A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows

Emmanuel Audusse[a], François Bouchut[b], Marie-Odile Bristeau[a], Rupert Klein[c] and Benoît Perthame[a,b] [1]

[a] INRIA Rocquencourt, projet BANG, Domaine de Voluceau, BP 105, 78153 Le Chesnay cedex, France
[b] Département de Mathématiques et Applications, CNRS & École Normale Supérieure, 45 rue d'Ulm, F 75230 Paris cedex 05, France
[c] Department of Mathematics and Computer Science, Freie Universität Berlin, D-14195 Berlin, Germany

Abstract

We consider the Saint-Venant system for shallow water flows with non-flat bottom. This is a hyperbolic system of conservation laws that approximately describes various geophysical flows, such as rivers, coastal areas, and oceans when completed with a Coriolis term, and granular flows when completed with friction. Numerical approximate solutions to this system may be generated using conservative finite volume methods, which are known to properly handle any shocks and contact discontinuities. Yet, these schemes prove to be problematic for near steady states, as the structure of their numerical truncation errors is generally not compatible with exact physical steady state conditions. This difficulty can be overcome by using so called well-balanced schemes. We describe a general strategy based on a local hydrostatic reconstruction that allows us to derive a well-balanced scheme from any solver for the homogeneous problem (Godunov, Roe, kinetic…). Whenever the initial solver satisfies some classical stability properties, it yields a simple and fast well-balanced scheme that preserves the nonnegativity of the water height and satisfies a semi-discrete entropy inequality.

Key-words: Shallow water equations, finite volume schemes, well-balanced schemes

Mathematics Subject Classification: 65M12, 76M12, 35L65

1 Introduction

The classical Saint-Venant system for shallow water has been widely validated. It assumes a slowly varying topography $z(x)$ (x denotes a coordinate in the horizontal direction) and describes the height of water $h(t,x)$, and the water velocity $u(t,x)$ in the direction parallel to the bottom. It uses the following equations in one space dimension,

$$\begin{cases} \partial_t h + \partial_x(hu) = 0, \\ \partial_t(hu) + \partial_x(hu^2 + gh^2/2) = -hgz_x, \end{cases} \tag{1.1}$$

where $g > 0$ denotes the gravity constant. For future reference we denote the flux by $F(U) = (hu, hu^2 + gh^2/2)$, with $U = (h, hu)$. This model is very robust, being hyperbolic and

[1]emails: emmanuel.audusse@inria.fr, fbouchut@dma.ens.fr, marie-odile.bristeau@inria.fr, rupert.klein@zib.de, perthame@dma.ens.fr,

admitting an entropy inequality (related to the physical energy)

$$\partial_t \widetilde{\eta}(U,z) + \partial_x \widetilde{G}(U,z) \le 0, \tag{1.2}$$

where

$$\begin{aligned} \eta(U) &= hu^2/2 + \frac{g}{2}h^2, \qquad & G(U) &= \Big(hu^2/2 + gh^2\Big)u, \\ \widetilde{\eta}(U,z) &= \eta(U) + hgz, \qquad & \widetilde{G}(U,z) &= G(U) + hgzu. \end{aligned} \tag{1.3}$$

Another advantageous property is that it preserves the steady state of a lake at rest

$$h + z = Cst, \qquad u = 0. \tag{1.4}$$

When solving numerically (1.1), it is very important to be able to preserve these steady states at the discrete level and to accurately compute the evolution of small deviations from them, because the majority of real-life applications resides in this flow regime.

Since the early works of Leroux and coauthors [14], [16], schemes satisfying such a property are called well-balanced. Several schemes have been proposed that satisfy this property, [23], [17], [13], [11], [30], [29], [3], but the difficulty is then to get schemes that also satisfy very natural properties such as conservativity of the water height h, nonnegativity of h, the ability to compute dry states $h = 0$ and transcritical flows when the jacobian matrix F' of the flux function becomes singular, and eventually to satisfy a discrete entropy inequality. Theoretically, the exact Godunov scheme satisfies these requirements [20], but it is in practice too computationally expensive, and not easily adaptable to more complex systems, such as for example the models proposed in [9]. The first attempt to derive an approximate solver satisfying all the requirements was performed in [4] for a scalar equation. A generalization to the case of the Saint-Venant system was obtained in [26], and another method by relaxation is also proposed in [7]. However, these approximate solver methods are still quite heavy in practice. The aim of this paper is to explain how it is possible by a very flexible approach involving a hydrostatic reconstruction, to obtain a well-balanced scheme satisfying all the above requirements, and that is computationally inexpensive. The present approach unifies and generalizes ideas developed independently in [5, 6] for nearly hydrostatic, multi-dimensional compressible flow, and in [1] for the Saint-Venant shallow water model. By opposition to the existing literature, it also gives a general method that can be used with any solver.

2 Well-balanced scheme with hydrostatic reconstruction

2.1 Semi-discrete scheme

Finite volume schemes for hyperbolic systems consist in using an upwinding of the fluxes. In the semi-discrete case they provide a discrete version of (1.1) under the form

$$\Delta x_i \frac{d}{dt} U_i(t) + F_{i+1/2} - F_{i-1/2} = S_i, \tag{2.1}$$

where Δx_i denotes a possibly variable mesh size $\Delta x_i = x_{i+1/2} - x_{i-1/2}$, and the cell-centered vector of discrete unknowns is

$$U_i(t) = \begin{pmatrix} h_i(t) \\ h_i(t)u_i(t) \end{pmatrix}. \tag{2.2}$$

In a basic first-order accurate scheme, the fluxes are classically computed as $F_{i+1/2} = \mathcal{F}(U_i(t), U_{i+1}(t))$ with a numerical flux $\mathcal{F}$ that is computed via an approximate resolution of the Riemann problem (a so-called *solver*), which provides stability of the method. We refer to [12] for the description of the most well-known solvers: Godunov, Roe, Kinetic...It is known since [14],[16] that cell-centered evaluations of the source term in (2.1) will generally not be able to maintain steady states of a lake at rest in time, which are characterized by

$$h_i + z_i = Cst, \qquad u_i = 0. \tag{2.3}$$

Following [1], [5, 6], we propose and analyze finite volume schemes according to (2.1) with flux functions

$$F_{i+1/2} = \mathcal{F}(U_{i+1/2-}, U_{i+1/2+}), \tag{2.4}$$

where the interface values $U_{i+1/2-}, U_{i+1/2+}$ are derived from a local hydrostatic reconstruction to be described shortly, which is similar to second-order reconstructions in higher-order methods. The source term is discretized as

$$S_i = \begin{pmatrix} 0 \\ \frac{g}{2}h^2_{i+1/2-} - \frac{g}{2}h^2_{i-1/2+} \end{pmatrix}. \tag{2.5}$$

This ansatz is motivated by the balancing requirement as follows. For nearly hydrostatic flows one has $u \ll \sqrt{gh}$. In the associated asymptotic limit the leading order water height $\underline{h}$ adjusts so as to satisfy the balance of momentum flux and momentum source terms, i.e.

$$\partial_x \left(\frac{g\underline{h}^2}{2} \right) = -\underline{h}\, g z_x\,. \tag{2.6}$$

Integrating over, say, the ith grid cell we obtain an approximation to the net source term as

$$-\int_{x_{i-1/2}}^{x_{i+1/2}} \underline{h}\, g z_x \; dx = \frac{g}{2}\underline{h}^2_{i+1/2-} - \frac{g}{2}\underline{h}^2_{i-1/2+}\,. \tag{2.7}$$

Thus we are able to locally represent the cell-averaged source term as the discrete gradient of the hydrostatic momentum flux, and this motivates the source term discretization in (2.5).

It is obvious now that any hydrostatic state is maintained exactly if, for such a state, the momentum fluxes in (2.1) and the locally reconstructed heights satisfy $F^{hu}_{i+1/2} = \frac{1}{2}gh^2_{i+1/2-} = \frac{1}{2}gh^2_{i+1/2+}$. This is the motivation for (2.4), which gives this property if for hydrostatic states we have $U_{i+1/2-} = U_{i+1/2+} = (h_{i+1/2-}, 0) = (h_{i+1/2+}, 0)$.

The hydrostatic balance in (2.6) is equivalent to the "lake at rest" equation (1.4), so that the reconstruction of the leading order heights is straightforward,

$$\underline{h}_{i+1/2-} = h_i + z_i - z_{i+1/2}, \qquad \underline{h}_{i+1/2+} = h_{i+1} + z_{i+1} - z_{i+1/2}. \tag{2.8}$$

The evaluation of the cell interface height $z_{i+1/2}$ is somewhat subtle since the scheme shall also robustly capture dry regions where $h \equiv 0$. The challenge is to design a scheme that guarantees nonnegativity of the water height even when cells begin to "dry out". We prove below that this can be achieved through a biased evaluation of the form

$$z_{i+1/2} = \max(z_i, z_{i+1}), \tag{2.9}$$

and with a nonnegativity-preserving truncation of the leading order heights in (2.8), $h_{i+1/2\pm} = \max(0, \underline{h}_{i+1/2\pm})$.

With these rules in place we can now summarize our first-order well-balanced finite volume scheme by

$$\Delta x_i \frac{d}{dt} U_i(t) + F_{i+1/2} - F_{i-1/2} = S_i, \tag{2.10}$$

where

$$F_{i+1/2} = \mathcal{F}(U_{i+1/2-}, U_{i+1/2+}), \tag{2.11}$$

$$U_{i+1/2-} = \begin{pmatrix} h_{i+1/2-} \\ h_{i+1/2-}\ u_i \end{pmatrix}, \qquad U_{i+1/2+} = \begin{pmatrix} h_{i+1/2+} \\ h_{i+1/2+}\ u_{i+1} \end{pmatrix}, \tag{2.12}$$

$$h_{i+1/2\pm} = \max(0, \underline{h}_{i+1/2\pm}), \tag{2.13}$$

and

$$S_i = S_{i+1/2-} + S_{i-1/2+} \equiv \begin{pmatrix} 0 \\ \frac{g}{2} h^2_{i+1/2-} - \frac{g}{2} h_i^2 \end{pmatrix} + \begin{pmatrix} 0 \\ \frac{g}{2} h_i^2 - \frac{g}{2} h^2_{i-1/2+} \end{pmatrix}. \tag{2.14}$$

The latter expression for the source is equivalent to the earlier (2.5), it shows that the source may be considered as being distributed to the cell interfaces. With this re-interpretation in mind, we may also rewrite the scheme as

$$\Delta x_i \frac{d}{dt} U_i(t) + \mathcal{F}_l(U_i, U_{i+1}, z_i, z_{i+1}) - \mathcal{F}_r(U_{i-1}, U_i, z_{i-1}, z_i) = 0, \tag{2.15}$$

with left and right numerical fluxes

$$\begin{aligned} \mathcal{F}_l(U_i, U_{i+1}, z_i, z_{i+1}) &= F_{i+1/2} - S_{i+1/2-} = \mathcal{F}(U_{i+1/2-}, U_{i+1/2+}) + \begin{pmatrix} 0 \\ \frac{g}{2} h_i^2 - \frac{g}{2} h^2_{i+1/2-} \end{pmatrix}, \\ \mathcal{F}_r(U_i, U_{i+1}, z_i, z_{i+1}) &= F_{i+1/2} + S_{i+1/2+} = \mathcal{F}(U_{i+1/2-}, U_{i+1/2+}) + \begin{pmatrix} 0 \\ \frac{g}{2} h^2_{i+1} - \frac{g}{2} h^2_{i+1/2+} \end{pmatrix}. \end{aligned} \tag{2.16}$$

Notice that (2.12), (2.13), (2.8) mean that we try to impose interface values satisfying some modified steady equations $h_{i+1/2-} + z_{i+1/2} = h_i + z_i$, $u_{i+1/2-} = u_i$, $h_{i+1/2+} + z_{i+1/2} = h_{i+1} + z_{i+1}$, $u_{i+1/2+} = u_{i+1}$, i.e. $h + z = cst$, $u = cst$ instead of Bernoulli's law $u^2/2 + g(h + z) = cst$, $hu = cst$.

Our construction, combined with a centered value of $z_{i+1/2}$, is not stable. The 'upwind' value proposed in (2.9), and the truncation of negative values in (2.13) have the advantage of giving nonnegative values of $h_{i+1/2\pm}$ and of being stable, as we state it now.

Theorem 2.1 *Consider a consistent numerical flux $\mathcal{F}$ for the homogeneous problem that preserves nonnegativity of $h_i(t)$ and satisfies an in-cell entropy inequality corresponding to the entropy η in (1.3). Then the finite volume scheme (2.8)-(2.14)*
(i) preserves the nonnegativity of $h_i(t)$,
(ii) preserves the steady state of a lake at rest (2.3),
(iii) is consistent with the Saint-Venant system (1.1),
(iv) satisfies an in-cell entropy inequality associated to the entropy $\widetilde{\eta}$ in (1.3),

$$\Delta x_i \frac{d}{dt} \widetilde{\eta}(U_i(t), z_i) + \widetilde{G}_{i+1/2} - \widetilde{G}_{i-1/2} \le 0. \tag{2.17}$$

Proof. The statement that $\mathcal{F}$ preserves the nonnegativity of $h_i(t)$ means exactly that $\mathcal{F}^h(h_i = 0, u_i, h_{i+1}, u_{i+1}) - \mathcal{F}^h(h_{i-1}, u_{i-1}, h_i = 0, u_i) \le 0$, for all choices of the other arguments. Since the sources in (2.14) have no contribution to the first component, $h_i(t)$ in our scheme satisfies a conservative equation with flux $\mathcal{F}^h(U_{i+1/2-}, U_{i+1/2+})$. Therefore we need to check that $\mathcal{F}^h(U_{i+1/2-}, U_{i+1/2+}) - \mathcal{F}^h(U_{i-1/2-}, U_{i-1/2+}) \le 0$ whenever $h_i = 0$. Our construction (2.8), (2.9), (2.13) (and this is the motivation for (2.9)), gives $h_{i+1/2-} = h_{i-1/2+} = 0$ when $h_i = 0$, and this gives (i).

Then we prove statement (ii). On a steady state of a lake at rest, we have $h_{i+1/2-} = h_{i+1/2+}$, $u_{i+1} = u_i = 0$, thus $U_{i+1/2-} = U_{i+1/2+}$ and by consistency of $\mathcal{F}$

$$F_{i+1/2} = F(U_{i+1/2-}) = F(U_{i+1/2+}) = \begin{pmatrix} 0 \\ \frac{g}{2} h^2_{i+1/2-} \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{g}{2} h^2_{i+1/2+} \end{pmatrix}. \tag{2.18}$$

Together with the expression of the source terms in (2.14), we get $F_{i+1/2} - S_{i+1/2-} = F(U_i)$, $F_{i+1/2} + S_{i+1/2+} = F(U_{i+1})$, and this proves (ii).

To prove (iii), we apply the criterion in [27], [7], and we need to check two properties related to the consistency with the exact flux F and the consistency with the source. The consistency with the exact flux $\mathcal{F}_l(U, U, z, z) = \mathcal{F}_r(U, U, z, z) = F(U)$ is obvious since $U_{i+1/2-} = U_i$ and $U_{i+1/2+} = U_{i+1}$ whenever $z_{i+1} = z_i$. For consistency with the source, the criterion becomes for the Saint-Venant system

$$\mathcal{F}_r^{hu}(U_i, U_{i+1}, z_i, z_{i+1}) - \mathcal{F}_l^{hu}(U_i, U_{i+1}, z_i, z_{i+1}) = -hg\Delta z_{i+1/2} + o(\Delta z_{i+1/2}) \tag{2.19}$$

as $U_i, U_{i+1} \to U$ and $\Delta z_{i+1/2} \to 0$, where $\Delta z_{i+1/2} = z_{i+1} - z_i$. In our case,

$$\mathcal{F}_r - \mathcal{F}_l = S_{i+1/2-} + S_{i+1/2+} = \begin{pmatrix} 0 \\ \frac{g}{2} h^2_{i+1/2-} - \frac{g}{2} h_i^2 + \frac{g}{2} h^2_{i+1} - \frac{g}{2} h^2_{i+1/2+} \end{pmatrix}. \tag{2.20}$$

Now, assuming $h > 0$, the maxima in (2.13) play no role if $h_i - h$, $h_{i+1} - h$ and $\Delta z_{i+1/2}$ are small enough. Thus we have $h^2_{i+1/2-}/2 - h_i^2/2 = h(z_i - z_{i+1/2}) + o(\Delta z_{i+1/2})$, $h^2_{i+1/2+}/2 - h^2_{i+1}/2 = h(z_{i+1} - z_{i+1/2}) + o(\Delta z_{i+1/2})$, which gives (2.19). In the special case $h = 0$, the maxima in (2.13) can play a role only when $h_i = O(\Delta z_{i+1/2})$, and we conclude that (2.19) always holds, proving (iii).

In order to prove (iv), we first write that the original numerical flux $\mathcal{F}$ satisfies a semi-discrete entropy inequality. According to [7], this means that we can find a numerical entropy flux $\mathcal{G}$ such that

$$\begin{aligned} &G(U_{i+1}) + \eta'(U_{i+1})(\mathcal{F}(U_i, U_{i+1}) - F(U_{i+1})) \\ &\le \mathcal{G}(U_i, U_{i+1}) \le G(U_i) + \eta'(U_i)(\mathcal{F}(U_i, U_{i+1}) - F(U_i)), \end{aligned} \tag{2.21}$$

where η' is the derivative of η with respect to $U = (h, hu)$, $\eta'(U) = (gh - u^2/2, u)$. Similarly, having an entropy inequality (2.17) for (1.1) with $\widetilde{G}_{i+1/2} = \widetilde{\mathcal{G}}(U_i, U_{i+1}, z_i, z_{i+1})$ is equivalent to finding some numerical entropy flux $\widetilde{\mathcal{G}}$ such that

$$\begin{aligned} &\widetilde{G}(U_{i+1}, z_{i+1}) + \widetilde{\eta}'(U_{i+1}, z_{i+1})(\mathcal{F}_r(U_i, U_{i+1}, z_i, z_{i+1}) - F(U_{i+1})) \\ &\le \widetilde{\mathcal{G}}(U_i, U_{i+1}, z_i, z_{i+1}) \le \widetilde{G}(U_i, z_i) + \widetilde{\eta}'(U_i, z_i)(\mathcal{F}_l(U_i, U_{i+1}, z_i, z_{i+1}) - F(U_i)). \end{aligned} \tag{2.22}$$

Let us prove that (2.22) holds with

$$\widetilde{\mathcal{G}}(U_i, U_{i+1}, z_i, z_{i+1}) = \mathcal{G}(U_{i+1/2-}, U_{i+1/2+}) + \mathcal{F}^h(U_{i+1/2-}, U_{i+1/2+}) g z_{i+1/2}. \tag{2.23}$$

Since both inequalities are obtained by the same type of estimates, let us prove only the upper inequality involving $\mathcal{F}_l$ in (2.22). By comparison to (2.21), it is enough to prove that

$$\begin{aligned} &G(U_{i+1/2-}) + \eta'(U_{i+1/2-})\big(\mathcal{F}(U_{i+1/2-}, U_{i+1/2+}) - F(U_{i+1/2-})\big) \\ &\quad + \mathcal{F}^h(U_{i+1/2-}, U_{i+1/2+}) g z_{i+1/2} \\ &\leq G(U_i) + \eta'(U_i)(\mathcal{F}_l - F(U_i)) + \mathcal{F}^h(U_{i+1/2-}, U_{i+1/2+}) g z_i. \end{aligned} \tag{2.24}$$

This inequality can be written, by denoting $\mathcal{F} = (\mathcal{F}^h, \mathcal{F}^{hu}) = \mathcal{F}(U_{i+1/2-}, U_{i+1/2+})$,

$$\begin{aligned} &(u_i^2/2 + g h_{i+1/2-}) h_{i+1/2-} u_i + (g h_{i+1/2-} - u_i^2/2)(\mathcal{F}^h - h_{i+1/2-} u_i) \\ &\quad + u_i(\mathcal{F}^{hu} - h_{i+1/2-} u_i^2 - g h_{i+1/2-}^2/2) + \mathcal{F}^h g (z_{i+1/2} - z_i) \\ &\leq (u_i^2/2 + g h_i) h_i u_i + (g h_i - u_i^2/2)(\mathcal{F}^h - h_i u_i) + u_i(\mathcal{F}_l^{hu} - h_i u_i^2 - g h_i^2/2), \end{aligned} \tag{2.25}$$

or after simplification

$$u_i(\mathcal{F}^{hu} - g h_{i+1/2-}^2/2) + \mathcal{F}^h g (h_{i+1/2-} - h_i + z_{i+1/2} - z_i) \leq u_i(\mathcal{F}_l^{hu} - g h_i^2/2). \tag{2.26}$$

Since $\mathcal{F}_l^{hu} - g h_i^2/2 = \mathcal{F}^{hu} - g h_{i+1/2-}^2/2$ by definition of $\mathcal{F}_l$ in (2.16), our inequality finally reduces to

$$\mathcal{F}^h(U_{i+1/2-}, U_{i+1/2+})(h_{i+1/2-} - h_i + z_{i+1/2} - z_i) \leq 0. \tag{2.27}$$

Now, according to (2.8), (2.13), when this quantity is nonzero, we have $h_{i+1/2-} = 0$ and the expression between parentheses is nonnegative. But since $\mathcal{F}$ preserves nonnegativity, we have $\mathcal{F}^h(h_{i+1/2-} = 0, u_i, h_{i+1/2+}, u_{i+1}) \leq 0$ and we conclude that (2.27) always holds. This completes the proof of (iv). □

2.2 Fully discrete scheme and CFL condition

When using the time-space fully discrete scheme

$$U_i^{n+1} - U_i^n + \frac{\Delta t}{\Delta x_i}\Big(\mathcal{F}_l(U_i, U_{i+1}, z_i, z_{i+1}) - \mathcal{F}_r(U_{i-1}, U_i, z_{i-1}, z_i)\Big) = 0, \tag{2.28}$$

the consistency and the well-balanced property are of course still valid. The question is then to obtain a CFL condition that guarantees stability.

One can prove that our hydrostatic reconstruction scheme does not satisfy a fully discrete entropy inequality. Indeed there exist some data with $h_i + z_i = cst$, $u_i = cst \neq 0$ such that for any $\Delta t > 0$, the fully discrete entropy inequality $\widetilde{\eta}(U_i^{n+1}, z_i) - \widetilde{\eta}(U_i^n, z_i) + \frac{\Delta t}{\Delta x_i}(\widetilde{G}_{i+1/2} - \widetilde{G}_{i-1/2}) \leq 0$ is violated. However, in practice we do not observe instabilities as long as the water height h_i remains nonnegative.

In order to preserve the nonnegativity of h_i, the CFL condition that needs to be used is not more restrictive than that of the homogeneous solver.

Proposition 2.2 *Assume that the homogeneous flux $\mathcal{F}$ preserves the nonnegativity of h by interface with a numerical speed $\sigma(U_i, U_{i+1}) \geq 0$, which means that whenever the CFL condition*

$$\sigma(U_i, U_{i+1})\Delta t \leq \min(\Delta x_i, \Delta x_{i+1}) \tag{2.29}$$

holds, we have

$$\begin{aligned} h_i - \frac{\Delta t}{\Delta x_i}(\mathcal{F}^h(U_i, U_{i+1}) - h_i u_i) \geq 0, \\ h_{i+1} - \frac{\Delta t}{\Delta x_{i+1}}(h_{i+1}u_{i+1} - \mathcal{F}^h(U_i, U_{i+1})) \geq 0. \end{aligned} \tag{2.30}$$

Then the fully discrete hydrostatic reconstruction scheme (2.28) also preserves the nonnegativity of h by interface,

$$\begin{aligned} h_i - \frac{\Delta t}{\Delta x_i}(\mathcal{F}^h(U_{i+1/2-}, U_{i+1/2+}) - h_i u_i) \geq 0, \\ h_{i+1} - \frac{\Delta t}{\Delta x_{i+1}}(h_{i+1}u_{i+1} - \mathcal{F}^h(U_{i+1/2-}, U_{i+1/2+})) \geq 0, \end{aligned} \tag{2.31}$$

under the CFL condition

$$\sigma(U_{i+1/2-}, U_{i+1/2+})\Delta t \leq \min(\Delta x_i, \Delta x_{i+1}). \tag{2.32}$$

Proof. Under the CFL condition (2.32), we have

$$\begin{aligned} h_{i+1/2-} - \frac{\Delta t}{\Delta x_i}(\mathcal{F}^h(U_{i+1/2-}, U_{i+1/2+}) - h_{i+1/2-}u_{i+1/2-}) \geq 0, \\ h_{i+1/2+} - \frac{\Delta t}{\Delta x_{i+1}}(h_{i+1/2+}u_{i+1/2+} - \mathcal{F}^h(U_{i+1/2-}, U_{i+1/2+})) \geq 0. \end{aligned} \tag{2.33}$$

Noticing that with the choice (2.8), (2.9), (2.13), we have $h_{i+1/2-} \leq h_i$ and $h_{i+1/2+} \leq h_{i+1}$, we conclude that (2.31) holds as soon as $1 + u_i\Delta t/\Delta x_i \geq 0$ and $1 - u_{i+1}\Delta t/\Delta x_{i+1} \geq 0$, which is necessarily the case from (2.32). □

3 Second-order extension

Starting from a given first-order method, an usual way to obtain a second-order extension is - as it was already mentioned before for the analogy with our hydrostatic reconstruction - to compute the fluxes from limited reconstructed values on both sides of each interface rather than cell-centered values, see [12], [22] or [28]. These new values are classically obtained with three ingredients: prediction of the gradients in each cell, linear extrapolation, and limitation procedure.

In the presence of a source and in the context of well-balanced schemes, this approach needs to be precised. In particular, according to [18], [19], [7], since not only the reconstructed values $U_{i,r}$ at $i + 1/2-$, and $U_{i+1,l}$ at $i + 1/2+$ need be defined but also $z_{i,r}$, $z_{i+1,l}$, a cell-centered source term S_{ci} must by added to preserve the consistency. We remark that even if z_i do not depend on time, the reconstructed values $z_{i,l}$, $z_{i,r}$ could depend on time via a coupling with U_i in the reconstruction step. Once known these second-order reconstructed values we apply the hydrostatic reconstruction scheme exposed in the previous section at each interface. This gives the second-order well-balanced scheme

$$\Delta x_i \frac{d}{dt} U_i(t) + F_{i+1/2} - F_{i-1/2} = S_i + S_{ci}, \tag{3.1}$$

where

$$F_{i+1/2} = \mathcal{F}(U_{i+1/2-}, U_{i+1/2+}), \tag{3.2}$$

$$U_{i+1/2-} = \begin{pmatrix} h_{i+1/2-} \\ h_{i+1/2-}\ u_{i,r} \end{pmatrix}, \qquad U_{i+1/2+} = \begin{pmatrix} h_{i+1/2+} \\ h_{i+1/2+}\ u_{i+1,l} \end{pmatrix}, \tag{3.3}$$

and the hydrostatic reconstruction is now

$$h_{i+1/2\pm} = \max(0, \underline{h}_{i+1/2\pm}), \tag{3.4}$$

$$\underline{h}_{i+1/2-} = h_{i,r} + z_{i,r} - z_{i+1/2}, \qquad \underline{h}_{i+1/2+} = h_{i+1,l} + z_{i+1,l} - z_{i+1/2}, \tag{3.5}$$

with

$$z_{i+1/2} = \max(z_{i,r}, z_{i+1,l}). \tag{3.6}$$

The source term is distributed as before at the interfaces,

$$S_i = S_{i+1/2-} + S_{i-1/2+}, \tag{3.7}$$

$$S_{i+1/2-} = \begin{pmatrix} 0 \\ \frac{g}{2} h^2_{i+1/2-} - \frac{g}{2} h^2_{i,r} \end{pmatrix}, \quad S_{i-1/2+} = \begin{pmatrix} 0 \\ \frac{g}{2} h^2_{i,l} - \frac{g}{2} h^2_{i-1/2+} \end{pmatrix}. \tag{3.8}$$

A simple well-balanced choice for the centered source term S_{ci} is

$$S_{ci} = \begin{pmatrix} 0 \\ g \frac{h_{i,l}+h_{i,r}}{2} (z_{i,l} - z_{i,r}) \end{pmatrix}. \tag{3.9}$$

Using the definitions of the left and right numerical fluxes $\mathcal{F}_l$, $\mathcal{F}_r$ in (2.16), a compact formulation of the scheme is

$$\Delta x_i \frac{d}{dt} U_i(t) + \mathcal{F}_l(U_{i,r}, U_{i+1,l}, z_{i,r}, z_{i+1,l}) - \mathcal{F}_r(U_{i-1,r}, U_{i,l}, z_{i-1,r}, z_{i,l}) = S_{ci}. \tag{3.10}$$

This formulation ensures that the second-order scheme inherits the stability properties of the first-order one.

Theorem 3.1 *Consider a consistent numerical flux $\mathcal{F}$ for the homogeneous problem that preserves nonnegativity of $h_i(t)$. Assume that the second-order reconstruction gives nonnegative values $h_{i,l}$, $h_{i,r}$, is well-balanced and is second-order-centered in z, which means by definition that whenever the sequences (U_i) and (z_i) are the cell averages of smooth functions $U(x)$, $z(x)$, we have*

$$\begin{aligned} z_{i+1,l} - z_{i,r} &= O\left((\Delta x_i + \Delta x_{i+1})^3\right), \\ \frac{z_{i,r} - z_{i,l}}{\Delta x_i} &= z_x(x_i) + O\left((\Delta x_{i-1} + \Delta x_i + \Delta x_{i+1})^2\right). \end{aligned} \tag{3.11}$$

Then the finite volume scheme (3.1)-(3.9) preserves the nonnegativity of $h_i(t)$, is well-balanced, i.e. it preserves the steady states of a lake at rest (2.3), and is second-order accurate.

Proof. It is well known that the second-order reconstruction strategy preserves the nonnegativity of the water height (under a half CFL condition in the fully discrete case). Here only the centered source term S_{ci} in (3.10) could cause difficulties, but it does not since its first component vanishes.

The preservation of the lake at rest steady states can be checked easily from the property of the second-order reconstruction to be well-balanced, which means by definition that if $u_i = 0$ and $h_i + z_i = h_{i+1} + z_{i+1}$ for all i, then $u_{i,l} = u_{i,r} = 0$ and $h_{i,l} + z_{i,l} = h_{i,r} + z_{i,r} = h_i + z_i$ for all i. Indeed we just have to notice that for a steady state, $S_{ci} = (0, g(h_{i,r}^2 - h_{i,l}^2)/2)$.

In order to prove the second-order accuracy, let us assume that (U_i) and (z_i) are realized as the cell averages of smooth functions $U(x)$ and $z(x)$, and denote by $\hbar$ the mesh size. Then, since we assumed implicitly that the second-order reconstruction is second-order, we have that $U_{i,r} = U(x_{i+1/2}) + O(\hbar^2)$, $U_{i+1,l} = U(x_{i+1/2}) + O(\hbar^2)$, $z_{i,r} = z(x_{i+1/2}) + O(\hbar^2)$, $z_{i+1,l} = z(x_{i+1/2}) + O(\hbar^2)$. It follows from (3.3)-(3.6) that $U_{i+1/2\pm} = U(x_{i+1/2}) + O(\hbar^2)$, thus by (3.2) $F_{i+1/2} = F(U(x_{i+1/2})) + O(\hbar^2)$. This proves the second-order accuracy in the weak sense of the flux difference in (3.1) since this part is in conservative form. For the right-hand side, there is no such cancellation thus we can only allow errors in $O(\Delta x_i \hbar^2)$ in (3.1). We have $(h_{i,l} + h_{i,r})/2 = h(x_i) + O(\hbar^2)$, and the second expansion in (3.11) yields with (3.9) that $S_{ci} = \big(0, -gh(x_i) z_x(x_i)\Delta x_i + O(\Delta x_i \hbar^2)\big) = \int_{x_{i-1/2}}^{x_{i+1/2}} (0, -gh(x) z_x(x))\, dx + O(\Delta x_i \hbar^2)$. Since $S_{i+1/2\pm} = O(z_{i+1,l} - z_{i,r}) = O(\hbar^3)$ by the first expansion in (3.11), this gives that $S_i = O(\hbar^3)$ and concludes the proof in the "regular" case when $\hbar = O(\Delta x_i)$, by just considering S_i as an error in (3.1). In the general case, we have to introduce a weighted average flux

$$\widetilde{F}_{i+1/2} = \frac{\Delta x_{i+1}\mathcal{F}_l + \Delta x_i \mathcal{F}_r}{\Delta x_i + \Delta x_{i+1}} = F_{i+1/2} + \frac{\Delta x_i S_{i+1/2+} - \Delta x_{i+1} S_{i+1/2-}}{\Delta x_i + \Delta x_{i+1}}. \tag{3.12}$$

Then by the first line in (3.11), we have $\widetilde{F}_{i+1/2} = F_{i+1/2} - S_{i+1/2-} + O(\Delta x_i \hbar^2)$ and also $\widetilde{F}_{i+1/2} = F_{i+1/2} + S_{i+1/2+} + O(\Delta x_{i+1}\hbar^2)$. Therefore,

$$\begin{aligned} \widetilde{F}_{i+1/2} - \widetilde{F}_{i-1/2} &= F_{i+1/2} - F_{i-1/2} - S_{i+1/2-} - S_{i-1/2+} + O(\Delta x_i \hbar^2) \\ &= F_{i+1/2} - F_{i-1/2} - S_i + O(\Delta x_i \hbar^2), \end{aligned} \tag{3.13}$$

and (3.1) can be rewritten as

$$\Delta x_i \frac{d}{dt} U_i(t) + \widetilde{F}_{i+1/2} - \widetilde{F}_{i-1/2} = S_{ci} + O(\Delta x_i \hbar^2), \tag{3.14}$$

which proves the second-order accuracy. □

Some important features arising in the second-order reconstruction must now be specified. First, the cell by cell reconstruction preserves the mass conservation property of the finite volume method. Second, the limitation procedure ensures the nonnegativity of the second-order reconstructed water heights. The third important point is that the second-order reconstruction has to preserve the lake at rest steady state. To ensure this property we reconstruct also the bottom topography $z(x)$ although it is a data. The idea to do so is not so new - see [21], [11] - but here we give details on the more stable way to do it. Indeed only two of the three quantities $h, z, h + z$ need be explicitly reconstructed, the last being necessarily a combination of the other two. This is consistent with the strategy for second order extensions of a well-balanced Godunov-type scheme for multi-dimensional compressible flow under gravity in [5, 6], in that the *deviations from the non-constant steady state* form the basis for reconstruction and slope limiting. A critical test to make the right choice is given by considering a lake at rest with non vertical shores, that is nothing else but considering an interface between a wet cell and a dry cell in the case where the bottom of the dry cell is higher than the free surface in the wet cell and where the fluid is at rest in the wet cell. As it appears in Figure 1, for the minmod reconstruction, the only choice which preserves the steady state - or even the nonnegativity of water height in the worst case, see the third subfigure - is to work with the quantities h and $h+z$. Notice that it follows that in some respect the bottom topography changes at each timestep. It is obvious then that the chosen second-order reconstruction preserves also the steady state in the classical case of wet-wet interfaces since we explicitly reconstruct the quantity $h+z$. The second-order-centered condition (3.11) can be realized with a second-order ENO reconstruction for example, but in practice we shall not do so because it becomes too complicate for 2d unstructured meshes, even if it necessarily means a slight loss of accuracy.

4 Numerical results

All numerical tests are computed with a kinetic solver. This solver is based on the kinetic theory developed in [25] and has the advantage - in the homogeneous case - to keep the water height nonnegative, to verify a discrete in-cell entropy inequality and to be able to compute problems with shocks or vacuum.

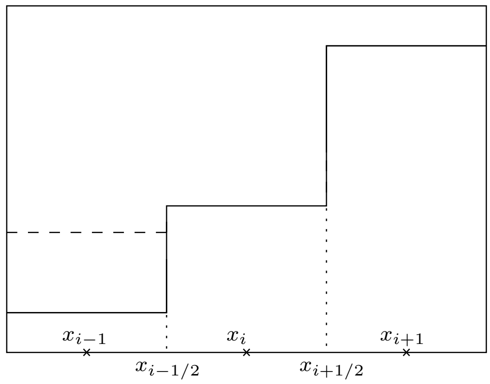
4.1 1d assessments

We first illustrate that the hydrostatic reconstruction does not affect the robustness of the homogeneous solver. We present a very classical numerical test of a constant discharge transcritical flow with shock over a bump - refer to [15] for a complete presentation. In Figure 3, where 101 points are used, we observe good first and second-order results for this test the stiffness of which is well-known. As we are far from a hydrostatic steady state the results of the well-balanced and standard schemes are quite similar. Notice however that the well-balanced version is less affected - whatever is the order of resolution - than the standard scheme where the derivative of the bottom topography presents strong variations.

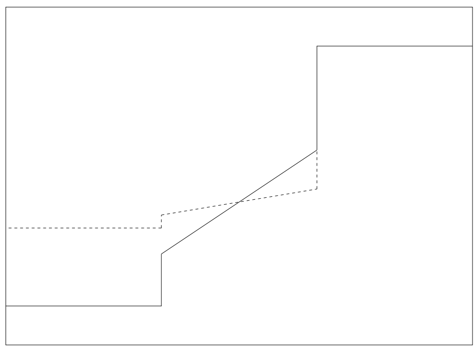
To exhibit the improvement due to the hydrostatic reconstruction we present now a quasi stationary case first proposed by Leveque in [23] which consists in computing small perturbations of the steady state of a lake at rest with a varying bottom topography,

$$\begin{aligned} z(x) &= \left(0.25\left(\cos\left(\pi(x-1.5)/0.1\right)+1\right)\right)_+ , \\ h(0,x) &= 1. + 0.001\mathbb{1}_{[1.1,1.2]}. \end{aligned}$$

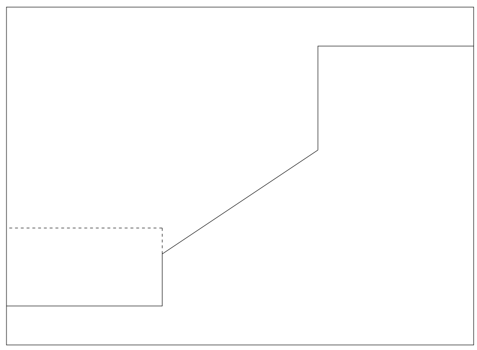
As we can see by considering the linearized equations, the small perturbation simply move to the right with a speed equal to $\sqrt{h(t,x)}$ i.e. $\sqrt{1-z(x)}$ at first-order approximation - the gravity is equal to one. We present in Figure 4 the results obtained - at $t = 0.7s$ and with 150 points - with the well-balanced scheme - on the right - and with the standard one - on



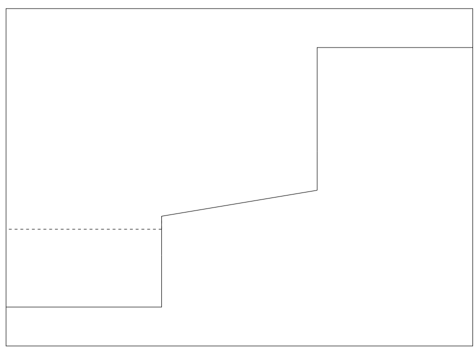
First-order solution



Reconstructed quantities : z and h+z



Reconstructed quantities : z and h



Reconstructed quantities : h and h+z

Figure 1: Second-order reconstruction strategy

Free surface (dotted line)
Bottom topography (continuous line)

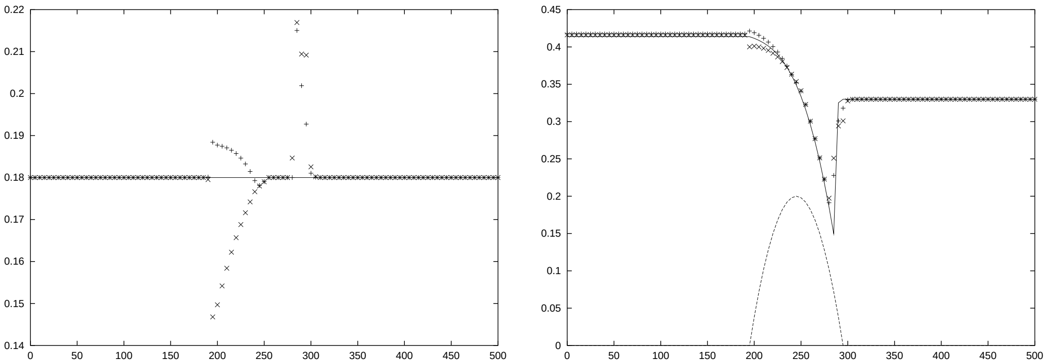


Figure 2: Constant discharge problem with shock - Discharge and water height

First-order standard scheme (times crosses)
First-order well-balanced scheme (plus crosses)
Exact solution and bottom topography (continuous and dotted lines)

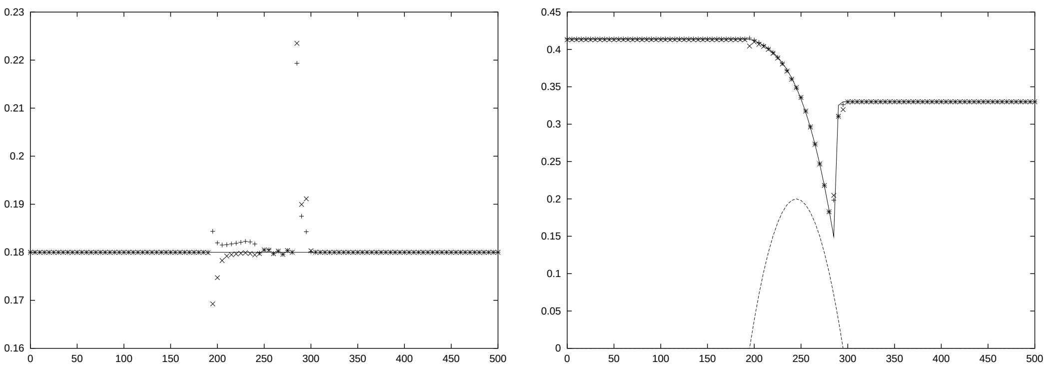
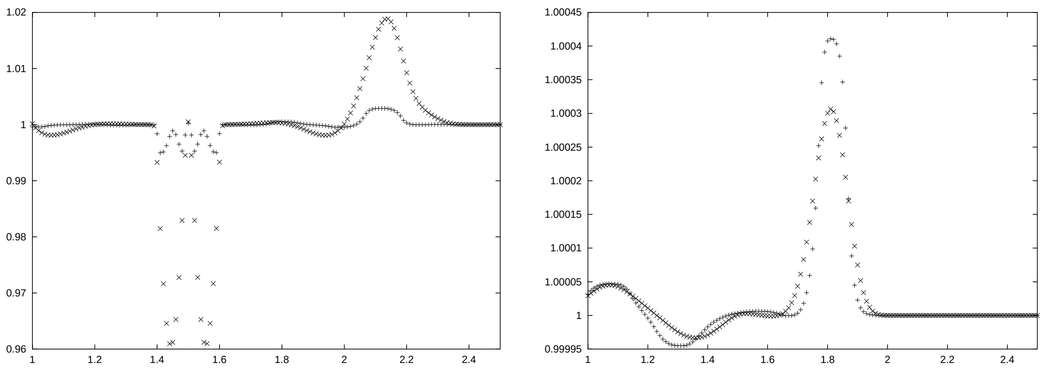


Figure 3: Constant discharge problem with shock - Discharge and water height

Second-order standard scheme (times crosses)
Second-order well-balanced scheme (plus crosses)
Exact solution and bottom topography (continuous and dotted lines)

the left. It appears that even for the second-order computation the unphysical perturbations induced by the standard scheme are larger than the initial perturbation of the free surface - the vertical scales differ significantly in the two subfigures... Moreover the standard scheme induced not only perturbation on the bump but also a perturbation which moves to the right with the same speed as the initial perturbation but which is more than one order of magnitude greater than the initial perturbation. On the contrary the results obtained with the well-balanced scheme are quite good, even for the first-order solution.



Standard scheme Well-balanced scheme

Figure 4: Quasi stationary problem with a small perturbation

First-order scheme (times crosses)
Second-order scheme (plus crosses)

To end with the 1d numerical assessments we present a test case which is determinant to test the robustness of a solver, since it needs to treat vacuum. It exhibits very clearly the improvement due to the second-order extension. We are interested in the case of an oscillating lake with a non flat bottom and non vertical shores. The lake is initially at rest but a small sinusoidal perturbation affect the free surface,

$$z(x) = .5(1 - .5(\cos(\pi(x - .5)/.5) + 1)),$$
$$h(0, x) = \max(0, .4 - z(x) + .04 \sin((x - .5)/.25) - \max(0, -.4 + z(x))).$$

Then the flow oscillates and at each timestep we have to treat an interface between a wet cell and a dry cell on each shore of the lake. We present in Figure 5 the results obtained with the well-balanced scheme - with 200 points and at $t = 19.87s$ because it corresponds to a time where the flow reaches its higher level on the left shore. Both first and second-order well-balanced schemes are robust but the first-order scheme damps the oscillations quickly, fifty oscillations are enough to get back to rest. On the other hand the second-order well-balanced scheme keeps the periodic regime up to the machine accuracy.

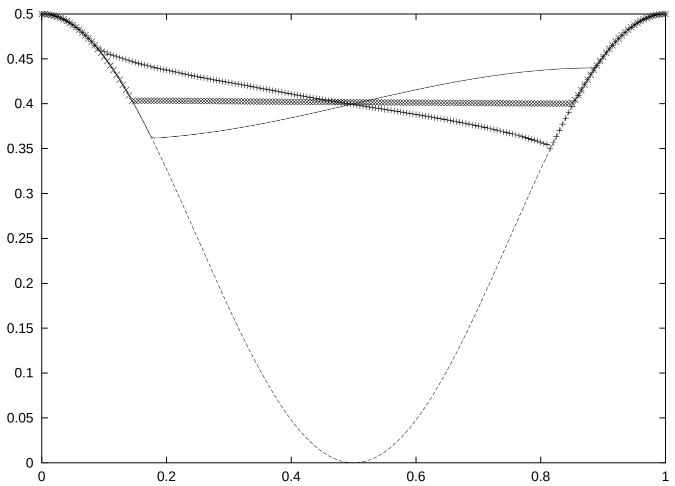


Figure 5: Oscillating lake - Well-balanced scheme

First-order scheme (times crosses)
Second-order scheme (plus crosses)
Initial solution and bottom topography (continuous and dotted lines)

4.2 2d assessments

As the extension to second-order accuracy, the extension to the bidimensional case does not modify the idea of the method. Nevertheless there exist some specific problems, especially for the construction of a 2d well-balanced second-order scheme. Then we just explain here the very basic background to fix the ideas and we refer to [2] for a more detailed description - and especially for the explanation of the 2d hydrostatic second-order reconstruction.
Starting from a triangulation of a 2d domain, the 2d finite volume method consists in constructing the so-called dual cells and then in using an upwinding of the fluxes on each cell interface with an interpolation of the normal component of the flux along each edge of the cell. Since the problem locally looks like a planar discontinuity at the cell interface, the interpolation can be performed using a 1d solver. For more details we refer to [1]. Then locally the *well-balanced* property is also a 1d problem and the hydrostatic reconstruction can be extended from the 1d case without major modification. By construction, the 2d *well-balanced* scheme preserves the lake at rest steady state and it preserves the water height nonnegativity if the original 1d solver does.

We first present the academic case of a dam break on a dry bed but containing a wet zone which consists in a small lake at rest. This case involves the vacuum and it allows to exhibit the effect of the hydrostatic reconstruction to preserve the initially at rest area. The first subfigure in Figure 6 presents the mesh and the bottom topography - the bottom topography of the lake we can see on the right is hemispheric. On the second subfigure we can see the initial water height: we see the dam on the middle and the small lake at rest on the right - the free surface level in the lake coincides with the reference level of the bottom topography of the river, i.e. $(h + z)(0, x, y) = 0$ everywhere on the right of the dam. On the third subfigure we can see the rarefaction wave. Since it does not yet reach the lake, the steady state is preserved. Then on the fourth subfigure the rarefaction wave reaches the lake and the water begins to move. On the last subfigure is presented the free surface level at this final time. We can notice strong variations on the lake area which lead to the formation of a hole - in the left part of the lake, in blue on the subfigure - and a bump - in the right part, in green in the subfigure - in the free surface.

Then we present in Figure 7 another 2d numerical test corresponding to the filling up of a river. This test still involves vacuum but also deals with complex realistic geometry and bottom topography since it takes into account a jetty in the transversal direction - on the upper part of the figures, a bridge pillar - the square on the lower part - and a small bump in the bottom topography. We start with an empty river and we prescribe a given water level as inflow condition. On the first subfigure are presented the mesh and the associated bottom topography. Then we can notice that the strong variations in the bottom topography due to the jetty or the pillar bridge does not affect the robustness of the computation. On the third and fourth subfigures we can see the bump since the water skirts it.

More results can be found in [1], [10], [26], [24], and in [8] with Coriolis force.

Acknowledgements

This work was partially supported by the ACI Modélisation de processus hydrauliques à surface libre en présence de singularités (http://www-rocq.inria.fr/m3n/CatNat/), by HYKE European programme HPRN-CT-2002-00282 (http://www.hyke.org), by EDF/LNHE (E.A., F.B., M.-O.B., B.P.), and by grant KL 611/6 of the Deutsche Forschungsgemeinschaft (R.K.).

References

[1] E. Audusse, M.-O. Bristeau, B. Perthame, Kinetic schemes for Saint-Venant equations with source terms on unstructured grids, INRIA Report RR-3989 (2000), http://www.inria.fr/RRRT/RR-3989.html

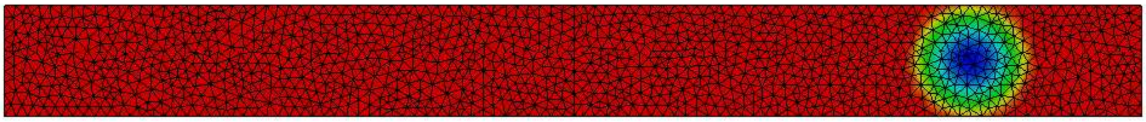
[2] E. Audusse, M.-O. Bristeau, B. Perthame, Second order kinetic scheme for Saint-Venant equations with source terms on unstructured grids, preprint.

[3] D.S. Bale, R.J. Leveque, S. Mitran, J.A. Rossmanith, A wave propagation method for conservation laws and balance laws with spatially varying flux functions, SIAM J. Sci. Comput. 24 (2002), 955-978.

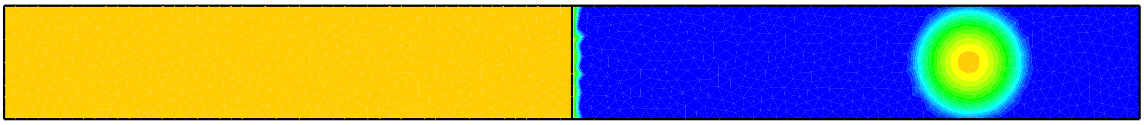
[4] R. Botchorishvili, B. Perthame, A. Vasseur, Equilibrium schemes for scalar conservation laws with stiff sources, Math. Comp. 72 (2003), 131-157.

[5] N. Botta, R. Klein, S. Langenberg, S. Lützenkirchen, Well-balanced finite volume methods for nearly hydrostatic flows, submitted to J. Comp. Phys., March (2002).

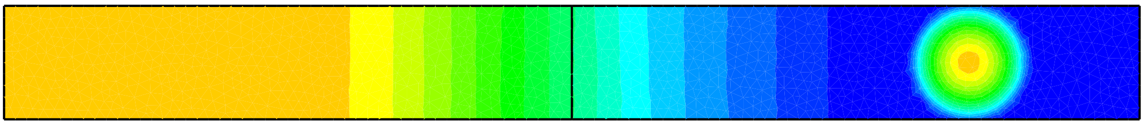
[6] N. Botta, R. Klein, A. Owinoh, Distinguished Limits, Multiple Scales Asymptotics, and Numerics for Atmospheric Flows, Amer. Meteorological Society, 13th Intl. Conf. on Atmosphere-Ocean Fluid Dynamics, Breckenridge, Colorado, July (2001)



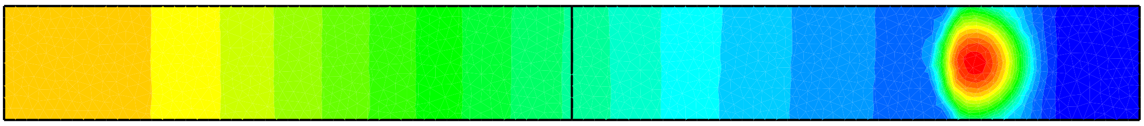
Mesh and bottom topography



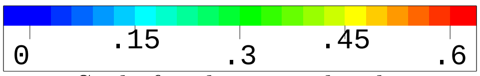
Water height - Initial Solution



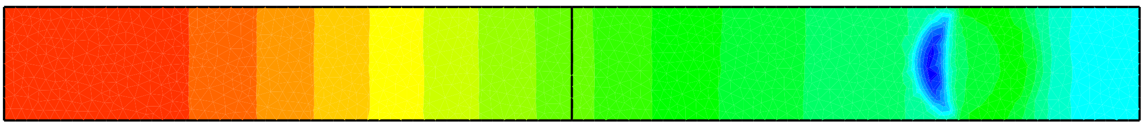
Water height - Solution at t=.08s



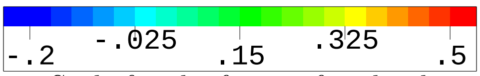
Water height - Solution at t=.16s



Scale for the water height



Free surface - Solution at t=.16s



Scale for the free surface level

Figure 6: 2D dam break on dry bed with a lake at rest area

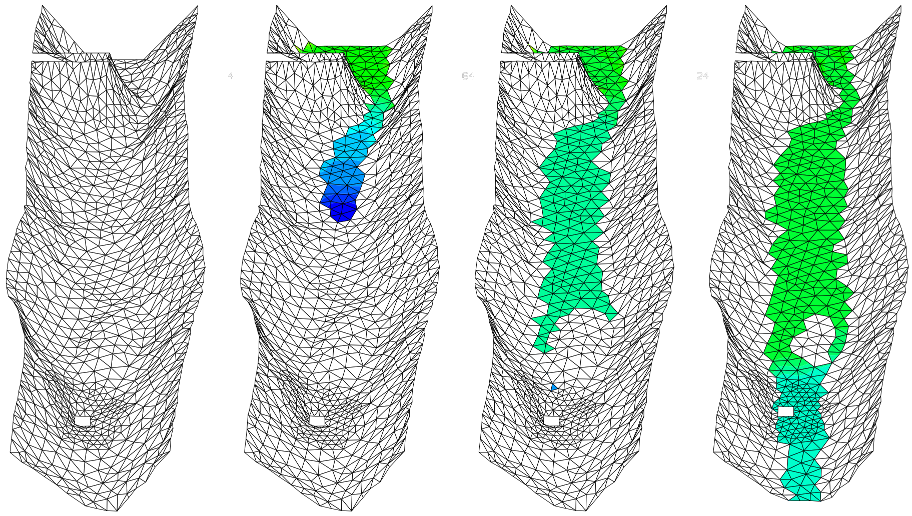


Figure 7: Filling up of a river - Bottom topography and free surface at different times

[7] F. Bouchut, An introduction to finite volume methods for hyperbolic systems of conservation laws with source, Ecole CEA - EDF - INRIA Ecoulements peu profonds à surface libre, 7-10 octobre 2002, INRIA Rocquencourt, available at http://www.dma.ens.fr/~fbouchut/publications/fvcours.ps.gz

[8] F. Bouchut, J. LeSommer, V. Zeitlin, Frontal geostrophic adjustment and nonlinear-wave phenomena in one dimensional rotating shallow water; Part 2 : high resolution numerical investigation, preprint.

[9] F. Bouchut, A. Mangeney-Castelnau, B. Perthame, J.-P. Vilotte, A new model of Saint-Venant and Savage-Hutter type for gravity driven shallow water flows, C. R. Acad. Sci. Paris Sér. I Math. 336 (2003), 531-536.

[10] M.-O. Bristeau, B. Coussin, Boundary conditions for the shallow water equations solved by kinetic schemes, INRIA Report RR-4282 (2001), http://www.inria.fr/RRRT/RR-4282.html

[11] T. Gallouët, J.-M. Hérard, N. Seguin, Some approximate Godunov schemes to compute shallow-water equations with topography, Computers and Fluids 32 (2003), 479-513.

[12] E. Godlewski, P.-A. Raviart, Numerical approximation of hyperbolic systems of conservation laws, Applied Mathematical Sciences, 118, Springer-Verlag, New York, 1996.

[13] L. Gosse, A well-balanced flux-vector splitting scheme designed for hyperbolic systems of conservation laws with source terms, Comp. Math. Appl. 39 (2000), 135-159.

[14] L. Gosse, A.-Y. LeRoux, A well-balanced scheme designed for inhomogeneous scalar conservation laws, C. R. Acad. Sci. Paris Sér. I Math. 323 (1996), 543-546.

[15] N. Goutal, F.Maurel, Proceedings of the second workshop on dam-break wave simulation, EDF-DER, (1997).

[16] 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.

[17] S. Jin, A steady-state capturing method for hyperbolic systems with geometrical source terms, Math. Model. Numer. Anal. 35 (2001), 631-645.

[18] T. Katsaounis, C. Simeoni, First and second order error estimates for the Upwind Interface Source method, preprint.

[19] T. Katsaounis, B. Perthame, C. Simeoni, Upwinding Sources at Interfaces in conservation laws, to appear in Appl. Math. Lett. (2003).

[20] A.-Y. LeRoux, Riemann solvers for some hyperbolic problems with a source term, ESAIM proc. 6 (1999), 75-90.

[21] A.-Y. LeRoux, Discrétisation des termes sources raides dans les problèmes hyperboliques. In : Systèmes hyperboliques : Nouveaux schémas et nouvelles applications. Ecoles CEA-EDF-INRIA 'problèmes non linéaires appliqués', INRIA Rocquencourt (France), March 1998 [in French]. Available from http://www-gm3.univ-mrs.fr/~leroux/publications/ay.le_roux.html.

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

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

[24] A. Mangeney, J.-P. Vilotte, M.-O. Bristeau, B. Perthame, C. Simeoni, S. Yernini, Numerical modeling of avalanches based on Saint-Venant equations using a kinetic scheme, preprint 2002.

[25] B. Perthame, Kinetic formulations of conservation laws, Oxford University Press (2002).

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

[27] B. Perthame, C. Simeoni, Convergence of the upwind Interface source method for hyperbolic conservation laws, to appear in Proc. of Hyp2002, T. Hou and E. Tadmor Editors.

[28] E.F. Toro, Riemann solvers and numerical methods for fluid dynamics. A practical introduction. Second edition. Springer-Verlag, Berlin, 1999.

[29] M.E. Vazquez-Cendon, Improved treatment of source terms in upwind schemes for the shallow water equations in channels with irregular geometry, J. Comput. Phys. 148 (1999), 497-526.

[30] K. Xu, A well-balanced gas-kinetic scheme for the shallow water equations with source terms, J. Comput. Phys. 178(2) (2002), 533-562.