A comprehensive explanation and exercise of the source terms in hyperbolic systems using Roe type solutions. Application to the 1D-2D shallow water equations.

J. Murillo*, A. Navas-Montilla

LIFTEC-EINA, CSIC-Universidad de Zaragoza, Spain

Abstract

Powerful numerical methods have to consider the presence of source terms of different nature, that intensely compete among them and may lead to strong spatiotemporal variations in the flow. When applied to shallow flows, numerical preservation of quiescent equilibrium, also known as the well-balanced property, is still nowadays the keystone for the formulation of novel numerical schemes. But this condition turns completely insufficient when applied to problems of practical interest. Energy balanced methods (E-schemes) can overcome all type of situations in shallow flows, not only under arbitrary geometries, but also with independence of the rheological shear stress model selected. They must be able to handle correctly transient problems including modeling of starting and stopping flow conditions in debris flow and other flows with a non-Newtonian rheological behavior. The numerical solver presented here satisfies these properties and is based on an approximate solution defined in a previous work. Given the relevant capabilities of this weak solution, it is fully theoretically derived here for a general set of equations. This useful step allows providing for the first time an E-scheme, where the set of source terms is fully exercised under any flow condition involving high slopes and arbitrary shear stress. With the proposed solver, a Roe type first order scheme in time and space, positivity conditions are explored under a general framework and numerical simulations can be accurately performed recovering an appropriate selection of the time step, allowed by a detailed analysis of the approximate solver. The use of case-dependent threshold values is unnecessary and exact mass conservation is preserved.

Keywords: Hyperbolic systems, Stopping conditions, Source terms, Well-balanced, Energy-balanced, Wet/dry front
2000 MSC: 35L65, 65M06, 65M12, 76M12, 76M20.

1. Introduction

There are a unaffordable number of processes over the earth surface where a common agent is present: water. Water can participate in different ways: as a result of a high porosity in landslide events or as almost pure water in rivers. Many geophysical or environmental flows in earth have another relevant characteristic: the geometrical scales presented in the problem allow us to define them mathematically as shallow type flows. The vertical scales can be considered very small if compared with the horizontal ones. This aspect ratio appears in channels, rivers, oceans or even the atmosphere,

*Corresponding author
Email address: Javier.Murillo@unizar.es (J. Murillo)

but also in debris flows, landslides and tsunamis. All these processes can be mathematically modeled and are defined as flows of hyperbolic nature. Their importance makes necessary the development of predictive tools. Predictive tools were first derived for gas dynamics, and the results were applied next to the shallow water equations. Shallow type flows are hyperbolic but not strictly hyperbolic. Their characteristics require the development of novel numerical techniques. Among them, most advanced numerical predictive methods consider partial results, e.g. the well-balanced property. This property is a particular case of energy-balanced or E-schemes that have a significant advantage: they provide accurate results when using a small amount of information. This amount of information can be measured as the number of computational cells where data is stored. When the number of cells decreases, the computational cost also does, allowing the integration of more processes in time and space. A complete understanding of solvers accounting for the presence of source terms also allows to successfully predict the behavior of sophisticated terms when applied to cases of not pure water floods, such as mud/debris floods, where unsteady flow phenomena includes stop and go mechanisms. In this way, it is possible to analyze the relative importance of the shear stresses versus bottom topography variations, allowing a correct tracking of the fluid moving boundaries.

Realistic applications of conservation laws involve the presence of source terms dominating the solution, where the flux gradients are nonzero but exactly balanced by source terms in steady situations [1]. When trying to reproduce numerical solutions with discontinuities in both the conserved variables and the source terms, the mathematical formulation of the governing equations and the selection of the numerical scheme is of utmost importance. Fractional step methods have been widely used to involve the presence of source terms in the solution [2], but from their earliest developments, well balanced numerical schemes [3, 4, 1] have gained maturity to progressively become methods of choice for the numerical simulation of conservation laws with source terms. When applied to the shallow water equations (SWE), the preservation of motionless steady state or quiescent equilibrium over irregular geometries has been the keystone for the construction of numerical schemes in the context of shallow flows [5, 6, 7, 8, 9].

Riemann solvers derived for the homogeneous case, in combination with a suitable treatment of the source terms, are able to ensure quiescent equilibrium when solving the SWE. This property can be ensured by expressing the equations following a deviatoric formulation [5, 6, 10], replacing the water depth by the water surface elevation as conserved variable. But changes in the selection of the variables have consequences. When selecting the deviatoric formulation, the approximate solution is single valued and even though in presence of bed discontinuities level surface is constant in cases of quiescent equilibrium, this solution is no longer valid in general problems over a bottom step, invalidating the use of classical Riemann solvers [11, 12, 13]. The most noticeable consequence is the inability of such type of solvers to ensure an exact preservation of the mass discharge in steady cases for any type of flow regime [14].

Also, as pointed out by [15] when solving the SWE two types of difficulties are often encountered: the preservation of steady state solutions and the preservation of water height positivity. The SWE admit the general moving water equilibrium and require exactly well-balanced methods in cases with moving water equilibrium [16, 17, 18, 19]. The well balanced numerical property in cases of quiescent equilibrium is a particular case. In [20, 21, 22, 23] exactly well-balanced methods, named energy balanced numerical schemes and hereafter referred as E-schemes, able to reproduce exactly steady solutions with independence of the mesh refinement, were presented. On the other hand, non-physical negative water height becomes problematic when computing simulations as the eigenvalues do not determine the time step size as a result of the not pure hyperbolic characteristic of the system of equations [24, 25, 15]. The use of case-dependent threshold values for the water depth, limiting the

computational domain in the computation of the flow advance over dry bed, is the current tendency to avoid numerical difficulties [26]. Other techniques involve velocity based limiters [27, 28] to control the stability at wet/dry fronts.

Riemann Problems (RP) in not strictly hyperbolic system of equations involve complex exact solutions, as the presence of the source terms may lead to resonant problems. These resonant problems include cases where characteristic speeds may coincide or cases where the total number of waves involved in the solution are larger than the number of characteristic fields [29]. An extensive review of approximate solutions to discontinuous problems in nonlinear hyperbolic systems can be found in [30]. Convergence to the exact solution can be ensured using appropriate Augmented solvers [1, 31, 32, 14]. Augmented solvers provide suitable explanations to the influence of the source terms in the numerical solution. They include an extra wave associated to the presence of the source terms in the approximate solution. The aforementioned possible computation of non-physical negative water height was explained and remedied in [32, 14] by the description of the internal structure of the associate approximate Riemann solution.

Even though a great variety of works that focus on the preservation of height positivity by exploring the effects of bed slope terms can be found in literature, when moving to realistic applications, the discretization of frictional source terms is essential to provide accurate results, independently of the friction stress model chosen. When the source term discretization of shear stress is not considered in the context of the approximate solution used, it can not only spoil the solution accuracy, but the numerical computation may become unstable and fail [33, 34]. Fractional explicit step methods lead to oversized discrete friction forces that ruin the simulation, and although an implicit treatment of the resistance source term ensures stability, an exact balance among fluxes and source terms is not generated [35], leading to undesirable non uniform discharge values. That is, convergence to the exact solution can never be provided. The upwind unified treatment of boundary shear stress ensures exact conservation of discharge in steady cases, but being an explicit treatment, the appearance of non-physical negative values of water depth and the selection of the time step size, become again problematic [34, 36], as in those numerical schemes where bed slope effects are only analyzed [25].

The definition of appropriate numerical schemes involving bed variations can be envisaged using families of paths [37] connecting the left and right states of the RP [38]. But even in cases where only discontinuous bed level is considered, the selection of families of paths is not a trivial task [39, 40, 29]. One commonly used strategy is based on supplementing the initial set of equations shaping the SWE with another extra equation, expressing a nil time derivative of the bed level surface [41, 18]. The results show that although Godunov-type path-consistent schemes do converge with mesh refinement, they do not necessarily converge to the physically relevant or correct solution [30]. Families of paths cannot be generalized when highly nonlinear relations appear [42, 43], and in the case of analyzing frictional source terms, the definition of an extra equation assuming nil time derivative of a specific variable makes no sense.

Approximate augmented solvers, as the ARoe (Augmented Roe) solver in [32], involve numerical strategies based on a direct discretization of all type of source terms, allowing to explain and correct if necessary, their impact in the solution. Augmented solvers allow to analyze approximate solutions involving variable density [44], one dimensional blood flow in arteries [47] or applications with complex rheologies, where the well-balanced property must be redefined to provide accurate stop-and-go triggering mechanisms [45, 46]. It is remarkable that although the ARoe solver uses a limited number of characteristic waves, it still ensures convergence in resonance regions [20] in the SWE.

Therefore, the ARoe solver provides a convenient way to evaluate source term discretization in situations far away from quasi-steady conditions, allowing the generation of verification tools for

any case where a source term is involved, such as discontinuous bed in combination with complex rheologies, avoiding computation of non-physical negative water height and restoring a clear time step selection. To achieve this goal it is not sufficient to define the solution of the approximate solver exclusively in the inter cell position but in the whole solution plane, as it does finally participate in the updating solution. This strategy, used to defined positively conservative methods in the homogeneous case by means of an entropy fix [48], can be extended to non-strictly hyperbolic system of equations with source terms, generating extra fix procedures that can be used not only to ensure positively conservative solutions, but also to define friction fix techniques able to ensure an accurate viscous dissipation rate. For steady cases, the resulting scheme allows to obtain the exact critical point at the cell with highest bed elevation, as transitions between subcritical and supercritical states are forbidden, as outlined by Alcrudo et al. [49].

Although the ARoe solver detailed here is only first order with an explicit Euler time-stepping, it is worth pointing out that a relevant feature of this solver is that it can be directly applied to flux-ADER schemes with arbitrary order [22, 23] without losing any of its desired properties.

The approximate solution underlying the ARoe solver was presented but not explicitly derived in [32], where a first investigation of the Riemann problem for the shallow water equations with source terms was performed. The present paper provides a significant improvement in the complete description of this approximate Riemann solver and it is now fully theoretically derived for the first time. In this work a useful, comprehensive description of the derivation of the ARoe solver is presented. In Section 2, we recall basic ideas regarding the scalar case with source terms, using them in Section 3 to construct solutions for non-strictly hyperbolic systems of equations of arbitrary size. The formal derivation of the approximate solutions proposed in previous works and its extension/generalization to hyperbolic systems of equations of arbitrary size with source terms, provides the necessary background for a successful application to a particular physically based model. Departing from these results, all possible types of flow transitions and the correct management of different types of source terms in any of these different cases are analyzed in depth.

Section 4 revisits the construction of solutions to the Riemann problem for the x-split two-dimensional SWE and the impact of the source terms in these solutions. Special emphasis is put here in the analysis of transcritical flow, that requires of a careful analysis and treatment of the entropy corrections in presence of source terms. In Section 5, numerical integration of the source terms in the SWE is revisited, recalling the energy balanced approach (E-property). The source term accounts for pressure and shear stress forces on the bed surface and the their effect over the approximate solution cannot be arbitrarily analyzed. In order to accurately reproduce a physically based solution, a suitable procedure is proposed, analyzing first the effect of an arbitrary shear stress and next the total non-conservative contributions. The impact of the shear stress evaluation in the solution is discussed in Section 6, proposing a friction fix that ensures a physically feasible numerical solution, that is, a solution where shear stress only acts as an energy dissipation mechanism. Positivity fix in the SWE is revisited in Section 7, where suitable modifications of the solution structures are provided preventing computation of non-physical negative water height while retaining an efficient selection of the time step based on the eigenvalues of the system. These modifications are presented by means of the definition of limiting functions over the values of the source terms.

In Section 8, the extension of the numerical scheme to two-dimensions is recalled. Numerical fluxes are constructed at each cell edge by solving the x-split two-dimensional RP according to Sections 6 and 7, ensuring appropriate values of friction while retaining positivity of the solution. By simply projecting these fluxes using a rotation matrix, the grid cells can be updated avoiding redefinition of the numerical fluxes depending on the type of mesh selected (structured/unstructured). Finally, Sec-

tion 9 is devoted to numerical applications involving unsteady problems with exact solution. Among the different test cases presented, comparisons between the proposed numerical scheme and a finite volume scheme based on the hydrostatic reconstruction technique [7] are presented.

2. Scalar conservative laws with source terms

The basic ideas underlying this work are first illustrated by examining the results of a nonlinear scalar equation applied to quasi-steady problems in [2]

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = s\,, \tag{1}$$

where $f(u)$ is a convex non linear flux of u and $s = s(u,x)$ is a source term. The source term, of geometric type, depends upon the position x and can be discontinuous. From $f(u)$ it is possible to find an advection, or transport velocity λ:

$$\lambda = \frac{df}{du}\,, \qquad \lambda = \lambda(u)\,. \tag{2}$$

The upwind method described is derived as a special case of the reconstruction, evolution and averaging steps method proposed originally by Godunov [50], that provides updated cell averages values for the conserved variables u_i^{n+1}. The Godunov method starts from piecewise constant data reconstructions. The theory of RPs can be applied, and does not require to determine the exact solution of the full wave structure of the RP, allowing the use of linearized approximate or weak solutions in the evolution step. The following RP is defined

$$\partial_t u + \partial_x f - s = 0\,, \qquad u(x,0) = \left\{ \begin{array}{ccc} u_i & if & x < 0 \\ u_{i+1} & if & x > 0 \end{array} \right. \tag{3}$$

and the initial value problem in (3) will be solved using the explicit conservative formula

$$u_i^{n+1} = u_i^n - \frac{\Delta t}{\Delta x}[f^-_{i+\frac{1}{2}} - f^+_{i-\frac{1}{2}}]\,, \tag{4}$$

where Δx is the cell size and Δt is the time step selected. Intercell numerical fluxes $f^{\pm}_{i+\frac{1}{2}}$ contain information regarding flux function f and source term s, and will be derived from approximate solutions of the RP.

2.1. Integral Relations in the Riemann Solution

Even when ignoring the exact solution of the RP in (3), it is possible to estimate its variation by integrating (3) over a suitable control volume. Figure 1 shows a right moving rarefaction wave, $\lambda > 0$, with initial values u_i, u_{i+1}, and a control volume given by the time interval $[0, \Delta t]$ and the space interval $[-x_L, x_R]$, with $x_R > \lambda\Delta t$ the position of the fastest wave at $t = \Delta t$. The initial solution is modified between x=0, where a point source term is present, and the position given by the fastest signal $\lambda\Delta t$. No signal propagates upstream of x=0. Integrating (3) over the control volume $[0, \Delta t] \times [-x_L, x_R]$

$$\int_{-x_L}^{x_R} \int_0^{\Delta t} \left(\frac{\partial u}{\partial t} + \frac{\partial f}{\partial x} - s \right)\, dtdx = 0\,, \tag{5}$$

the first term in (5) is

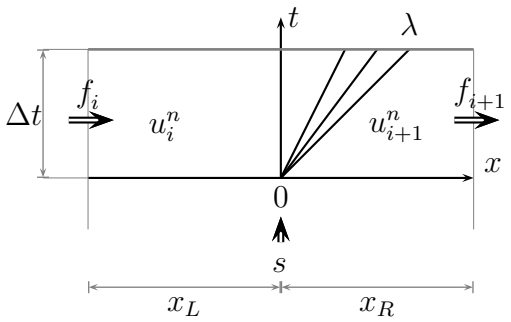

Figure 1: Integration control volume defined by a time interval $[0, \Delta t]$ and a space interval $[-x_L, x_R]$

$$\int_{-x_L}^{x_R}\int_0^{\Delta t}\left(\frac{\partial u}{\partial t}\right)\, dtdx = \int_{-x_L}^{x_R} u(x,\Delta t)\, dx - \left(x_R u_{i+1}^n + x_L u_i^n\right) \tag{6}$$

and the second term in (5) becomes

$$\int_0^{\Delta t}\int_{-x_L}^{x_R}\left(\frac{\partial f}{\partial x}\right)\, dxdt = \delta f_{i+1/2}\Delta t\,, \tag{7}$$

with $\delta(\cdot)_{i+1/2} = (\cdot)_{i+1} - (\cdot)_i$, $f_{i+1} = f(u_{i+1})$ and $f_i = f(u_i)$. Considering that source terms are not necessarily constant in time, the following time linearization of the nonconservative term is applied [32]

$$\int_{-x_L}^{x_R}\int_0^{\Delta t} s\, dtdx \approx \Delta t \int_{-x_L}^{x_R} s(u_{i+1}^n, u_i^n, t=0)\, dx = \Delta t \bar{s}_{i+\frac{1}{2}}\,, \tag{8}$$

retaining $\bar{s}$ as a singular source function constant in time [51], as the source term that becomes a Dirac delta. A suitable evaluation of $\bar{s}$ has to be provided depending of the type of source term present. In the case of a geometric source term [23], it is possible to define discrete average values able to reproduce correct steady solutions as well as provide convergence to the exact solution in RP [32].

This approximation is in agreement with the assumption of a constant solution in time of the RP in the intercell region, $x = 0$, for Godunov methods. Reordering, the following expression for the integral volume of $u(x,t)$ is obtained

$$\int_{-x_L}^{x_R} u(x,\Delta t)\, dx = x_R u_{i+1} + x_L u_i - (\delta f - \bar{s})_{i+1/2}\Delta t\,. \tag{9}$$

On the other hand, the integral average of the exact solution of the RP at time Δt between the fastest signal and $x = 0$, that will be referred to as $\bar{u}$, can be derived by simply setting $x_R = \lambda\Delta t$ and $x_L = 0$

$$\bar{u} = \frac{\int_{-x_L}^{x_R} u(x,\Delta t)\, dx}{\Delta t\lambda} = \frac{\lambda u_{i+1} - (\delta f - \bar{s})_{i+1/2}}{\lambda}\,. \tag{10}$$

2.2. Approximate solution. A two wave approximate Riemann solver

The definition of an approximate solver must be consistent with the results given by the integral average of the exact solution in (9) or (10). The approximate solution $\hat{u}(x,t)$ is constructed defining the following constant coefficient linear RP

$$\frac{\partial \hat{u}}{\partial t} + \widetilde{\lambda}_{i+\frac{1}{2}} \frac{\partial \hat{u}}{\partial x} = s$$
$$\hat{u}(x,0) = \begin{cases} u_i & if \quad x < 0 \\ u_{i+1} & if \quad x > 0 \end{cases} \tag{11}$$

and includes a propagating wave with celerity $\widetilde{\lambda}_{i+\frac{1}{2}}$, yet to be defined. Integrating over the same control volume we obtain

$$\int_{-x_L}^{x_R} \hat{u}(x, \Delta t)\, dx = x_R u_{i+1} + x_L u_i - (\widetilde{\lambda}\delta u - \bar{s})_{i+\frac{1}{2}} \Delta t \tag{12}$$

and in order to enforce consistency with result in (9) the constraint that follows is

$$\widetilde{\lambda}_{i+\frac{1}{2}} = \frac{\delta f_{i+1/2}}{\delta u_{i+1/2}}\,. \tag{13}$$

For both scalar or systems of conservation laws, when using Roe type solvers, any characteristic speed $\widetilde{\lambda}_{i+\frac{1}{2}} = \widetilde{\lambda}(u_i, u_{i+1})$ is defined ensuring that in case that $u_i = u_{i+1}$, the average value returns the analytical value provided by the differential form, that is, $\widetilde{\lambda}_{i+\frac{1}{2}} = \lambda_i$, with $\lambda = \partial f / \partial u$ in the scalar case, avoiding divison by zero [32].

Being the approximate solution a constant coefficient linear problem, it is assumed that the presence of the discontinuous source term $\bar{s}$ introduces a variation in $\hat{u}$ in the solution across $x = 0$, leading to two constant values u_i^-, u_{i+1}^+ at the left and right side of the (x,t) plane solution respectively, where

$$u_i^- = \lim_{x \to 0^-} \hat{u}_i(x, t > 0)\,, \qquad u_{i+1}^+ = \lim_{x \to 0^+} \hat{u}_{i+1}(x, t > 0)\,. \tag{14}$$

Figure 2 shows a sketch of the approximate solution in the particular case with $\widetilde{\lambda}_{i+\frac{1}{2}} > 0$, where $u_i^- = u_i$, as no signal propagates upstream $x = 0$. In this case, by simply setting $x_R = \widetilde{\lambda}_{i+\frac{1}{2}} \Delta t$ and $x_L = 0$ we have that

$$u_{i+1}^+ = \frac{\int_0^{x_R} \hat{u}(x, \Delta t)\, dx}{\Delta t \widetilde{\lambda}_{i+\frac{1}{2}}} = u_i + \left(\frac{\bar{s}}{\widetilde{\lambda}} \right)_{i+\frac{1}{2}}\,. \tag{15}$$

On the other hand, the solution for u_{i+1}^+ in (15) can also be derived by defining a RH (Rankine-Hugoniot) relation at $x = 0$. The presence of the source term leads to a discontinuous intercell approximate flux function, f_i^- and f_{i+1}^+, at the left and right side of the (x,t) plane solution,

$$f_i^- = \lim_{x \to 0^-} \hat{f}_i(x, t > 0)\,, \qquad f_{i+1}^+ = \lim_{x \to 0^+} \hat{f}_{i+1}(x, t > 0)\,, \tag{16}$$

with $\hat{f}(x,t)$ the approximate flux function linked to approximate solution $\hat{u}(x,t)$, with $\hat{f} = \hat{f}(\hat{u})$. When $\widetilde{\lambda}_{i+\frac{1}{2}} > 0$, the following RH condition across this right moving wave is satisfied

$$f_{i+1} - f_{i+1}^+ = \widetilde{\lambda}_{i+\frac{1}{2}} (u_{i+1} - u_{i+1}^+)\,, \tag{17}$$

while equation (13) can be rewritten as

$$f_{i+1} - f_i = \widetilde{\lambda}_{i+\frac{1}{2}} (u_{i+1} - u_i)\,. \tag{18}$$

Subtracting (17) to (18) the following expression is obtained

$$f^+_{i+1} - f_i = \widetilde{\lambda}_{i+\frac{1}{2}}(u^+_{i+1} - u_i)\,, \tag{19}$$

equivalent to:

$$f^+_{i+1} - f^-_i = \widetilde{\lambda}_{i+\frac{1}{2}}(u^+_{i+1} - u^-_i)\,, \tag{20}$$

The following RH relation for the steady contact wave [41, 52], of speed $S = 0$, at $x = 0$ is defined

$$f^+_{i+1} - f^-_i - \bar{s}_{i+\frac{1}{2}} = S(u^+_{i+1} - u^-_i) = 0 \tag{21}$$

and using result in (20), condition (21) becomes

$$u^+_{i+1} = u^-_i + \begin{pmatrix} \bar{s} \\ \widetilde{\lambda} \end{pmatrix}_{i+\frac{1}{2}}\,, \tag{22}$$

recovering expression for (15). This means that, if the approximate wave speed $\widetilde{\lambda}$ in (13) can be defined, original wave speed λ in (10) can be appropriately represented.

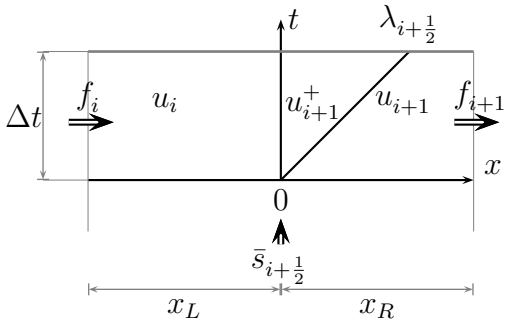

Figure 2: Integration control volume defined by a time interval $[0, \Delta t]$ and a space interval $[-x_L, x_R]$. The solution includes an inner constant state separated by a stationary contact wave at $x = 0$.

Also, when $\widetilde{\lambda}_{i+\frac{1}{2}} > 0$, the following RH condition across this right moving wave is satisfied

$$f_{i+1} - f^+_{i+1} = \widetilde{\lambda}_{i+\frac{1}{2}}(u_{i+1} - u^+_{i+1})\,. \tag{23}$$

If definition of u^+_{i+1} in (22) is inserted in (23) we obtain

$$f^+_{i+1} = f_i + \bar{s}_{i+\frac{1}{2}} \tag{24}$$

and the associated approximate intercell flux solution in the (x, t) plane is now completely defined. Figure 3 shows a sketch of the approximate solution in the particular case $\widetilde{\lambda}_{i+\frac{1}{2}} > 0$, $u_{i+1} > u_i$ and $\bar{s}_{i+\frac{1}{2}} < 0$.

The weak solution of the RP in (3) in case that $\widetilde{\lambda}_{i+\frac{1}{2}} > 0$ is given by

$$\hat{u}(x,t) = \begin{cases} u_i & if \quad x < 0 \\ u_{i+1} - (\theta\delta u)_{i+\frac{1}{2}} = u_i + \left(\begin{smallmatrix} \bar{s} \\ \widetilde{\lambda} \end{smallmatrix}\right)_{i+\frac{1}{2}} & if \quad 0 < x < \widetilde{\lambda}_{i+\frac{1}{2}}\, t \\ u_{i+1} & if \quad x > \widetilde{\lambda}_{i+\frac{1}{2}}\, t \end{cases} \tag{25}$$

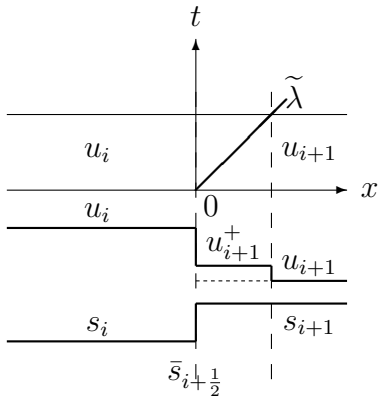

Figure 3: Approximate solution for $\hat{u}(x,t)$.

with

$$\theta_{i+\frac{1}{2}} = 1 - \left(\frac{\bar{s}}{\delta f}\right)_{i+\frac{1}{2}} . \tag{26}$$

The same procedure can be easily extended in case that $\widetilde{\lambda}_{i+\frac{1}{2}} < 0$, and the resulting solution is:

$$\hat{u}(x,t) = \begin{cases} u_i & if \quad x < \widetilde{\lambda}_{i+\frac{1}{2}}\, t \\ u_i + (\theta\delta u)_{i+\frac{1}{2}} = u_{i+1} - \left(\frac{\bar{s}}{\widetilde{\lambda}}\right)_{i+\frac{1}{2}} & if \quad \widetilde{\lambda}_{i+\frac{1}{2}}\, t < x < 0 \\ u_{i+1} & if \quad x > 0 \end{cases} \tag{27}$$

Using previous definitions, the values of the approximate solutions for the RP, u_i^- and u_{i+1}^+ are given by

$$\begin{aligned} u_i^- &= \begin{cases} u_i & if \quad \widetilde{\lambda}_{i+\frac{1}{2}} > 0 \\ u_i + (\theta\delta u)_{i+\frac{1}{2}} & if \quad \widetilde{\lambda}_{i+\frac{1}{2}} < 0 \end{cases} \\ u_{i+1}^+ &= \begin{cases} u_{i+1} - (\theta\delta u)_{i+\frac{1}{2}} & if \quad \widetilde{\lambda}_{i+\frac{1}{2}} > 0 \\ u_{i+1} & if \quad \widetilde{\lambda}_{i+\frac{1}{2}} < 0 \end{cases} \end{aligned} \tag{28}$$

and the values of the approximate fluxes for the RP, f_i^- and f_{i+1}^+ are given by

$$f_i^- = f_i + (\widetilde{\lambda}^- \theta\delta u)_{i+\frac{1}{2}} , \qquad f_{i+1}^+ = f_{i+1} - (\widetilde{\lambda}^+ \theta\delta u)_{i+\frac{1}{2}} , \tag{29}$$

with

$$\widetilde{\lambda}^{\pm} = \frac{1}{2}\left(\widetilde{\lambda} \pm |\widetilde{\lambda}|\right) . \tag{30}$$

Therefore, the corresponding intercell flux for the approximate first order Godunov method in (4) is given by two functions,

$$f_{i+\frac{1}{2}}^- = f_i^- , \qquad f_{i-\frac{1}{2}}^+ = f_i^+ , \tag{31}$$

resulting in

$$u_i^{n+1} = u_i^n - \frac{\Delta t}{\Delta x}[f_i^- - f_i^+] \tag{32}$$

By inserting definitions in (29), Godunov's method can be expressed in fluctuation form [30],

$$u_i^{n+1} = u_i^n - \frac{\Delta t}{\Delta x}[(\delta m)^-_{i+\frac{1}{2}} + (\delta m)^+_{i-\frac{1}{2}}]\,, \tag{33}$$

with $(\delta m)^-_{i+\frac{1}{2}} = (\widetilde{\lambda}^- \theta \delta u)_{i+\frac{1}{2}}$ and $(\delta m)^+_{i-\frac{1}{2}} = (\widetilde{\lambda}^+ \theta \delta u)_{i-\frac{1}{2}}$. In steady state problems, numerical updating of the initial data results in solutions governed by

$$(\delta m)^\pm_{i\mp\frac{1}{2}} = 0\,, \tag{34}$$

ensuring a path consistent scheme [30], where fluxes and source terms are equilibrated at each RP. Depending on the approximations done in the integration of the source term, $\bar{s}$, the numerical solution will converge to different solutions with mesh refinement. The selection of appropriate estimates of $\bar{s}$ allows to converge to the exact solution [32] and must be based on convergence analysis to exact solutions. When possible, the discretization should be based on discrete equilibrium between the numerical flux and the numerical source term.

3. 1D Systems of conservation laws with source terms

The discussion is next extended to hyperbolic nonlinear systems of equations with source terms in 1D, composed of N_λ hyperbolic conservation laws and expressed as

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial \mathbf{F}(\mathbf{U})}{\partial x} = \mathbf{S}(\mathbf{U})\,. \tag{35}$$

From this formulation it is possible to define a Jacobian matrix for the convective part $\mathbf{J}$

$$\mathbf{J} = \frac{d\mathbf{F}}{d\mathbf{U}}\,. \tag{36}$$

Assuming that the convective part in (35) is strictly hyperbolic with N_λ distinct real eigenvalues λ and eigenvectors $\mathbf{e}$, it is possible to define two matrices $\mathbf{P} = (\mathbf{e}^1, ..., \mathbf{e}^{N_\lambda})$ and $\mathbf{P}^{-1}$ with the property that they diagonalize the Jacobian $\mathbf{J}$ as

$$\mathbf{J} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^{-1}\,. \tag{37}$$

The system of equations in (35) will be solved using approximate linear solutions of an initial value problem, by means of an explicit conservative formula, that according to the Godunov first order method is written as

$$\mathbf{U}_i^{n+1} = \mathbf{U}_i^n - (\mathbf{F}^-_{i+\frac{1}{2}} - \mathbf{F}^+_{i-\frac{1}{2}})\frac{\Delta t}{\Delta x}\,, \tag{38}$$

where $\mathbf{F}^\pm_{i+\frac{1}{2}}$ is the intercell flux at each RP and $\mathbf{U}_i^n$ is a piecewise constant approximation of the solution at time t^n that will be updated using approximate solutions of local RPs with the following initial conditions:

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial \mathbf{F}}{\partial x} = \mathbf{S}$$
$$\mathbf{U}(x,0) = \begin{cases} \mathbf{U}_i & if \quad x < 0 \\ \mathbf{U}_{i+1} & if \quad x > 0 \end{cases} \tag{39}$$

3.1. Integral Relations in the Riemann Solution

Even when ignoring the exact solution of the RP $\mathbf{U}(x,t)$ in (39), it is possible to estimate its variation by integrating (39) over a suitable control volume. Figure 4 shows a RP, with initial values $\mathbf{U}_L, \mathbf{U}_R$, and a control volume given by the time interval $[0, \Delta t]$ and the space interval $[-x_L, x_R]$, where

$$-x_L \leq \lambda_L \Delta t, \quad x_R \geq \lambda_R \Delta t\,, \tag{40}$$

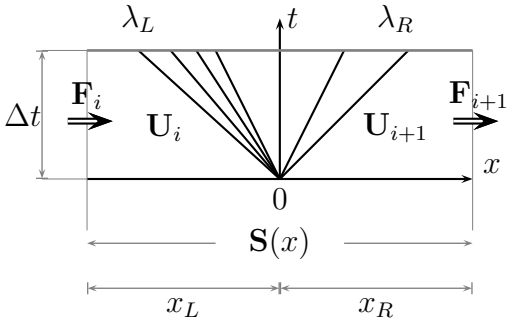

Figure 4: Integration control volume defined by a time interval $[0, \Delta t]$ and a space interval $[-x_L, x_R]$

being λ_L, λ_R the minimum and maximum wave velocities respectively in the domain given by the eigenvalues of (36) at $t = \Delta t$. Integrating (39) over the control volume $[0, \Delta t] \times [-x_L, x_R]$,

$$\int_{-x_L}^{x_R} \int_0^{\Delta t} \left(\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial \mathbf{F}}{\partial x} - \mathbf{S} \right) \, dt dx = 0\,, \tag{41}$$

the following expression for the integral volume of $\mathbf{U}(x, \Delta t)$ is obtained

$$\int_{-x_L}^{x_R} \mathbf{U}(x, \Delta t) \, dx = x_R \mathbf{U}_{i+1} + x_L \mathbf{U}_i - (\delta \mathbf{F} - \bar{\mathbf{S}})_{i+\frac{1}{2}} \Delta t\,, \tag{42}$$

with $\mathbf{F}_{i+1} = \mathbf{F}(\mathbf{U}_{i+1})$, $\mathbf{F}_i = \mathbf{F}(\mathbf{U}_i)$ and the source term included at the discontinuity point $x = 0$, assuming the following time linearization

$$\int_{-x_L}^{x_R} \int_0^{\Delta t} \mathbf{S} \, dt dx \approx \Delta t \int_{-x_L}^{x_R} \mathbf{S}(\mathbf{U}_{i+1}, \mathbf{U}_i, t = 0) \, dx = \Delta t \bar{\mathbf{S}}_{i+\frac{1}{2}}\,, \tag{43}$$

consistent with the definition of a similarity solution only dependent on the ratio (x/t).

3.2. Approximate solution

The RP in (39) is now approximated by using the following constant coefficient linear RP [48]

$$\begin{gathered}\frac{\partial \hat{\mathbf{U}}}{\partial t}+\widetilde{\mathbf{J}}_{i+\frac{1}{2}}\frac{\partial \hat{\mathbf{U}}}{\partial x}=\mathbf{S}\\ \hat{\mathbf{U}}(x,0)=\left\{\begin{array}{ccc}\mathbf{U}_i & if & x<0\\ \mathbf{U}_{i+1} & if & x>0\end{array}\right.\end{gathered}\tag{44}$$

where $\widetilde{\mathbf{J}}_{i+\frac{1}{2}}=\widetilde{\mathbf{J}}_{i+\frac{1}{2}}(\mathbf{U}_i,\mathbf{U}_{i+1})$ is a constant matrix yet to be defined. Integrating (44) over the same control volume,

$$\int_{-x_L}^{x_R}\hat{\mathbf{U}}(x,\Delta t)\,dx=x_R\mathbf{U}_{i+1}+x_L\mathbf{U}_L-\left(\widetilde{\mathbf{J}}\delta\mathbf{U}-\bar{\mathbf{S}}\right)_{i+\frac{1}{2}}\Delta t\,,\tag{45}$$

and performing the same approaches over the source term, the following constraint involving conservation across discontinuities appears to ensure (42)

$$\delta\mathbf{F}_{i+\frac{1}{2}}=\widetilde{\mathbf{J}}_{i+\frac{1}{2}}\delta\mathbf{U}_{i+\frac{1}{2}}\,,\tag{46}$$

where $\widetilde{\mathbf{J}}_{i+\frac{1}{2}}$ is defined using the approximate Roe's solver [11], being diagonalizable with N_λ approximate real eigenvalues

$$\widetilde{\lambda}^1_{i+\frac{1}{2}}<\ldots<\widetilde{\lambda}^I_{i+\frac{1}{2}}<0<\widetilde{\lambda}^{I+1}_{i+\frac{1}{2}}<...<\widetilde{\lambda}^{N_\lambda}_{i+\frac{1}{2}}\tag{47}$$

and N_λ eigenvectors $\widetilde{\mathbf{e}}^1,...,\widetilde{\mathbf{e}}^{N_\lambda}$. With them, two approximate matrices, $\widetilde{\mathbf{P}}_{i+\frac{1}{2}}=(\widetilde{\mathbf{e}}^1,...,\widetilde{\mathbf{e}}^{N_\lambda})_{i+\frac{1}{2}}$ and $\widetilde{\mathbf{P}}^{-1}_{i+\frac{1}{2}}$ are built with the following property

$$\widetilde{\mathbf{J}}_{i+\frac{1}{2}}=(\widetilde{\mathbf{P}}\widetilde{\mathbf{\Lambda}}\widetilde{\mathbf{P}}^{-1})_{i+\frac{1}{2}},\qquad \widetilde{\mathbf{\Lambda}}_{i+\frac{1}{2}}=\left(\begin{array}{ccc}\widetilde{\lambda}^1 & & 0\\ & \ddots & \\ 0 & & \widetilde{\lambda}^{N_\lambda}\end{array}\right)_{i+\frac{1}{2}},\tag{48}$$

where $\widetilde{\mathbf{\Lambda}}_{i+\frac{1}{2}}$ is a diagonal matrix with approximate eigenvalues in the main diagonal. One result of Roe's linearization is that the resulting approximate Riemann solution consists of only discontinuities and $\hat{\mathbf{U}}(x,t)$ is constructed as a sum of jumps or shocks. The solutions for $\hat{\mathbf{U}}(x,t)$ are governed by the celerities in $\widetilde{\mathbf{\Lambda}}_{i+\frac{1}{2}}$ and each one consists of N_λ regions connected by $N_\lambda+1$ waves, one of them steady, with celerity $S=x/t=0$.

According to the Godunov method, it is sufficient to provide the solution for $\hat{\mathbf{U}}(x,t)$ at the intercell position $x=0$ in order to derive the updating numerical fluxes in (38). In order to recover the value of the approximate intermediate states $\mathbf{U}_i^-$ and $\mathbf{U}_{i+1}^+$ at the left and right side of the (x,t) plane solution respectively,

$$\mathbf{U}_i^-=\lim_{x\to 0^-}\hat{\mathbf{U}}_i(x,t>0)\,,\qquad \mathbf{U}_{i+1}^+=\lim_{x\to 0^+}\hat{\mathbf{U}}_{i+1}(x,t>0)\,,\tag{49}$$

system in (44) is transformed by using $\widetilde{\mathbf{P}}^{-1}$ matrix as follows

$$\widetilde{\mathbf{P}}^{-1}_{i+\frac{1}{2}}\left(\frac{\partial \hat{\mathbf{U}}}{\partial t}+\widetilde{\mathbf{J}}_{i+\frac{1}{2}}\frac{\partial \hat{\mathbf{U}}}{\partial x}\right)=\widetilde{\mathbf{P}}^{-1}_{i+\frac{1}{2}}\mathbf{S}\,,\tag{50}$$

expressing (44) in terms of the characteristic variables $\mathbf{V} = \widetilde{\mathbf{P}}^{-1}_{i+\frac{1}{2}}\hat{\mathbf{U}}$, with $\mathbf{V} = (V^1, ..., V^{N_\lambda})^T$. This transformation leads to a decoupled system [48] that generates the following linear RP

$$\begin{gathered}\frac{\partial \mathbf{V}}{\partial t} + \widetilde{\mathbf{\Lambda}}_{i+\frac{1}{2}}\frac{\partial \mathbf{V}}{\partial x} = \mathbf{B}_{i+\frac{1}{2}} \\ \mathbf{V}(x,0) = \begin{cases} \mathbf{V}_i = \widetilde{\mathbf{P}}^{-1}_{i+\frac{1}{2}}\mathbf{U}_i & if \quad x < 0 \\ \mathbf{V}_{i+1} = \widetilde{\mathbf{P}}^{-1}_{i+\frac{1}{2}}\mathbf{U}_{i+1} & if \quad x > 0 \end{cases}\end{gathered} \tag{51}$$

with $\mathbf{B}_{i+\frac{1}{2}} = \widetilde{\mathbf{P}}^{-1}_{i+\frac{1}{2}}\mathbf{S} = (\beta^1, ..., \beta^{N_\lambda})^T_{i+\frac{1}{2}}$, where each equation

$$\frac{\partial V^m}{\partial t} + \widetilde{\lambda}^m_{i+\frac{1}{2}}\frac{\partial V^m}{\partial x} = \beta^m_{i+\frac{1}{2}}, \qquad m = 1, ..., N_\lambda \tag{52}$$

involves the variable V^m and the source term $\beta^m_{i+\frac{1}{2}}$.

At this stage matrix $\mathbf{B}_{i+1/2}$ does not require the bar symbol, as has not yet been integrated, but has been locally generated by projecting the initial source term $\mathbf{S}$ on the approximate Jacobian eigenvectors basis in $\mathbf{P}_{i+1/2}$. Equation in (52) allows to generate a set of independent equations that can be solved exactly for each characteristic variable V^m. The solution for each V^m characteristic variable is given by the solution of the scalar case in (28) and consists of three regions. Figure 5 shows the solution for a characteristic variable V^m with a wave of speed $\widetilde{\lambda}^m_{i+\frac{1}{2}} > 0$. The solution can be expressed as a piecewise constant function depending on x and t as

$$V^m(x,t) = \begin{cases} V^m_i & if \quad x < \widetilde{\lambda}^m_{i+\frac{1}{2}}t \\ V^m_i + (\theta\delta V)^m_{i+\frac{1}{2}} & if \quad \widetilde{\lambda}^m_{i+\frac{1}{2}}t < x < 0 \\ V^m_{i+1} & if \quad 0 < x \end{cases} \tag{53}$$

when $\widetilde{\lambda}^m_{i+\frac{1}{2}} < 0$, and

$$V^m(x,t) = \begin{cases} V^m_i & if \quad x < 0 \\ V^m_{i+1} - (\theta\delta V)^m_{i+\frac{1}{2}} & if \quad 0 < x < \widetilde{\lambda}^m_{i+\frac{1}{2}}t \\ V^m_{i+1} & if \quad \widetilde{\lambda}^m_{i+\frac{1}{2}}t < x \end{cases} \tag{54}$$

when $\widetilde{\lambda}^m_{i+\frac{1}{2}} > 0$, where

$$\theta^m_{i+\frac{1}{2}} = 1 - \left(\frac{\bar{\beta}^m}{\widetilde{\lambda}^m\alpha^m}\right)_{i+\frac{1}{2}}, \tag{55}$$

with

$$\mathbf{A}_{i+\frac{1}{2}} = (\alpha^1, ..., \alpha^{N_\lambda})^T_{i+\frac{1}{2}} = \delta\mathbf{V}_{i+\frac{1}{2}} = \widetilde{\mathbf{P}}^{-1}_{i+\frac{1}{2}}\delta\mathbf{U}_{i+\frac{1}{2}} \tag{56}$$

the set of wave strengths and

$$\bar{\mathbf{B}}_{i+\frac{1}{2}} = (\bar{\beta}^1, ..., \bar{\beta}^{N_\lambda})^T_{i+\frac{1}{2}} = (\widetilde{\mathbf{P}}^{-1}\bar{\mathbf{S}})_{i+\frac{1}{2}} \tag{57}$$

the set of source strengths. It is worth mentioning that α^m wave strengths allow to express simple linear relations for both conserved variables and flux vector differences as follow

$$\delta \mathbf{U}_{i+\frac{1}{2}} = \sum (\alpha \widetilde{\mathbf{e}})^m_{i+\frac{1}{2}} \ , \qquad \delta \mathbf{F}_{i+\frac{1}{2}} = \sum \left(\widetilde{\lambda} \alpha \widetilde{\mathbf{e}} \right)^m_{i+\frac{1}{2}} \ . \tag{58}$$

The value of V^m at the left and right side of the intercell position is given by

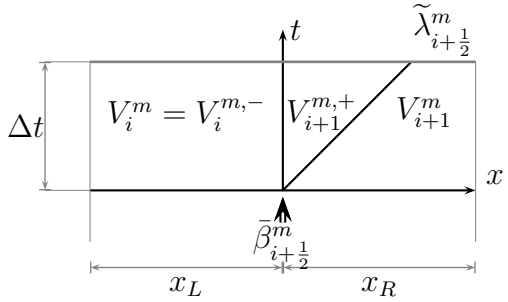

Figure 5: Integration control volume defined by a time interval $[0, \Delta t]$ and a space interval $[-x_L, x_R]$. The solution for V^m involves an inner constant state separated by a stationary contact wave at $x = 0$ and by a wave of speed $\widetilde{\lambda}^m_{i+\frac{1}{2}}$.

$$\begin{aligned} V^{m,-}_i &= \begin{cases} V^m_i & if \quad \widetilde{\lambda}^m_{i+\frac{1}{2}} > 0 \\ V^m_i + (\theta \delta V)^m_{i+\frac{1}{2}} & if \quad \widetilde{\lambda}^m_{i+\frac{1}{2}} < 0 \end{cases} \\ V^{m,+}_{i+1} &= \begin{cases} V^m_{i+1} - (\theta \delta V)^m_{i+\frac{1}{2}} & if \quad \widetilde{\lambda}^m_{i+\frac{1}{2}} > 0 \\ V^m_{i+1} & if \quad \widetilde{\lambda}^m_{i+\frac{1}{2}} < 0 \end{cases} \end{aligned} \tag{59}$$

and can be expressed as

$$\begin{aligned} V^{m,-}_i &= V^m_i + \left(\frac{\widetilde{\lambda}^-}{\widetilde{\lambda}} \right)^m_{i+\frac{1}{2}} (\theta \delta V)^m_{i+\frac{1}{2}} \, , \\ V^{m,+}_{i+1} &= V^m_{i+1} - \left(\frac{\widetilde{\lambda}^+}{\widetilde{\lambda}} \right)^m_{i+\frac{1}{2}} (\theta \delta V)^m_{i+\frac{1}{2}} \, , \end{aligned} \tag{60}$$

or in matrix form

$$\begin{aligned} \mathbf{V}^-_i &= \mathbf{V}_i + \left(\widetilde{\mathbf{\Lambda}}^{-1} \widetilde{\mathbf{\Lambda}}^- \mathbf{\Theta} \delta \mathbf{V} \right)_{i+\frac{1}{2}} \, , \\ \mathbf{V}^+_{i+1} &= \mathbf{V}_{i+1} - \left(\widetilde{\mathbf{\Lambda}}^{-1} \widetilde{\mathbf{\Lambda}}^+ \mathbf{\Theta} \delta \mathbf{V} \right)_{i+\frac{1}{2}} \, , \end{aligned} \tag{61}$$

where $\mathbf{\Theta}$ is a diagonal matrix with θ's in the main diagonal

$$\mathbf{\Theta}_{i+\frac{1}{2}} = \begin{pmatrix} \theta^1 & & 0 \\ & \ddots & \\ 0 & & \theta^{N_\lambda} \end{pmatrix}_{i+\frac{1}{2}} \, . \tag{62}$$

Now, the intermediate states $\mathbf{U}^-_i$ and $\mathbf{U}^+_{i+1}$ can be directly obtained by using $\widetilde{\mathbf{P}}$ matrix. Vector solutions $\mathbf{U}^-_i = \widetilde{\mathbf{P}} \mathbf{V}^-_i$ and $\mathbf{U}^+_{i+1} = \widetilde{\mathbf{P}} \mathbf{V}^+_{i+1}$ are recovered from (61) as follows

$$\begin{aligned}\mathbf{U}_i^- &= \mathbf{U}_i + (\widetilde{\mathbf{P}}\widetilde{\mathbf{\Lambda}}^{-1}\widetilde{\mathbf{\Lambda}}^-\mathbf{\Theta}\widetilde{\mathbf{P}}^{-1})_{i+\frac{1}{2}}\delta\mathbf{U}_{i+\frac{1}{2}} = \mathbf{U}_i + \sum_{\widetilde{\lambda}^m<0}(\alpha\theta\widetilde{\mathbf{e}})^m_{i+\frac{1}{2}}\,,\\ \mathbf{U}_{i+1}^+ &= \mathbf{U}_{i+1} - (\widetilde{\mathbf{P}}\widetilde{\mathbf{\Lambda}}^{-1}\widetilde{\mathbf{\Lambda}}^+\mathbf{\Theta}\widetilde{\mathbf{P}}^{-1})_{i+\frac{1}{2}}\delta\mathbf{U}_{i+\frac{1}{2}} = \mathbf{U}_{i+1} - \sum_{\widetilde{\lambda}^m>0}(\alpha\theta\widetilde{\mathbf{e}})^m_{i+\frac{1}{2}}\,,\end{aligned} \tag{63}$$

with the following property

$$\begin{aligned}\mathbf{U}_{i+1}^+ - \mathbf{U}_i^- &= \mathbf{U}_{i+1} - \mathbf{U}_i - \sum_{m=1}^{N_\lambda}(\alpha\theta\widetilde{\mathbf{e}})^m_{i+\frac{1}{2}}\\ &= \mathbf{U}_{i+1} - \mathbf{U}_i - \sum_{m=1}^{N_\lambda}\alpha\left(\widetilde{\mathbf{e}} - \frac{\bar{\beta}}{\alpha\widetilde{\lambda}}\widetilde{\mathbf{e}}\right)^m_{i+\frac{1}{2}}\\ &= \sum_{m=1}^{N_\lambda}\left(\frac{\bar{\beta}}{\widetilde{\lambda}}\widetilde{\mathbf{e}}\right)^m_{i+\frac{1}{2}} = (\widetilde{\mathbf{P}}\widetilde{\mathbf{\Lambda}}^{-1}\bar{\mathbf{B}})_{i+\frac{1}{2}}.\end{aligned} \tag{64}$$

Relation in (63) provides the solution in the vicinity of $x = 0$. As mentioned above, traditionally, when applying Godunov's method on system it is only necessary to compute the value of the solution state along $x/t = 0$. But when focusing on the impact of the source terms in the updating step, it is mandatory to define the full wave structure of the solution. Next, relation in (64) is recovered based on the wave separation.

The definition of N_λ celerities in $\widetilde{\mathbf{\Lambda}}_{i+\frac{1}{2}}$ plus the existence of a stationary contact wave, with celerity $S = 0$ at $x = 0$, gives as a result an approximate solution that involves the two initial conditions and N_λ inner states. Figure 6 illustrates the linear approximate solution. The inner constant states on the left side of the (x, t) plane state will be named $\mathbf{U}_i^{m,-}$, where $1 \le m \le I$, with $\mathbf{U}_i^{I,-} = \mathbf{U}_i^-$ and $\mathbf{U}_i^{0,-} = \mathbf{U}_i$. On the right side of the (x, t) plane solution, inner constant states are labeled as $\mathbf{U}_{i+1}^{m,+}$, where $I + 1 \le m \le N_\lambda$, with $\mathbf{U}_i^{I+1,+} = \mathbf{U}_{i+1}^+$ and $\mathbf{U}_{i+1}^{N_\lambda+1,+} = \mathbf{U}_{i+1}$.

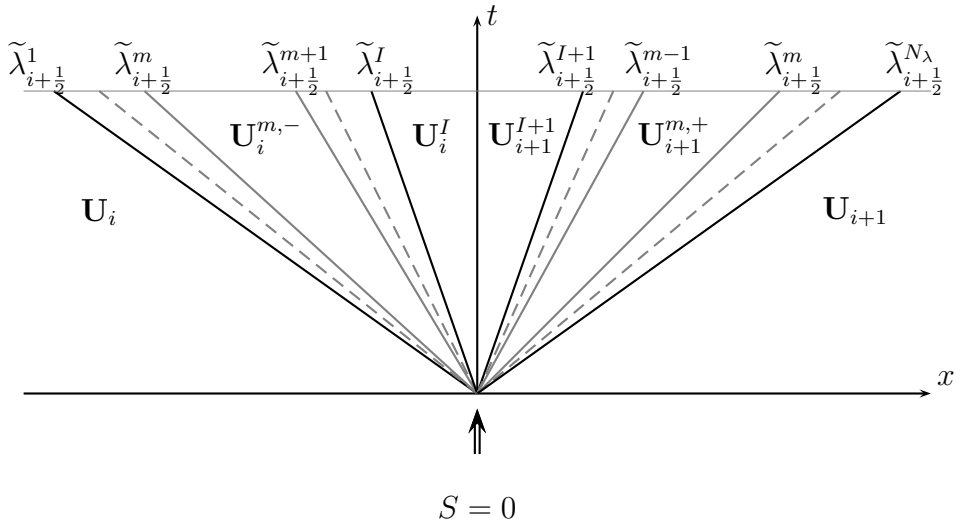

Figure 6: Approximate solution $\hat{\mathbf{U}}(x, t)$. The solution consist of N_λ inner constant states separated by a stationary contact wave, with celerity $S = 0$ at $x = 0$.

The derivation of the general solution $\hat{\mathbf{U}}(x,t)$ for a linear system is based on the expansion of the solution as a linear combination of the vectors that compose the Jacobian's eigenvectors basis [48], using the relation $\mathbf{U} = \widetilde{\mathbf{P}}\mathbf{V}$, as follows

$$\hat{\mathbf{U}}(x,t) = \sum_{m_1=1}^{N_\lambda} V^{m_1}(x,t)\, \widetilde{\mathbf{e}}^{m_1}_{i+\frac{1}{2}}\,, \tag{65}$$

where the scalar values $V^{m_1}(x,t)$ are the solution of the characteristic variables at the sought point and represent the strength of each wave. Recall that the solution for a certain V^{m_1} in RP (51) is computed as a scalar RP only involving V^{m_1}.

If focusing on a constant state on the left hand side of the t-axis, $\mathbf{U}^{m,-}$, defined between characteristic lines $\widetilde{\lambda}^m t$ and $\widetilde{\lambda}^{m+1} t$, the solution is given by the combination of the characteristic solutions in the spatial domain $[\widetilde{\lambda}^m t, \widetilde{\lambda}^{m+1} t]$. Following expansion in (65), $\mathbf{U}^{m,-}$ is given by

$$\mathbf{U}_i^{m,-} = \sum_{\widetilde{\lambda}^{m_1} \le \widetilde{\lambda}^m} V^{m_1}(x,t)\widetilde{\mathbf{e}}^{m_1}_{i+\frac{1}{2}} + \sum_{\widetilde{\lambda}^{m_1} \ge \widetilde{\lambda}^{m+1}} V^{m_1}(x,t)\widetilde{\mathbf{e}}^{m_1}_{i+\frac{1}{2}}\,. \tag{66}$$

The solutions for the characteristic variables are given by (53) and (54). The first term of the right hand side of equation (66) is then equal to

$$\sum_{\widetilde{\lambda}^{m_1} \le \widetilde{\lambda}^m} V^{m_1}(x,t)\widetilde{\mathbf{e}}^{m_1}_{i+\frac{1}{2}} = \sum_{m_1=1}^{m} \left(V_i^{m_1} + (\theta\delta V)^{m_1}_{i+\frac{1}{2}} \right) \widetilde{\mathbf{e}}^{m_1}_{i+\frac{1}{2}} \tag{67}$$

and the second term becomes

$$\sum_{\widetilde{\lambda}^{m_1} \ge \widetilde{\lambda}^{m+1}} V^{m_1}(x,t)\widetilde{\mathbf{e}}^{m_1}_{i+\frac{1}{2}} = \sum_{m_1=m+1}^{I} V_i^{m_1}\widetilde{\mathbf{e}}^{m_1}_{i+\frac{1}{2}} + \sum_{m_1=I+1}^{N_\lambda} V_i^{m_1}\widetilde{\mathbf{e}}^{m_1}_{i+\frac{1}{2}}\,. \tag{68}$$

Primitive vector solution in (66) can be expressed as

$$\mathbf{U}_i^{m,-} = \sum_{m_1=1}^{N_\lambda} V_i^{m_1}\widetilde{\mathbf{e}}^{m_1}_{i+\frac{1}{2}} + \sum_{m_1=1}^{m} (\theta\delta V\widetilde{\mathbf{e}})^{m_1}_{i+\frac{1}{2}} \tag{69}$$

and considering that $\mathbf{U}_i = \sum_{m_1=1}^{N_\lambda} V_i^{m_1}\widetilde{\mathbf{e}}^{m_1}_{i+\frac{1}{2}}$ and $\delta V^{m_1}_{i+\frac{1}{2}} = \alpha^{m_1}_{i+\frac{1}{2}}$, equation (69) can be rewritten as

$$\mathbf{U}_i^{m,-} = \mathbf{U}_i + \sum_{m_1=1}^{m} (\theta\alpha\widetilde{\mathbf{e}})^{m_1}_{i+\frac{1}{2}}\,. \tag{70}$$

By separating the $\widetilde{\lambda}^m$-wave contribution from the summation,

$$\mathbf{U}_i^{m,-} = \mathbf{U}_i + \sum_{m_1=1}^{m-1} (\theta\alpha\widetilde{\mathbf{e}})^{m_1}_{i+\frac{1}{2}} + (\theta\alpha\widetilde{\mathbf{e}})^{m}_{i+\frac{1}{2}}\,, \tag{71}$$

it is noticed that $\mathbf{U}_i^{m,-}$ in (70) can be expressed in terms of its left adjacent state, $\mathbf{U}_i^{m-1,-}$, leading to the following jump between vector solutions

$$\mathbf{U}_i^{m,-} - \mathbf{U}_i^{m-1,-} = (\alpha\theta\widetilde{\mathbf{e}})^{m}_{i+\frac{1}{2}}\,, \tag{72}$$

for $1 \leq m \leq I$. Remark that equation (71) can only provide solutions in the spatial domain $[\widetilde{\lambda}^1 t, 0]$.

When seeking the primitive vector solution for a state defined on the right hand side of the t-axis, $\mathbf{U}_{i+1}^{m,+}$, it has to be defined between characteristic lines $\widetilde{\lambda}^{m-1}t$ and $\widetilde{\lambda}^m t$. Following expansion in (65), the combination of the characteristic solutions in the spatial domain $[\widetilde{\lambda}^{m-1}t, \widetilde{\lambda}^m t]$ provides

$$\mathbf{U}_{i+1}^{m,+} = \sum_{\widetilde{\lambda}^{m_1} \leq \widetilde{\lambda}^{m-1}} V^{m_1}(x,t)\widetilde{\mathbf{e}}_{i+\frac{1}{2}}^{m_1} + \sum_{\widetilde{\lambda}^{m_1} \geq \widetilde{\lambda}^{m}} V^{m_1}(x,t)\widetilde{\mathbf{e}}_{i+\frac{1}{2}}^{m_1} \tag{73}$$

and using characteristic solutions in (53) and (54), the first term of the primitive vector solution becomes

$$\sum_{\widetilde{\lambda}^{m_1} \leq \widetilde{\lambda}^{m-1}} V^{m_1}(x,t)\widetilde{\mathbf{e}}_{i+\frac{1}{2}}^{m_1} = \sum_{m_1=1}^{I} V_{i+1}^{m_1}\widetilde{\mathbf{e}}_{i+\frac{1}{2}}^{m_1} + \sum_{m_1=I+1}^{m-1} V_{i+1}^{m_1}\widetilde{\mathbf{e}}_{i+\frac{1}{2}}^{m_1} \tag{74}$$

and the second one

$$\sum_{\widetilde{\lambda}^{m_1} \geq \widetilde{\lambda}^{m}} V^{m_1}(x,t)\widetilde{\mathbf{e}}_{i+\frac{1}{2}}^{m_1} = \sum_{m_1=m}^{N_\lambda} \left(V_{i+1}^{m_1} - (\theta\delta V)_{i+\frac{1}{2}}^{m_1}\right)\widetilde{\mathbf{e}}_{i+\frac{1}{2}}^{m_1}\,, \tag{75}$$

allowing to express $\mathbf{U}_{i+1}^{m,+}$ as follows

$$\mathbf{U}_{i+1}^{m,+} = \sum_{m_1=1}^{N_\lambda} V_{i+1}^{m_1}\widetilde{\mathbf{e}}_{i+\frac{1}{2}}^{m_1} - \sum_{m_1=m}^{N_\lambda} (\theta\delta V\widetilde{\mathbf{e}})_{i+\frac{1}{2}}^{m_1}\,. \tag{76}$$

As done for (69), considering that $\mathbf{U}_{i+1} = \sum_{m_1=1}^{N_\lambda} V_{i+1}^{m_1}\widetilde{\mathbf{e}}_{i+\frac{1}{2}}^{m_1}$, equation (76) can be rewritten as

$$\mathbf{U}_{i+1}^{m,+} = \mathbf{U}_{i+1} - \sum_{m_1=m}^{N_\lambda} (\theta\alpha\widetilde{\mathbf{e}})_{i+\frac{1}{2}}^{m_1} \tag{77}$$

and by separating the $\widetilde{\lambda}^m$-wave contribution from the summation as

$$\mathbf{U}_{i+1}^{m,+} = \mathbf{U}_{i+1} - \sum_{m_1=m+1}^{N_\lambda} (\theta\alpha\widetilde{\mathbf{e}})_{i+\frac{1}{2}}^{m_1} - (\theta\alpha\widetilde{\mathbf{e}})_{i+\frac{1}{2}}^{m}\,, \tag{78}$$

it can be expressed in terms of its right adjacent state, $\mathbf{U}_{i+1}^{m+1,+}$, as follows

$$\mathbf{U}_{i+1}^{m+1,+} - \mathbf{U}_{i+1}^{m,+} = (\theta\alpha\widetilde{\mathbf{e}})_{i+\frac{1}{2}}^{m}\,, \tag{79}$$

leading to a jump between vector solutions. Now, equation (79) provides exclusively solutions in the spatial domain $[0, \widetilde{\lambda}^{N_\lambda} t]$.

In the vicinity of $x = 0$, left and right states denoted by $\mathbf{U}_i^-$ and $\mathbf{U}_{i+1}^+$ are defined inside spatial domains $[\widetilde{\lambda}^I t, 0]$ and $[0, \widetilde{\lambda}^{I+1} t]$ respectively. Expression in (63) can be derived from the previous results, setting $m = I$ in (70) and $m = I + 1$ in (77) for the left and right states respectively, leading to

$$\begin{aligned}\mathbf{U}_i^- &= \mathbf{U}_i + \sum_{m=1}^{I} (\theta\alpha\widetilde{\mathbf{e}})_{i+\frac{1}{2}}^m \,, \\ \mathbf{U}_{i+1}^+ &= \mathbf{U}_{i+1} - \sum_{m=I+1}^{N_\lambda} (\theta\alpha\widetilde{\mathbf{e}})_{i+\frac{1}{2}}^m \,,\end{aligned} \tag{80}$$

recovering solutions in (63).

3.3. Approximate numerical flux

Being the solution defined as a sum of jumps or shocks between the different intermediate states, the solution for the approximate flux function $\hat{\mathbf{F}}(x,t)$ involves the two initial unaltered fluxes, $\mathbf{F}_i$ and $\mathbf{F}_{i+1}$, and N_λ inner states. The structure follows the pattern illustrated in Figure 6, and now, each intermediate constant state involves an intermediate constant flux function. Approximate flux function $\hat{\mathbf{F}}(x,t)$ provides the intercell fluxes at the left and right side of the initial discontinuity at $x = 0$, labeled as $\mathbf{F}_i^-$ and $\mathbf{F}_{i+1}^+$ respectively in a general RP, with

$$\mathbf{F}_i^- = \lim_{x\to 0^-} \hat{\mathbf{F}}(x, t > 0) \,, \qquad \mathbf{F}_{i+1}^+ = \lim_{x\to 0^+} \hat{\mathbf{F}}(x, t > 0) \,. \tag{81}$$

As pointed out by [48], the definition of linear system in (44) would suggest that intercell fluxes in a general RP are given by $\mathbf{F}_i^- = \widetilde{\mathbf{J}}_{i+\frac{1}{2}} \mathbf{U}_i^-$ and $\mathbf{F}_{i+1}^+ = \widetilde{\mathbf{J}}_{i+\frac{1}{2}} \mathbf{U}_{i+1}^+$, which is incorrect.

The relation between the intercell approximate fluxes $\mathbf{F}_i^-$ and $\mathbf{F}_{i+1}^+$, in a general RP, can be analyzed using the RH (Rankine-Hugoniot) relation at $x = 0$, that includes a steady contact wave [41, 52] between approximate solutions $\mathbf{U}_i^-$ and $\mathbf{U}_{i+1}^+$

$$\mathbf{F}_{i+1}^+ - \mathbf{F}_i^- - \bar{\mathbf{S}}_{i+\frac{1}{2}} = S(\mathbf{U}_{i+1}^+ - \mathbf{U}_i^-) = 0 \,. \tag{82}$$

Interestingly, as in the scalar case, we can assume the following relation among fluxes and conserved variables in the inner regions

$$\mathbf{F}_{i+1}^+ - \mathbf{F}_i^- = \widetilde{\mathbf{J}}_{i+\frac{1}{2}} (\mathbf{U}_{i+1}^+ - \mathbf{U}_i^-) \,, \tag{83}$$

where the approximate Jacobian can be substituted by $\widetilde{\mathbf{P}}\widetilde{\mathbf{\Lambda}}^{-1}\widetilde{\mathbf{P}}^{-1}$ according to (48), which allows to express $(\mathbf{U}_{i+1}^+ - \mathbf{U}_i^-)$ as follows

$$\mathbf{U}_{i+1}^+ - \mathbf{U}_i^- = (\widetilde{\mathbf{P}}\widetilde{\mathbf{\Lambda}}^{-1}\widetilde{\mathbf{P}}^{-1}\bar{\mathbf{S}})_{i+\frac{1}{2}} = (\widetilde{\mathbf{P}}\widetilde{\mathbf{\Lambda}}^{-1}\bar{\mathbf{B}})_{i+\frac{1}{2}} \,, \tag{84}$$

recovering (64), and confirming condition (83).

In order to provide a complete description of the approximate flux function $\hat{\mathbf{F}}(x,t)$, the inner constant fluxes on the left side of the (x,t) plane will be named $\mathbf{F}_i^{m,-}$, where $1 \le m \le I$. On the right side of the (x,t) plane solution, inner constant states are labeled as $\mathbf{F}_{i+1}^{m,+}$, where $I+1 \le m \le N_\lambda$.

Following the linear case, the approximate solution for the fluxes can be constructed defining appropriate RH condition across each moving wave, that will be given by

$$\mathbf{F}_i^{m,-} - \mathbf{F}_i^{m-1,-} = \widetilde{\lambda}^m (\mathbf{U}_i^m - \mathbf{U}_i^{m-1}) = \left(\widetilde{\lambda}\alpha\theta\widetilde{\mathbf{e}}\right)_{i+\frac{1}{2}}^m \,, \tag{85}$$

for $1 \le m \le I$, where $\mathbf{F}_i^{I,-} = \mathbf{F}_i^-$, $\mathbf{F}_i^{0,-} = \mathbf{F}_i$, and

$$\mathbf{F}_{i+1}^{m,+} - \mathbf{F}_{i+1}^{m+1,+} = \widetilde{\lambda}^m(\mathbf{U}_{i+1}^m - \mathbf{U}_{i+1}^{m+1}) = -\left(\widetilde{\lambda}\alpha\theta\widetilde{\mathbf{e}}\right)_{i+\frac{1}{2}}^m , \tag{86}$$

for $I+1 \leq m \leq N_\lambda$, with $\mathbf{F}_{i+1}^{I+1,+} = \mathbf{F}_{i+1}^+$ and $\mathbf{F}_{i+1}^{N_\lambda+1,+} = \mathbf{F}_{i+1}$.

The telescopic properties of the linear solutions for the approximate flux function provide the definition of fluxes at $x = 0$, $\mathbf{F}_i^-$ and $\mathbf{F}_{i+1}^+$,

$$\begin{aligned} \mathbf{F}_i^- &= \mathbf{F}_i + (\widetilde{\mathbf{P}}\widetilde{\mathbf{\Lambda}}^-\mathbf{\Theta}\widetilde{\mathbf{P}}^{-1})_{i+\frac{1}{2}}\delta\mathbf{U}_{i+\frac{1}{2}} = \mathbf{F}_i + \sum_{m=1}^{I}\left(\widetilde{\lambda}^-\alpha\theta\widetilde{\mathbf{e}}\right)_{i+\frac{1}{2}}^m , \\ \mathbf{F}_{i+1}^+ &= \mathbf{F}_{i+1} - (\widetilde{\mathbf{P}}\widetilde{\mathbf{\Lambda}}^+\mathbf{\Theta}\widetilde{\mathbf{P}}^{-1})_{i+\frac{1}{2}}\delta\mathbf{U}_{i+\frac{1}{2}} = \mathbf{F}_{i+1} - \sum_{m=I+1}^{N_\lambda}\left(\widetilde{\lambda}^+\alpha\theta\widetilde{\mathbf{e}}\right)_{i+\frac{1}{2}}^m , \end{aligned} \tag{87}$$

with

$$\mathbf{F}_{i+1}^+ - \mathbf{F}_i^- = \sum_{m=1}^{N_\lambda}\left(\bar{\beta}\widetilde{\mathbf{e}}\right)_{i+\frac{1}{2}}^m = \bar{\mathbf{S}}_{i+\frac{1}{2}} , \tag{88}$$

recovering the RH condition in (82).

At this point, the corresponding intercell flux for the approximate first order Godunov method in (38) is given by

$$\mathbf{F}_{i+\frac{1}{2}}^- = \mathbf{F}_i^- , \qquad \mathbf{F}_{i-\frac{1}{2}}^+ = \mathbf{F}_i^+ , \tag{89}$$

resulting in

$$\mathbf{U}_i^{n+1} = \mathbf{U}_i^n - (\mathbf{F}_i^- - \mathbf{F}_i^+)\frac{\Delta t}{\Delta x} . \tag{90}$$

It is worth mentioning that it is not longer possible to define a general intercell flux function independent of the side of the solution considered, due to the presence of source terms, as in the homogeneous case.

3.4. Fluctuation form

Numerical schemes for nonconservative systems can be written in the following fluctuation form [30, 23]

$$\mathbf{U}_i^{n+1} = \mathbf{U}_i^n - (\delta\mathbf{M}_{i+\frac{1}{2}}^- + \delta\mathbf{M}_{i-\frac{1}{2}}^+)\frac{\Delta t}{\Delta x} , \tag{91}$$

where if functions $\delta\mathbf{M}_{i+\frac{1}{2}}^{\pm}$ satisfy

$$\delta\mathbf{M}_{i\mp\frac{1}{2}}^{\pm} = 0 \tag{92}$$

in steady cases, convergence to a solution with mesh refinement is guaranteed. Depending on the approximations made to evaluate source term integral $\bar{\mathbf{S}}_{i+\frac{1}{2}}$, the solution may not converge to the exact solution. In cases where the numerical solution converges to the exact solution, the numerical scheme is exactly balanced.

The equivalent fluctuation form of numerical scheme in (38) can be straightforward derived by simply using the intercell flux definitions (87) in (90), as

$$\mathbf{U}_i^{n+1} = \mathbf{U}_i^n - \left(\left[\mathbf{F}_i + \sum_m \left(\widetilde{\lambda}^- \alpha \theta \widetilde{\mathbf{e}} \right)_{i+\frac{1}{2}}^m \right] - \left[\mathbf{F}_i - \sum_m \left(\widetilde{\lambda}^+ \alpha \theta \widetilde{\mathbf{e}} \right)_{i-\frac{1}{2}}^m \right] \right) \frac{\Delta t}{\Delta x} , \tag{93}$$

leading to

$$\delta \mathbf{M}_{i+\frac{1}{2}}^- = \sum_m \left(\widetilde{\lambda}^- \alpha \theta \widetilde{\mathbf{e}} \right)_{i+\frac{1}{2}}^m , \quad \delta \mathbf{M}_{i-\frac{1}{2}}^+ = \sum_m \left(\widetilde{\lambda}^+ \alpha \theta \widetilde{\mathbf{e}} \right)_{i-\frac{1}{2}}^m . \tag{94}$$

The fluctuation form can also be derived by assuming that the linearized solution $\hat{\mathbf{U}}$ satisfies the following constant coefficient linear problem [32]

$$\frac{\partial \hat{\mathbf{U}}}{\partial t} + \mathbf{L}_{i+1/2} \frac{\partial \hat{\mathbf{U}}}{\partial x} = 0 . \tag{95}$$

Integration over the same control volume in (45) leads to the following constraint

$$\delta \mathbf{M}_{i+\frac{1}{2}} = \mathbf{L}_{i+1/2} \delta \mathbf{U}_{i+\frac{1}{2}} = \delta \mathbf{F}_{i+\frac{1}{2}} - \bar{\mathbf{S}}_{i+\frac{1}{2}} , \tag{96}$$

providing the following value for $\mathbf{L}_{i+1/2}$ matrix [32]

$$\mathbf{L}_{i+\frac{1}{2}} = (\widetilde{\mathbf{P}} \widetilde{\mathbf{\Lambda}} \mathbf{\Theta} \widetilde{\mathbf{P}}^{-1})_{i+\frac{1}{2}} . \tag{97}$$

Now, difference $\delta \mathbf{M}_{i+\frac{1}{2}}$ can be splitted in two terms

$$\delta \mathbf{M}_{i+\frac{1}{2}}^- + \delta \mathbf{M}_{i+\frac{1}{2}}^+ = \mathbf{L}_{i+1/2}^- \delta \mathbf{U}_{i+\frac{1}{2}} + \mathbf{L}_{i+1/2}^+ \delta \mathbf{U}_{i+\frac{1}{2}} , \tag{98}$$

allowing to express the numerical scheme (38) in fluctuation form, following the quasi-steady wave-propagation algorithm of LeVeque [1]

$$\mathbf{U}_i^{n+1} = \mathbf{U}_i^n - (\mathbf{L}_{i+1/2}^- \delta \mathbf{U}_{i+\frac{1}{2}} + \mathbf{L}_{i-1/2}^+ \delta \mathbf{U}_{i-\frac{1}{2}}) \frac{\Delta t}{\Delta x} , \tag{99}$$

with $\mathbf{L}_{i+1/2}^{\pm} = (\widetilde{\mathbf{P}} \widetilde{\mathbf{\Lambda}}^{\pm} \mathbf{\Theta} \widetilde{\mathbf{P}}^{-1})_{i+1/2}$ equivalent to (91).

3.5. Cell averaging and stability.

When using the intercell flux form in (38) or the fluctuation form in (99) when defining the updating numerical scheme, the information regarding the evaluation of the inner states defined by the approximate solver behind the numerical scheme remains hidden. It is well known that in the homogeneous case linearized solvers may fail, generating unphysical values of the conserved variables, as negative values of density for the Euler equations or negative values of water depth for the SWE. In those cases, positively conservative methods can be ensured by the generation of entropy fix algorithms, as the Harten-Hyman entropy fix [53, 48]. The Harten-Hyman entropy fix is based on the reconstruction of the approximate solution and results in an adequate modification of the constant coefficients used to define the intercell flux function. When moving to non-strictly hyperbolic system of equations with source terms, conservative methods require to generate extra fix procedures that can be used not only to ensure positively conservative methods, but also friction fix techniques able to ensure an accurate viscous dissipation rate. In order to clearly explore how all the inner states defined by the approximate solver are involved in the solution, the numerical scheme in fluctuation form (99)

is directly defined here by cell-averaging the approximate solutions involved in the updating step. This procedure is of great importance when formulating appropriate source term fixes in a specific system of equations.

Numerical scheme in (91) can be constructed following Godunov's method, by cell-averaging the piecewise constant solutions of the adjacent RP solutions evolved for a time equal to the time step. The integral volume $[0, \Delta x] \times [0, \Delta t]$ in cell i is illustrated in Figure 7. Focusing on the updating rule for cell i, three different regions define the integral cell volume

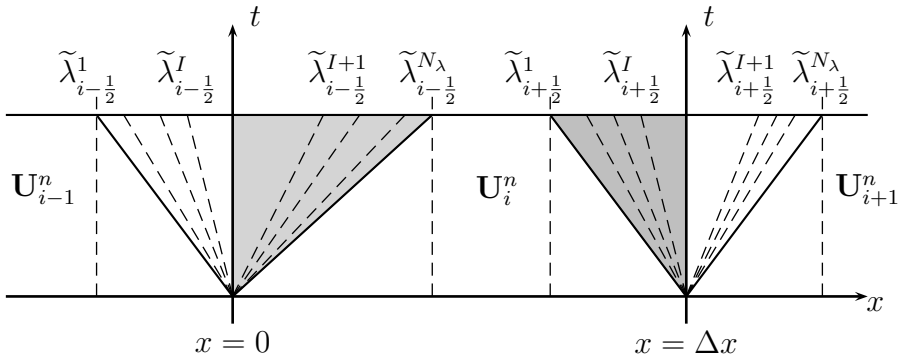

Figure 7: Cell average control volume for cell i.

$$\mathbf{U}_i^{n+1}\Delta x = V_1 + V_2 + V_3\,, \tag{100}$$

where V_1 is the average of the approximate solution of the $i - \frac{1}{2}$ RP in cell i, given by

$$V_1 = \mathbf{U}_i^{I+1,+}\widetilde{\lambda}^{I+1}_{i-\frac{1}{2}}\Delta t + \sum_{m=I+2}^{N_\lambda} \mathbf{U}_i^{m+}(\widetilde{\lambda}^m - \widetilde{\lambda}^{m-1})_{i-\frac{1}{2}}\Delta t\,, \tag{101}$$

V_3 is the average of the approximate solution of the $i + \frac{1}{2}$ RP developed in cell i, with

$$V_3 = \sum_{m=1}^{I-1} \mathbf{U}_i^{m-}(\widetilde{\lambda}^{m+1} - \widetilde{\lambda}^m)_{i+\frac{1}{2}}\Delta t - \mathbf{U}_i^{I,-}\widetilde{\lambda}^I_{i+\frac{1}{2}}\Delta t \tag{102}$$

and V_2 is the unaltered initial region not affected by the surrounding RPs, defined by

$$V_2 = \mathbf{U}_i^n(\Delta x - \widetilde{\lambda}^{N_\lambda}_{i-\frac{1}{2}}\Delta t + \widetilde{\lambda}^1_{i+\frac{1}{2}}\Delta t)\,. \tag{103}$$

The cell average of the different regions can be reformulated as follows

$$\mathbf{U}_i^{n+1}\Delta x = V_{i-\frac{1}{2}} + V_{i+\frac{1}{2}} + \mathbf{U}_i^n\Delta x\,, \tag{104}$$

with $V_{i-\frac{1}{2}}$ and $V_{i+\frac{1}{2}}$ involving all celerities in the $i - \frac{1}{2}$ and $i + \frac{1}{2}$ RPs respectively. Average quantity $V_{i-\frac{1}{2}}$

$$V_{i-\frac{1}{2}} = \left[\mathbf{U}_i^{I+1,+}\widetilde{\lambda}^{I+1} + \sum_{m=I+2}^{N_\lambda} \mathbf{U}_i^{m,+}(\widetilde{\lambda}^m - \widetilde{\lambda}^{m-1}) - \mathbf{U}_i^n\widetilde{\lambda}^{N_\lambda}\right]_{i-\frac{1}{2}} \Delta t \tag{105}$$

can be expressed as

$$V_{i-\frac{1}{2}} = \left[\sum_{m=I+1}^{N_\lambda} \mathbf{U}_i^{m,+} \widetilde{\lambda}^m - \sum_{m=I+2}^{N_\lambda+1} \mathbf{U}_i^{m,+} \widetilde{\lambda}^{m-1} \right]_{i-\frac{1}{2}} \Delta t \,. \tag{106}$$

By translating index m in the second term, as

$$V_{i-\frac{1}{2}} = \left[\sum_{m=I+1}^{N_\lambda} \mathbf{U}_i^{m,+} \widetilde{\lambda}^m - \sum_{m=I+1}^{N_\lambda} \mathbf{U}_i^{m+1,+} \widetilde{\lambda}^m \right]_{i-\frac{1}{2}} \Delta t \,, \tag{107}$$

all terms can be brought together and if using (79)

$$V_{i-\frac{1}{2}} = - \left[\sum_{m=I+1}^{N_\lambda} (\mathbf{U}_i^{m+1,+} - \mathbf{U}_i^{m,+}) \widetilde{\lambda}^m \right]_{i-\frac{1}{2}} \Delta t = - \sum_{m=I+1}^{N_\lambda} \left(\widetilde{\lambda}^+ \alpha\theta \widetilde{\mathbf{e}} \right)_{i-\frac{1}{2}}^m \Delta t \tag{108}$$

expressed in fluctuation form. The procedure is similar for average quantity $V_{i+\frac{1}{2}}$

$$V_{i+\frac{1}{2}} = \left[\mathbf{U}_i^n \widetilde{\lambda}^1 + \sum_{m=1}^{I-1} \mathbf{U}_i^{m,-} (\widetilde{\lambda}^{m+1} - \widetilde{\lambda}^m) - \mathbf{U}_i^{I,-} \widetilde{\lambda}^I \right]_{i+\frac{1}{2}} \Delta t \,, \tag{109}$$

that can also be expressed as follows

$$V_{i+\frac{1}{2}} = \left[\sum_{m=0}^{I-1} \mathbf{U}_i^{m,-} \widetilde{\lambda}^{m+1} - \sum_{m=1}^{I} \mathbf{U}_i^{m,-} \widetilde{\lambda}^m \right]_{i+\frac{1}{2}} \Delta t \,. \tag{110}$$

By translating index m in the first term, as

$$V_{i+\frac{1}{2}} = \left[\sum_{m=1}^{I} \mathbf{U}_i^{m-1,-} \widetilde{\lambda}^m - \sum_{m=1}^{I} \mathbf{U}_i^{m,-} \widetilde{\lambda}^m \right]_{i+\frac{1}{2}} \Delta t \,, \tag{111}$$

all terms can be unified and by using definition in (72)

$$V_{i+\frac{1}{2}} = - \left[\sum_{m=1}^{I} (\mathbf{U}_i^{m,-} - \mathbf{U}_i^{m-1,-}) \widetilde{\lambda}^m \right]_{i+\frac{1}{2}} \Delta t = - \sum_{m=1}^{I} \left(\widetilde{\lambda}^- \alpha\theta \widetilde{\mathbf{e}} \right)_{i+\frac{1}{2}}^m \Delta t \tag{112}$$

expressed in fluctuation form. Cell integral average value $\Delta x \mathbf{U}_i^{n+1}$ is given by

$$\Delta x \mathbf{U}_i^{n+1} = \Delta x \mathbf{U}_i^n - \left[\sum_{m=I+1}^{N_\lambda} \left(\widetilde{\lambda}^+ \alpha\theta \widetilde{\mathbf{e}} \right)_{i-\frac{1}{2}}^m + \sum_{m=1}^{I} \left(\widetilde{\lambda}^- \alpha\theta \widetilde{\mathbf{e}} \right)_{i+\frac{1}{2}}^m \right] \Delta t \,, \tag{113}$$

recovering the fluctuation form, equivalent to the intercell flux formulation.

Derivation of expression in (113) makes clear that although approximate flux function at $x = 0$ is the key-stone of the Godunov method, the whole solution participates in the updating step. In Figure 7 the time step is small enough so that there is no interaction of waves from neighboring Riemann problems. This would be necessary if we wanted to construct the solution at $\mathbf{U}_i^{n+1}$ in order to explicitly calculate the cell average. According to [2], in order to use the flux formula it is only necessary that the edge values $\hat{\mathbf{U}}(x,t)$ remain constant in time over the entire time step, which allows a time step roughly twice as large and the time step is limited by

$$\Delta t \leq \min\left[\Delta t^{\tilde{\lambda}}_{i+\frac{1}{2}}\right], \qquad \Delta t^{\tilde{\lambda}}_{i+\frac{1}{2}} = \frac{\Delta x}{\max_{m=1,N_\lambda} |\widetilde{\lambda}^m|_{i+\frac{1}{2}}}. \tag{114}$$

But this choice of the time step may not be sufficient. If we are demanding, for instance, positivity over a particular conserved variable, only if positivity of all inner states is ensured, integral of the conserved variable in the cell in (113) will remain positive. Source term integration is a difficult task that in some cases may lead to the definition of unphysical values in the inner states. When trying to preserve the conservative character of the numerical scheme, it may be necessary to reduce the time step to ensure a positive cell average solution. On the other hand, the knowledge of the inner solution allows to redefine unrealistic predictions allowing to recover the time step selected by (114). Further discussion about the positivity for discretization of the shallow water equations is provided in Section 7.1.

4. The shallow water equations

In this section the x-split two-dimensional shallow water equations are analyzed using the Augmented Roe approach detailed in the previous section. This analysis can be straightforward translated to two-dimensions without loss of generality, avoiding redefinition of the numerical fluxes depending on the mesh characteristics. In a Cartesian coordinate system, the Reynolds transport for mass and momentum conservation leads to a mathematical model where components in (35) are given by

$$\mathbf{U} = \begin{pmatrix} h \\ hu \\ hv \end{pmatrix}, \quad \mathbf{F} = \begin{pmatrix} hu \\ hu^2 + \frac{1}{2}gh^2 \\ huv \end{pmatrix}, \tag{115}$$

where h represents the water depth, (u, v) are the depth averaged components of the velocity vector $\mathbf{u}$ along the (x, y) coordinates respectively and g is the acceleration of gravity. The source term of the system is split in two kind of terms, $\mathbf{S} = \mathbf{S}_z + \mathbf{S}_\tau$, where

$$\mathbf{S}_z = \begin{pmatrix} 0 \\ \frac{p_{bx}}{\rho_w} \\ 0 \end{pmatrix}, \quad \mathbf{S}_\tau = \begin{pmatrix} 0 \\ -\frac{\tau_{bx}}{\rho_w} \\ 0 \end{pmatrix}. \tag{116}$$

The terms p_{bx} and τ_{bx} are the pressure along the bottom and the shear stress in the x direction respectively, with ρ_w the density of water. The above formulation is written in terms of the unit discharge and not valid for arbitrary cross sections.

4.1. Approximate solutions for the shallow water equations

The approximate Jacobian $\widetilde{\mathbf{J}}$ for the homogeneous part,

$$\widetilde{\mathbf{J}} = \begin{pmatrix} 0 & 1 & 0 \\ \widetilde{c}^2 - \widetilde{u}^2 & 2\widetilde{u} & 0 \\ -\widetilde{u}\widetilde{v} & \widetilde{v} & \widetilde{u} \end{pmatrix}, \tag{117}$$

is constructed with the following Roe averaged variables [11]

$$\widetilde{u}_{i+\frac{1}{2}} = \frac{u_i\sqrt{h_i} + u_{i+1}\sqrt{h_{i+1}}}{\sqrt{h_i} + \sqrt{h_{i+1}}}, \qquad \widetilde{v}_{i+\frac{1}{2}} = \frac{v_i\sqrt{h_i} + v_{i+1}\sqrt{h_{i+1}}}{\sqrt{h_i} + \sqrt{h_{i+1}}},$$
$$\widetilde{c}_{i+\frac{1}{2}} = \sqrt{g\bar{h}_{i+1/2}}, \tag{118}$$

with $\bar{h}_{i+1/2} = (h_i + h_{i+1})/2$, leading to the following set of approximate eigenvalues

$$\widetilde{\lambda}^1_{i+\frac{1}{2}} = (\widetilde{u} - \widetilde{c})_{i+\frac{1}{2}}\,, \qquad \widetilde{\lambda}^2_{i+\frac{1}{2}} = \widetilde{u}_{i+\frac{1}{2}}\,, \qquad \widetilde{\lambda}^3_{i+\frac{1}{2}} = (\widetilde{u} + \widetilde{c})_{i+\frac{1}{2}}\,, \tag{119}$$

and eigenvectors

$$\widetilde{\mathbf{e}}^1_{i+\frac{1}{2}} = \begin{pmatrix} 1 \\ \widetilde{\lambda}^1 \\ \widetilde{v} \end{pmatrix}_{i+\frac{1}{2}}\,, \qquad \widetilde{\mathbf{e}}^2_{i+\frac{1}{2}} = \begin{pmatrix} 0 \\ 0 \\ \widetilde{c} \end{pmatrix}_{i+\frac{1}{2}}\,, \qquad \widetilde{\mathbf{e}}^3_{i+\frac{1}{2}} = \begin{pmatrix} 1 \\ \widetilde{\lambda}^3 \\ \widetilde{v} \end{pmatrix}_{i+\frac{1}{2}}\,. \tag{120}$$

The wave strengths are given by

$$\begin{gathered} \alpha^1_{i+\frac{1}{2}} = \left(\frac{\delta h}{2} + \frac{\widetilde{u}\delta h - \delta(hu)}{2\widetilde{c}}\right)_{i+\frac{1}{2}}\,, \qquad \alpha^2_{i+\frac{1}{2}} = \left(\frac{\delta(hv) - \widetilde{v}\delta h}{\widetilde{c}}\right)_{i+\frac{1}{2}}\,, \\ \alpha^3_{i+\frac{1}{2}} = \left(\frac{\delta h}{2} - \frac{\widetilde{u}\delta h - \delta(hu)}{2\widetilde{c}}\right)_{i+\frac{1}{2}} \end{gathered} \tag{121}$$

and source strengths are given by

$$\bar{\beta}^1_{i+\frac{1}{2}} = \left(-\frac{\bar{S}_2}{2\widetilde{c}}\right)_{i+\frac{1}{2}}\,, \qquad \bar{\beta}^2_{i+\frac{1}{2}} = 0\,, \qquad \bar{\beta}^3_{i+\frac{1}{2}} = \left(\frac{\bar{S}_2}{2\widetilde{c}}\right)_{i+\frac{1}{2}}\,, \tag{122}$$

with $(\bar{S}_2) = \bar{S}_z + \bar{S}_\tau$, the second component of the source term vector, yet to be defined.

Depending on the flow regime, given by the average Froude number, $\widetilde{Fr} = \widetilde{u}/\widetilde{c}$, four approximate solutions appear. They are revisited next focusing on the solutions of water depth h and unit discharge hu. Considering that unit discharge hv acts as a passive scalar, that is, it does not participate directly in the hydrodynamics of the approximate solution for h or hu, it is possible to ignore the jump in hv, simplifying the analysis.

4.1.1. Subcritical RP with $\widetilde{u} > 0$.

In the case that $0 < \widetilde{Fr} < 1$, the weak solution, illustrated in Figure 8, is given by

$$\begin{gathered} \mathbf{U}^{1,-}_i = \mathbf{U}_i + (\theta\alpha\widetilde{\mathbf{e}})^1_{i+\frac{1}{2}}\,, \\ \mathbf{U}^{3,+}_{i+1} = \mathbf{U}_{i+1} - (\theta\alpha\widetilde{\mathbf{e}})^3_{i+\frac{1}{2}}\,, \quad \mathbf{U}^{2,+}_{i+1} = \mathbf{U}^{3,+}_{i+1} - (\theta\alpha\widetilde{\mathbf{e}})^2_{i+\frac{1}{2}}\,. \end{gathered} \tag{123}$$

In this case, the solution for the water depth varies along $x = 0$ as follows

$$h^{1,-}_i = h^\star_{i+\frac{1}{2}} - \left(\frac{\bar{\beta}}{\widetilde{\lambda}}\right)^1_{i+\frac{1}{2}}\,, \qquad h^{3,+}_{i+1} = h^{2,+}_{i+1} = h^\star_{i+\frac{1}{2}} + \left(\frac{\bar{\beta}}{\widetilde{\lambda}}\right)^3_{i+\frac{1}{2}}\,, \tag{124}$$

with $h^\star$ the approximate solution for the homogeneous case without source terms, that can be expressed as

$$h^\star_{i+\frac{1}{2}} = h_i + \alpha^1_{i+\frac{1}{2}} = h_{i+1} - \alpha^3_{i+\frac{1}{2}} \tag{125}$$

if using definition in (56). Contrary to the case with source terms, the intermediate solution $h^\star_{i+1/2}$ is continuous in the region of solutions given by $\widetilde{\lambda}^1 t < x < \widetilde{\lambda}^3 t$ (gray region in Figure 8). Unit water discharge in x is constant in the region of solutions,

$$(hu)_i^{1,-} = (hu)_{i+1}^{2,+} = (hu)_{i+1}^{3,+} = (hu)_{i+\frac{1}{2}}^{\downarrow}\,, \tag{126}$$

that written in terms of the solution for the homogeneous case, $(hu)^\star$, leads to

$$(hu)_{i+\frac{1}{2}}^{\downarrow} = (hu)_{i+\frac{1}{2}}^{\star} + \bar{\beta}_{i+\frac{1}{2}}^{3} = (hu)_{i+\frac{1}{2}}^{\star} - \bar{\beta}_{i+\frac{1}{2}}^{1}\,, \tag{127}$$

also continuous in the region of solutions, and given by

$$(hu)_{i+\frac{1}{2}}^{\star} = (hu)_{i+1} - (\alpha\widetilde{\lambda})_{i+\frac{1}{2}}^{3} = (hu)_i + (\alpha\widetilde{\lambda})_{i+\frac{1}{2}}^{1}\,. \tag{128}$$

That is, the inner solutions in (123) conserve the same value of hu at $x = 0$ from left to right, and from right to left, but this symmetry disappears for h. The constant value of hu at $x = 0$ is a consequence of the conservation properties of the flux between cells, hence ensuring mass conservation in the final updating scheme.

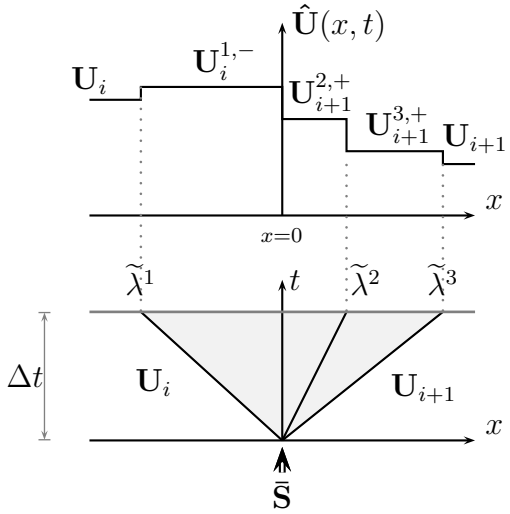

Figure 8: Values of the solution $\hat{\mathbf{U}}(x,t)$ in each wedge of the (x,t) plane for the subcritical case, $\widetilde{u} > 0$.

4.1.2. Subcritical RP with $\widetilde{u} < 0$

When $-1 < \widetilde{Fr} < 0$, the weak solution, illustrated in Figure 9, is given by

$$\begin{gathered}\mathbf{U}_i^{1,-} = \mathbf{U}_i + (\theta\alpha\widetilde{\mathbf{e}})_{i+\frac{1}{2}}^{1}\,, \quad \mathbf{U}_i^{2,-} = \mathbf{U}_i^{1,-} + (\theta\alpha\widetilde{\mathbf{e}})_{i+\frac{1}{2}}^{2}\,, \\ \mathbf{U}_{i+1}^{3,+} = \mathbf{U}_{i+1} - (\theta\alpha\widetilde{\mathbf{e}})_{i+\frac{1}{2}}^{3}\end{gathered} \tag{129}$$

and the solution for the water depth varies along $x = 0$ as follows

$$h_i^{1,-} = h_i^{2,-} = h_{i+\frac{1}{2}}^{\star} - \left(\frac{\bar{\beta}}{\widetilde{\lambda}}\right)_{i+\frac{1}{2}}^{1}\,, \qquad h_{i+1}^{3,+} = h_{i+\frac{1}{2}}^{\star} + \left(\frac{\bar{\beta}}{\widetilde{\lambda}}\right)_{i+\frac{1}{2}}^{3}\,, \tag{130}$$

with $h^\star$ defined as in (125). The solution for the unit water discharge in x is

$$(hu)_i^{1,-} = (hu)_i^{2,-} = (hu)_{i+1}^{3,+} = (hu)_{i+\frac{1}{2}}^{\downarrow}\,, \tag{131}$$

with $(hu)^{\downarrow}$ as in (127), and constant in the region of solutions given by $\widetilde{\lambda}^1 t < x < \widetilde{\lambda}^3 t$ (gray region in Figure 9).

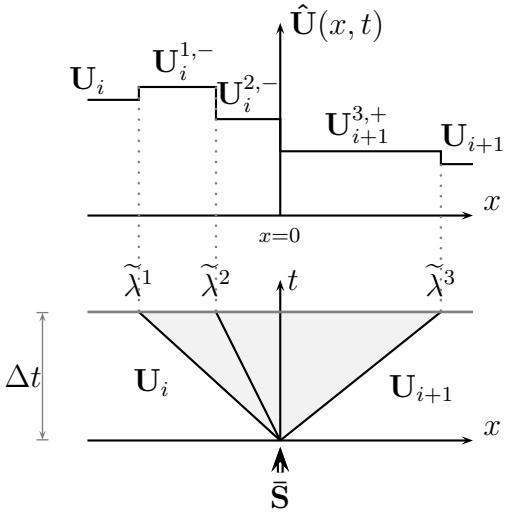

Figure 9: Values of the solution $\hat{\mathbf{U}}(x,t)$ in each wedge of the (x,t) plane for the subcritical case, $\widetilde{u} < 0$.

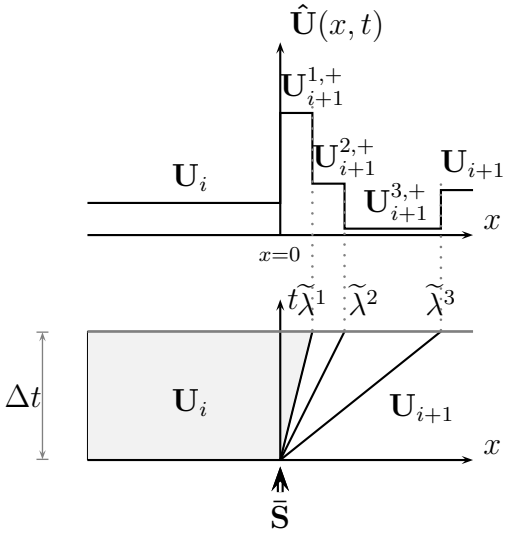

Figure 10: Values of the solution $\hat{\mathbf{U}}(x,t)$ in each wedge of the (x,t) plane for the supercritical case, $\widetilde{u} > 0$.

4.1.3. Supercritical RP with $\widetilde{u} > 0$

Figure 10 illustrates the approximate solution when $\widetilde{Fr} > 1$. On the left side of the RP, it is

$$\mathbf{U}_i^- = \mathbf{U}_i \tag{132}$$

and on the right side

$$\begin{aligned} \mathbf{U}_{i+1}^{3,+} = \mathbf{U}_{i+1} - (\theta\alpha\widetilde{\mathbf{e}})_{i+\frac{1}{2}}^3 \ , \quad \mathbf{U}_{i+1}^{2,+} = \mathbf{U}_{i+1}^{3,+} - (\theta\alpha\widetilde{\mathbf{e}})_{i+\frac{1}{2}}^2 \ , \\ \mathbf{U}_{i+1}^{1,+} = \quad \mathbf{U}_{i+1}^{2,+} - (\theta\alpha\widetilde{\mathbf{e}})_{i+\frac{1}{2}}^1 \ , \end{aligned} \tag{133}$$

Water depth varies along x generating the following states

$$h_i^- = h_i\,, \qquad h_{i+1}^{3,+} = h_{i+1}^{2,+} \neq h_{i+1}^{1,+}\,, \tag{134}$$

with

$$h_{i+1}^{1,+} = h_i - 2\left(\frac{\widetilde{c}\bar{\beta}^3}{\widetilde{\lambda}^1\widetilde{\lambda}^3}\right)_{i+\frac{1}{2}}\,, \qquad h_{i+1}^{3,+} = h^\star_{i+\frac{1}{2}} + \left(\frac{\bar{\beta}}{\widetilde{\lambda}}\right)^3_{i+\frac{1}{2}}\,, \tag{135}$$

with $h^\star$ as in (125). Unit water discharge in x is characterized by

$$(hu)_i^- = (hu)_{i+1}^{1,+} = (hu)^\downarrow_{i+\frac{1}{2}}, \qquad (hu)_{i+1}^{3,+} = (hu)_{i+1}^{2,+}\,, \tag{136}$$

The solution for unit discharge is continuous in the region $\widetilde{\lambda}^1 t < x < \widetilde{\lambda}^3 t$ and $(hu)^\downarrow$ is constant in $x < \widetilde{\lambda}^1 t$ (gray region in Figure 10).

4.1.4. Supercritical RP with $\widetilde{u} < 0$
When $\widetilde{Fr} < -1$ the solution $\hat{\mathbf{U}}(x,t)$ provides

$$\begin{gathered}\mathbf{U}_i^{1,-} = \mathbf{U}_i + (\theta\alpha\widetilde{\mathbf{e}})^1_{i+\frac{1}{2}}\,, \qquad \mathbf{U}_i^{2,-} = \mathbf{U}_i^{1,-} + (\theta\alpha\widetilde{\mathbf{e}})^2_{i+\frac{1}{2}}\,,\\ \mathbf{U}_i^{3,-} = \mathbf{U}_i^{2,-} + (\theta\alpha\widetilde{\mathbf{e}})^3_{i+\frac{1}{2}}\,,\\ \mathbf{U}_{i+1}^{+} = \mathbf{U}_{i+1}\,.\end{gathered} \tag{137}$$

Water depth varies along x generating the following states

$$h_i^{1,-} = h_i^{2,-} \neq h_i^{3,-}\,, \qquad h_{i+1}^{+} = h_{i+1}\,, \tag{138}$$

where

$$h_i^{1,-} = h^\star_{i+\frac{1}{2}} - \left(\frac{\bar{\beta}}{\widetilde{\lambda}}\right)^1_{i+\frac{1}{2}}\,, \qquad h_i^{3,-} = h_{i+1} - 2\left(\frac{\widetilde{c}\bar{\beta}^1}{\widetilde{\lambda}^1\widetilde{\lambda}^3}\right)_{i+\frac{1}{2}}\,, \tag{139}$$

with $h^\star$ as in (125). Unit water discharge (hu) is characterized by

$$(hu)_i^{1,-} = (hu)_i^{2,-}\,, \qquad (hu)_i^{3,-} = (hu)_{i+1}^{+} = (hu)^\downarrow_{i+\frac{1}{2}}\,, \tag{140}$$

where now

$$(hu)_i^{1,-} = (hu)^\star_{i+\frac{1}{2}} - \bar{\beta}^1_{i+\frac{1}{2}}\,, \qquad (hu)^\downarrow_{i+\frac{1}{2}} = (hu)_{i+1}\,, \tag{141}$$

with $(hu)^\star$ as in (128). Unit discharge is again continuous in the solution region $\widetilde{\lambda}^1 t < x < \widetilde{\lambda}^3 t$ and $(hu)^\downarrow$ is constant in $x > \widetilde{\lambda}^3 t$ (gray region in Figure 11).

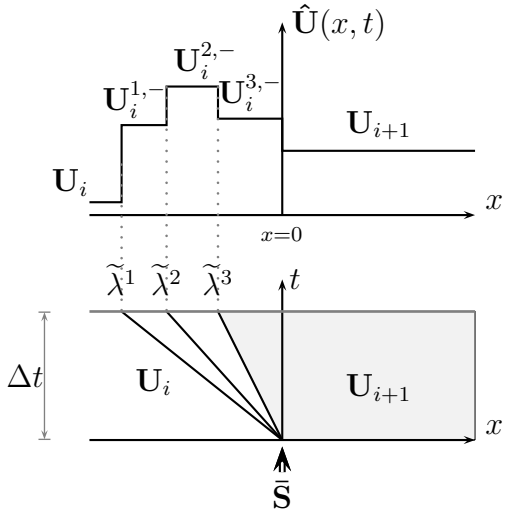

Figure 11: Values of the solution $\hat{\mathbf{U}}(x,t)$ in each wedge of the (x,t) plane for the supercritical case, $\widetilde{u} < 0$.

4.2. Transcritical RPs

One result of Roe's linearization is that the resulting approximate Riemann solution consists of only discontinuities and is constructed as a sum of jumps or shocks, leading to a good approximation for contact and shock waves. In rarefaction waves, a continuous change in flow variables appears, and approximations based on shocks may lead to inaccurate results, especially in cases of transcritical flow where the flow passes from subcritical to supercritical conditions.

To improve the accuracy of the numerical predictions for the SWE, the Harten-Hyman entropy fix in [53, 48] is applied. The entropy fix ensures accurate results in transcritical dam break problems as well as in steady solutions, allowing to reproduce exactly the sonic point in presence of source terms [20]. The numerical method is modified accordingly while retaining its conservative character.

4.2.1. Left transcritical rarefaction

The left transcritical rarefaction is characterized by $\lambda_i^1 < 0 < \lambda_{i+1}^1$, with $\lambda_i = \lambda(\mathbf{U}_i)$ and $\lambda_{i+1} = \lambda(\mathbf{U}_{i+1})$. In this case, the initial solution provided by the Roe approach and driven by the average value $\widetilde{\lambda}^1$ produces inaccurate results. The solution proposed in [53, 48] for the homogeneous case without source terms is based on the definition of a virtual intermediate state defined between celerities λ_i^1 and λ_{i+1}^1. In the case without source term, the numerical flux has a single value in the star region, hence the full description of the overall approximate solution can be bypassed by defining a new wave speed, $\check{\lambda}_{i+1/2}^1$, leading to a simple and effective modification of the general flux formulation. This new celerity is given by

$$\check{\lambda}_{i+\frac{1}{2}}^1 = \lambda_i^1 \frac{(\lambda_{i+1}^1 - \widetilde{\lambda}_{i+\frac{1}{2}}^1)}{(\lambda_{i+1}^1 - \lambda_i^1)}, \tag{142}$$

with $\check{\lambda}^1 < 0$ by definition.

The approximate solution proposed is based on the preservation of the conservative character of the numerical scheme, involving the flux jump associated $\widetilde{\lambda}_{i+1/2}^1$. Following Roe approximation two new jumps are involved in the solution. Flux splitting in (98) is redefined considering the new wave and source wave strengths as follows

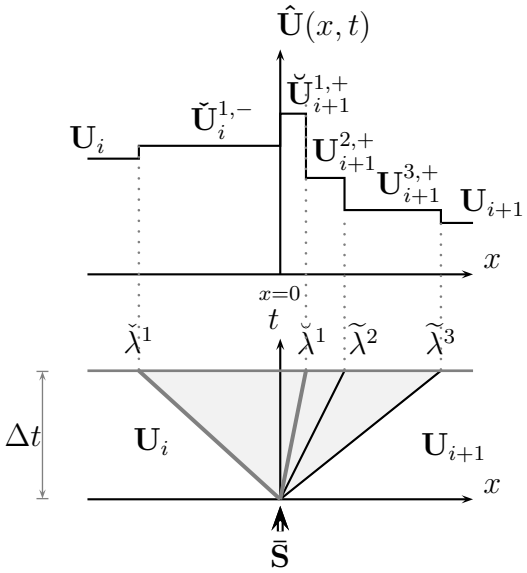

Figure 12: Approximate solution in a left transcritical rarefaction.

$$\delta\mathbf{M}_{i+\frac{1}{2}} = ((\breve{\lambda}\alpha - \breve{\beta})\widetilde{\mathbf{e}})^1_{i+\frac{1}{2}} + ((\breve{\lambda}\alpha - \breve{\beta})\widetilde{\mathbf{e}})^1_{i+\frac{1}{2}} + \sum_{m=2}^{3} \left(\widetilde{\lambda}\alpha\theta\widetilde{\mathbf{e}}\right)^m_{i+\frac{1}{2}} , \tag{143}$$

with

$$\breve{\beta}^1 + \breve{\beta}^1 = \bar{\beta}^1, \qquad \breve{\lambda}^1 + \breve{\lambda}^1 = \widetilde{\lambda}^1 , \tag{144}$$

to preserve flux difference or conservation across discontinuities in (46). Novel wave $\breve{\lambda}^1$ is expressed as

$$\breve{\lambda}^1_{i+\frac{1}{2}} = \lambda^1_{i+1} \frac{(\widetilde{\lambda}^1 - \lambda^1_i)}{(\lambda^1_{i+1} - \lambda^1_i)} , \tag{145}$$

where $\breve{\lambda}^1 > 0$ by definition.

The new approximate solution is illustrated in Figure 12 and involves the novel left state $\breve{\mathbf{U}}^{1,-}_i$ and the novel right state $\breve{\mathbf{U}}^{1,+}_{i+1}$. Using flux splitting in (143), the new jump across negative celerity $\breve{\lambda}$ leads to the following RH condition

$$\breve{\mathbf{F}}^{1,-}_i - \mathbf{F}_i = (\breve{\lambda}\breve{\theta}\alpha\widetilde{\mathbf{e}})^1_{i+\frac{1}{2}} = \breve{\lambda}^1_{i+\frac{1}{2}} \left(\breve{\mathbf{U}}^{1,-}_i - \mathbf{U}_i\right) , \tag{146}$$

with dimensionless parameter $\breve{\theta}^1 = 1 - \left(\breve{\beta}/\breve{\lambda}\alpha\right)^1$. On the right side of the solution plane, new jump across positive celerity $\breve{\lambda}$ leads to the following RH condition

$$\breve{\mathbf{F}}^{1,+}_{i+1} - \mathbf{F}^{2,+}_{i+1} = \breve{\lambda}^1_{i+\frac{1}{2}} \left(\breve{\mathbf{U}}^{1,+}_{i+1} - \mathbf{U}^{2,+}_{i+1}\right) , \tag{147}$$

with $\breve{\theta}^1 = 1 - (\breve{\beta}/\breve{\lambda}\alpha)^1$.

Definition of fluxes at $x = 0$, $\mathbf{F}^-_i$ and $\mathbf{F}^+_{i+1}$ in the intercell flux form of the updating scheme in (38) are modified as follows

$$\mathbf{F}_i^- = \mathbf{F}_i + \left(\check{\lambda}\check{\theta}\alpha\widetilde{\mathbf{e}}\right)_{i+\frac{1}{2}}^1 ,$$
$$\mathbf{F}_{i+1}^+ = \mathbf{F}_{i+1} - \left(\check{\lambda}\alpha\breve{\theta}\widetilde{\mathbf{e}}\right)_{i+\frac{1}{2}}^1 - \sum_{m=2}^{3}\left(\widetilde{\lambda}\alpha\theta\widetilde{\mathbf{e}}\right)_{i+\frac{1}{2}}^m \tag{148}$$

and the fluctuation form of the numerical scheme in (94) has to be extended to consider the new wave

$$\delta\mathbf{M}_{i+\frac{1}{2}}^- = \left(\check{\lambda}^1\check{\theta}\alpha\widetilde{\mathbf{e}}\right)_{i+\frac{1}{2}}^1 , \quad \delta\mathbf{M}_{i+\frac{1}{2}}^+ = \left(\check{\lambda}^1\alpha\breve{\theta}\widetilde{\mathbf{e}}\right)_{i+\frac{1}{2}}^1 + \sum_{m=2}^{3}\left(\widetilde{\lambda}\alpha\theta\widetilde{\mathbf{e}}\right)_{i+\frac{1}{2}}^m , \tag{149}$$

preserving the conservative character of the numerical scheme.

While inner states $\mathbf{U}_{i+1}^{3,+}$ and $\mathbf{U}_{i+1}^{2,+}$ are defined as in (133), by projecting the solution in the basis of eigenvectors of the approximate Jacobian, novel states $\check{\mathbf{U}}_i^{1,-}$ and $\breve{\mathbf{U}}_{i+1}^{1,+}$ can not be defined using the same strategy, as the novel wave celerities $\check{\lambda}^1$ and $\breve{\lambda}^1$ are not eigenvalues of $\widetilde{\mathbf{J}}$. From the first component of (146) the variation in the discharge (hu) on the left side and the variation of water depth are related by

$$(\check{hu})_i^{1,-} - (hu)_i = \check{\lambda}_{i+\frac{1}{2}}^1\left(\check{h}_i^{1,-} - h_i\right) , \tag{150}$$

leading to

$$(\check{hu})_i^{1,-} = (hu)_i + \left(\alpha\check{\lambda}\right)_{i+\frac{1}{2}}^1 - \check{\beta}_{i+\frac{1}{2}}^1 \tag{151}$$

and

$$\check{h}_i^{1,-} = h_{i+\frac{1}{2}}^\star - \left(\frac{\check{\beta}}{\check{\lambda}}\right)_{i+\frac{1}{2}}^1 . \tag{152}$$

On the right side, the values of (hu) and h in the novel state are defined departing from (147) with

$$(\breve{hu})_{i+1}^{1,+} = (hu)_{i+1}^{2,+} - (\alpha\breve{\lambda})_{i+\frac{1}{2}}^1 + \breve{\beta}_{i+\frac{1}{2}}^1 \tag{153}$$

and

$$\breve{h}_{i+1}^{1,+} = h_i + \left(\frac{\bar{\beta}}{\bar{\lambda}}\right)_{i+\frac{1}{2}}^3 + \left(\frac{\breve{\beta}}{\breve{\lambda}}\right)_{i+\frac{1}{2}}^1 . \tag{154}$$

The definition of novel inner states ensures the exact conservation property of mass,

$$(\check{hu})_i^{1,-} = (\breve{hu})_{i+1}^{1,+} . \tag{155}$$

Right unit mass discharge in (153) can be redefined as follows

$$(\breve{hu})_{i+1}^{1,+} = (hu)_i + \left(\alpha\check{\lambda}\right)_{i+\frac{1}{2}}^1 + \breve{\beta}_{i+\frac{1}{2}}^1 + \bar{\beta}_{i+\frac{1}{2}}^3 \tag{156}$$

and if compared with (151) the following relation appears

$$-\check{\beta}_{i+\frac{1}{2}}^1 = \breve{\beta}_{i+\frac{1}{2}}^1 + \bar{\beta}_{i+\frac{1}{2}}^3 , \tag{157}$$

satisfying condition in (144).

In any case the values of $\check{\beta}^1$ and $\breve{\beta}^1$ are not yet defined. The resulting condition derived from (155) would suggest that any possible relation between $\check{\beta}^1$ and $\breve{\beta}^1$ would be plausible, but this hypothesis is wrong.

Entropy correction must ensure convergence to the exact solution in presence of sonic points. When convergence to a sonic point is reproduced, conditions of the associate RP evolve to $\lambda_i < 0$ and $\lambda_{i+1} = \epsilon$. When steady state is achieved, $\epsilon \to 0$ to ensure $Fr_i < 1$ and $Fr_{i+1} = 1$. In this particular case, the novel waves given by the entropy correction are

$$\check{\lambda}^1(\epsilon \to 0) = \widetilde{\lambda}^1 \qquad \breve{\lambda}^1(\epsilon \to 0) = 0\,. \tag{158}$$

In the limit $\epsilon = 0$, if setting $\breve{\beta}^1 \neq 0$, the inner state $\breve{\mathbf{U}}^{1,+}_{i+1}$ is linked not to a region, but to the solution ray $x/t = 0$. In that case, this approximate state can not participate either on the left or on the right side of the solution. Consequently, if an exact splitting of the source term can not be provided, inaccurate solutions appear and convergence to the exact solution can not be guaranteed. Therefore, only if setting

$$\check{\beta}^1_{i+\frac{1}{2}} = \bar{\beta}^1_{i+\frac{1}{2}}\,, \qquad \breve{\beta}^1_{i+\frac{1}{2}} = 0\,, \tag{159}$$

convergence to the exact sonic point or its preservation is ensured, as the balance of fluxes and source terms is exactly reproduced.

4.2.2. Right transcritical rarefaction

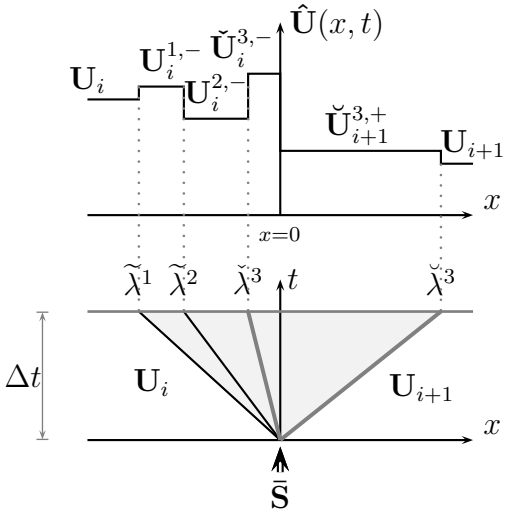

Figure 13: Approximate solution in a right transcritical rarefaction.

For a right transcritical rarefaction, with $\lambda^3_i < 0 < \lambda^3_{i+1}$, the entropy fix procedure is entirely analogous to the left rarefaction case. The single jump in $\widetilde{\lambda}^3$ is split into two smaller jumps $\check{\lambda}^3$ and $\breve{\lambda}^3$. Flux splitting in (98) is redefined as follows

$$\delta\mathbf{M}_{i+\frac{1}{2}} = \sum_{m=1}^{2} \left(\widetilde{\lambda}\alpha\theta\widetilde{\mathbf{e}}\right)^m_{i+\frac{1}{2}} + ((\check{\lambda}\alpha - \check{\beta})\widetilde{\mathbf{e}})^3_{i+\frac{1}{2}} + ((\breve{\lambda}\alpha - \breve{\beta})\widetilde{\mathbf{e}})^3_{i+\frac{1}{2}}\,, \tag{160}$$

preserving conservation across discontinuities in (46), with $\check{\lambda}^3 + \check{\lambda}^3 = \widetilde{\lambda}^3$, given by

$$\check{\lambda}^3_{i+\frac{1}{2}} = \lambda^3_i \frac{(\lambda^3_{i+1} - \widetilde{\lambda}^3_{i+\frac{1}{2}})}{(\lambda^3_{i+1} - \lambda^3_i)}\,, \qquad \check{\lambda}^3_{i+\frac{1}{2}} = \lambda^3_{i+1} \frac{(\widetilde{\lambda}^3_{i+\frac{1}{2}} - \lambda^3_i)}{(\lambda^3_{i+1} - \lambda^3_i)}\,, \tag{161}$$

with $\check{\lambda}^3 < 0$ and $\check{\lambda}^3 > 0$ by definition, and $\check{\beta}^3 + \check{\beta}^3 = \bar{\beta}^3$, with

$$\check{\beta}^3_{i+\frac{1}{2}} = 0\,, \qquad \check{\beta}^3_{i+\frac{1}{2}} = \bar{\beta}^3_{i+\frac{1}{2}}\,, \tag{162}$$

to ensure exact reproduction of sonic points.

The new approximate solution is depicted in Figure 13. The novel left state $\check{\mathbf{U}}^{3,-}_i$ and the novel right state $\check{\mathbf{U}}^{3,+}_{i+1}$ appear. The approximate solution is completed by using

$$\check{\mathbf{F}}^{3,-}_i - \mathbf{F}^{2,-}_i = \check{\lambda}^3_{i+\frac{1}{2}} \left(\check{\mathbf{U}}^{3,-}_i - \mathbf{U}^{2-}_i\right) \tag{163}$$

and

$$\mathbf{F}_{i+1} - \check{\mathbf{F}}^{3,+}_{i+1} = \check{\lambda}^3_{i+\frac{1}{2}} \left(\mathbf{U}_{i+1} - \check{\mathbf{U}}^{3,+}_{i+1}\right)\,. \tag{164}$$

Definition of the fluxes in the intercell flux form updating scheme in (38), $\mathbf{F}^-_i$ and $\mathbf{F}^+_{i+1}$, are modified as follows

$$\begin{aligned} \mathbf{F}^-_i &= \mathbf{F}_i + \sum_{m=1}^{2} \left(\widetilde{\lambda}\alpha\theta\widetilde{\mathbf{e}}\right)^m_{i+\frac{1}{2}} + \left(\check{\lambda}\alpha\check{\theta}\widetilde{\mathbf{e}}\right)^3_{i+\frac{1}{2}}\,, \\ \mathbf{F}^+_{i+1} &= \mathbf{F}_{i+1} - \left(\check{\lambda}\alpha\check{\theta}\widetilde{\mathbf{e}}\right)^3_{i+\frac{1}{2}}\,, \end{aligned} \tag{165}$$

with

$$\check{\theta}^3 = 1 - \left(\frac{\check{\beta}}{\check{\lambda}\alpha}\right)^3, \quad \check{\theta}^3 = 1 - \left(\frac{\check{\beta}}{\check{\lambda}\alpha}\right)^3 \tag{166}$$

and the fluctuation form of the numerical scheme in (94) is extended to consider the new wave

$$\delta\mathbf{M}^-_{i+\frac{1}{2}} = \sum_{m=1}^{2} \left(\widetilde{\lambda}\alpha\theta\widetilde{\mathbf{e}}\right)^m_{i+\frac{1}{2}} + \left(\check{\lambda}\alpha\check{\theta}\widetilde{\mathbf{e}}\right)^3_{i+\frac{1}{2}}\,, \quad \delta\mathbf{M}^+_{i+\frac{1}{2}} = \left(\check{\lambda}^3\alpha\check{\theta}\widetilde{\mathbf{e}}\right)^3_{i+\frac{1}{2}}\,, \tag{167}$$

preserving the conservative character of the numerical scheme.

Inner states $\mathbf{U}^{1,-}_i$ and $\mathbf{U}^{2,-}_i$ are defined as in (137). Water discharge (hu) is described on the right side using (164)

$$(\check{hu})^{3,+}_{i+1} = (hu)_{i+1} - (\alpha\check{\lambda})^3_{i+\frac{1}{2}} + \check{\beta}^3_{i+\frac{1}{2}} \tag{168}$$

and water depth on the right side of the solution is given by

$$\check{h}^{3,+}_{i+1} = h^\star_{i+\frac{1}{2}} + \left(\frac{\check{\beta}}{\check{\lambda}}\right)^3_{i+\frac{1}{2}}\,. \tag{169}$$

On the left side of the solution, discharge (hu) is

$$(\check{hu})_i^{3,-} = (\breve{hu})_{i+1}^{3,+}\,, \tag{170}$$

guaranteeing exact conservation, with

$$\check{h}_i^{3,-} = h_{i+1} - \left(\frac{\bar{\beta}}{\widetilde{\lambda}}\right)_{i+\frac{1}{2}}^{1} - \left(\frac{\check{\beta}}{\breve{\lambda}}\right)_{i+\frac{1}{2}}^{3} \tag{171}$$

the water depth at the same inner constant state.

5. Numerical integration of the source term

Source term component $S_2 = S_z + S_\tau$, accounts for pressure and shear stress forces on the bed surface, respectively. Depending on the approximations made over S_2, the solutions given by the consistent numerical scheme in (38), even convergent with mesh refinement, may not necessarily reproduce the exact or physically based solution.

When a piecewise constant data reconstruction of the bed elevation, z, is defined, pressure along the bottom can be integrated leading to a term that accounts for the thrust exerted by the bed step. If assuming that the pressure distribution is hydrostatic over the step and depends only on the free-surface level on the side of the discontinuity where the bottom elevation is lower, the source term $\bar{S}_z$ evaluated explicitly at $t = 0$ following (43) can be approached in a RP by [32]

$$\left(\bar{S}_{z,1}\right)_{i+\frac{1}{2}} = -g\left(h_j - \frac{|\delta z'|}{2}\right)_{i+\frac{1}{2}} \delta z'_{i+\frac{1}{2}}\,, \tag{172}$$

where z is the bed level surface, and j and $\delta z'$ are given by

$$j = \begin{cases} i & \text{if } \delta z_{i+\frac{1}{2}} \geq 0 \\ i+1 & \text{if } \delta z_{i+\frac{1}{2}} < 0 \end{cases} \qquad \delta z' = \begin{cases} h_i & \text{if } \delta z_{i+\frac{1}{2}} \geq 0 \text{ and } d_i < z_{i+1} \\ h_{i+1} & \text{if } \delta z_{i+\frac{1}{2}} < 0 \text{ and } d_{i+1} < z_i \\ \delta z & \text{otherwise} \end{cases} \tag{173}$$

and $d = (h + z)$ is the water level surface. Another explicit integration rule, namely the trapezoidal rule, can be adopted assuming smooth variation of the functions involved in the RP, leading to [32]

$$\left(\bar{S}_{z,2}\right)_{i+\frac{1}{2}} = -(g\bar{h}\delta z)_{i+\frac{1}{2}}\,, \tag{174}$$

based on the differential form of the thrust term, $S_z = -gh\partial_x z$, the so called bed slope source term.

The SWE is a set of depth-averaged equations and the equations that describe the tangential forces generated by the stresses carry into the momentum equations the main rheological traits of the fluid in motion. Depending on the case considered, different types of shear stresses appear: turbulent, dispersive, Coulomb-type, yield and viscous stresses. They can also act simultaneously. When the concentration of transported materials in the water column is negligible, an empirical friction coefficient c_f can be used to describe dispersive and turbulent effects near the bed [45]

$$S_\tau = -c_f u|\widetilde{\mathbf{u}}|\,. \tag{175}$$

The approximate space and time integral of the source term $\bar{S}_\tau$ in (43) can be expressed as [45]

$$(\bar{S}_\tau)_{i+\frac{1}{2}} = -\Delta x(c_f u_{min}|\widetilde{\mathbf{u}}|)_{i+\frac{1}{2}} \tag{176}$$

where $|u_{min}| = \min(|u|_i, |u|_{i+1})$ avoids an unrealistic estimation of the friction when the water depth approaches zero.

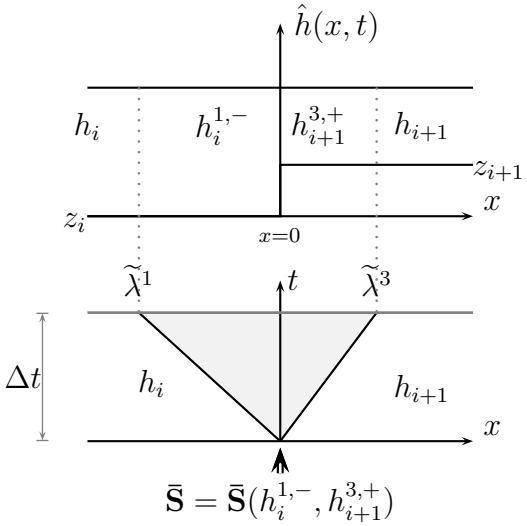

Figure 14: Solution $\hat{h}(x,t)$ in case of static equilibrium, continuous water level surface, using options (173) and (172).

The integration of the source terms is therefore based on approximations and has an impact in the solution. One of the most relevant consequences is the possible loss of positivity of the water depth. When trying to converge to physically admissible solutions, most of the current numerical developments focus on the presence of the bed slope, but when the SWE are applied to realistic scenarios, approximate solvers have to be able to completely exercise and explain the influence of the whole set of source terms. Under this perspective, the definition of well balanced numerical schemes in cases of quiescent equilibrium is a particular case that limits the range of application of the numerical solver. In the unified discretization followed here, the especial characteristics can be analyzed and the necessary strategies can be clearly envisaged.

The appearance of negative values of water depth is commonly associated to the presence of wetting/drying fronts. It is remarkable that all source terms can be responsible of this bad behavior. In what follows, it is shown how numerical modeling of source terms can be performed using appropriate strategies. In order to provide a clear perspective, the analysis of the quiescent equilibrium is performed first, moving next to more complex cases.

Bed level source term, $\bar{S}_z$, is a geometrical source term that also depends on the flow conditions. In cases of quiescent equilibrium, pressure exerted over the bed is certainly hydrostatic. In this particular case the integral approach in (43) involving the variation in time of the water depth is exact when using hydrostatic distribution of pressure in (172), equivalent to evaluation in (174). The definition of the approximate solutions enables a correct understanding of the source terms. Figure 14 plots the approximate solution $\hat{h}(x,t)$ in case of static equilibrium with continuous water level surface, using (172). Considering that in this particular case, $\delta h = -\delta z$, inner states $h_i^{1,-}$ and $h_{i+1}^{3,+}$ in (124) are given by

$$h_i^{1,-} = h_i + \frac{1}{2}\delta d_{i+\frac{1}{2}} = h_i \tag{177}$$

and

$$h_{i+1}^{3,+} = h_{i+1} - \frac{1}{2}\delta d_{i+\frac{1}{2}} = h_{i+1}\,, \tag{178}$$

while the solution for the unit discharge, $(hu)^{\downarrow}$ becomes

$$(hu)^{\downarrow}_{i+\frac{1}{2}} = -\frac{1}{2}\widetilde{c}_{i+\frac{1}{2}}\delta d_{i+\frac{1}{2}} = 0\,. \tag{179}$$

Therefore, the inner solution of the approximate solver guarantees equilibrium as coherent values for the intermediate conserved variables, reproducing the exact solution. The source term in (43) has been solved exactly, considering the bed pressure as a discontinuous function, where thrust term is computed using the exact information around the step, given by

$$\bar{S}_z = \bar{S}_z(h_i^{1,-}, h_{i+1}^{3,+}) = \frac{1}{\Delta t}\int_{-\Delta x/2}^{\Delta x/2}\int_0^{\Delta t} S_z(h_i^{1,-}, h_{i+1}^{3,+})\; dxdt\,. \tag{180}$$

Figure 15 illustrates a wet/dry quiescent equilibrium situation that produces a discontinuity in the water surface elevation, smaller than the bed level. In this case option (172) evaluates exactly the solution in (180), reproducing the expected result,

$$h_i^{1,-} = h_i, \quad h_{i+1}^{3,+} = 0\,, \quad (hu)^{\downarrow}_{i+\frac{1}{2}} = 0\,. \tag{181}$$

In this case option in (174) fails, as hydrostatic forces are grossly estimated,

$$h_i^{1,-} > h_i\,, \quad h_{i+1}^{3,+} < 0\,, \quad (hu)^{\downarrow}_{i+\frac{1}{2}} < 0 \tag{182}$$

leading to unphysical results. From this result, it has only been proved that numerical integration of the thrust term $\bar{S}_{z,1}$ in (172) is exact in any situation involving quiescent equilibrium, but this result can not be extrapolated to other flow conditions.

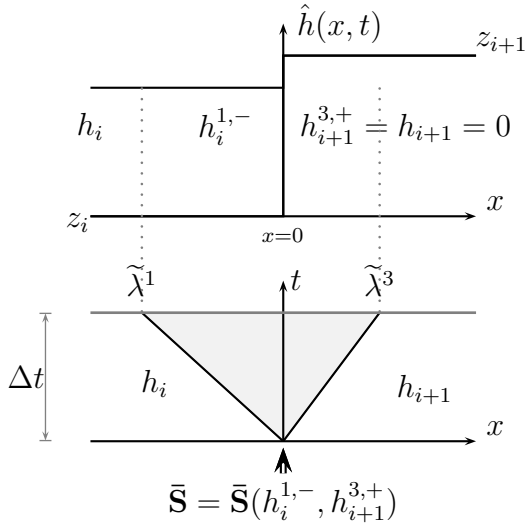

Figure 15: Solution $\hat{h}(x,t)$ in case of static equilibrium, discontinuous water level surface, using option (173).

Though the preservation of quiescent equilibrium is of importance, non zero steady and unsteady state flows must be correctly predicted by the numerical scheme. The definition of the integral source

term in (180) would suggest that, in general, source terms can be adequately integrated by using, for instance, iterative/implicit algorithms. Nevertheless, numerical results evidence that the approximate solution does not experience any improvement if using this technique.

In any case, in the search of numerical solutions for steady cases with moving water equilibrium, it is still possible to find solutions involving exact equilibrium between fluxes and source terms, reproducing a constant value of unit discharge equal to the exact solution. Exact balance can be provided by using both options, $\bar{S}_{z,1}$ and $\bar{S}_{z,2}$. Being both estimations plausible, no one will reproduce the physically based solution.

Shallow water equations are here described in terms of mass and momentum conservation equations, but it is also possible to describe them using mass and energy conservation equations. Weak solutions of conservation laws need not be unique, so an additional criterion is imposed. Energy can be used as the entropy function. This criterion can be used to ensure that the numerical scheme converges to a solution of reference with mesh refinement, and that this solution of reference is the exact physically based solution. Expanding derivatives in the momentum conservation equation for hu, and expressing pressure forces over the bed in differential form, it is possible to obtain

$$u\left[\partial_t h + \partial_x(hu)\right] + h\left[\partial_t u + g\partial_x(h+z) + u\partial_x(u)\right] = S_\tau\,. \tag{183}$$

Use of mass conservation followed by the division through h gives

$$\partial_t u + \partial_x\left[g(h+z) + \frac{u^2}{2}\right] = \frac{1}{h}S_\tau\,, \tag{184}$$

that involves the variation of the head energy, $H = \frac{u^2}{2g} + h + z$, for smooth variations, therefore excluding the presence of hydraulic jumps where mechanical energy is dissipated and only mass and momentum equations in (115) are valid. Now the bed variation participates accounting for the potential energy. Conservation of total head energy is a principle that is applied along a stream line. In the one-dimensional case the grid is exactly aligned with the stream line, and the differential form of the total energy equation can be used. Conservation of total head H in steady cases is therefore the entropy function selected that will provide exactly balanced numerical schemes.

Apart from integral approaches in (172) and (174) it is possible to evaluate the source term using the following linear combination at each $i+1/2$ RP

$$\bar{S}_z = (1-\Pi)\bar{S}_{z,2} + \Pi\bar{S}_{z,1}\,, \tag{185}$$

with Π to be defined. With independence of the approximate solver selected, momentum conservation in steady conditions at the discrete level can be written as

$$\delta\left(hu^2\right) + g\bar{h}\delta(h+z) = \Pi(\bar{S}_{z,1} - \bar{S}_{z,2}) + \bar{S}_\tau\,. \tag{186}$$

In absence of hydraulic jumps, equation in (184) can be expressed in discrete form as

$$\bar{h}\delta\left[\frac{1}{2}u^2 + g(h+z)\right] = \bar{S}_\tau\,, \tag{187}$$

consistent with the differential form in smooth cases. In order to satisfy both energy and momentum conservation laws under steady conditions, we set $\Pi = \Pi_E$, where the latter is given by

$$\Pi_E = \frac{\delta\left(hu^2\right) - \bar{h}\;\delta\left(\frac{1}{2}u^2\right)}{\bar{S}_{z,1} - \bar{S}_{z,2}}\,, \tag{188}$$

limited by $0 \leq \Pi_E \leq 1$.

The value of weight coefficient Π_E in (188) was defined in [20] to ensure energy conservation away from hydraulic jumps, discriminating between smooth solutions and discontinuous solutions where energy must be dissipated. For more details, see [20].

It is worth mentioning that approach in (188) allows to integrate exactly the slope source term, with independence of the nature of the frictional source terms. The bed slope source term may produce deceleration, acceleration or a change in the flow direction. On the other hand, the friction source term acts dissipating energy and therefore, when numerically integrated, it can not change the sign of the flow velocity within one time step and, in cases of equilibrium with zero velocity, it must preserve the initial state.

Exact solutions involving shear stress under steady state conditions may involve static equilibrium or equilibrium with velocity. For clear water, static equilibrium is produced in cases of zero velocity where the turbulent friction is nil. But for more complex fluids as Bingham type fluids, or fluids presenting Coulomb-type stresses, frictional terms produce a static variable free surface elevation. By means of an appropriate discretization of $\bar{S}_\tau$, they can be exactly integrated, ensuring the well balanced property in quiescent equilibrium with variable free surface elevation [45].

Therefore, even in some simplified situations, such as quiescent equilibrium, it is possible to integrate exactly all source terms. When moving to realistic cases that exercise the complete discretization of all terms in the governing equations, the integration of the source term is anything but trivial. Large discontinuities in the bed elevation and empirical closure relations have to be considered far away from quasi-steady problems, so smooth conditions or small perturbations as in [1] can not be assumed. This means that gross estimations of the source terms may be produced.

The simplicity of the quiescent equilibrium case with discontinuous water level surface in (182) when using an inadequate integral approach of the source term illustrates this effect. As in this case $h^n_{i+1} = 0$, the cell averaging of the approximate solution in cell $i+1$ in the updating step would lead to an unphysical result with independence of the time step selected. In general if the source term $S_2 = S_z + S_\tau$ is not accurately computed, negative values of water depth can appear in the inner states of the approximate solution, requiring the generation of a positivity fix. Also, frictional stress S_τ may be overestimated leading to unphysical oscillations in the solution, requiring thus, generation of a friction fix.

Another important concern when trying to enforce the positivity of the solution is that the augmented solutions, based on an expansion of the linear solution provided for the homogeneous part, may result in unphysical approximate solutions due to the appearance of negative values of $h^\star$ [2]. Even this undesirable behavior is commonly associated to strong expansions involving rarefaction waves [2], it is not exclusively produced in transcritical flows as a result of the strong linearization of the RP, and must be always considered.

Following the philosophy of the entropy fix in [53] to overcome the generation of unphysical approximate solutions in strong expansions, in what follows, a friction fix and a positivity fix of the source terms are presented. Special attention is put when analyzing transcritical flow conditions, as the application of the entropy fix has to be analyzed in coexistence with the positivity fix of the source terms.

6. Friction Fix

While the current tendency is to explore positivity conditions focusing on the bed slope terms, when exercising the complete set of source terms, relations between frictional source terms and bed

source terms have to be first examined, as both terms participate together in the approximate solution through source term integral $\bar{S}_2$.

The friction fix is used to avoid gross estimations of the shear stresses and can be applied with independence of the shear stress model selected. The formulation of the turbulent stress in (175) is a clear example of how friction fixes are mandatory. The empirical coefficient c_f in (175) calibrated for steady flow in channels [54, 55, 56], depends on the water depth in SWE applications, following a power-law model of the form $c_f \propto h^{-\epsilon}$. When applied to unsteady flows involving thin layers of water, friction stress in the bed is overestimated leading to unrealistic values in the integration step. This result has motivated the appearance of semi-implicit formulations [34, 36], unable to preserve exactly balanced solutions. A suitable limitation technique is therefore required.

The definition of the approximate water discharge solution can be used to provide correct values of the total flow resistance in presence of discontinous bed elevation, allowing to ensure equilibrium when necessary. The initial analysis in [45] is extended to subcritical, supercritical and sonic RPs here.

6.1. Friction fix in a subcritical RP

In subcritical RPs the approximate solution for (hu) is given by a constant state, $(hu)^{\downarrow}$ in (127). The definition of the average value of the approximate solution on left side of the RP between $x = \widetilde{\lambda}^1 \Delta t$ and $x = 0$, $(\bar{hu})_i$ and the average value of the approximate solution on right side of the RP between $x = 0$ and $x = \widetilde{\lambda}^3 \Delta t$ are trivial,

$$(\bar{hu})_i = (\bar{hu})_{i+1} = (hu)^{\downarrow}_{i+\frac{1}{2}} \,, \tag{189}$$

reducing the complexity of the problem and allowing to focus exclusively on this value.

Inner constant state $(hu)^{\downarrow}$ considers the effect of both friction and bed slope terms. If neglecting frictional terms in the RP, the following equivalent constant inner state appears

$$(hu)^{\downarrow}_{z,i+\frac{1}{2}} = (hu)^{\star}_{i+\frac{1}{2}} + \frac{1}{2}\left(\frac{\bar{S}_z}{\widetilde{c}}\right)_{i+\frac{1}{2}} \,, \tag{190}$$

as this approximate solution only considers the acceleration effects due to the presence of the source slope term.

Considering that friction terms must work exerting a resistance to the acceleration of the flow, and that the fluid is driven by the existence of gravitational forces, suitable integral evaluations of $\bar{S}_\tau$ ensure

$$\left[(hu)^{\downarrow}_z (hu)^{\downarrow}\right]_{i+\frac{1}{2}} \geq 0 \,, \tag{191}$$

as shear stress acts as an energy dissipation mechanism.

In the limit, friction source term absorbs all the kinetic energy, which results in a nil value of $(hu)^{\downarrow}$. When condition in (191) is not satisfied, $(hu)^{\downarrow}_z (hu)^{\downarrow} < 0$, frictional stress has been overestimated and has to be reduced. This can be done enforcing $(hu)^{\downarrow} = 0$, leading to the following value of $\bar{S}_2$

$$\bar{S}^q_2 = -2\widetilde{c}_{i+\frac{1}{2}} (hu)^{\star}_{i+\frac{1}{2}} \,. \tag{192}$$

In order to avoid incorrect evaluations of the friction source terms, the following friction fix is proposed for the source term

$$\bar{S}_2 = \begin{cases} \bar{S}_2^q & \text{if} \quad \left[(hu)_z^{\downarrow}(hu)^{\downarrow}\right]_{i+\frac{1}{2}} \leq 0 \\ \bar{S}_2 & \text{otherwise} \end{cases} \tag{193}$$

limiting the energy dissipation rate with independence of the shear stress model.

When replacing $\bar{S}_2$ by $\bar{S}_2^q$, the approximate solution changes, $(\bar{hu})_i = (\bar{hu})_{i+1} = 0$, and

$$h_i^{1,-} = h_i \left(1 - \frac{u_i}{\widetilde{\lambda}^1_{i+\frac{1}{2}}}\right) , \qquad h_{i+1}^{3,+} = h_{i+1} \left(1 - \frac{u_{i+1}}{\widetilde{\lambda}^3_{i+\frac{1}{2}}}\right) . \tag{194}$$

In the special case of static initial equilibrium, initial equilibrium is preserved without requiring a zero surface slope, with independence of the rheology model.

6.2. Friction fix in a supercritical RP

Consider a supercritical RP where $\widetilde{u} > 0$. In this RP the discharge is constant in the region defined by $x < \widetilde{\lambda}^1 t$ and given by $(hu)_i$, not affected by the presence of source terms. In the region between $\widetilde{\lambda}^1 t$ and $\widetilde{\lambda}^3 t$, unit discharge is

$$(hu)_{i+1}^{3,+} = (hu)_{i+1}^{2,+} = (hu)^{\star}_{i+\frac{1}{2}} + \frac{1}{2}\left(\frac{\bar{S}_2}{\widetilde{c}}\right)_{i+\frac{1}{2}} . \tag{195}$$

Instead of defining an average value of the discharge on the right side of the plane solution, it is possible to focus on the inner state $(hu)_{i+1}^{3,+}$. Considering that friction terms must work exerting a limited resistance to the acceleration of the flow, it is necessary to enforce that

$$\left[(hu)_{i+1}^{3,+}(hu)_z^{3,+}\right]_{i+\frac{1}{2}} \geq 0 \, , \tag{196}$$

with

$$(hu)_{z,i+\frac{1}{2}}^{3,+} = (hu)^{\star}_{i+\frac{1}{2}} + \frac{1}{2}\left(\frac{\bar{S}_z}{\widetilde{c}}\right)_{i+\frac{1}{2}} \tag{197}$$

only involving gravitational forces, that leads to the same condition in (193). The procedure for a supercritical RP with $\widetilde{u} < 0$ is entirely analogous and generates again condition in (193).

6.3. Friction fix in a sonic RP

In a left transcritical rarefaction the strategy applied avoids reversal water discharge (hu) in $x = 0$, by enforcing

$$\bar{S}_2 = \begin{cases} \bar{S}_2^{q1} & \text{if} \quad (\check{hu})_i^{1,-}(\check{hu})_{z,i}^{1,-} \leq 0 \\ \bar{S}_2 & \text{otherwise} \end{cases} \tag{198}$$

where

$$(\check{hu})_{z,i}^{1,-} = (hu)_i + \left(\alpha\check{\lambda}\right)^1_{i+\frac{1}{2}} + \frac{1}{2}\left(\frac{\bar{S}_z}{\widetilde{c}}\right)_{i+\frac{1}{2}} \tag{199}$$

and

$$\bar{S}_2^{q1} = -2\widetilde{c}_{i+\frac{1}{2}} \left[(hu)_i + \left(\alpha\breve{\lambda}\right)^1_{i+\frac{1}{2}} \right] . \tag{200}$$

In a right transcritical rarefaction the same procedure is repeated,

$$\bar{S}_2 = \begin{cases} \bar{S}_2^{q3} & \text{if} \quad (\breve{hu})^{3,+}_{i+1} (\breve{hu})^{3,+}_{z,i+1} \leq 0 \\ \bar{S}_2 & \text{otherwise} \end{cases} \tag{201}$$

where

$$(\breve{hu})^{3,+}_{i+1} = (hu)_{i+1} - (\alpha\breve{\lambda})^3_{i+\frac{1}{2}} + \frac{1}{2} \left(\frac{\bar{S}_z}{\widetilde{c}} \right)_{i+\frac{1}{2}} \tag{202}$$

and

$$\bar{S}_2^{q3} = -2\widetilde{c}_{i+\frac{1}{2}} \left[(hu)_{i+1} - (\alpha\breve{\lambda})^3_{i+\frac{1}{2}} \right] . \tag{203}$$

This simple and cheap modification of the source term $\bar{S}_2$, not only avoids instabilities when dealing with empirical coefficients modeling dispersive and turbulent effects in the bed and fluid column, but also allows a correct simulation of the stop and go mechanism in debris and granular flows.

It is remarkable that the friction fix provides a lumped modification of the total contributions of the source terms, and does not require to separately quantify the presence of bed level and friction slope effects.

7. Positivity fix in the SWE

In order to avoid unphysical results while restoring a clear selection of the time step size, the strategy proposed here is based on enforcing positive values of the updated solution using the information provided by the approximate solution. Depending on the flow conditions, the positivity fix is enforced by means of a suitable control volume in the cell averaging step, or by analyzing each inner state of the approximate solution separately. In both cases the solution is analyzed focusing on the impact of integral of the total source term $\bar{S}_2$ in $\hat{h}(x,t)$.

As seen in section 4.1 four approximate solutions appear. In the SWE $\widetilde{e}_1^2 = \widetilde{e}_2^2 = 0$, celerity $\widetilde{\lambda}^2$ does not have any impact on the cell averaging of the water depth in the subcritical case, reducing the number of cases to study to three.

7.1. Positivity fix in a subcritical RP

In this case, with independence of the sign of $\widetilde{\lambda}^2$, the solution evolves in both sides of the (x,t) plane solution as illustrated in Figures 8 and 9. In the subcritical case, the following results can be derived:

- Positive values of $h_i^{1,-} \geq 0$ require the following limit over source term integral $\bar{S}_2$ (derived expressing the approximate solution in terms of the source tem)

$$\bar{S}_2 \leq \bar{S}^{sub}_{2,max} \qquad \bar{S}^{sub}_{2,max} = -2(h^\star \widetilde{c} \widetilde{\lambda}^1)_{i+\frac{1}{2}} > 0 \, , \tag{204}$$

  with $h^\star > 0$. In case that $h_i^{1,-}$ becomes negative, $\bar{S}_2$ can be replaced by $\bar{S}^{sub}_{2,max}$ with the following consequences

$$h_i^{1,-} = 0 \qquad h_{i+1}^{3,+} = h_{i+\frac{1}{2}}^{\star} \left(1 - \frac{\widetilde{\lambda}^1}{\widetilde{\lambda}^3} \right)_{i+\frac{1}{2}} > 0 \,, \tag{205}$$

as $\widetilde{\lambda}^1 < 0$ and $\widetilde{\lambda}^3 > 0$, ensuring positive values of water depth on the right side of the plane solution. In case that the left side of the initial solution is dry, $h_i = 0$, if setting $\bar{S}_2 = \bar{S}_{2,max}^{sub}$ the discharge across $x = 0$ becomes nil,

$$(hu)_{i+\frac{1}{2}}^{\downarrow} = 0 \tag{206}$$

and $\hat{\mathbf{U}}(x < 0, t) = \mathbf{U}_i = \mathbf{0}$, with independence of the sign of $\widetilde{\lambda}^2$. Therefore, the approximate solution only varies on the right side of the (x, t) plane, and the modification of the source strengths acts as the imposition of a reflecting or solid wall when $h_i = 0$.

- Positive values of $h_{i+1}^{3,+}$ lead to the following limit over $\bar{S}_2$

$$\bar{S}_2 \geq \bar{S}_{2,min} \qquad \bar{S}_{2,min} = -2(h^{\star}\widetilde{c}\widetilde{\lambda}^3)_{i+\frac{1}{2}} < 0 \,, \tag{207}$$

provide that $h^{\star} > 0$. In case that $h_{i+1}^{3,+}$ becomes negative, $\bar{S}_2$ can be replaced by $\bar{S}_{2,min}^{sub}$ with the following consequences

$$h_{i+1}^{3,+} = 0 \qquad h_i^{1,-} = h_{i+\frac{1}{2}}^{\star} \left(1 - \frac{\widetilde{\lambda}^3}{\widetilde{\lambda}^1} \right)_{i+\frac{1}{2}} > 0 \,, \tag{208}$$

ensuring positive values of water depth in both sides of the solution. In the particular case, with the left side of the initial solution dry, $h_{i+1} = 0$, if enforcing $\bar{S}_2 = \bar{S}_{2,min}^{sub}$, we have

$$(hu)_{i+\frac{1}{2}}^{\downarrow} = 0 \tag{209}$$

and $\hat{\mathbf{U}}(x > 0, t) = \mathbf{U}_{i+1} = \mathbf{0}$, with independence of the sign of $\widetilde{\lambda}^2$. The modification of the source strengths acts as the imposition of a reflecting or solid wall in the wet/dry RP with $h_{i+1} = 0$.

Therefore positive values of $h_i^{1,-}$ and $h_{i+1}^{3,+}$ can be ensured if

$$\bar{S}_{2,min}^{sub} \leq \bar{S}_2 \leq \bar{S}_{2,max}^{sub} \tag{210}$$

when $h^{\star} > 0$. When integral approaches of the source terms are modified following condition in (210) to ensure positivity in the water depth in one side of the RP, positive values of the space solution $\hat{h}(x, t)$ are ensured. When one side of the initial solution is dry, a reflexion boundary condition is generated if limits in (210) are imposed.

7.2. Positivity fix in a supercritical RP with $\widetilde{\lambda}^2 > 0$

In this case, the solution evolves exclusively on the right side of the RP, as depicted in Figure 10. This makes possible to analyze the average approximate solution $\bar{h}^+_{i+1}$ between $x = 0$ and $x = \widetilde{\lambda}^3 \Delta t$ or the inner states $h^{1,+}_{i+1}$ and $h^{3,+}_{i+1}$ separately.

- The average approximate solution $\bar{h}^+_{i+1}$ in the domain $[0, \Delta t] \times [0, \widetilde{\lambda}^3 \Delta t]$ is

$$\bar{h}^+_{i+1} = \frac{h^{1,+}_{i+1} \widetilde{\lambda}^1_{i+\frac{1}{2}} + h^{2,+}_{i+1} (\widetilde{\lambda}^2 - \widetilde{\lambda}^1)_{i+\frac{1}{2}} + h^{3,+}_{i+1} (\widetilde{\lambda}^3 - \widetilde{\lambda}^2)_{i+\frac{1}{2}}}{\widetilde{\lambda}^3_{i+\frac{1}{2}}} \tag{211}$$

  and can be expressed as

$$\bar{h}^+_{i+1} = h_{i+1} - \left(\frac{\delta(hu)}{\widetilde{\lambda}^3} \right)_{i+\frac{1}{2}}, \tag{212}$$

  that is, it is independent of the source term, as in the homogeneous case. Depending on the initial conditions of the RP positive values of $\bar{h}^+_{i+1}$ may not be guaranteed. A positive cell averaging of the solution in the control volume $[0, \Delta t] \times [0, \Delta x]$, given by

$$\bar{h}^+_{i+1} \widetilde{\lambda}^3_{i+\frac{1}{2}} \Delta t + h_{i+1} (\Delta x - \widetilde{\lambda}^3_{i+\frac{1}{2}} \Delta t) \geq 0\,, \tag{213}$$

  can be enforced, leading to the following time step restriction

$$\Delta t \leq \frac{h_{i+1} \Delta x}{|\delta(hu)|_{i+\frac{1}{2}}}, \qquad h_{i+1} > 0\,. \tag{214}$$

- If the updating contributions are analyzed separately, positive values of $h^{3,+}_{i+1}$ lead to the following limit over $\bar{S}_2$

$$\bar{S}_2 \geq \bar{S}^{sup+}_{2,min}\,, \qquad \bar{S}^{sup+}_{2,min} = -2 (h^\star \widetilde{c} \widetilde{\lambda}^3)_{i+\frac{1}{2}} < 0\,, \tag{215}$$

  with $h^\star > 0$. In case that the solution $h^{3,+}_{i+1}$ becomes negative, if $\bar{S}_2$ is replaced by $\bar{S}^{sup+}_{2,min}$,

$$h^{3,+}_{i+1} = 0\,, \qquad h^{1,+}_{i+1} = h_i + 2 \left(\frac{\widetilde{c} h^\star}{\widetilde{\lambda}^1} \right)_{i+\frac{1}{2}} > 0\,, \tag{216}$$

  ensuring positive values of $h^{1,+}_{i+1} > 0$.

- Positive values of $h^{1,+}_{i+1}$ require that

$$h^{1,+}_{i+1} = h_i + 2 \left(\frac{\bar{\beta}^1 \widetilde{c}}{\widetilde{\lambda}^1 \widetilde{\lambda}^3} \right)_{i+\frac{1}{2}} \geq 0\,, \tag{217}$$

  leading to the following limit over $\bar{S}_2$

$$\bar{S}_2 \leq \bar{S}^{sup+}_{2,max}\,, \qquad \bar{S}^{sup+}_{2,max} = h_i(\widetilde{\lambda}^1\widetilde{\lambda}^3)_{i+\frac{1}{2}} > 0\,. \tag{218}$$

In case that the solution $h^{1,+}_{i+1}$ becomes negative, when $\bar{S}_2$ is replaced by $\bar{S}^{sup+}_{2,max}$ the resulting states are given by

$$h^{1,+}_{i+1} = 0 \qquad h^{3,+}_{i+1} = h^{\star}_{i+\frac{1}{2}} + h_i \left(\frac{\widetilde{\lambda}^1}{2\widetilde{c}}\right)_{i+\frac{1}{2}}\,, \tag{219}$$

positive in all cases, provided that $h^\star > 0$.

Summarizing, positive values of all inner states, $h^{1,+}_{i+1}$ and $h^{3,+}_{i+1}$, can be ensured if

$$\bar{S}^{sup+}_{2,min} \leq \bar{S}_2 \leq \bar{S}^{sup+}_{2,max}\,, \tag{220}$$

when $h^\star > 0$. In case that $h^\star$ becomes negative, condition over the time step in (214), derived from the cell averaging procedure, has to be considered additionally to the time step Courant condition in (114), to ensure positivity of the solution.

On the other hand, and contrary to the subcritical case, the modification of the source terms does not generate changes in the value of the discharge across the RP edge at $x = 0$, $(hu)^{\downarrow}_{i+1/2}$, being in this particular case $(hu)^{\downarrow}_{i+1/2} = (hu)_i$. Considering that the left side of the RP remains unchanged, a reflexion condition would lead to the lost of the conservative character of the numerical scheme.

7.3. Positivity fix in a supercritical RP with $\widetilde{\lambda}^2 < 0$

In this case, the average of the approximate solution, $\bar{h}^-_i$ in the domain $[0, \Delta t] \times [\widetilde{\lambda}^1\Delta t, 0]$, or inner states $h^{1,-}_i$ $h^{3,-}_i$ on the left side of the RP, can become negative.

- The average of the approximate solution, $\bar{h}^-_i$, on the left side of the RP is given by

$$\bar{h}^-_i = \frac{-h^{3,-}_i\widetilde{\lambda}^3_{i+\frac{1}{2}} - h^{2,-}_i(\widetilde{\lambda}^2 - \widetilde{\lambda}^3)_{i+\frac{1}{2}} - h^{1,-}_i(\widetilde{\lambda}^1 - \widetilde{\lambda}^2)_{i+\frac{1}{2}}}{-\widetilde{\lambda}^1_{i+\frac{1}{2}}} \tag{221}$$

and reduces to

$$\bar{h}^-_i = h_i + \left(\frac{\delta(hu)}{\widetilde{\lambda}^1}\right)_{i+\frac{1}{2}}\,. \tag{222}$$

Positivity of averaging solution $\bar{h}^-_i$ only depends on the initial conditions of the RP. A positive cell averaging of the control volume given by $[0, \Delta t] \times [-\Delta x, 0]$ requires

$$\bar{h}^-_i(-\widetilde{\lambda}^1_{i+1/2}\Delta t) + h_i(\Delta x + \widetilde{\lambda}^1\Delta t) \geq 0\,, \tag{223}$$

linked to the following time step restriction

$$\Delta t \leq \frac{h_i\Delta x}{|\delta(hu)|_{i+1/2}}, \qquad h_i > 0\,. \tag{224}$$

- If focusing on positive values of $h_i^{1,-}$, the following limit over $\bar{S}_2$ appears

$$\bar{S}_2 \leq \bar{S}_{2,max}^{sup-}\,, \qquad \bar{S}_{2,max}^{sup-} = -2(h^\star \widetilde{c}\widetilde{\lambda}^1)_{i+1/2} > 0\,, \tag{225}$$

  with $h^\star > 0$. In case that the solution $h_i^{1,-}$ becomes negative, if $\bar{S}_2$ is replaced by $\bar{S}_{2,max}^{sup-}$, approximate states are given by

$$h_i^{1,-} = 0\,, \qquad h_i^{3,-} = h_{i+1} - 2\left(\frac{h^\star \widetilde{c}}{\widetilde{\lambda}^3}\right)_{i+1/2} > 0\,, \tag{226}$$

  ensuring positivity.

- Positive values of $h_i^{3,-}$ lead to the following limit over $\bar{S}_2$

$$\bar{S}_2 \geq \bar{S}_{2,min}^{sup-}\,, \qquad \bar{S}_{2,min}^{sup-} = -h_{i+1}(\widetilde{\lambda}^1\widetilde{\lambda}^3)_{i+\frac{1}{2}} < 0\,, \tag{227}$$

  In case that the solution $h_i^{3,-}$ becomes negative, if setting $\bar{S}_2 = \bar{S}_{2,min}^{sup-}$ the approximate solution is given by

$$h_i^{3,-} = 0\,, \qquad h_i^{1,-} = h_{i+\frac{1}{2}}^\star - h_{i+1}\left(\frac{\widetilde{\lambda}^3}{2\widetilde{c}}\right)_{i+\frac{1}{2}}\,, \tag{228}$$

  ensuring positive values in all cases, if $h^\star > 0$.

In general, positive values of $h_i^{1,-}$ and $h_i^{3,-}$ can only be obtained if

$$\bar{S}_{2,min}^{sup-} \leq \bar{S}_2 \leq \bar{S}_{2,max}^{sup-}\,, \tag{229}$$

provide that intermediate value $h_{i+1/2}^\star$ is greater than zero. If $h^\star < 0$, condition over the time step in (224) has to be considered additionally to the Courant condition in (114). The flow being supercritical, the value of the discharge across the RP edge remains constant $(hu)_{i+1/2}^{\downarrow} = (hu)_{i+1}^n$, not affected by the presence or the possible modification of the source terms.

7.4. Positivity fix in the left transcritical rarefaction

If enforcing positivity in the solution, the average solution and the inner states can be analyzed.

- On the left hand side of the solution, positive values of $\check{h}_i^-$ lead to the following limit over $\bar{S}_2$

$$\bar{S}_2 \leq \bar{S}_{2,max}^{e1}\,, \qquad \bar{S}_{2,max}^{e1} = -2(h^\star \widetilde{c}\check{\lambda}^1)_{i+\frac{1}{2}} > 0\,, \tag{230}$$

  in case that $h^\star > 0$.

- The positivity conditions over the problem are explored averaging the approximate solution on the right region. Average value $\bar{h}^{+}_{i+1}$ is given by

$$\bar{h}^{+}_{i+1} = \frac{\breve{h}^{+}_{i+1}\breve{\lambda}^{1}_{i+\frac{1}{2}} + h^{3,+}_{i+1}(\widetilde{\lambda}^{3} - \breve{\lambda}^{1})_{i+\frac{1}{2}}}{\widetilde{\lambda}^{3}_{i+\frac{1}{2}}} \tag{231}$$

and leads to the following limit over $\bar{S}_2$

$$\bar{S}_2 \geq \bar{S}^{e1}_{2,min}\,, \qquad \bar{S}^{e1}_{2,min} = 2(\widetilde{c}\widetilde{\lambda}^{3})_{i+\frac{1}{2}} \left(\alpha^{1}\frac{\breve{\lambda}^{1}}{\widetilde{\lambda}^{3}} - h^{\star} \right)_{i+\frac{1}{2}}\,, \tag{232}$$

that depends on the initial conditions of the RP.

Therefore source terms must be limited by

$$\bar{S}^{e1}_{2,min} \leq \bar{S}_2 \leq \bar{S}^{e1}_{2,max} \tag{233}$$

in cases where initial conditions ensure $\bar{S}^{e1}_{2,min} \leq \bar{S}^{e1}_{2,max}$ and $h^{\star} > 0$. Otherwise it is necessary to enforce a time step restriction to ensure a positive cell average value of water depth.

Unphysical approximate solutions of water depth, with $h^{\star} < 0$, may appear in strong expansions involving rarefaction waves [2]. When exercising the complete set of source terms involving bed level variations, rarefaction waves linked to drying processes may lead to transcritical flows. In that case, erroneous estimations of the source terms may be combined with negative predictions of $h^{\star}$. Numerical practice shows how these cases are related with drying processes involving very thin layers of water and high velocity. Considering that inner states may be grossly estimated in this situation, the alternative strategy is to avoid the entropy correction when condition (233) fails, and handle the RP as subcritical or a supercritical RP, depending on the initial conditions.

It is worth mentioning that when trying to reproduce the physics of thin layers, the effects of surface tension must be considered, using for instance, the model of thin liquid films [30]. The physically relevant scales considered in the SWE model are much larger, and the reproduction of the movement of very thin layers of water only makes sense in theoretical problems. Therefore, when solving the SWE in their application range, the strategy proposed does not affect to the quality of the numerical results. Also, a clear and efficient selection of the time step is restored.

7.5. Positivity fix in a right transcritical rarefaction

The properties of the average solution and the inner states are analyzed.

- Positive values of water depth on the right side of the solution, $\breve{h}^{3,+}_{i+1}$, lead to the following limit over $\bar{S}_2$

$$\bar{S}_2 \geq \bar{S}^{e3}_{2,min}\,, \qquad \bar{S}^{e3}_{2,min} = -2(h^{\star}\widetilde{c}\breve{\lambda}^{3})_{i+\frac{1}{2}} \leq 0\,, \tag{234}$$

provide that $h^{\star} > 0$.

- Instead of enforcing positive values of both $\check{h}_i^{3,-}$ and $h_i^{1,-}$ (as $h_i^{2,-} = \check{h}_i^{3,-}$), and considering that the solution is cell averaged, it is sufficient to explore the average right solution $\bar{h}_{i+1}^{+}$ given by

$$\bar{h}_{i+1}^{+} = \frac{-\check{h}_i^{3,-}\check{\lambda}_{i+\frac{1}{2}}^{3} - h_i^{1,-}(\widetilde{\lambda}^1 - \check{\lambda}^3)_{i+\frac{1}{2}}}{-\widetilde{\lambda}_{i+\frac{1}{2}}^{1}}\,, \tag{235}$$

leading to the following limit over $\bar{S}_2$

$$\bar{S}_2 \leq \bar{S}_{2,max}^{e3}\,, \qquad \bar{S}_{2,max}^{e3} = -2(\widetilde{c}\widetilde{\lambda}^1)_{i+\frac{1}{2}}\left(\alpha^3\frac{\check{\lambda}^3}{\widetilde{\lambda}^1} + h^\star\right)_{i+\frac{1}{2}}\,. \tag{236}$$

As a conclusion, in cases with a right transcritical rarefaction, when initial conditions ensure $\bar{S}_{2,min}^{e} \leq \bar{S}_{2,max}^{e}$ and $h^\star > 0$, source terms may be limited by

$$\bar{S}_{2,min}^{e3} \leq \bar{S}_2 \leq \bar{S}_{2,max}^{e3}\,, \tag{237}$$

otherwise, the limits defined for the source term integral $\bar{S}_2$ in (237) lead to contradictory results. If condition (237) can not be applied, positivity in the solution may be provided by a reduction of the time step. As done for the left transcritical rarefaction, another possible strategy is to avoid entropy correction and use the subcritical or supercritical approximate solutions.

8. Extension to multi-dimensions

The 2D extensions of the model is formulated as follows

$$\frac{\partial \mathbf{W}}{\partial t} + \frac{\partial \mathbf{F}_1(\mathbf{W})}{\partial x_1} + \frac{\partial \mathbf{F}_2(\mathbf{W})}{\partial x_2} = \mathbf{T}\,, \tag{238}$$

where

$$\mathbf{W} = \begin{pmatrix} h \\ hu_1 \\ hu_2 \end{pmatrix}\,, \quad \mathbf{F}_1 = \begin{pmatrix} hu_1 \\ hu_1^2 + \frac{1}{2}gh^2 \\ hu_1u_2 \end{pmatrix}\,, \quad \mathbf{F}_2 = \begin{pmatrix} hu_2 \\ hu_2u_1 \\ hu_2^2 + \frac{1}{2}gh^2 \end{pmatrix}\,, \tag{239}$$

with

$$\mathbf{T} = \begin{pmatrix} 0, & -gh\frac{\partial z}{\partial x_1} - c_f u_1|\mathbf{u}|, & -gh\frac{\partial z}{\partial x_2} - c_f u_2|\mathbf{u}| \end{pmatrix}^T\,, \tag{240}$$

with (u_1, u_2) the depth averaged components of the velocity along the x_1 and x_2 coordinates respectively and $|\mathbf{u}| = \sqrt{u_1^2 + u_2^2}$.

The computational domain is divided in 2D cells, shaped by NE edges, where NE stands for number of edges. Vector $\mathbf{n}_{i,k} = (n_1, n_2)$ indicates the outward unit normal vector to the cell i at edge k and l_k is the corresponding edge length. To introduce the finite volume scheme, (238) is integrated in a grid cell Ω using Gauss-Ostrogradsky's theorem [57]

$$\frac{\partial}{\partial t}\int_{\Omega_i} \mathbf{W}d\Omega + \sum_{k=1}^{NE}(\mathbf{F}_1 n_1 + \mathbf{F}_2 n_2)_k l_k = \int_{\Omega_i} \mathbf{T}d\Omega\,. \tag{241}$$

The approximate solution of each RP is reduced at each k edge to a 1D Riemann problem projected onto the direction $\mathbf{n}$, defined along the reference coordinate, x, parallel to vector $\mathbf{n}_{i,k}$. Thus, the following normal flux is defined

$$\mathbf{F}_1 n_1 + \mathbf{F}_2 n_2 \,, \tag{242}$$

while bed and friction slope source terms are expressed in the normal direction as follows [14]

$$\mathbf{T_n} = \begin{pmatrix} 0, & T_{\mathbf{n}} n_1, & T_{\mathbf{n}} n_2 \end{pmatrix}^T \,, \tag{243}$$

with $T_{\mathbf{n}} = (-gh\partial z_x - c_f \mathbf{un}|\mathbf{u}|)$.

The SWE's satisfy the rotational invariance property [57],

$$\mathbf{F}_1 n_1 + \mathbf{F}_2 n_2 = \mathbf{R}^{-1}\mathbf{F}_1(\mathbf{RW}) \,, \tag{244}$$

where $\mathbf{R}$ is the rotation matrix

$$\mathbf{R} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & n_1 & n_2 \\ 0 & -n_2 & n_1 \end{pmatrix} \,, \tag{245}$$

that provides the conserved variable $\mathbf{U} = \mathbf{RW}$ and flux $\mathbf{F} = \mathbf{F}_1(\mathbf{U}) = \mathbf{F}_1(\mathbf{RW})$, recovering the x-split two-dimensional SWE's in (115).

The normal velocity through the interface is given now by $u = u_1 n_1 + u_2 n_2$ and the tangential velocity is $v = -u_1 n_2 + u_2 n_1$. If the rotation matrix is applied to the source term, the equivalent source term vector appears [14]

$$\mathbf{S} = \mathbf{RT_n} = \begin{pmatrix} 0, & S_2, & 0 \end{pmatrix}^T \,. \tag{246}$$

Thus, the 2D-SWE's are written in a one-dimensional form and the numerical scheme is written using the equivalent form of the corresponding intercell flux for the approximate first order Godunov method in (38) as follows

$$\mathbf{W}_i^{n+1} = \mathbf{W}_i^n - \sum_{k=1}^{NE} \mathbf{R}^{-1}\mathbf{F}_{i,k}^- \frac{\Delta t \; l_k}{A_i} \,, \tag{247}$$

where the numerical flux $\mathbf{F}_{i,k}^-$ can be computed using the solver presented previously, allowing to express the numerical scheme in fluctuation form

$$\mathbf{W}_i^{n+1} = \mathbf{W}_i^n - \sum_{k=1}^{NE} \mathbf{R}^{-1}(\delta\mathbf{M}_{i,k}^-) \frac{\Delta t \; l_k}{A_i} \,, \tag{248}$$

as done for the 1D case in (91).

As in the 1D case, the updated value is defined by cell averaging the contributions of the local RPs shaping the contour cell, and condition in (114) has to be modified. In the 2D framework, considering unstructured meshes, the relevant distance equivalent to Δx, will be referred to as χ_i in each cell i. This distance considers the volume of the cell and the length of the shared k edges

$$\chi_i = \frac{A_i}{\max_{k=1,NE} l_k} \tag{249}$$

Considering that each k RP is used to deliver information to a pair of neighboring cells of different size, the distance $\min(A_i, A_j)/l_k$ is relevant. The time step is limited by

$$\Delta t \leq CFL\ \Delta t^{\tilde{\lambda}} \qquad \Delta t^{\tilde{\lambda}} = \frac{\min(\chi_i, \chi_j)}{\max(\widetilde{\lambda}_k^m)} \tag{250}$$

with CFL=1/2, as the construction of finite volume schemes from direct application of one-dimensional fluxes leads to reduced stability ranges [48]. Friction fixes and positivity constraints over the source terms remain unaltered.

9. Applications.

9.1. Accelerating supercritical flow over a downward slope

The numerical techniques presented here ensure the positivity of the water depth while preserving energy when necessary, with independence of the grid size. Convergence to the exact solution in unsteady RPs even in resonant cases, and in steady problems with sub/supercritical conditions and hydraulic jumps as shown in previous works [20, 21] is provided when using the numerical improvements presented here and are not recalled for the sake of brevity. Also, the performance of the approximate solver used here, was analyzed using a set of realistic 1D open channel flow test cases with analytical solution very well suited to validate the numerical schemes involving bed variations and bed friction [59]. Numerical solutions for frictional dominant problems was compared with experimental data in [45]. As the purpose of this paper is to focus on 2D problems with transient wet/dry boundaries those cases are not repeated here.

Although the well-balanced property is a particular case of the E-property, numerical schemes based on hydrostatic reconstructions for quiescent equilibrium are still widely used. In order to evidence the consequences in the choice of the type of solver selected, in this section, numerical solutions for steady supercritical flow over an inclined plane with constant slope provided by an augmented E-scheme are compared to the solutions provided by the popular finite volume scheme of Audusse et al., based on the so-called hydrostatic reconstruction approach [7]. Such technique is designed to ensure the well-balanced property, that is, to ensure quiescent equilibrium for water at rest. Moreover, it satisfies other physical properties such as the positivity of the water depth and the semidiscrete entropy inequality. Recently, Audusse et al. [60] showed that this scheme, when used with the classical kinetic solver, also satisfies a fully discrete entropy inequality, with an error term, which proves that the hydrostatic reconstruction should converge with mesh refinement. Numerical schemes using the hydrostatic reconstruction are designed to provide accurate results when dealing with near-hydrostatic situations and are easy to implement, for this reason, they have been widely used. However, as reported in [61], first order schemes using the hydrostatic reconstruction are unable to provide accurate results in some particular cases with moving water, such as steady flow over a constant slope. On the other hand, it is worth pointing out that the accuracy of such schemes can be importantly improved by increasing the order of the numerical scheme to second order.

In the test case proposed in this section, numerical fluxes for the scheme in [7] are computed by means of the traditional Roe solver applied to the hydrostatic variables.The bed profile consists in an inclined plane given by

$$z(x) = -0.01\alpha x \tag{251}$$

defined inside the domain $[0, 10]$ m, where the coefficient α is a constant value that represents the slope of the plane as a percentage. Inflow conditions are given by

$$h(0,t) = 0.02\,\mathrm{m} \qquad q(0,t) = 0.01\,\mathrm{m}^2/\mathrm{s} \tag{252}$$

and outflow conditions are not required as the flow remains supercritical along the whole domain. The computational domain is discretized in 100 computational cells of $\Delta x = 0.1$ m and CFL number is set to 0.8. The solution is presented for $\alpha = 1.5\%$, 3%, 6%, 9%, 12%, 15% and 18% at time $t = 600$ s, which is sufficient time for the numerical schemes to reach the steady state. Numerical results for water depth and discharge provided by the ARoe E-scheme and the well-balanced hydrostatic scheme in [7] are presented in Figure 16, on left and right position respectively, and are compared with the exact solution for the different combinations of slopes.

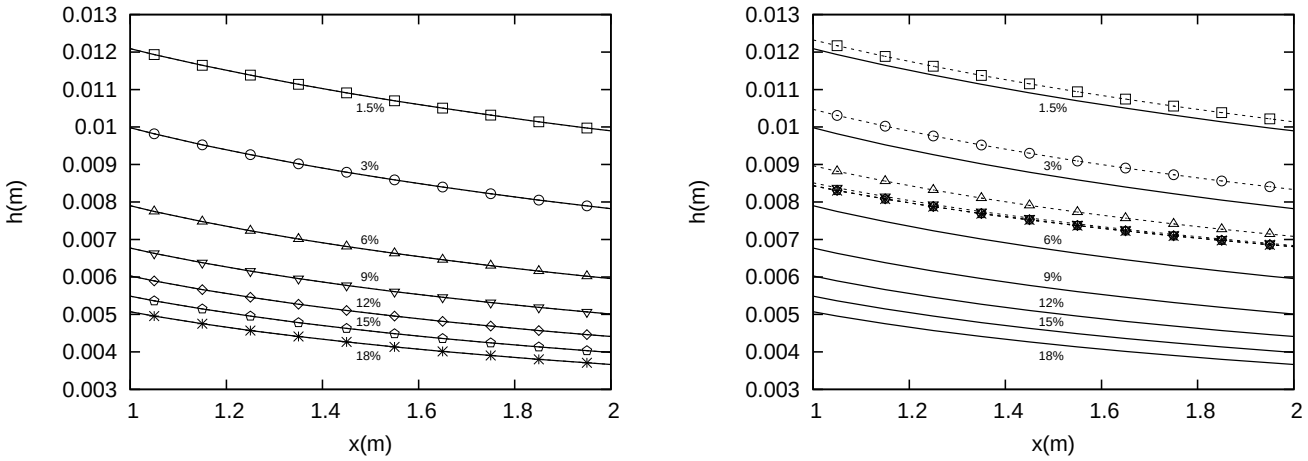

Figure 16: Section 9.1. Numerical solution for water depth setting $\alpha = 1.5\%$, 3%, 6%, 9%, 12%, 15% and 18% using the ARoe E-scheme (left) and the well-balanced hydrostatic scheme in [7] (right), with $\Delta x = 0.1$. Exact solutions are represented by continuous lines.

Results in Figure 16 evidence that the scheme from [7] may work well when dealing with very low slopes ($< 1.5\%$) but it exhibits a significant lack of accuracy when increasing the slope, unlike the ARoe E-scheme that provides the exact solution for any slope and with independence of the grid. Figure 17 depicts the numerical solution provided by the scheme in [7] when setting $\alpha = 3\%$ (left) and 12% (right) for different mesh sizes, namely $\Delta x = 0.1$, 0.05, 0.025 and 0.0125 m. From this figure, we notice first that the greater the slope, the less accurate the solution. It is also worth noticing that, even with grid refinement, the numerical scheme is unable to converge to the exact solution in this particular case. Numerical results for the second order version of this hydrostatic scheme are provided in [61] for the same test case, showing a enhanced accuracy and convergence to the exact solution. Moreover, theoretical results recently presented in [60] prove that the hydrostatic reconstruction, in combination with the classical kinetic solver, should converge to the exact solution as it satisfies a fully discrete entropy inequality with an error term.

9.2. Planar surface in a circular parabolic frictionless basin

The following test case has an exact solution and allows to compare the performance of the numerical improvements presented here. The solution includes transient boundaries dominated by the non-conservative terms and has a periodic solution that depending on the region involves subcritical and supercritical conditions as well as sonic transitions. A frictionless parabolic topography is defined by the depth function [62]

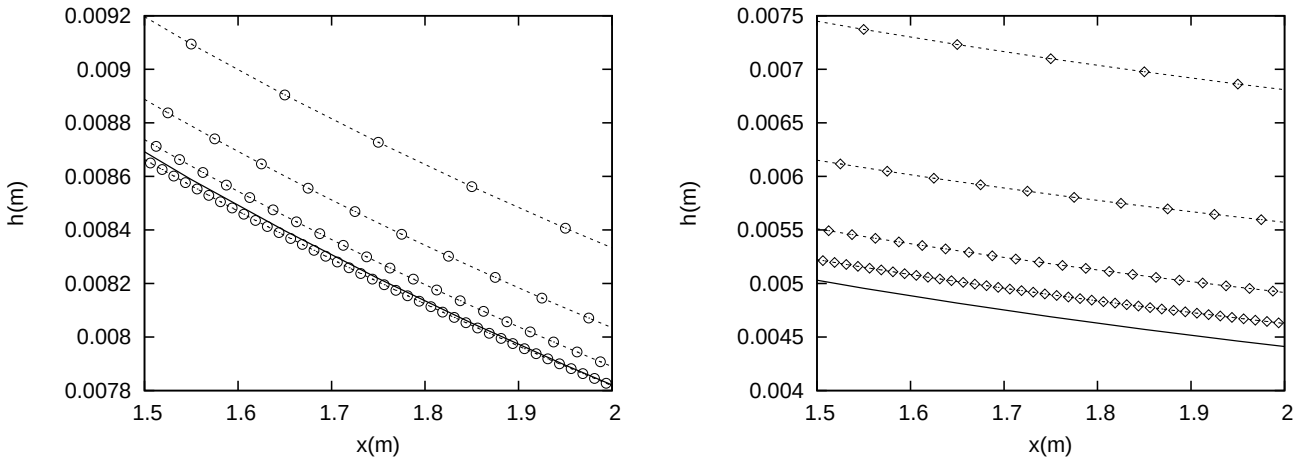

Figure 17: Section 9.1. Numerical solution for water depth for $\alpha = 3\%$ (left) and 12% (right) using the well-balanced hydrostatic scheme in [7] and four different cell sizes $\Delta x = 0.1$, 0.05, 0.025 and 0.0125 m.

$$z(x,y) = -z_0 \left(1 - \frac{x^2+y^2}{a^2}\right) , \tag{253}$$

together with the periodic analytical solution

$$\begin{array}{c} h(x,y,t) = \max\left(0, \frac{\sigma z_0}{a^2}(2x\cos(\omega t) + 2y\sin(\omega t) - \sigma) - z(x,y)\right) , \\ u(x,y,t) = \sigma\omega\sin(\omega t) , \qquad v(x,y,t) = \sigma\omega\cos(\omega t) , \end{array} \tag{254}$$

where u and v are the velocities in the x and y directions respectively and $\omega = \sqrt{2gz_0}/a$. This test case is useful to explore the limits of the approximated Riemann solver in 2D problems where the dry/wet boundaries are present and can be considered among the most difficult cases for a numerical model [26]. Numerical results for this test case involving the proposed improvements are compared with the numerical results provided by the initial solver presented in [32]. In [32], the influence of the source terms in the positivity of the solution was considered only in subcritical RP problems as under supercritical conditions the average solution for the water depth becomes unaltered by their presence. Analysis of sonic transitions was omitted in this previous work. In this test case, sonic transitions strongly affect the stability region in vacuum or drying RPs in wet/dry edges. The stability region in [32] was recovered by imposing systematically solid walls conditions in wet/dry RPs with discontinuous adverse slope. With the numerical strategies presented here, the stability region is recovered avoiding extra solid walls conditions.

Following [14] constant parameters are $a = 1000$ m, $\sigma = 300$ m, $g = 9.80665$ ms^{-2} and $z_0 = 10$ m. Rotation period is $T = 448.57$ s. A squared domain 3000×3000 m^2 is divided in 2×200^2 triangular cells built by drawing the diagonals on a quadrilateral grid and using Cartesian discretizations of the domain. CFL is set equal to 0.8 in all simulations. Figure 18 shows 3D contour plots of the computed water level surface at times (a) $t = T/4$, (b) $t = T/2$, (c) $t = 3T/4$ and (d) at $t = T$ using the numerical techniques described in the present work.

The level surface is kept planar throughout the computation as shown in Figure 18 and the moving shoreline is correctly captured without oscillations. Numerical results for the water level surface, $h + z$, given by the present solver and solver in [32] are compared with the exact solution in Figure 19. Numerical solutions are plotted along a section crossing the center of the domain keeping constant the y coordinate. In Figure 20 and Figure 21, comparisons for both the u and v velocities

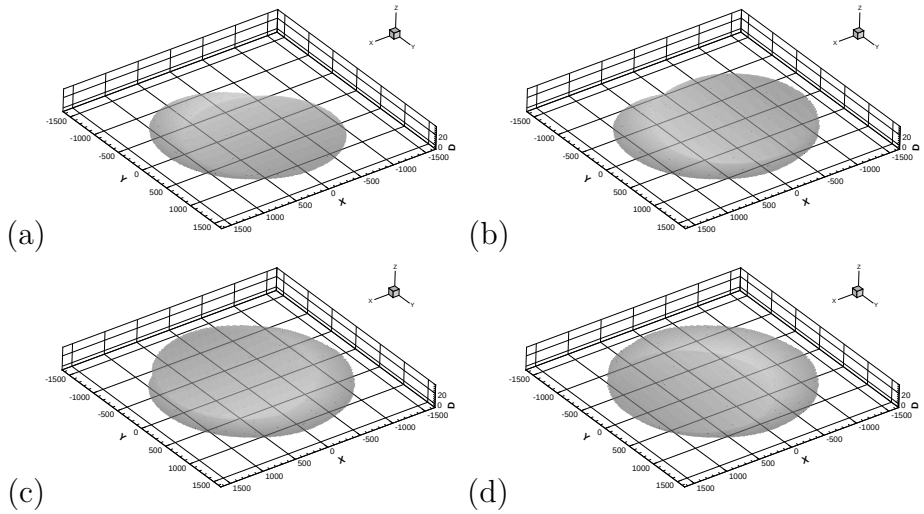

Figure 18: Section 9.2. Water level surface at (a) $t = T/4$, (b) $t = T/2$, (c) $t = 3T/4$ and (d) at $t = T$, with a rotation period T=448.57s.

are shown. Computational time step along the simulation for the present solver and solver in [32] are compared in Figure 22. No significant differences are found between two solvers regarding the selection of the time step. Considering that the initial velocity and water depth fields are displaced following a rotation movement, the novel solver is able to retain an approximate constant value of time step, as expected.

In the present solver solid wall conditions are neglected, but numerical simulation remains stable and surface variation in time is accurately followed. No remarkable differences appear in the prediction of the water level surface if compared with solver in [32]. On the other hand, velocity predictions are better estimated by the present solver, being much closer to the exact solution as shown in Figures 20 and 21. The differences become easily observable if examining the resulting Froude number. Figure 23 shows a 2D contour plot of (a) the exact solution and computed solutions using (b) present solver and (c) solver in [32], after a rotation period. Despite of the presence of extremely large variations of Froude number in the vicinity of the shoreline, the present solver provides a much more accurate solution than initial solver in [32]. Numerical solutions for the water depth evolution in time at different locations (x=0, 150, 300, 450, 600, 750, 900, 1050, 1200, 1350 m and $y = 0$ m) are plotted in Figure 24 and compared with the exact solutions for four periods. Although after each period numerical diffusion attenuates the water depth, the wet/dry transient boundary is accurately tracked in time. This result confirms that the influence of the source terms in the inner states of the approximate solution has to be analyzed in sonic and supercritical conditions. Supercritical conditions may lead to misleading conclusions, as the average approximate solution for the water depth remains unaltered by the presence of source terms in this particular case. Initial water volume is exactly conserved along the whole simulation without requiring any threshold parameter.

9.3. Experimental spreading of granular mass over a fixed rough inclined plane

A dense granular flow over a rough inclined plane is considered here [63, 64]. The experiment was carried out over a rough inclined plane and the initial condition was defined by a spherical surface with a maximum depth of 3.1 cm. Granular material was composed of glass beads with 0.5 mm $\pm$ 0.04 in diameter. Shear stress is modeled here involving two components: an internal friction angle equal to 23^o and a turbulent dissipation factor provided by a Manning coefficient equal to n=0.03

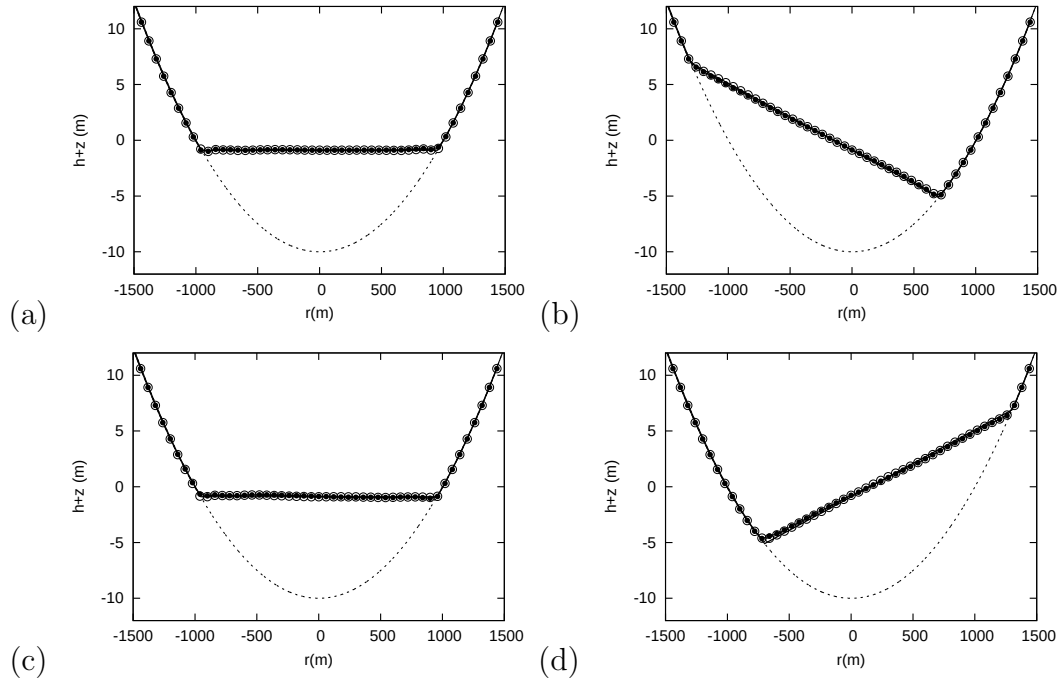

Figure 19: Section 9.2. Exact solution along the cross-section $y = 0$ (——) and computed solution using present solver $(- \circ -)$ and solver in [32] $(- \bullet -)$ for the water level surface at (a) $t = T/4$, (b) $t = T/2$, (c) $t = 3T/4$ and (d) at $t = T$, with a rotation period T=448.57s.

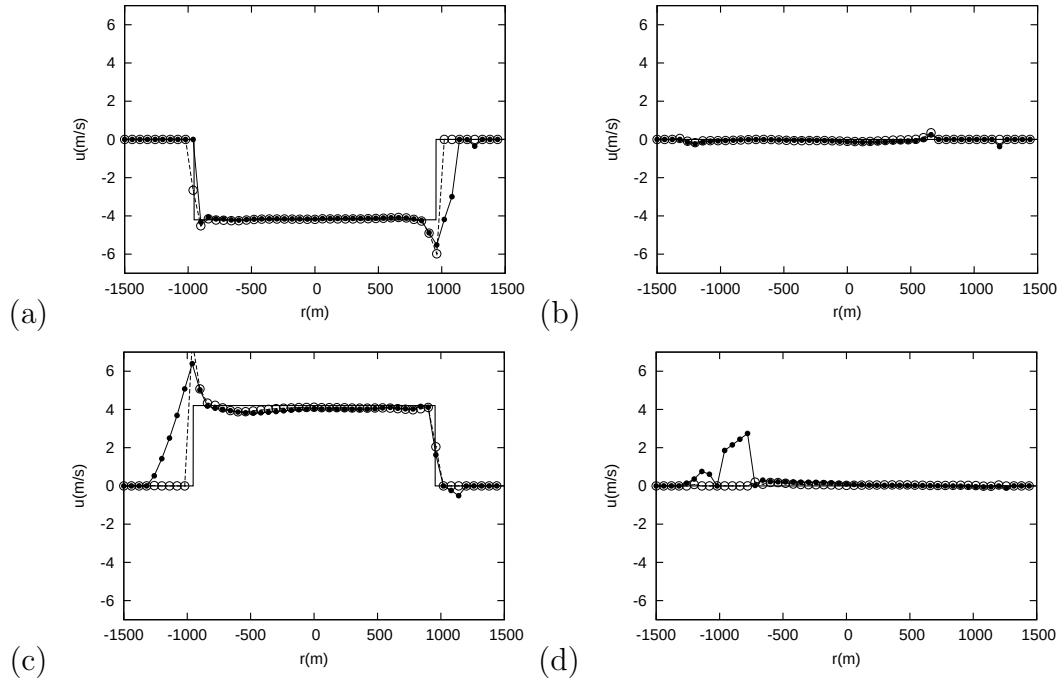

Figure 20: Section 9.2. Exact solution along the cross-section $y = 0$ (——) and computed solution using present solver $(- \circ -)$ and solver in [32] $(- \bullet -)$ for velocity u at (a) $t = T/4$, (b) $t = T/2$, (c) $t = 3T/4$ and (d) at $t = T$, with a rotation period T=448.57s.

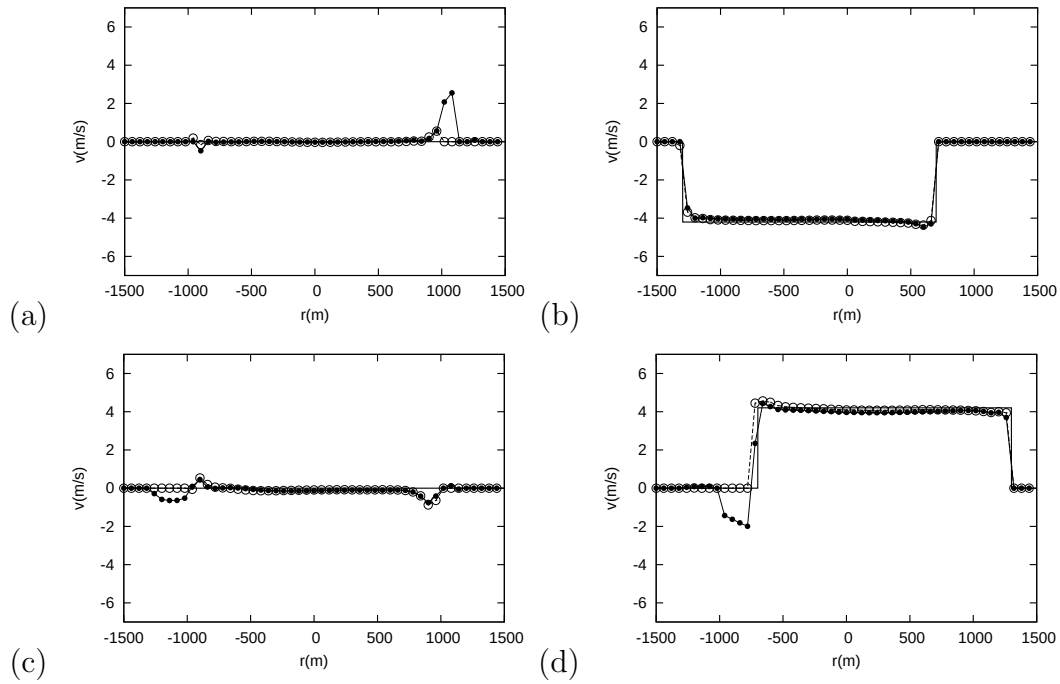

Figure 21: Section 9.2. Exact solution along the cross-section $y = 0$ (——) and computed solution using present solver $(-\circ-)$ and solver in [32] $(-\bullet-)$ for velocity v at (a) $t = T/4$, (b) $t = T/2$, (c) $t = 3T/4$ and (d) at $t = T$, with a rotation period T=448.57s.

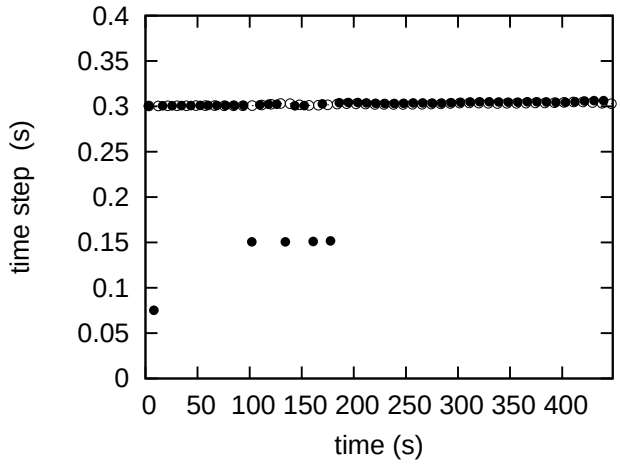

Figure 22: Section 9.2. Time step using present solver $(-\circ-)$ and solver in [32] $(-\bullet-)$ in a rotation period T=448.57s.

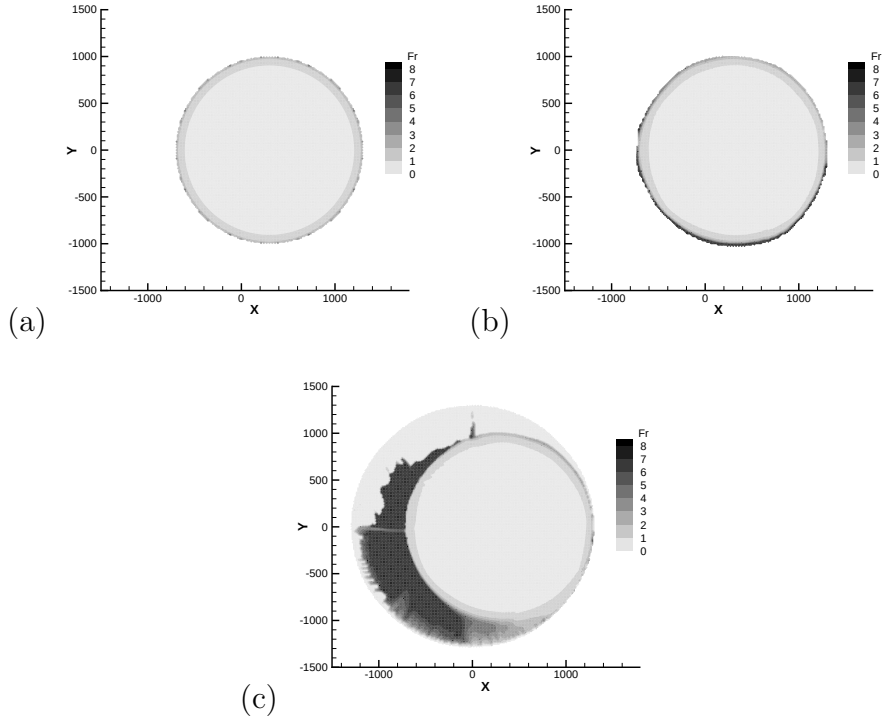

Figure 23: Section 9.2. Exact solution (a) and computed solution using present solver (b) and solver in [32] (c) for Froude number at $t = T$, with a rotation period T=448.57s.

$s^2m^{-2/3}$. Numerical results are compared with the recorded temporal evolution of the free surface level within a period of 6 s.

This numerical test case allows a complete exercise of all source terms. Numerical simulation involves transient RPs ranging from subcritical to supercritical conditions. Correct modeling of start/stop flow conditions is mandatory. An unstructured mesh, generated by triangular cells with an area of 2 mm^2, is used to discretize the numerical domain. The numerical scheme in [64] based in [32] and the present solver will be used to predict the spreading of the granular flow influenced by a longitudinal slope. The CFL is set equal to 0.8 in all cases. Figure 25 shows 3D contour plots of the numerical results for the free surface level when using the present solver. Initially the mass is put in motion and spreads over the longitudinal and transversal direction. Then, flow is oriented to the steeper direction and the mass flow is stretched longitudinally. The tail of the flow keeps at rest whereas the front of the flow propagates down the initial deposit.

Figure 26 shows the calculated free surface depth at different instants of time, with a slope angle equal to 23^o using present solver $(-\circ-)$ and solver in [64]$(-\bullet-)$. Numerical results are in agreement with the experimental data $(-\blacktriangle-)$. Although both solvers provide almost identical results, important differences between them arise. Previous solver in [64] does not involve all type of possible transitions considered in the present solver, as only subcritical conditions were considered in [32]. Also, solver in [64] required the use of limiting values of energy dissipation generated by frictional terms, provided by the maximum kinetic energy allowable in each cell. Solid wall conditions were also supplied when necessary to improve stability as in [32]. In the present solver, both conditions are omitted without losing accuracy neither compromising numerical stability. The spreading of the granular flow is accurately tracked in time by the present numerical scheme, allowing to provide good predictions of the final thickness layer and of the maximum run out of the flow. Also, numerical solutions provide a

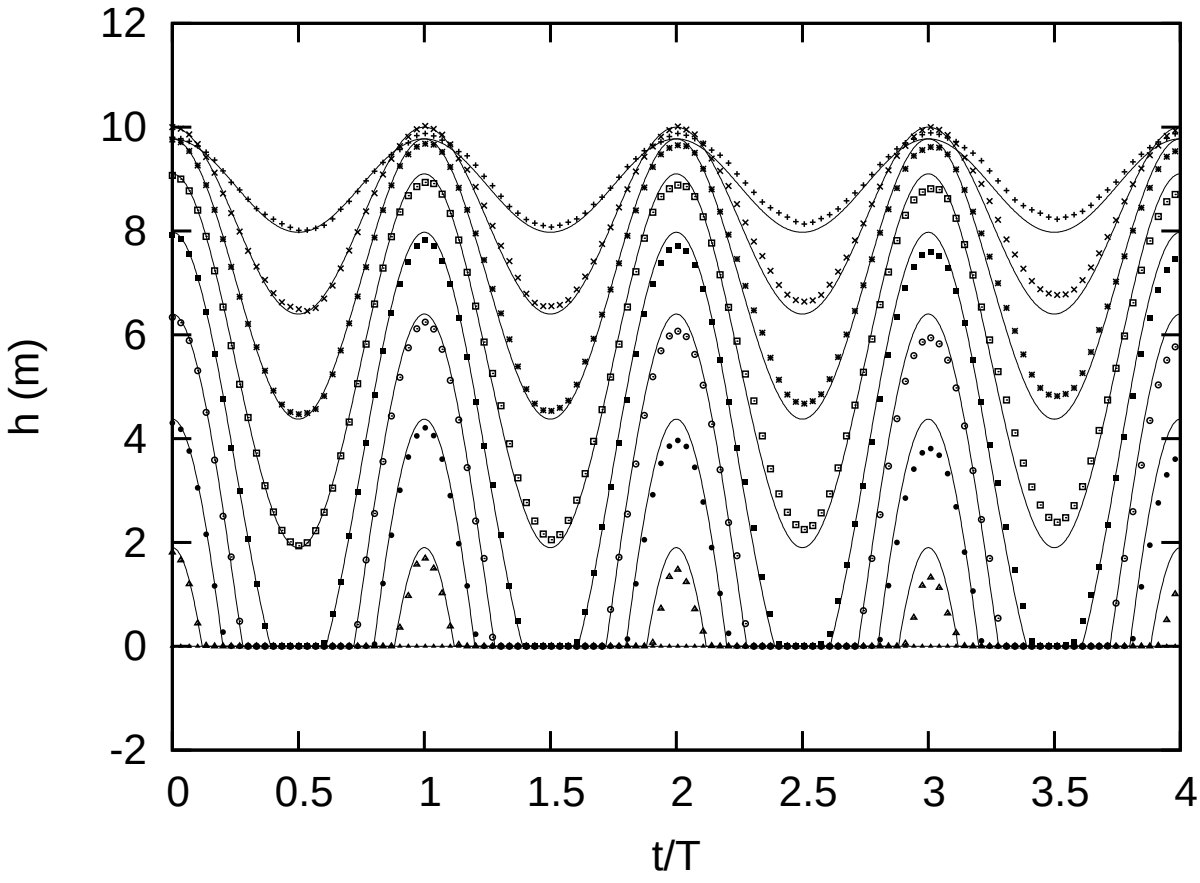

Figure 24: Section 9.2. Numerical solutions (dots) for the water depth evolution in time at different locations (x=0, 150, 300, 450, 600, 750, 900, 1050, 1200, 1350 m and $y = 0$ m) and exact solutions (continuous line) for four periods.

correct tracking of the front celerity. Numerical results point out that the present solver can be used to predict faithfully the overall behavior of complex phenomena.

10. Conclusions

In the present work the ARoe solver proposed in [32] is revisited, presenting significant improvements for the complete description of this approximate Riemann solver.

With the ARoe solver, the solution is not only provided at cell interfaces, but the full wave structure of the solution is defined, as it does participate in the updating step. The updating scheme is presented in this work in both the intercell flux form and in the fluctuation form, including a proper description of the approximate numerical flux based on the definition of RH conditions across each wave and using the inner states.

It is shown that the use of well balanced numerical schemes limits the range of application of the numerical solver when moving to realistic scenarios. Here it is proposed to use energy as entropy function in the discretization of the bed level source, ensuring that the numerical scheme converges to a physically based solution of reference with mesh refinement. As a result, an E-scheme is derived, where the complete set of source terms is fully exercised under any flow condition involving high slopes and arbitrary shear stress.

The present ARoe solver provides a convenient way to evaluate source term discretization in transient situations and now includes extra fix procedures that can be used not only to ensure positively conservative solutions, but also to define friction fix techniques able to ensure an accurate viscous dissipation rate. Positivity conditions are explored under a general framework and numerical simulations can be accurately performed recovering an appropriate selection of the time step, allowed by a detailed analysis of the approximate solver. The use of case-dependent threshold values is unnecessary and exact mass conservation is preserved.

As a result of including the proposed improvements in the original ARoe solver, the novel solver can be used as a verification tool for any case where a source term is involved. The present work has focused on its application to shallow flows, showing clear improvements in the numerical results when comparing with the original solver and a faithful performance in complex test cases that involve subcritical and supercritical conditions as well as sonic transitions.

Acknowledgment

This work has been funded by the Spanish Ministerio de Economía y Competitividad under research project CGL2015–66114–R.

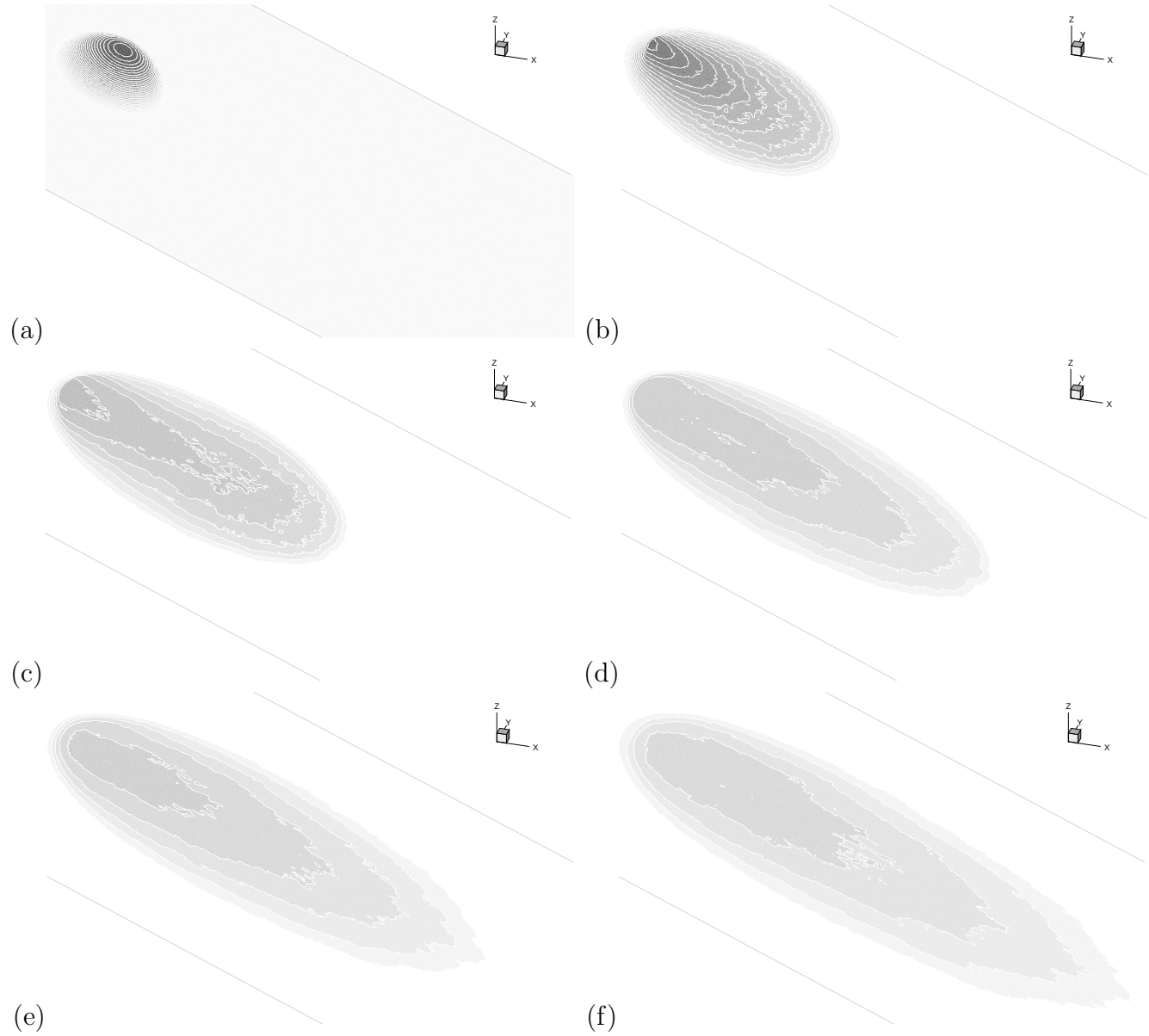

Figure 25: Section 9.3. Temporal evolution: contours of constant thickness every 1.0 mm at times (a) $t = 0$, (b) $t = 0.24$, (c) $t = 0.48$, (d) $t = 0.96$, (e) $t = 2.40$ s and (f) $t = 6$ seconds with a slope angle of 23^o.

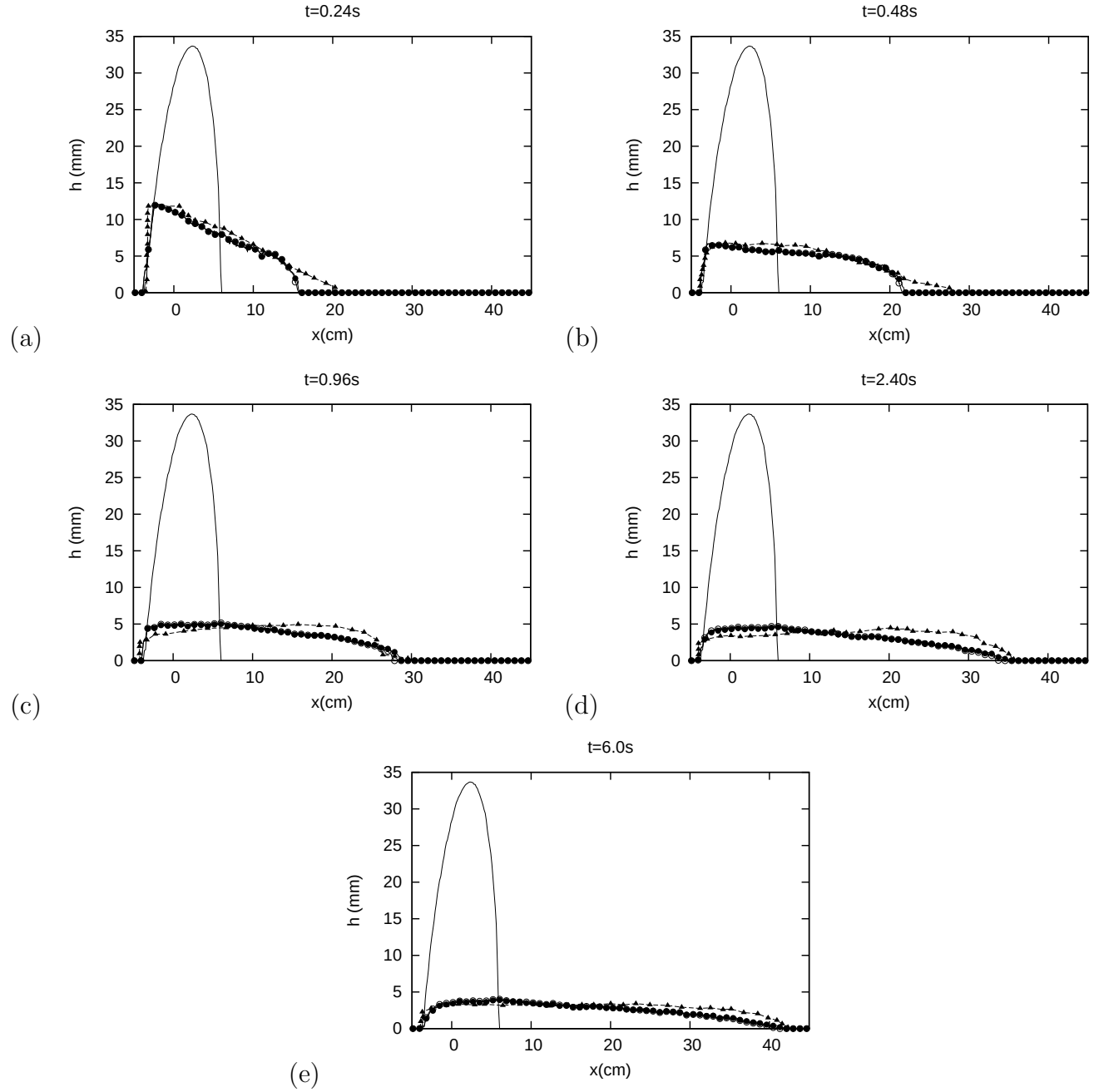

Figure 26: Section 9.3. Initial conditions (——), measured (−▲−) and computed thickness profiles along $y = 0$ using present solver (− ∘ −) and solver in [64](− • −) at times $t = 0.24$, 0.48, 0.96, 2.40 and 6.0 s.

[1] R.J. LeVeque, Balancing source terms and flux gradients in high-resolution Godunov methods: The quasi-steady wave-propagation algorithm, J. Comput. Phys. 146 (1998) 346–365.

[2] R.J. LeVeque, Finite Volume Methods for Hyperbolic Problems, Cambridge University Press, (2002) 311–2002.

[3] A. Bermudez, M. E. Vázquez-Cendón, Upwind methods for hyperbolic conservation laws with source terms, Comput. Fluids. 23 (1994) 1049–1071.

[4] 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 (1996) 1–16.

[5] J.G. Zhou, D.M. Causon, C.G. Mingham, D.M. Ingram, The surface gradient method for the treatment of source terms in the shallow-water equations, J. Comput. Phys. 168 (2001) 1–25.

[6] Q. Liang, A.G.L. Borthwick, Adaptive quadtree simulation of shallow flows with wet-dry fronts over complex topography, Comput. and Fluids. 38 (2009) 221–234.

[7] E. Audusse, F. Bouchut, M.O. Bristeau, R. Klein, B. Perthame, A Fast and Stable Well-Balanced Scheme with Hydrostatic Reconstruction for Shallow Water Flows, SIAM J. Sci. Comput. 25 (2004) 2050–2065.

[8] P. García-Navarro, M. E. Vázquez-Cendón, On the numerical treatment of the source terms in the shallow water equations, Comput. Fluids. 29 (2000) 951–979.

[9] M. E. Hubbard, P. García-Navarro, Flux difference splitting and the balancing of source terms and flux gradients, J. Comp. Phys. 165 (2000) 89–125.

[10] A. Bollermann, G. Chen, A. Kurganov and S. Noelle, A well-balanced reconstruction of wet/dry fronts for the shallow water equations, SIAM J. Sci. Comput. 56 (2013) 267–290.

[11] P.L. Roe, Approximate Riemann solvers, parameter vectors, and difference schemes, J. Comput. Phys. 43 (1981) 357–372.

[12] A. Harten, P. Lax and B. van Leer, On upstream differencing and Godunov type methods for hyperbolic conservation laws, SIAM review. 25 (1983) 35–61.

[13] E.F. Toro, M. Spruce, W. Spears, Restoration of the contact surface in the HLL Riemann solver, Shock Waves. 4 (1994) 25–34.

[14] J. Murillo, P. García-Navarro, Augmented versions of the HLL and HLLC Riemann Solvers including source terms in one and two dimensions for shallow flow applications, J. Comput. Phys. 231 (2012) 6861–6906.

[15] Y. Xing, C. W. Shu, A survey of high order schemes for the shallow water equations, J. Math. Study, 47 (2014) 221–249.

[16] S. Noelle, Y. Xing and C. Shu, High-order well-balanced finite volume WENO schemes for shallow water equation with moving water, J. Comput. Phys. 226 (2007) 29–58.

[17] U.S. Fjordholm, S. Mishra, E. Tadmor, Well-balanced and energy stable schemes for the shallow water equations with discontinuous topography, J. Comput. Phys. 230 (2011) 5587–5609.

[18] M.J. Castro Díaz, J.A. López-García, Carlos Parés, High order exactly well-balanced numerical methods for shallow water systems, J. Comput. Phys. 246 (2013) 242–264.

[19] Y. Xing, Exactly well-balanced discontinuous Galerkin methods for the shallow water equations with moving water equilibrium, J. Comput. Phys. 257 (2014) 536–553.

[20] J. Murillo, P. García-Navarro, Energy balance numerical schemes for shallow water equations with discontinuous topography, J. Comput. Phys. 236 (2012) 119–142.

[21] J. Murillo, P. García-Navarro, Accurate numerical modeling of 1D flow in channels with arbitrary shape. Application of the energy balanced property, J. Comput. Phys. 260 (2014) 222–248.

[22] A. Navas-Montilla, J. Murillo, Energy balanced numerical schemes with very high order. The Augmented Roe Flux ADER scheme. Application to the shallow water equations, J. Comput Phys. 290 (2015) 188–218.

[23] A. Navas-Montilla, J. Murillo, Asymptotically and exactly energy balanced augmented flux-ADER schemes with application to hyperbolic conservation laws with geometric source terms, J. Comput Phys. 317 (2016) 108147.

[24] Y. Xing, X. Zhang and C.-W. Shu, Positivity-preserving high order well-balanced discontinuous Galerkin methods for the shallow water equations, Adv. Water Resour. 33 (2010) 1476–1493.

[25] Y. Xing, X. Zhang, Positivity-preserving well-balanced discontinuous Galerkin methods for the shallow water equations on unstructured triangular meshes, J. Sci. Comput. 57 (2013) 19–41.

[26] A. I. Delis, I. K. Nikolos, M. Kazolea, Performance and Comparison of Cell-Centered and Node-Centered Unstructured Finite Volume Discretizations for Shallow Water Free Surface Flows, Arch. of Comput. Methods Eng. 18 (2011) 57–118.

[27] A. I. Delis, I. K. Nikolos, M. Kazolea, A robust well-balanced finite volume model for shallow water flows with wetting and drying over irregular terrain, Adv. Water Resour. 34 (2011) 915–932.

[28] S. Vater, N. Beisiegela, J. Behrensb, A limiter-based well-balanced discontinuous Galerkin method for shallow-water flows with wetting and drying: One-dimensional case, Adv. Water Resour. 85 (2015) 1–13.

[29] P.G. LeFloch, M.D. Thanh, A Godunov-type method for the shallow water equations with discontinuous topography in the resonant regime, J. Comput. Phys. 230 (2011) 7631–7660.

[30] P. G. LeFloch and S. Mishra, Numerical methods with controlled dissipation for small-scale dependent shocks, Acta Numer. 23 (2014), 1–72.

[31] D.L. George, Augmented Riemann solvers for the shallow water equations over variable topography with steady states and inundation, J. Comput. Phys. 227 (2008) 3089–3113.

[32] J. Murillo, P. García-Navarro, Weak solutions for partial differential equations with source terms: application to the shallow water equations, J. Comput. Phys. 229 (2010) 4327–4368.

[33] J. Murillo, P. García-Navarro, J. Burguete, P. Brufau. A conservative 2D model of inundation flow with solute transport over dry bed, Int. J. Numer. Meth. Fluids 52 (2006) 1059–1592.

[34] J. Burguete, P. García-Navarro, J. Murillo, Friction term discretization and limitation to preserve stability and conservation in the 1D shallow-water model: Application to unsteady irrigation and river flow, Int. J. Numer. Meth. Fluids 54 (2008) 403–425.

[35] J. Murillo, P. García-Navarro, J. Burguete, Conservative numerical simulation of multicomponent transport in two-dimensional unsteady shallow water flow, J. Comput. Phys. 228 (2009) 5539–5573.

[36] J. Murillo, P. García-Navarro, J. Burguete, Time step restrictions for well balanced shallow water solutions in non-zero velocity steady states, Int. J. Numer. Meth. Fluids 56 (2008) 661–686.

[37] C. Parés, Numerical methods for nonconservative hyperbolic systems: a theoretical framework. SIAM J. Num. Anal. 44 (2006) 300–321.

[38] G. Dal Maso, P.G. LeFloch, F. Murat, Definition and weak stability of nonconservative products, J. Math. Pures Appl. 74 (1995) 483–548.

[39] T.Y. Hou, P.G. LeFloch, Why nonconservative schemes converge to wrong solutions: error analysis, Math. Comput. 62 (1994) 497–530.

[40] M.J. Castro, P.G. LeFloch, M.L. Muñoz-Ruiz, C. Parés, Why many theories of shock waves are necessary. Convergence error in formally path-consistent schemes, J. Comput. Phys. 227 (2008) 8107–8129.

[41] G. Rosatti, L. Begnudelli, The Riemann Problem for the one-dimensional, free-surface Shallow Water Equations with a bed step: theoretical analysis and numerical simulations, J. Comput. Phys. 229 (2010) 760–787.

[42] G. Rosatti, J. Murillo, L. Fraccarollo. Generalized Roe schemes for 1D two-phase, free-surface flows over a mobile bed, J. Comput. Phys. 227 (2008) 10058–10077.

[43] J. Murillo, P. García-Navarro, An Exner-based coupled model for two-dimensional transient flow over erodible bed, J. Comput. Phys. 229 (2010) 8704–8732.

[44] J. Murillo, B. Latorre, P. García-Navarro, A Riemann solver for unsteady computation of 2D shallow flows with variable density, J. Comput. Phys. 231 (2012) 4775–4807.

[45] J. Murillo, P. García-Navarro, Wave Riemann description of friction terms in unsteady shallow flows: application to water and mud/debris floods, J. Comput. Phys. 231 (2011) 1963–2001.

[46] C. Juez, D. Caviedes-Voullieme, J. Murillo, P. García-Navarro, 2D dry granular free-surface transient flow over complex topography with obstacles. Part II: Numerical predictions of fluid structures and benchmarking, Comput. Geosci. 73 (2014) 142–163.

[47] J. Murillo, P. García-Navarro, A Roe type energy balanced solver for 1D arterial blood flow and transport, Comput. Fluids. 117 (2015) 149–167.

[48] E.F. Toro, Riemann solvers and numerical methods for fluid dynamics: a practical introduction, third ed., Springer-Verlag, Berlin, Heidelberg, 2009.

[49] F. Alcrudo, F. Benkhaldoun, Exact solutions to the Riemann problem of the shallow water equations with a bottom step, Comput. Fluids 30 (2001) 643–671.

[50] S.K. Godunov, A finite difference method for the computation of discontinuous solutions of the equations of fluid dynamics, Mat. Sb. 47 (1959) 357–393

[51] S. Sahmim, F. Benkhaldoun, F. Alcrudo, A sign matrix based scheme for non-homogeneous PDEs with an analysis of the convergence stagnation phenomenon, J. Comput. Phys. 26 (2007) 17531783.

[52] J. Li and G. Chen, The generalized Riemann problem method for the shallow water equations with bottom topography, Int J. Numer. Meth. Eng 65 (2006), 834–862.

[53] A. Harten and J. M. Hyman, Self adjusting grid methods for one-dimensional hyperbolic conservation laws, J. Comput. Phys. 50 (1983) 235–269.

[54] H. P. G. Darcy, Recherches expérimentales relatives aux mouvements de l'eau dans les tuyaux, Mémoires Présentés à l'Académie des Sciences, Paris, 1858.

[55] P. G. Gauckler, Études théoriques et pratiques sur l'écoulement et le mouvement des eaux, Comptes Rendus de l'Académie des Sciences, Paris, 1867.

[56] R. Manning, On the flow of water in open channels and pipes, Trans. Inst. Civil Engineers, 20 (1890) 161–207.

[57] E. Godlewski and P.A. Raviart, Numerical approximation of hyperbolic systems of conservation laws, Springer-Verlag, New York, 1996.

[58] M. Morales-Hernandez, P. Garcia-Navarro, J. Murillo, A large time step 1D upwind explicit scheme (CFL ¿ 1): Application to shallow water equations. Journal of Computational Physics 231 (2012) 65326557.

[59] I. MacDonald, M.J. Baines, N.K. Nichols, P.G. Samuels, Analytical benchmark solutions for open-channel flows, ASCE J. Hydraulic Eng. 123 (11) (1997) 10411045.

[60] E. Audusse, F. Bouchut, M. O. Bristeau, J. Sainte-Marie, Kinetic entropy inequality and hydrostatic reconstruction scheme for the Saint-Venant system, Math. Comp. 85 (2016), 2815–2837

[61] O. Delestre, S. Cordier, F. Darboux and F. James, A limitation of the hydrostatic reconstruction technique for Shallow Water equations, C. R. Acad. Sci. Paris, Ser. I, 350 (2012) 677-681.

[62] W.C. Thacker, Some exact solutions to the non linear shallow water equations, J. Fluid Mech. 107 (1981) 499-508.

[63] O. Pouliquen, Y. Forterre, Friction law for dense granular flows: application to the motion of a mass down a rough inclined plane, J. of Fluid Mech. 453 (2002) 133–151.

[64] C. Juez, J. Murillo, P. García-Navarro, 2D simulation of granular flow over irregular steep slopes using global and local coordinates, J. Comput. Phys. 255 (2013) 166–204.