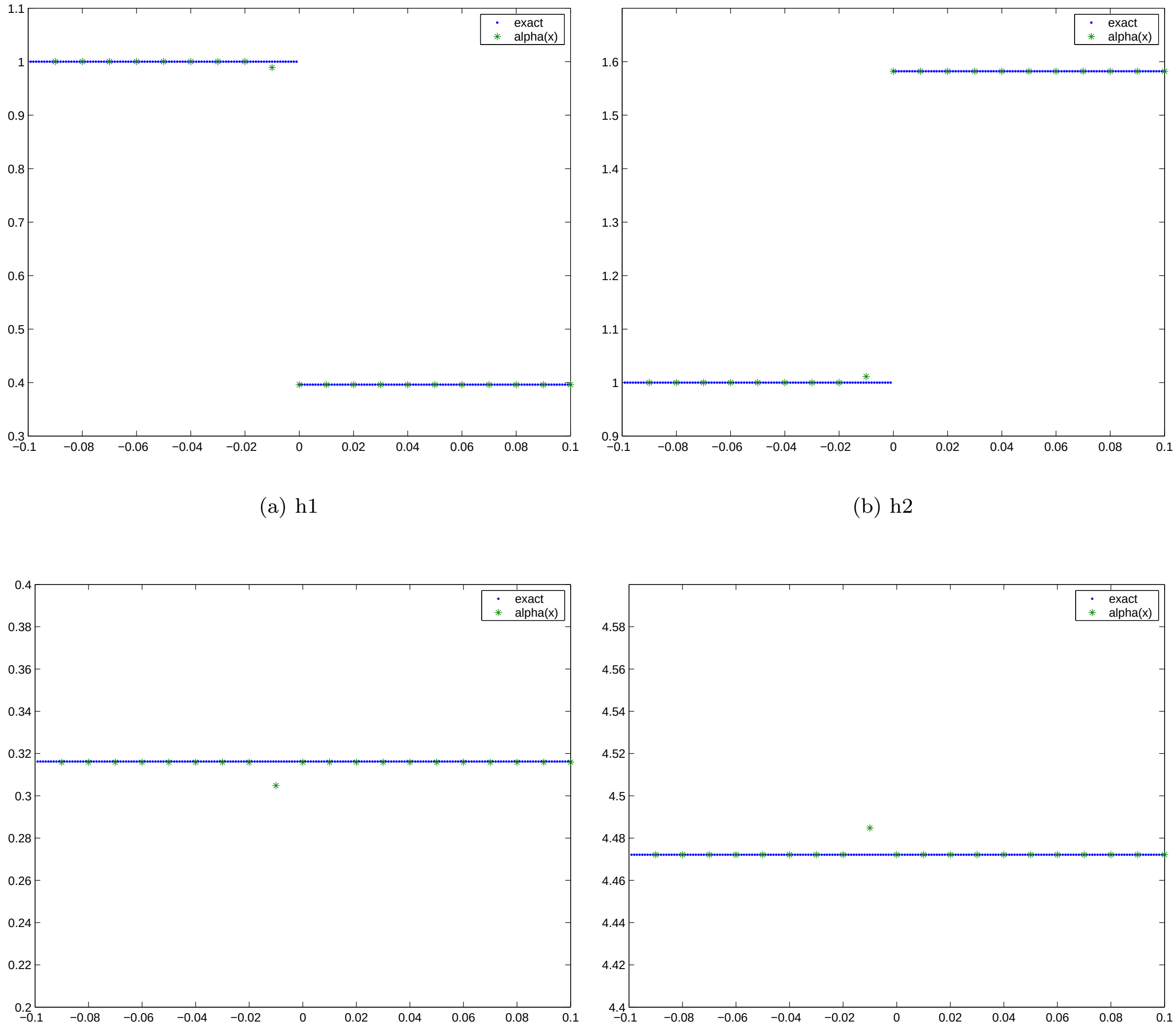
(a) h1

(b) h2

(c) q1

(d) q2

Figure 5: Internal hydraulic jump: water depths and discharges at each layer of fluid for $\bar{\alpha} = 0$, 1/8, 1/4 and 3/8.



(a) h1

(b) h2

(c) q1

(d) q2

Figure 6: Internal hydraulic jump: water depths and discharges at each layer of fluid for $\alpha(x)$ with $\bar{\alpha} = 1/8$ defined in remark 7.

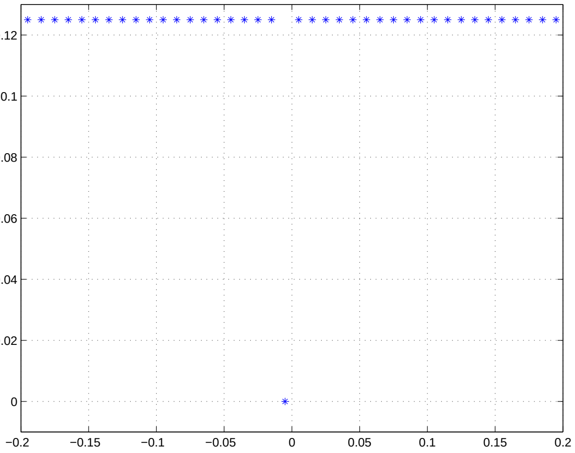


Figure 7: $\alpha(x)$ near the discontinuity placed at $x = 0$.