# Staggered and well-balanced discretization of shallow-water equations

Nguyen Quan Ba Hong

May 9, 2019

**Abstract**

In this context, we consider a class of finite volume schemes for the shallow water equations with variable bottom topography.

## Contents



We follow the notations in [4].

# 1 Introduction

The shallow water equations are a nonlinear hyperbolic system of conservation laws with a source term due to the variable bottom topography.

*Balance laws* often consist of the conservation laws for the vector $U(x,t)$ of *mass* and *momentum*, accelerated by *conservation advection* and *pressure forces* (denoted by $-\frac{\partial F(U)}{\partial x}$ in (1.1) below), and by additional nonconservative forces $S(U,x)$, also called *source terms*.

The *equations of motion* may be written as

$$\frac{\partial U}{\partial t}+\frac{\partial F(U)}{\partial x}=S(U,x). \tag{1.1}$$

In this context, we consider the case where there is no source term in the equation of mass, so $S$ can be written as $S=(0,s)^T$.

The *residuum* $R$ is defined as

$$R(x,t):=-\frac{\partial F(U)}{\partial x}+S(U,x), \tag{1.2}$$

which indicates *near-equilibrium flows* when it nearly vanishes.

A *semidiscrete, first order accurate finite volume scheme* (abbr., se-dis $1^{st}$ FVM) may be written as a method of lines

$$\frac{d}{dt}U_i(t)=R_i(t):=-\frac{1}{\Delta x}\left(F_{i+\frac{1}{2}}-F_{i-\frac{1}{2}}\right)+S_i, \tag{1.3}$$

where $U_i$ approximates the *cell average* over the $i^{\text{th}}$ cell $C_i:=\left[x_{i-\frac{1}{2}},x_{i+\frac{1}{2}}\right]$ at time $t$, $R_i(t)$ is the cell average of the residuum, $\Delta x:=x_{i+\frac{1}{2}}-x_{i-\frac{1}{2}}$ is the *spatial grid size*, $F_{i\pm+\frac{1}{2}}$ is a *conservative numerical flux function*, and $S_i$ approximates the cell average of the source term.

In this context, we focus on one-dimensional (1-D) shallow water equations, given by

$$U=\begin{pmatrix}h\\hu\end{pmatrix},\quad F(U)=\begin{pmatrix}hu\\hu^2+\frac{1}{2}gh^2\end{pmatrix},\text{ and }S(U,x)=-\begin{pmatrix}0\\ghb'\end{pmatrix}. \tag{1.4}$$

More explicitly,

$$\begin{cases}\partial_t h+\partial_x(hu)=0,\quad g=\text{const},\\ \partial_t(hu)+\partial_x\left(hu^2+\frac{1}{2}gh^2\right)=-ghb'(x),\end{cases} \tag{1.5}$$

Here $b(x)$ is the *bottom topography*[1], $h(x,t)$ the *water height*, $u(x,t)$ the *water velocity*, and $g=9.8m/s^2$ the *gravitational acceleration*. Thus the source term $S$ models the force of gravity tangential to a sloped bottom[2]. The residuum can be rewritten as

$$R=-\partial_x\begin{pmatrix}hu\\hu^2\end{pmatrix}-gh\partial_x\begin{pmatrix}0\\w\end{pmatrix}, \tag{1.6}$$

[1]Here the bottom is assumed *rigid* under the effect of gravity, i.e., the bottom is time-independent $\partial_t b=0$ on $\Gamma_{\text{bot}}$, see [4, pp. 2–3] for more details.

[2]The bottom topography $b$ is assumed to be of class $C_1$ for simplicity, but well-balanced schemes, which preserves the lake at rest discretely, use either continuous or discontinuous topography, see [2, p. 760].

where $w := h + b$ is the *water level*.

For such a problem, where shocks can form in the solution, finite volume methods have proved to be very effective.

We recall two important *equilibria* stated in [2, p. 759]:

i) *still water*, where

$$u = 0 \text{ and } \partial_x w = 0, \text{ and} \tag{1.7}$$

ii) the *lake at rest*, which is stil water together with dry boundaries:

$$u = 0 \text{ and } h\partial_x w = 0. \tag{1.8}$$

Thus, the lake at rest residuum combines the dry shore ($h = 0$) with the flat water level ($\partial_x w = 0$) in a single product, which suggests a natural splitting of the nonconservative product at the wet-dry front.

The goal of this context is to modify numerical schemes on staggered grids for a precise consideration of these equilibrium states.

These second equilibria make it possible to take into account the transitions between dry zones and wet areas, whereas the former, defined by relation (**??**), presuppose water everywhere. We hope, and in face we see, that preserving exactly these states at the discrete level are also more accurate in the transient case.

We are interested in numerical schemes whose water height and velocity have shifted discretizations. In addition, we restrict ourselves to schemes that allow to capture shock waves. We consider $N_x$ points (or $N_x$ meshes), and the unknown notations from [5].

- water level: $h_i^n$, $i \in \{1, \dots, N_x\}$,
- velocity: $u_{i+\frac{1}{2}}$, $i \in \{0, \dots, N_x\}$,

and $n$ denotes the number of the iteration in time.

In addition to equation (1.5), considered on $[0, T] \times (0, L)$, $T$ being the final time and $L$ the length of the domain which has dry shore at its both boundaries, we give ourselves an initial condition

$$(h, hu)(0, x) = (h_0, q_0)(x), \quad x \in (0, L),$$

and conditions at the boundaries

$$(hu)(t, 0) = (hu)(t, L) = 0, \quad t \in [0, T].$$

We further define the space step $\Delta x = \frac{L}{N_x}$ and the time step $\Delta t = \frac{T}{N_t}$ which will be subjected to a stability condition.

## 2 A staggered scheme in 1-D by Doyen and Gunawan

We reuse the notations in Sec. 2, [3, p. 229]: The left end, the center and the right end of the $i$-th cell are denoted by $x_{i-\frac{1}{2}}$, $x_i$ and $x_{i+\frac{1}{2}}$, respectively. We set $C_i := \left(x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}\right)$ for all $i \in \mathcal{M} := \{1, \ldots, N_x\}$, $\mathcal{E}_{\text{int}} := \{1, \ldots, N_x - 1\}$, $\mathcal{E}_b := \{0, N_x\}$, and $\mathcal{E} := \mathcal{E}_{\text{int}} \cup \mathcal{E}_b$. The water height $h$ and the topography $b$ are discretized at the center of the cells. The approximation of $h$ at point $x_i$ and at time $t^n$ is denoted by $h_i^n$. The approximation of $b$ at point $x_i$ is denoted by $b_i$. The velocity $u$ is discretized at the interfaces between the cells. The approximation of $u$ at point $x_{i+\frac{1}{2}}$ and at time $t^n$ is denoted by $u_{i+\frac{1}{2}}^n$.

The mass conservation equation is discretized with an explicit upwind scheme:

$$h_i^{n+1} = h_i^n - \frac{\Delta t}{\Delta x}\left(F_{i+\frac{1}{2}}^n - F_{i-\frac{1}{2}}^n\right), \quad \forall i \in \mathcal{M}, \tag{2.1}$$

where

$$F_{i+\frac{1}{2}}^n := h_{i+\frac{1}{2}}^n u_{i+\frac{1}{2}}^n, \quad \forall i \in \mathcal{M},$$

and where $h_{i+\frac{1}{2}}^n$ is calculated by an upwind shift according to the sign of $u_{i+\frac{1}{2}}^n$:

$$\forall i \in \mathcal{E}, \quad h_{i+\frac{1}{2}}^n = \begin{cases} h_i^n, & \text{if } u_{i+\frac{1}{2}}^n \geq 0, \\ h_{i+1}^n, & \text{otherwise.} \end{cases}$$

**Proposition 1** (Conservation of the total water height)**.** *The explicit upwind scheme* (2.1) *preserves the total water height.*

*Proof.* Define

$$\mathcal{H}_n := \Delta x \sum_{i=1}^{N_x} h_i^n, \quad \forall n \in \{1, \ldots, N_t\},$$

we claim that $\mathcal{H}_{n+1} = \mathcal{H}_n$ for all $n \in \{0, \ldots, N_t - 1\}$. Indeed,

$$\begin{aligned}
\mathcal{H}_{n+1} &= \Delta x \sum_{i=1}^{N_x} h_i^{n+1} \\
&= \Delta x \sum_{i=1}^{N_x} \left[h_i^n - \frac{\Delta t}{\Delta x}\left(F_{i+\frac{1}{2}}^n - F_{i-\frac{1}{2}}^n\right)\right] \\
&= \Delta x \sum_{i=1}^{N_x} h_i^n - \Delta t \sum_{i=1}^{N_x} \left(F_{i+\frac{1}{2}}^n - F_{i-\frac{1}{2}}^n\right) \\
&= \mathcal{H}_n - \Delta t \left(F_{N_x+\frac{1}{2}}^n - F_{\frac{1}{2}}^n\right) \\
&= \mathcal{H}_n, \quad \forall n \in \{1, \ldots, N_t - 1\},
\end{aligned}$$

and then $\mathcal{H}_n = \mathcal{H}_1$ for all $n \in \{1, \ldots, N_t\}$. This completes our proof. □

The momentum balance equation in (1.5) is discretized with explicit upwind fluxes for the convection term and implicit centered fluxes for the pressure term and topography term:

$$\bar{h}^{n+1}_{i+\frac{1}{2}} u^{n+1}_{i+\frac{1}{2}} = \bar{h}^{n}_{i+\frac{1}{2}} u^{n}_{i+\frac{1}{2}} - \frac{\Delta t}{\Delta x}\left[G^n_{i+1} - G^n_i + \frac{g}{2}\left(\left(h^{n+1}_{i+1}\right)^2 - \left(h^{n+1}_i\right)^2\right) + g\bar{h}^{n+1}_{i+\frac{1}{2}}\left(b_{i+1} - b_i\right)\right], \quad \forall i \in \mathcal{E}_{\text{int}}, \tag{2.2}$$

where

$$\bar{h}^n_{i+\frac{1}{2}} := \frac{1}{2}\left(h^n_i + h^n_{i+1}\right), \quad \forall i \in \mathcal{E}_{\text{int}},$$

and

$$G^n_i = \frac{1}{2} u^n_i \left(F^n_{i-\frac{1}{2}} + F^n_{i+\frac{1}{2}}\right),$$

where $u^n_i$ is calculated by an upwind shift according to the sign of $F^n_{i-\frac{1}{2}} + F^n_{i+\frac{1}{2}}$:

$$\forall i \in \mathcal{M}, \quad u^n_i = \begin{cases} u^n_{i-\frac{1}{2}}, \text{ if } F^n_{i-\frac{1}{2}} + F^n_{i+\frac{1}{2}} \geq 0, \\ u^n_{i+\frac{1}{2}}, \text{ otherwise.} \end{cases}$$

The discrete boundary conditions are

$$u^{n+1}_{i+\frac{1}{2}} = 0, \quad \forall i \in \mathcal{E}_b. \tag{2.3}$$

The computation of the discrete unknowns at each time step is completely explicit. First the discrete water heights $\left\{h^{n+1}_i\right\}$ are computed with (2.1), then the discrete velocities $\left\{u^{n+1}_{i+\frac{1}{2}}\right\}$ are computed with (2.2) (if $\bar{h}^{n+1}_{i+\frac{1}{2}} = 0$ , by convention, $u^{n+1}_{i+\frac{1}{2}}$ is set to zero).

**Proposition 2** (Preserved quantities)**.** *This scheme conserves the mass, and, for a flat topography, the total momentum, provided the space step is small enough for the latter.*

*Proof.* We will prove the conservation properties for the following quantities:

- *Mass*: Recall in [4] that the mass is defined by

$$\mathcal{Z} := \int_{\mathbb{R}} \zeta\left(t, x\right) dx = \int_{\mathbb{R}} \left(h\left(t, x\right) + b\left(x\right) - H\right) dx,$$

  this quantity is preserved (see [4, Sec. 3.1, p. 21]):

$$\frac{d}{dt}\mathcal{Z} = 0. \tag{2.4}$$

  Now we prove this property in the discrete level. To do this, we define

$$\mathcal{Z}_n := \Delta x \sum_{i=1}^{N_x} \left(h^n_i + b_i - H\right), \quad \forall n \in \{0, \ldots, N_t\}.$$

We claim that $\mathcal{Z}_{n+1} = \mathcal{Z}_n$ for all $n \in \{0, \dots, N_t - 1\}$. Indeed,

$$\begin{aligned}
\mathcal{Z}_{n+1} &= \Delta x \sum_{i=1}^{N_x} \left(h_i^{n+1} + b_i - H\right) \\
&= \Delta x \sum_{i=1}^{N_x} \left(h_i^n - \frac{\Delta t}{\Delta x}\left(F^n_{i+\frac{1}{2}} - F^n_{i-\frac{1}{2}}\right) + b_i - H\right) \\
&= \Delta x \sum_{i=1}^{N_x} \left(h_i^n + b_i - H\right) - \Delta t \left(\sum_{i=1}^{N_x} F^n_{i+\frac{1}{2}} - \sum_{i=1}^{N_x} F^n_{i-\frac{1}{2}}\right) \\
&= \mathcal{Z}_n - \Delta t \left(\sum_{i=1}^{N_x} F^n_{i+\frac{1}{2}} - \sum_{i=0}^{N_x-1} F^n_{i+\frac{1}{2}}\right) \\
&= \mathcal{Z}_n - \Delta t \left(F^n_{N_x+\frac{1}{2}} - F^n_{\frac{1}{2}}\right) \\
&= \mathcal{Z}_n, \quad \forall n \in \{0, \dots, N_t - 1\},
\end{aligned}$$

where the last equality is deduced from (2.3). We then have $\mathcal{Z}_n = \mathcal{Z}_0$, for all $n \in \{0, \dots, N_t\}$, i.e., $\mathcal{Z}_n$ is independent in time. This is the discrete version of (2.4).

- *Total momentum*: Recall that the total momentum is defined by

$$\mathfrak{M} := \int_{\mathbb{R}} (hu)(t, x)\, dx,$$

this quantity is preserved (see [4, Sec 3.1, p.21]):

$$\frac{d}{dt}\mathfrak{M} = 0. \tag{2.5}$$

Similarly, we now look at this property in the discrete level. Define

$$\mathfrak{M}_n := \Delta x \sum_{i=1}^{N_x-1} \bar{h}^n_{i+\frac{1}{2}} u^n_{i+\frac{1}{2}}, \quad \forall n \in \{0, \dots, N_t\},$$

we claim that $\mathfrak{M}_{n+1} = \mathfrak{M}_n$ for all $n \in \{0, \dots, N_t - 1\}$ provided that the bottom topography is flat. Indeed,

$$\begin{aligned}
\mathfrak{M}_{n+1} &= \Delta x \sum_{i=1}^{N_x-1} \bar{h}^{n+1}_{i+\frac{1}{2}} u^{n+1}_{i+\frac{1}{2}} \\
&= \Delta x \sum_{i=1}^{N_x-1} \left[\bar{h}^n_{i+\frac{1}{2}} u^n_{i+\frac{1}{2}} - \frac{\Delta t}{\Delta x}\left[G^n_{i+1} - G^n_i + \frac{g}{2}\left(\left(h^{n+1}_{i+1}\right)^2 - \left(h^{n+1}_i\right)^2\right)\right]\right] \\
&= \Delta x \sum_{i=1}^{N_x-1} \bar{h}^n_{i+\frac{1}{2}} u^n_{i+\frac{1}{2}} - \Delta t \sum_{i=1}^{N_x-1} \left[G^n_{i+1} - G^n_i + \frac{g}{2}\left(\left(h^{n+1}_{i+1}\right)^2 - \left(h^{n+1}_i\right)^2\right)\right] \\
&= \mathfrak{M}_n - \Delta t \left[G^n_{N_x} - G^n_1 + \frac{g}{2}\left(\left(h^{n+1}_{N_x}\right)^2 - \left(h^{n+1}_1\right)^2\right)\right]
\end{aligned}$$

$$
\begin{aligned}
&= \mathfrak{M}_n - \Delta t \left[ \begin{array}{l} \frac{1}{2} u^n_{N_x} \left( F^n_{N_x - \frac{1}{2}} + F^n_{N_x + \frac{1}{2}} \right) - \frac{1}{2} u^n_1 \left( F^n_{\frac{1}{2}} + F^n_{\frac{3}{2}} \right) \\ + \frac{g}{2} \left( \left( h^n_{N_x} - \frac{\Delta t}{\Delta x} \left( F^n_{N_x + \frac{1}{2}} - F^n_{N_x - \frac{1}{2}} \right) \right)^2 - \left( h^n_1 - \frac{\Delta t}{\Delta x} \left( F^n_{\frac{3}{2}} - F^n_{\frac{1}{2}} \right) \right)^2 \right) \end{array} \right] \\
&= \mathfrak{M}_n - \Delta t \left[ \frac{1}{2} u^n_{N_x} F^n_{N_x - \frac{1}{2}} - \frac{1}{2} u^n_1 F^n_{\frac{3}{2}} + \frac{g}{2} \left( \left( h^n_{N_x} + \frac{\Delta t}{\Delta x} F^n_{N_x - \frac{1}{2}} \right)^2 - \left( h^n_1 - \frac{\Delta t}{\Delta x} F^n_{\frac{3}{2}} \right)^2 \right) \right] \\
&= \mathfrak{M}_n, \quad \forall n \in \{0, \ldots, N_t - 1\},
\end{aligned}
$$

where the last inequality is obtain by choosing $N_x$ large enough for which

$$
h^n_1 = h^n_{\frac{3}{2}} = h^n_{N_x - \frac{1}{2}} = h^n_{N_x} = 0
$$

holds. We then have $\mathfrak{M}_n = \mathfrak{M}_0$, for all $n \in \{0, \ldots, N_t\}$, i.e., $\mathfrak{M}_n$ is independent in time. This is the discrete version of (2.5).

This completes our proof. □

We apply the Finite Volume Method for 1-D scalar conservation laws for the mass conservation equation

$$
\partial_t h + \partial_x f(h) = 0, \quad \text{in } [0, T] \times (0, L),
$$

where $f(h) = (hu)(x, t)$.

We consider the numerical flux $g : \mathbb{R}^2 \to \mathbb{R}$ defined piecewisely by

$$
g(h, k) := \begin{cases} h u^n_{i+\frac{1}{2}}, & \text{if } u^n_{i+\frac{1}{2}} \geq 0, \\ k u^n_{i+\frac{1}{2}}, & \text{otherwise,} \end{cases} \quad \text{on } C_i \times M_{i+1}.
$$

This numerical flux $g$ is not consistent but *monotone* $g(\nearrow, \searrow)$, i.e., $g$ is non-decreasing w.r.t. its first variable and non-increasing w.r.t. its second variable.

**Proposition 3.** *If the Courant-Friedrichs-Levy-like (CFL-like) condition:*

$$
\Delta t \leq \frac{\Delta x}{\left( -u^n_{i\frac{1}{2}} \right)_- + \left( u^n_{i+\frac{1}{2}} \right)_+}, \quad \forall i \in \mathcal{M} \tag{2.6}
$$

*holds (where, for any $a \in \mathbb{R}$, $(a)_+ := \max(a, 0)$ and $(a)_- := \min(a, 0)$) then the numerical scheme (2.1) is monotone: Let $H : \mathbb{R}^3 \to \mathbb{R}$ defined by $h^{n+1}_i = H\left(h^n_{i-1}, h^n_i, h^n_{i+1}\right)$ then $H$ is non-decreasing w.r.t. its three variables $H(\nearrow, \nearrow, \nearrow)$. Therefore*

$$
A \leq h_0 \leq B \Rightarrow A \leq h_\Delta \leq B \ \text{ a.e.},
$$

*where*

$$
h_\Delta(t, x) := \sum_{i=1}^{N_x} \sum_{n=1}^{N_t} h^n_i \mathbf{1}_{[t^n, t^{n+1}] \times C_i}(t, x), \ \text{ in } [0, T] \times (0, L).
$$

*Proof.* Define

$$H(h,k,l) := k - \frac{\Delta t}{\Delta x}\left(g(k,l) - g(h,k)\right), \quad \text{in } \mathbb{R}^3,$$

its first-order partial derivatives are given by

$$\begin{aligned}
\partial_1 H &= \frac{\Delta t}{\Delta x}\partial_1 g(h,k) \geq 0,\\
\partial_2 H &= 1 - \frac{\Delta t}{\Delta x}\left(\partial_1 g(k,l) - \partial_2 g(h,k)\right),\\
\partial_3 H &= -\frac{\Delta t}{\Delta x}\partial_2 g(k,l) \geq 0,
\end{aligned}$$

Notice that the weak first-order partial derivatives of $g$ are given by

$$\partial_1 g(h,k) = \begin{cases} u^n_{i+\frac{1}{2}}, & \text{if } u^n_{i+\frac{1}{2}} \geq 0,\\ 0, & \text{otherwise,} \end{cases} \text{ on } C_i \times M_{i+1},$$

and

$$\partial_2 g(h,k) = \begin{cases} 0, & \text{if } u^n_{i+\frac{1}{2}} \geq 0,\\ u^n_{i+\frac{1}{2}}, & \text{otherwise,} \end{cases} \text{ on } C_i \times M_{i+1},$$

i.e.,

$$\begin{aligned}
\partial_1 g(h,k) &= \sum_{i=1}^{N_x-1} \left(u^n_{i+\frac{1}{2}}\right)_+ \mathbf{1}_{M_i \times M_{i+1}}(h,k),\\
\partial_2 g(h,k) &= \sum_{i=1}^{N_x-1} \left(u^n_{i+\frac{1}{2}}\right)^- \mathbf{1}_{M_i \times M_{i+1}}(h,k),
\end{aligned}$$

then we can estimate $\partial_2 H$ as follows,

$$\begin{aligned}
\partial_2 H &= 1 - \frac{\Delta t}{\Delta x}\left(\partial_1 g(k,l) - \partial_2 g(h,k)\right)\\
&= 1 - \frac{\Delta t}{\Delta x}\left(\sum_{i=1}^{N_x-1} \left(u^n_{i+\frac{1}{2}}\right)_+ \mathbf{1}_{M_i \times M_{i+1}}(k,l) - \sum_{i=1}^{N_x-1} \left(u^n_{i+\frac{1}{2}}\right)^- \mathbf{1}_{M_i \times M_{i+1}}(h,k)\right)\\
&= 1 - \frac{\Delta t}{\Delta x}\left(\sum_{i=1}^{N_x-1} \left(u^n_{i+\frac{1}{2}}\right)_+ \mathbf{1}_{C_i}(k) \times \mathbf{1}_{M_{i+1}}(l) - \sum_{i=1}^{N_x-1} \left(u^n_{i+\frac{1}{2}}\right)^- \mathbf{1}_{C_i}(h) \times \mathbf{1}_{M_{i+1}}(k)\right)\\
&= 1 - \frac{\Delta t}{\Delta x}\left(\sum_{i=1}^{N_x-1} \left(u^n_{i+\frac{1}{2}}\right)_+ \mathbf{1}_{C_i}(k) \times \mathbf{1}_{M_{i+1}}(l) - \sum_{i=2}^{N_x} \left(u^n_{i-\frac{1}{2}}\right)^- \mathbf{1}_{M_{i-1}}(h) \times \mathbf{1}_{C_i}(k)\right)\\
&= 1 - \frac{\Delta t}{\Delta x}\left[\begin{array}{l} \sum\limits_{i=2}^{N_x-1}\left[\mathbf{1}_{C_i}(k) \times \left(\left(u^n_{i+\frac{1}{2}}\right)_+ \mathbf{1}_{M_{i+1}}(l) - \left(u^n_{i-\frac{1}{2}}\right)^- \mathbf{1}_{M_{i-1}}(h)\right)\right]\\ +\left(u^n_{\frac{3}{2}}\right)_+ \mathbf{1}_{M_1}(k) \times \mathbf{1}_{M_2}(l) - \left(u^n_{N_x-\frac{1}{2}}\right)^- \mathbf{1}_{M_{N_x-1}}(h) \times \mathbf{1}_{M_{N_x}}(k) \end{array}\right].
\end{aligned}$$

Hence, if the CFL-like condition (2.6) holds, then $\partial_2 H \geq 0$. Thus, $H(\nearrow, \nearrow, \nearrow)$.

Now, we assume that $A \leq h_i^0 \leq B$, for all $i \in \{1, \ldots, N_x\}$. Let us remark that

$$\forall \kappa \in \mathbb{R}, \quad H(\kappa, \kappa, \kappa) = \kappa - \frac{\Delta t}{\Delta x}\left(g(\kappa, \kappa) - g(\kappa, \kappa)\right) = \kappa.$$

As a consequence: if $h_0 \equiv \kappa$ then $h_\Delta \equiv \kappa$.

Now, if we take $\kappa = A$ and $\kappa = B$, we can obtain the $L^\infty$-stability by monotonicity: Assume that $\forall i \in \mathbb{Z}$, $A \leq h_i^n \leq B$, then

$$\begin{aligned} h_i^{n+1} &= H\left(h_{i-1}^n, h_i^n, h_{i+1}^n\right) \leq H\left(B, h_i^n, h_{i+1}^n\right) \leq H\left(B, B, h_{i+1}^n\right) \leq H(B, B, B) = B, \\ h_i^{n+1} &= H\left(h_{i-1}^n, h_i^n, h_{i+1}^n\right) \geq H\left(A, h_i^n, h_{i+1}^n\right) \geq H\left(A, A, h_{i+1}^n\right) \geq H(A, A, A) = A, \end{aligned}$$

This ends our proof. □

Applying this proposition with $A = 0$, we obtain immediately the following direct consequence.

**Corollary 1** (Nonnegativity of water height)**.** *The water height remains nonnegative at time $t^{n+1}$ under the CFL-like condition (2.6).*

The next proposition indicates the well-balanced property of the proposed numerical scheme.

**Proposition 4** (Well-balanced property)**.** *The steady states at rest are preserved under the numerical scheme (2.1)-(2.2), i.e., if $u_{i+\frac{1}{2}}^n = 0$ and $h_i^n + b_i = const$ for all $i \in \mathcal{M}$, then $u_{i+\frac{1}{2}}^{n+1} = 0$ and $h_i^{n+1} + b_i = const$ for all $i \in \mathcal{M}$.*

*Proof.* We recall (2.1) and rewrite (2.2) as

$$\begin{aligned} h_i^{n+1} &= h_i^n - \frac{\Delta t}{\Delta x}\left(h_{i+\frac{1}{2}}^n u_{i+\frac{1}{2}}^n - h_{i-\frac{1}{2}}^n u_{i-\frac{1}{2}}^n\right), \quad \forall i \in \mathcal{M}, \\ \bar{h}_{i+\frac{1}{2}}^{n+1} u_{i+\frac{1}{2}}^{n+1} &= \bar{h}_{i+\frac{1}{2}}^n u_{i+\frac{1}{2}}^n - \frac{\Delta t}{\Delta x}\left[G_{i+1}^n - G_i^n + \frac{g}{2}\left(h_i^{n+1} + h_{i+1}^{n+1}\right)\left(\left(h_{i+1}^{n+1} + b_{i+1}\right) - \left(h_i^{n+1} + b_i\right)\right)\right], \quad \forall i \in \mathcal{E}_{\text{int}}. \end{aligned}$$

Suppose that $u_{i+\frac{1}{2}}^n = 0$ and $h_i^n + b_i = \text{const}$ for all $i \in \mathcal{E}$, then these equations become

$$\begin{aligned} h_i^{n+1} &= h_i^n, \quad \forall i \in \mathcal{M}, \\ \frac{h_i^{n+1} + h_{i+1}^{n+1}}{2} u_{i+\frac{1}{2}}^{n+1} &= -\frac{\Delta t}{\Delta x}\left[\frac{g}{2}\left(h_i^{n+1} + h_{i+1}^{n+1}\right)\left(\left(h_{i+1}^{n+1} + b_{i+1}\right) - \left(h_i^{n+1} + b_i\right)\right)\right], \quad \forall i \in \mathcal{E}_{\text{int}}. \end{aligned} \tag{2.7}$$

As a direct consequence, $h_i^{n+1} + b_i = h_i^n + b_i = \text{const}$.

Plugging the former into the latter yields

$$\frac{h_i^n + h_{i+1}^n}{2} u_{i+\frac{1}{2}}^{n+1} = -\frac{\Delta t}{\Delta x}\left[\frac{g}{2}\left(h_i^n + h_{i+1}^n\right)\left(\left(h_{i+1}^n + b_{i+1}\right) - \left(h_i^n + b_i\right)\right)\right], \quad \forall i \in \mathcal{E}_{\text{int}}.$$

Since $h_i^n + b_i = \text{const}$ for all $i \in \mathcal{E}$, the last equation implies $u_{i+\frac{1}{2}}^{n+1} = 0$ ($h_i^n \geq 0$ and $h_{i+1}^n \geq 0$ by Corollary 1). This finishes our proof. □

Consider the case which dry shore are permitted in our domain $\Omega$, we decompose $\mathcal{M}$ as $\mathcal{M} = \mathcal{M}^n_{\text{dry}} \cup \mathcal{M}^n_{\text{wet}}$. The *dry-component* $\mathcal{M}_{\text{dry}}$ consists of all indices $i$ such that $h^n_i = 0$ and the *wet-component* $\mathcal{M}_{\text{wet}}$ consists of all indices $i$ such that $h^n_i > 0$. This is reasonable because the water height at a particular cell obviously indicates whether that cell is "dry" or "wet".

With allowing dry shore in our domain $\Omega$, a well-balanced property still holds in this scenario, with a little modification.

**Proposition 5** (Ill-balanced property with dry shore)**.** *When the domain $\Omega$ contains dry shore, the above well-balanced property does not hold in general.*

*Proof.* Given $n$, since the domain $\Omega$ contains dry shore, there exists an index $i_0 \in \mathcal{M}$ such that $i_0 \in \mathcal{M}_{\text{dry}}$ and $i_0 + 1 \in \mathcal{M}_{\text{wet}}$, or $i_0 \in \mathcal{M}_{\text{wet}}$ and $i_0 + 1 \in \mathcal{M}_{\text{dry}}$. Without loss of generality, suppose that the former holds, i.e., $h^n_{i_0} = 0$ and $h^n_{i_0+1} > 0$, we follow the proof of Proposition 4 to arrive at $h^{n+1}_{i_0} = h^n_{i_0} = 0$, $h^{n+1}_{i_0+1} = h^n_{i_0+1} > 0$, and (2.7) gives us

$$\frac{h^n_{i_0+1}}{2} u^{n+1}_{i_0+\frac{1}{2}} = -\frac{g\Delta t}{2\Delta x} h^n_{i_0+1} \left( h^n_{i_0+1} + b_{i_0+1} - b_{i_0} \right).$$

So if $h^n_{i_0+1} + b_{i_0+1} < b_{i_0}$ (a figure of this case can be easily drawn) then this equality implies $u^{n+1}_{i_0+\frac{1}{2}} > 0$. Therefore, the previous well-balanced property falls in this case. □

At each time step, the Courant number is defined by

$$\nu := \frac{\Delta t}{\Delta x} \max_{i \in \mathcal{M}} \left( \frac{\left| q^n_{i+\frac{1}{2}} + q^n_{i-\frac{1}{2}} \right|}{2h^n_i} + \sqrt{g h^n_i} \right).$$

The numerical simulations in [3, pp. 232–234] show that the staggered scheme is stable under the CFL condition $\nu < 1$.

# 3 Three hydrostatic reconstruction schemes based on subcell reconstructions

We briefly rewrite the three hydrostatic reconstruction schemes in [2] in our notations. All of them introduce reconstructed values $U_{i+\frac{1}{2}\pm}$ of the unknowns to the left and right of the interface $x_{i+\frac{1}{2}}$, and define the numerical flux via a Riemann solver $\mathcal{F}$:

$$F_{i+\frac{1}{2}} = \mathcal{F}\left( U_{i+\frac{1}{2}-}, U_{i+\frac{1}{2}+} \right). \tag{3.1}$$

Those HR schemes split the singular source term at the interface into a left and a right part, $S_{i+\frac{1}{2}-}$ and $S_{i+\frac{1}{2}+}$, and compute the source term in (1.3) as

$$S_i = S_{i-\frac{1}{2}+} + S_{i+\frac{1}{2}-} = \left(0, s_{i-\frac{1}{2}+}\right)^T + \left(0, s_{i+\frac{1}{2}-}\right)^T. \tag{3.2}$$

There are two main types of interface, depending on how the water covers the bottom to the left and right sides of the interfaces (see Fig. 1[3]):

- *Fully wet interface*: where the water level on each side is higher than the higher side of the bottom topography:

$$\min(w_i, w_{i+1}) > \max(b_i, b_{i+1}), \tag{3.3}$$

  where $b_i$ and $w_i$ are the bottom topography and the water level, respectively, in cell $C_i$.

- *Partially wet interface*: where the water level on one side is equal to or below the topography on the other side:

$$\min(w_i, w_{i+1}) \leq \max(b_i, b_{i+1}). \tag{3.4}$$

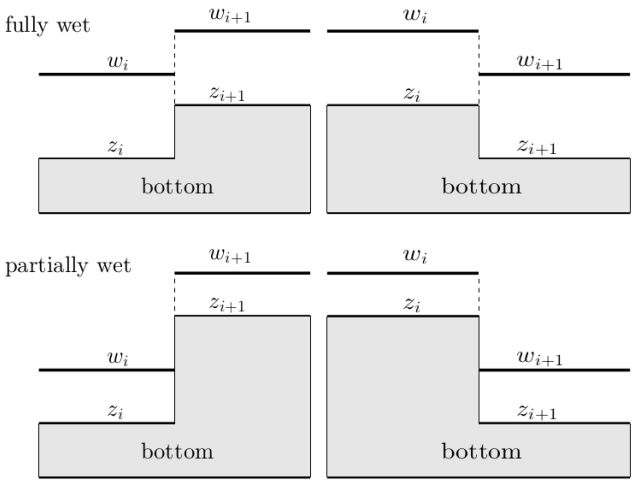


Figure 1: Examples of interfaces: fully wet (top), and partially wet (bottom).

After the water heights $h_{i-\frac{1}{2}+}$ and $h_{i+\frac{1}{2}-}$ are reconstructed in the following three HR schemes, the conservative variables are defined as

$$U_{i+\frac{1}{2}-} := \begin{pmatrix} h_{i+\frac{1}{2}-} \\ h_{i+\frac{1}{2}-} u_{i+\frac{1}{2}-} \end{pmatrix}, \quad U_{i-\frac{1}{2}+} := \begin{pmatrix} h_{i-\frac{1}{2}+} \\ h_{i-\frac{1}{2}+} u_{i-\frac{1}{2}+} \end{pmatrix} \text{ with } u_{i+\frac{1}{2}-} = u_{i-\frac{1}{2}+} = u_i. \tag{3.5}$$

## 3.1 The original HR method

Audusse et al. [1] introduce their first order HR scheme by choosing the *intermediate bottom* as

$$b^{\mathrm{AUD}}_{i+\frac{1}{2}} := \max(b_i, b_{i+1}), \tag{3.6}$$

[3]This figure is extracted in [2, p. 761].

and the *interface water height* as

$$h^{\mathrm{AUD}}_{i+\frac{1}{2}-} := \left(w_i - b^{\mathrm{AUD}}_{i+\frac{1}{2}}\right)^+, \quad h^{\mathrm{AUD}}_{i+\frac{1}{2}+} := \left(w_{i+1} - b^{\mathrm{AUD}}_{i+\frac{1}{2}}\right)^+. \tag{3.7}$$

The source term is discretized as

$$s^{\mathrm{AUD}}_{i+\frac{1}{2}-} := \frac{g}{2\Delta x}\left(\left(h^{\mathrm{AUD}}_{i+\frac{1}{2}-}\right)^2 - h_i^2\right), \tag{3.8}$$

$$s^{\mathrm{AUD}}_{i+\frac{1}{2}+} := \frac{g}{2\Delta x}\left(h_{i+1}^2 - \left(h^{\mathrm{AUD}}_{i+\frac{1}{2}+}\right)^2\right). \tag{3.9}$$

### 3.2 The HR method of Morales et al.

The HR scheme of Morales et al. [6] is identical to that of Audusse's scheme,

$$b^{\mathrm{MOR}}_{i+\frac{1}{2}} := b^{\mathrm{AUD}}_{i+\frac{1}{2}}, \quad h^{\mathrm{MOR}}_{i+\frac{1}{2}-} := h^{\mathrm{AUD}}_{i+\frac{1}{2}-}, \quad h^{\mathrm{MOR}}_{i+\frac{1}{2}+} := h^{\mathrm{AUD}}_{i+\frac{1}{2}+}, \tag{3.10}$$

and the source term is defined as

$$s^{MOR}_{i+\frac{1}{2}-} := s^{\mathrm{AUD}}_{i+\frac{1}{2}-}, \quad s^{MOR}_{i+\frac{1}{2}+} := s^{\mathrm{AUD}}_{i+\frac{1}{2}+}, \tag{3.11}$$

except for the partially wet interfaces (3.4), where water either flows downhill or flows uphill with enough kinetic energy to climb the jump of the bottom at the interface. This results in the following two cases.

- $b_i < b_{i+1}$ *(ascending bottom).* If $u_i < 0$, or $u_i > 0$ and

$$\frac{|u_i|^2}{2} + g\left(w_i - b_{i+1}\right) \geq \frac{3}{2}\sqrt[3]{g(h_i\,|u_i|)^3}, \tag{3.12}$$

  then the left interface source term is redefined as

$$s^{\mathrm{MOR}}_{i+\frac{1}{2}-} := -\frac{g}{\Delta x}\frac{h_i}{2}\left(b^{\mathrm{MOR}}_{i+\frac{1}{2}} - b_i\right). \tag{3.13}$$

- $b_i > b_{i+1}$ *(descending bottom).* If $u_{i+1} > 0$, or $u_{i+1} < 0$ and

$$\frac{|u_{i+1}|^2}{2} + g\left(w_{i+1} - b_i\right) \geq \frac{3}{2}\sqrt[3]{g(h_{i+1}\,|u_{i+1}|)^3}, \tag{3.14}$$

  then the right interface source term is redefined as

$$s^{\mathrm{MOR}}_{i+\frac{1}{2}+} := -\frac{g}{\Delta x}\frac{h_{i+1}}{2}\left(b_{i+1} - b^{\mathrm{MOR}}_{i+\frac{1}{2}}\right). \tag{3.15}$$

### 3.3 The third HR method

For the third HR method, the intermediate bottom is defined as

$$b^{\mathrm{CN}}_{i+\frac{1}{2}} := \min\left(\max\left(b_i, b_{i+1}\right), \min\left(w_i, w_{i+1}\right)\right). \tag{3.16}$$

The interface water heights are given by

$$h^{\mathrm{CN}}_{i+\frac{1}{2}-} := \min\left(w_i - b^{\mathrm{CN}}_{i+\frac{1}{2}}, h_i\right), \quad h^{\mathrm{CN}}_{i+\frac{1}{2}+} := \min\left(w_{i+1} - b^{\mathrm{CN}}_{i+\frac{1}{2}}, h_{i+1}\right), \tag{3.17}$$

and the interface source terms are defined as

$$s^{CN}_{i+\frac{1}{2}-} := -\frac{g}{\Delta x}\frac{h_i + h^{CN}_{i+\frac{1}{2}-}}{2}\left(b^{CN}_{i+\frac{1}{2}} - b_i\right), \tag{3.18}$$

$$s^{CN}_{i+\frac{1}{2}+} := -\frac{g}{\Delta x}\frac{h^{CN}_{i+\frac{1}{2}+} + h_{i+1}}{2}\left(b_{i+1} - b^{CN}_{i+\frac{1}{2}}\right). \tag{3.19}$$

**Remark 1.** *For fully wet interfaces, all schemes are identical. The differences for the partially wet case are rather subtle.*

### 3.4 The numerical flux

In [2, p. 763], the authors used Harten-Lax-van Leer (HLL)-type Riemann solvers,

$$\mathcal{F}_{\mathrm{HLL}}\left(U_-, U_+\right) = \frac{s^+ F\left(U_-\right) - s^- F\left(U_+\right) + s^+ s^-\left(U_+ - U_-\right)}{s^+ - s^-}, \tag{3.20}$$

where the smallest and largest wave speed $s^-$ and $s^+$ are chosen as

$$s^- = \min\left(u_- - a_-, u_+ - a_+, 0\right), \quad s^+ = \max\left(u_- + a_-, u_+ + a_+, 0\right), \tag{3.21}$$

with gravitational wave speed $a = \sqrt{gh}$.

The following inequalities satisfied by the HLL flux are crucial for the stability analysis. The first one states that there is no numerical mass flux out of an empty cell, and is used to prove positivity of the water height later.

**Lemma 1.** *The first component of the HLL flux satisfies*

$$\mathcal{F}^h\left((0,0),(h,hu)\right) \le 0, \quad \mathcal{F}^h\left((h,hu),(0,0)\right) \ge 0. \tag{3.22}$$

*Proof.* From the definition of the wave speeds (3.21), the value of HLL flux at $U_- = (0,0)$ and $U_+ = (h,hu)$ is

$$\mathcal{F}_{\mathrm{HLL}}\left((0,0),(h,hu)\right) = \frac{-s^-\left(hu, hu^2 + \frac{1}{2}gh^2\right) + s^+ s^-\left(h, hu\right)}{s^+ - s^-}. \tag{3.23}$$

Hence, the first component of the HLL flux is given by

$$\mathcal{F}^h_{HLL}\left((0,0),(h,hu)\right) = \frac{-s^- hu + s^+ s^- h}{s^+ - s^-} = hs^- \frac{s^+ - u}{s^+ - s^-}. \tag{3.24}$$

Since $h \geq 0$, $s^- \leq 0$, $s^+ \geq u$, $s^+ > s^-$, the last equality gives us $\mathcal{F}^h_{HLL} \leq 0$.

Similarly,

$$\mathcal{F}_{\text{HLL}}\left((h,hu),(0,0)\right) = \frac{s^+\left(hu, hu^2 + \frac{1}{2}gh^2\right) - s^+ s^-\left(h,hu\right)}{s^+ - s^-}, \tag{3.25}$$

and thus

$$\mathcal{F}^h_{\text{HLL}}\left((h,hu),(0,0)\right) = s^+ h \frac{u - s^-}{s^+ - s^-}. \tag{3.26}$$

Since $h \geq 0$, $s^+ \geq 0$, $u \geq s^-$, $s^+ > s^-$, the last equality implies that $\mathcal{F}^h_{\text{HLL}} \geq 0$. □

The second inequality is used to prove the semidiscrete entropy inequality. It states that the numerical mass flux into a vacuum cell is at least as large as the physical mass flux.

**Lemma 2.** *The first component of the HLL flux satisfies*

$$\mathcal{F}^h_{\text{HLL}}\left((h,hu),(0,0)\right) - hu \geq 0, \quad hu - \mathcal{F}^h_{\text{HLL}}\left((0,0),(h,hu)\right) \geq 0. \tag{3.27}$$

*Proof.* The proof follows from (3.20) and (3.21) via

$$\mathcal{F}^h_{\text{HLL}}\left((h,hu),(0,0)\right) - hu = \frac{s^+ hu - s^+ s^- h}{s^+ - s^-} - hu = -\frac{s^- h\left(s^+ - u\right)}{s^+ - s^-} \geq 0,$$
$$hu - \mathcal{F}^h_{\text{HLL}}\left((0,0),(h,hu)\right) = hu - \frac{-s^- hu + s^+ s^- h}{s^+ - s^-} = \frac{s^+ h\left(u - s^-\right)}{s^+ - s^-} \geq 0.$$

This completes our proof. □

### 3.5 Interpretation via subcell reconstructions

Recall that in [2] the *singular layers* (or *internal boundary layers*) are defined by

$$\widehat{C}^{\varepsilon}_{i+\frac{1}{2}} := \left[x_{i+\frac{1}{2}} - \varepsilon, x_{i+\frac{1}{2}} + \varepsilon\right], \forall i \in \mathcal{E}. \tag{3.28}$$

Over each of these infinitesimal layers the bottom is reconstructed continuously by a function $b_\varepsilon(x)$. The flow variables are reconstructed by piecewise continuous functions $h_\varepsilon(x)$, $w_\varepsilon(x)$, and $u_\varepsilon(x)$ over the singular subcells

$$\widehat{C}^{\varepsilon}_{i+\frac{1}{2}-} := \left[x_{i+\frac{1}{2}} - \varepsilon, x_{i+\frac{1}{2}}\right] \text{ and } \widehat{C}^{\varepsilon}_{i+\frac{1}{2}+} := \left[x_{i+\frac{1}{2}}, x_{i+\frac{1}{2}} + \varepsilon\right], \quad \forall i \in \mathcal{E}. \tag{3.29}$$

These reconstructions provide the data of the Riemann problem at the interface, with an approximate Riemann solver $F_\varepsilon\left(x_{i+\frac{1}{2}}\right)$. The source term is computed over the singular subcells. Together, this gives the *residuum*

$$R^{\varepsilon}_i := -\frac{1}{\Delta x}\left(F_\varepsilon\left(x_{i+\frac{1}{2}}\right) - F_\varepsilon\left(x_{i-\frac{1}{2}}\right)\right) + \frac{1}{\Delta x}\int_{C_i} S\left(U_\varepsilon(x), b_\varepsilon(x)\right) dx. \tag{3.30}$$

### 3.5.1 Splitting the cells into subcells

Let us denote the *interior subcell* by $C_i^\varepsilon := \left[x_{i-\frac{1}{2}} + \varepsilon, x_{i+\frac{1}{2}} - \varepsilon\right]$. Then

$$C_i = \widehat{C}^\varepsilon_{i-\frac{1}{2}+} \cup C_i^\varepsilon \cup \widehat{C}^\varepsilon_{i+\frac{1}{2}-}. \tag{3.31}$$

The *piecewise continuous reconstruction* is defined as follows.

**Definition 1** (subcell reconstruction)**.** *Given values $\varphi_i$ and $\varphi_{i+\frac{1}{2}\pm}$ for $i \in \mathbb{Z}$, let $\widehat{\varphi}^\varepsilon_{i+\frac{1}{2}\pm} : \widehat{C}^\varepsilon_{i+\frac{1}{2}\pm} \to \mathbb{R}$ be Lipschitz continuous functions with boundary values*

$$\varphi_{i+\frac{1}{2}\pm}\left(x_{i+\frac{1}{2}}\right) = \varphi_{i+\frac{1}{2}\pm}, \;\; \varphi_{i+\frac{1}{2}\pm}\left(x_{i+\frac{1}{2}} \pm \varepsilon\right) = \varphi_{i+\frac{1\pm1}{2}}. \tag{3.32}$$

*Then $\varphi_\varepsilon : \mathbb{R} \to \mathbb{R}$ is the piecewise continuous function given by*

$$\varphi_\varepsilon(x) := \begin{cases} \varphi_i, \;\; \text{if } x \in C_i^\varepsilon, \\ \widehat{\varphi}^\varepsilon_{i+\frac{1}{2}\pm}, \;\; \text{if } x \in \widehat{C}^\varepsilon_{i+\frac{1}{2}\pm}. \end{cases}$$

**Remark 2.** *If $\widehat{\varphi}^\varepsilon_{i+\frac{1}{2}-}$ and $\widehat{\varphi}^\varepsilon_{i+\frac{1}{2}+}$ are both linear, we call them the* standard subcell reconstruction at interface $x_{i+\frac{1}{2}}$*. The only exception from the standard subcell reconstruction will occur in the definition of the water level for Audusse's and Morales' schemes for partially wet interfaces.*

### 3.5.2 Reconstruction of the bottom $b_\varepsilon(x)$

For all three HR schemes, a continuous bottom is defined by the standard subcell reconstruction (Def. 1) with

$$b_{i-\frac{1}{2}+} := b_{i-\frac{1}{2}}, \;\; b_{i+\frac{1}{2}-} := b_{i+\frac{1}{2}}. \tag{3.33}$$

The reconstructed bottom is globally continuous for fixed $\varepsilon > 0$, but will have steep layers in $\widehat{C}^\varepsilon_{i+\frac{1}{2}-}$ and $\widehat{C}^\varepsilon_{i+\frac{1}{2}-}$.

### 3.5.3 Infinitesimal HR

The water level and height are now reconstructed. Several modern well-balanced schemes, as well as the third HR schemes, use the fact that the water level $w(x)$ is constant for still water. Thus, the piecewise constant reconstruction becomes exact for this equilibrium state.

Given the bottom topography $b_\varepsilon(x)$, these schemes reconstruct the water level $w_\varepsilon(x)$. The reconstructed water height is then reconstructed simply as

$$h_\varepsilon(x) := w_\varepsilon(x) - b_\varepsilon(x). \tag{3.34}$$

The *conservative variables* are given by

$$U_\varepsilon(x) := \sum_{i=1}^{N_x} \begin{pmatrix} h_\varepsilon(x) \\ h_\varepsilon(x)\, u_i \end{pmatrix} \mathbf{1}_{C_i}(x), \tag{3.35}$$

The following integral averages of $h_\varepsilon(x)$ in subcells $\widehat{C}^\varepsilon_{i+\frac{1}{2}-}$ and $\widehat{C}^\varepsilon_{i+\frac{1}{2}+}$ are used to calculate the flux and the source term later.

$$\bar{h}_{i+\frac{1}{2}-} := \frac{1}{\varepsilon}\int_{\widehat{C}^\varepsilon_{i+\frac{1}{2}-}} h_\varepsilon(x)\,dx, \quad \bar{h}_{i+\frac{1}{2}+} := \frac{1}{\varepsilon}\int_{\widehat{C}^\varepsilon_{i+\frac{1}{2}+}} h_\varepsilon(x)\,dx. \tag{3.36}$$

**3.5.3.1. The original HR method.** Define $w_\varepsilon(x)$ as in Def. 1, with

$$w_i^{\mathrm{AUD}} := h_i + b_i, \quad w_{i+1}^{\mathrm{AUD}} := h_{i+1} + b_{i+1}, \tag{3.37}$$

$$w_{i+\frac{1}{2}-}^{\mathrm{AUD}} := \max\left(b_{i+\frac{1}{2}}^{\mathrm{AUD}}, \ w_i\right), \quad w_{i+\frac{1}{2}+}^{\mathrm{AUD}} := \max\left(b_{i+\frac{1}{2}}^{\mathrm{AUD}}, w_{i+1}\right). \tag{3.38}$$

In the fully wet case (see Fig. 2),

$$\widehat{w}_{i+\frac{1}{2}-}^{\mathrm{AUD}} \equiv w_i, \quad \widehat{w}_{i+\frac{1}{2}+}^{\mathrm{AUD}} \equiv w_{i+1}, \tag{3.39}$$

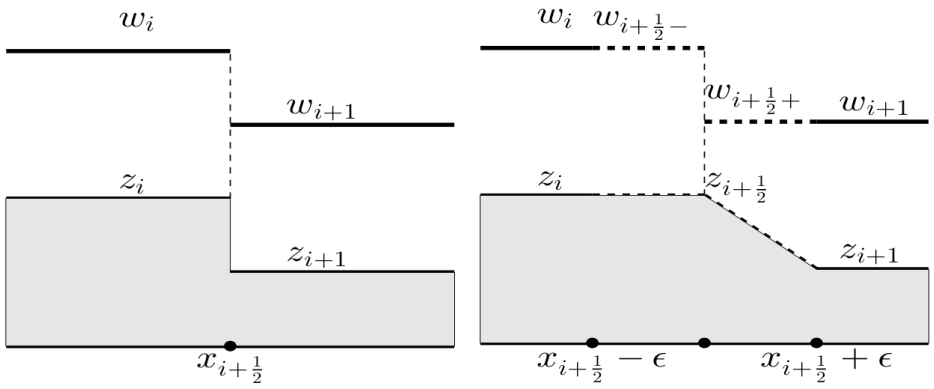


Figure 2: Subcell reconstruction of topography and water level for the fully wet case. Left: Riemann data. Right: reconstructed $b_\varepsilon(x)$ and $w_\varepsilon(x)$ for all three HR schemes.

while in the partially wet case (see Fig. 3),

$$\widehat{w}_{i+\frac{1}{2}-}^{\mathrm{AUD}}(x) := \max\left(b_\varepsilon^{\mathrm{AUD}}(x), w_i\right), \quad \widehat{w}_{i+\frac{1}{2}+}^{\mathrm{AUD}}(x) := \max\left(b_\varepsilon^{\mathrm{AUD}}(x), w_{i+1}\right). \tag{3.40}$$

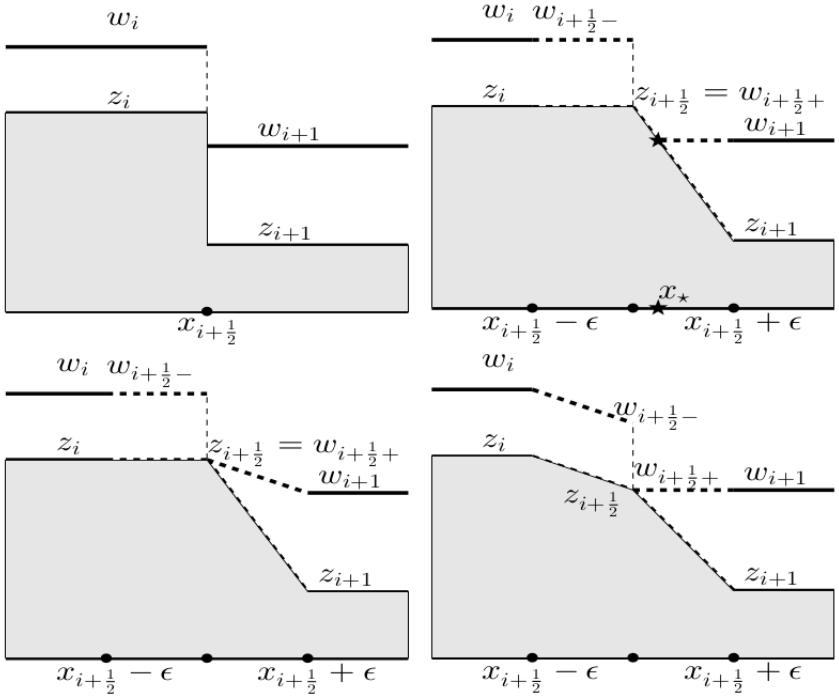


Figure 3: Subcell reconstruction of topography and water level for the partially wet case. Top left: Riemann data. Top right: Audusse or Morales scheme for slow uphill flow (the vacuum region $\left[x_{i+\frac{1}{2}}, x_\star\right]$). Bottom left: Morales scheme for fast uphill flow. Bottom right: CN scheme.

This is the only instance where the subcell reconstruction may differ from the standard definition. This will happen if and only if the wet-dry front is contained in one of the cells $\widehat{C}^{\varepsilon}_{i+\frac{1}{2}-}$ or $\widehat{C}^{\varepsilon}_{i+\frac{1}{2}-}$.

The average of $h^{\mathrm{AUD}}_{\varepsilon}(x)$ over subcell $\widehat{C}^{\varepsilon}_{i+\frac{1}{2}-}$ is defined as follows (that over subcell $\widehat{C}^{\varepsilon}_{i+\frac{1}{2}+}$ is defined similarly). There are two cases to be discussed.

i) In the fully wet case (in which all three schemes coincide) the water height $\widehat{h}^{\mathrm{AUD}}_{i+\frac{1}{2}+}$ is linear, so

$$\bar{h}^{\mathrm{AUD}}_{i+\frac{1}{2}+} = \frac{h^{\mathrm{AUD}}_{i+\frac{1}{2}+} + h_{i+1}}{2}. \tag{3.41}$$

ii) In the partially wet case, $h^{\mathrm{AUD}}_{i+\frac{1}{2}+} = 0$ (see top right of Fig. 3). Assume that the wet front is located at $x_\star \in \widehat{C}^{\varepsilon}_{i+\frac{1}{2}+}$. Then the average water height is

$$\bar{h}^{\mathrm{AUD}}_{i+\frac{1}{2}+} = \frac{x_{i+\frac{1}{2}} + \varepsilon - x_*}{\varepsilon} \frac{h(x_\star) + h_{i+1}}{2} = \frac{h_{i+1}}{b_{i+\frac{1}{2}} - b_{i+1}} \frac{h_{i+1}}{2}, \tag{3.42}$$

where we have used the intercept theorem in the last inequality.

Summarizing (3.41) and (3.42), we obtain

$$\bar{h}^{\mathrm{AUD}}_{i+\frac{1}{2}+} = \begin{cases} \dfrac{h_{i+1} + h^{\mathrm{AUD}}_{i+\frac{1}{2}+}}{2}, & \text{if } h^{\mathrm{AUD}}_{i+\frac{1}{2}+} > 0, \\ \dfrac{h_{i+1}}{2} \dfrac{h_{i+1}}{b^{\mathrm{AUD}}_{i+\frac{1}{2}} - b_{i+1}}, & \text{if } h^{\mathrm{AUD}}_{i+\frac{1}{2}+} = 0. \end{cases} \tag{3.43}$$

Similarly,

$$\bar{h}^{\mathrm{AUD}}_{i+\frac{1}{2}-} = \begin{cases} \dfrac{h_i + h^{\mathrm{AUD}}_{i+\frac{1}{2}-}}{2}, & \text{if } h^{\mathrm{AUD}}_{i+\frac{1}{2}-} > 0, \\ \dfrac{h_i}{2} \dfrac{h_i}{b^{\mathrm{AUD}}_{i+\frac{1}{2}} - b_i}, & \text{if } h^{\mathrm{AUD}}_{i+\frac{1}{2}-} = 0. \end{cases} \tag{3.44}$$

**3.5.3.2. The HR scheme of Morales et al.** The continuous bottom of Morales' scheme coincides with that of the original HR scheme:

$$b^{\mathrm{MOR}}_{\varepsilon}(x) \equiv b^{\mathrm{AUD}}_{\varepsilon}(x). \tag{3.45}$$

The water level coincides with that of the original scheme, except for the partially wet interface (3.4). If the water flows downhill, or uphill with enough kinetic energy to climb the discrete jump of the bottom, i.e., (3.12) (resp., (3.14)) holds, then over $\widehat{C}_{i+\frac{1}{2}-}$ (resp. $\widehat{C}_{i+\frac{1}{2}+}$) the reconstructed water level $w^{\mathrm{MOR}}_{\varepsilon}(x)$ is given by the standard subcell reconstruction (see Def. 1) instead of Audusse's piecewise linear reconstruction (3.38) (see Fig. 3). Then the local averages of $h^{\mathrm{MOR}}_{\varepsilon}(x)$ over subcells $\widehat{C}^{\varepsilon}_{i+\frac{1}{2}-}$ (resp., $\widehat{C}^{\varepsilon}_{i+\frac{1}{2}+}$) are simply

$$\bar{h}^{\mathrm{MOR}}_{i+\frac{1}{2}-} = \frac{h_i + h^{\mathrm{MOR}}_{i+\frac{1}{2}-}}{2} = \frac{h_i}{2}, \quad \bar{h}^{\mathrm{MOR}}_{i+\frac{1}{2}+} = \frac{h_{i+1} + h^{\mathrm{MOR}}_{i+\frac{1}{2}+}}{2} = \frac{h_{i+1}}{2}, \tag{3.46}$$

since $h^{MOR}_{i+\frac{1}{2}-} = 0$ (resp., $h^{MOR}_{i+\frac{1}{2}+} = 0$).

**3.5.3.3. The third HR scheme.** The continuous bottom of the third HR scheme in [2], denoted by $b^{\mathrm{CN}}_{\varepsilon}$ is defined by the standard subcell reconstruction with

$$b^{\mathrm{CN}}_{i+\frac{1}{2}-} = b^{\mathrm{CN}}_{i+\frac{1}{2}+} = b^{\mathrm{CN}}_{i+\frac{1}{2}}. \tag{3.47}$$

The reference values for the water surface are given by

$$w^{\mathrm{CN}}_{i+\frac{1}{2}-} := \min\left(w_i, b^{\mathrm{CN}}_{i+\frac{1}{2}} + h_i\right), \quad w^{\mathrm{CN}}_{i+\frac{1}{2}+} := \min\left(w_{i+1}, b^{\mathrm{CN}}_{i+\frac{1}{2}} + h_{i+1}\right), \tag{3.48}$$

and the reference values for the water depth are given by (3.17). Due to the linearity of $\widehat{h}_{i+\frac{1}{2}-}$ (resp., $\widehat{h}_{i+\frac{1}{2}+}$), the average values of $h_\varepsilon$ over the singular subcells $\widehat{C}^{\varepsilon}_{i+\frac{1}{2}-}$ (resp., $\widehat{C}^{\varepsilon}_{i+\frac{1}{2}+}$) are

$$\bar{h}^{\mathrm{CN}}_{i+\frac{1}{2}-} = \frac{h_i + h^{\mathrm{CN}}_{i+\frac{1}{2}-}}{2}, \quad \bar{h}^{\mathrm{CN}}_{i+\frac{1}{2}+} = \frac{h_i + h^{\mathrm{CN}}_{i+\frac{1}{2}+}}{2}. \tag{3.49}$$

**Remark 3.** *From (3.43), (3.44), (3.46) and (3.49), the subcell averages* $\bar{h}_{i+\frac{1}{2}-}$ *and* $\bar{h}_{i+\frac{1}{2}+}$ *of* $h_\varepsilon(x)$ *over subcells* $\widehat{C}^\varepsilon_{i+\frac{1}{2}-}$ *and* $\widehat{C}^\varepsilon_{i+\frac{1}{2}+}$ *obtained by the three schemes are in fact independent of* $\varepsilon$*.*

*Proof.* . . . . . □

### 3.5.4 Fluxes and source terms based on subcell reconstructions

For all three hydrostatic schemes the flux vector $F_\varepsilon(x)$ is reconstructed by the standard subcell reconstruction (see Def. 1) with reference values

$$F_i := F(U_i), \quad F_{i+\frac{1}{2}-} := F_{i+\frac{1}{2}+} := F_{i+\frac{1}{2}}, \tag{3.50}$$

where

$$F_{i+\frac{1}{2}} := \mathcal{F}_{\mathrm{HLL}}\left(U_{i+\frac{1}{2}-}, U_{i+\frac{1}{2}+}\right) \tag{3.51}$$

is an approximate Riemann solver. Note that $F_\varepsilon(x)$ is globally continuous. The definition of the reconstructed source term $S_\varepsilon(x) := (0, s_\varepsilon(x))^T$ takes the natural form

$$s_\varepsilon(x) := -g h_\varepsilon(x)\, b_\varepsilon{}'(x), \tag{3.52}$$

and hence corresponds directly to (1.4).

Given (3.50)-(3.52), we now introduce the reconstructed, cell-averaged residuum by

$$R_i^\varepsilon := -\frac{1}{\Delta x}\left(F_\varepsilon\left(x_{i+\frac{1}{2}}\right) - F_\varepsilon\left(x_{i-\frac{1}{2}}\right)\right) + \frac{1}{\Delta x}\int_{C_i} S\left(U_\varepsilon(x), b_\varepsilon(x)\right) dx. \tag{3.53}$$

Depending on the choice of hydrostatic scheme, we denote the residuum by $R_i^{\varepsilon,\mathrm{AUD}}$, $R_i^{\varepsilon,\mathrm{MOR}}$, and $R_i^{\varepsilon,\mathrm{CN}}$.

**Theorem 1.** *For each of the three hydrostatic schemes, and for each cell* $C_i$*, the reconstructed residuums are independent of* $\varepsilon$*,* $R_i^\varepsilon = \bar{R}_i$ *for all* $\varepsilon > 0$*, and coincide with the original definitions given above:*

$$\bar{R}_i^{\mathrm{AUD}} = R_i^{\mathrm{AUD}}, \bar{R}_i^{\mathrm{MOR}} = R_i^{\mathrm{MOR}}, \bar{R}_i^{\mathrm{CN}} = R_i^{\mathrm{CN}}. \tag{3.54}$$

*Proof.* See [2, pp. 768–769] □

**Remark 4.** *The main advantage of the third HR scheme is its accuracy for shallow downhill flows.* . . . .

### 3.6 Comparison of the HR schemes

The third HR scheme only differs from the previous methods in the partially wet case (3.4). Additionally, it differs only in $\bar{h}_{i+\frac{1}{2}\pm}$ and $b_{i+\frac{1}{2}}$.

**Proposition 6.** *i) For all interfaces* $x_{i+\frac{1}{2}}$,

$$h^{\mathrm{AUD}}_{i+\frac{1}{2}\pm} = h^{\mathrm{MOR}}_{i+\frac{1}{2}\pm} = h^{\mathrm{CN}}_{i+\frac{1}{2}\pm} \tag{3.55}$$

*and*

$$0 \le h_{i+\frac{1}{2}-} \le h_i, \;\; 0 \le h_{i+\frac{1}{2}+} \le h_{i+1}. \tag{3.56}$$

*ii) For fully wet interfaces* $x_{i+\frac{1}{2}}$,

$$b^{\mathrm{AUD}}_{i+\frac{1}{2}} = b^{\mathrm{MOR}}_{i+\frac{1}{2}} = b^{\mathrm{CN}}_{i+\frac{1}{2}} \text{ and } \bar{h}^{\mathrm{AUD}}_{i+\frac{1}{2}\pm} = \bar{h}^{\mathrm{MOR}}_{i+\frac{1}{2}\pm} = \bar{h}^{\mathrm{CN}}_{i+\frac{1}{2}\pm}. \tag{3.57}$$

*iii) For partially wet interfaces* $x_{i+\frac{1}{2}}$ *(see (3.4)), and if water flows downhill, or if it flows uphill with too little kinetic energy to climb the discrete jump of the bottom (i.e., neither (3.12) nor (3.14) holds), then*

$$b^{\mathrm{AUD}}_{i+\frac{1}{2}} = b^{\mathrm{MOR}}_{i+\frac{1}{2}} \text{ and } \bar{h}^{\mathrm{AUD}}_{i+\frac{1}{2}\pm} = \bar{h}^{\mathrm{MOR}}_{i+\frac{1}{2}\pm}. \tag{3.58}$$

*Proof.* The proof is a direct computation based on the definition of the interface values (3.6), (3.7), (3.16), and (3.17), and the integral averages of the subcell water heights defined in (3.36).

. . . . . □

### 3.7 Stability analysis

Recall that the well-known convex decomposition of the semidiscrete finite volume scheme (1.3) is defined by

$$\frac{d}{dt}U_i(t) = R_{i-\frac{1}{2}+} + R_{i-\frac{1}{2}+} \tag{3.59}$$

$$:= -\frac{1}{\Delta x}\left(F(U_i) - F_{i-\frac{1}{2}}\right) + S_{i-\frac{1}{2}+} - \frac{1}{\Delta x}\left(F_{i+\frac{1}{2}} - F(U_i)\right) + S_{i+\frac{1}{2}-}. \tag{3.60}$$

The following theorem states that the third HR scheme preserves the positivity of the water height under the same condition as Audusse's scheme.

**Theorem 2** (Positivity of water height)**.** *Under condition (3.22), the new semidiscrete HR scheme guarantees nonnegative water height for the homogeneous shallow water equations.*

*Proof.* . . . . . □

Before proving that the third HR scheme is well-balanced, we would like to distinguish the follow two classes of equilibria.

**Definition 2.** *i) Given a constant water level $w_{eq}$, the* still water equilibrium *is given by $u(x) \equiv 0$ and*

$$h(x) + b(x) \equiv w_{eq}. \tag{3.61}$$

*The cell averages are consistent with the still water equilibrium, if for all $i$, $u_i = 0$ and*

$$h_i + b_i = w_{eq}. \tag{3.62}$$

*ii) The* lake at rest equilibrium *is given by $u(x) \equiv 0$ and*

$$h(x)\,\partial_x\left(h(x) + b(x)\right) \equiv 0, \tag{3.63}$$

*for some constant $w_{eq} \geq \max_{x\in\mathbb{R}} b(x)$. Moreover, near a wet-dry interface, the dry part of $b$ should not be lower than the adjacent water level.*

*The cell averages are* locally *(at interface $x_{i+\frac{1}{2}}$)* consistent with the lake at rest*, if $u_i = u_{i+1} = 0$ and either $x_{i+\frac{1}{2}}$ is an interior interface (the still water case)*

$$h_i > 0, h_{i+1} > 0, \text{ and } h_i + b_i = h_{i+1} + b_{i+1}, \tag{3.64}$$

*or a dry-wet front*

$$h_i = 0, h_{i+1} > 0, \text{ and } b_i \geq h_{i+1} + b_{i+1}, \tag{3.65}$$

*or a wet-dry front*

$$h_i > 0, h_{i+1} = 0, \text{ and } b_{i+1} \geq h_i + b_i, \tag{3.66}$$

*or dry*

$$h_i = h_{i+1} = 0. \tag{3.67}$$

*The cell averages are* globally consistent with the lake at rest*, if they are locally consistent with the lake at rest for all interfaces $x_{i+\frac{1}{2}}$.*

*iii) Suppose that the cell averages of the semidiscrete finite volume scheme (1.3) are consistent with a given equilibrium state. Then we call the scheme* well balanced *for this equilibrium state if $R_i = 0$ for all $i$.*

## 4 Offset equilibrium schemes

It suffices to modify in the upwind decenter scheme (2.1) only the discretization of velocity (unlike collocated schemes).

In [3], the author proposed the following modification:

$$\bar{h}_{i+\frac{1}{2}}^{n+1}u_{i+\frac{1}{2}}^{n+1} = \bar{h}_{i+\frac{1}{2}}^{n}u_{i+\frac{1}{2}}^{n} - \frac{\Delta t}{\Delta x}\left[G_{i+1}^{n} - G_{i}^{n} + \frac{g}{2}\left(\left(h_{i+1}^{n}\right)^{2} - \left(h_{i}^{n}\right)^{2}\right) + \frac{g}{2}\left(h_{i}^{n} + h_{i+1}^{n}\right)\left(b_{i+1} - b_{i}\right)\right],$$

which can be rewritten as

$$\bar{h}_{i+\frac{1}{2}}^{n+1}u_{i+\frac{1}{2}}^{n+1} = \bar{h}_{i+\frac{1}{2}}^{n}u_{i+\frac{1}{2}}^{n} - \frac{\Delta t}{\Delta x}\left[G_{i+1}^{n} - G_{i}^{n} + \frac{g}{2}\left(h_{i}^{n} + h_{i+1}^{n}\right)\left(\left(h_{i+1}^{n} + b_{i+1}\right) - \left(h_{i}^{n} + b_{i}\right)\right)\right]. \quad (4.1)$$

We can easily verify that this numerical scheme preserves the balance-type states *still water*.

**Proposition 7** (Well-balanced property)**.** *The numerical scheme* (2.1)-(4.1) *preserved the steady states at rest stated in Proposition 4.*

*Proof.* Suppose that $u_{i+\frac{1}{2}}^{n} = 0$ and $h_{i}^{n} + b_{i} = \text{const}$ for all $i \in \mathcal{E}$, as the proof of Proposition 4, we arrive at

$$\begin{aligned} h_{i}^{n+1} &= h_{i}^{n}, \quad \forall i \in \mathcal{M}, \\ \frac{h_{i}^{n} + h_{i+1}^{n}}{2}u_{i+\frac{1}{2}}^{n+1} &= -\frac{\Delta t}{\Delta x}\frac{g}{2}\left(h_{i}^{n} + h_{i+1}^{n}\right)\left(\left(h_{i+1}^{n} + b_{i+1}\right) - \left(h_{i}^{n} + b_{i}\right)\right), \quad \forall i \in \mathcal{E}_{\text{int}}. \end{aligned}$$

It is straightforward to obtain $h_{i}^{n+1} + b_{i} = \text{const}$ and $u_{i+\frac{1}{2}}^{n+1} = 0$ for all $i \in \mathcal{M}$. □

In the case where the computational domain includes transitions between zones dry and wet areas, we can draw inspiration from the scheme developed in [2].

We define

$$\begin{cases} w_{i}^{n} = h_{i}^{n} + b_{i}, \\ b_{i+\frac{1}{2}}^{n} = \min\left(\max\left(b_{i}, b_{i+1}\right), \min\left(w_{i}^{n}, w_{i+1}^{n}\right)\right), \\ h_{i+\frac{1}{2}-}^{n} = \min\left(w_{i}^{n} - b_{i+\frac{1}{2}}^{n}, h_{i}^{n}\right), \\ h_{i+\frac{1}{2}+}^{n} = \min\left(w_{i+1}^{n} - b_{i+\frac{1}{2}}^{n}, h_{i+1}^{n}\right), \end{cases} \quad (4.2)$$

from which we deduce the discretization of the equation in velocity

$$\bar{h}_{i+\frac{1}{2}}^{n+1}u_{i+\frac{1}{2}}^{n+1} = \bar{h}_{i+\frac{1}{2}}^{n}u_{i+\frac{1}{2}}^{n} - \frac{\Delta t}{\Delta x}\left[G_{i+1}^{n} - G_{i}^{n} + \frac{g}{2}\left(h_{i}^{n} + h_{i+1}^{n}\right)\left(h_{i+\frac{1}{2}+}^{n} - h_{i+\frac{1}{2}-}^{n}\right)\right]. \quad (4.3)$$

We can check that this new scheme preserves the steady states of *the lake at rest*. We can assimilate the calculations of water depths $h_{i+\frac{1}{2}\pm}^{n}$ to a local *runup* algorithm. It remains to be seen whether this scheme practice on unsteady cases.

## References


[1] Emmanuel Audusse et al. "A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows". In: *SIAM J. Sci. Comput.* 25.6 (2004), pp. 2050–2065. ISSN: 1064-8275. DOI: 10.1137/S1064827503431090. URL: https://doi.org/10.1137/S1064827503431090.

[2] Guoxian Chen and Sebastian Noelle. "A new hydrostatic reconstruction scheme based on subcell reconstructions". In: *SIAM J. Numer. Anal.* 55.2 (2017), pp. 758–784. ISSN: 0036-1429. DOI: 10.1137/15M1053074. URL: https://doi.org/10.1137/15M1053074.

[3] D. Doyen and P. H. Gunawan. "An explicit staggered finite volume scheme for the shallow water equations". In: *Finite volumes for complex applications VII. Methods and theoretical aspects*. Vol. 77. Springer Proc. Math. Stat. Springer, Cham, 2014, pp. 227–235. DOI: 10.1007/978-3-319-05684-5_21. URL: https://doi.org/10.1007/978-3-319-05684-5_21.

[4] Vincent Duchêne. *Shallow-water models for water waves (Lectures in Master 2 Fundamental Mathematics and Applications, Université de Rennes 1)*. 2019.

[5] R. Herbin, J.-C. Latché, and T. T. Nguyen. "Explicit staggered schemes for the compressible Euler equations". In: *Applied mathematics in Savoie—AMIS 2012: Multiphase flow in industrial and environmental engineering*. Vol. 40. ESAIM Proc. EDP Sci., Les Ulis, 2013, pp. 83–102. DOI: 10.1051/proc/201340006. URL: https://doi.org/10.1051/proc/201340006.

[6] T. Morales de Luna, M. J. Castro Díaz, and C. Parés. "Reliability of first order numerical schemes for solving shallow water system over abrupt topography". In: *Appl. Math. Comput.* 219.17 (2013), pp. 9012–9032. ISSN: 0096-3003. DOI: 10.1016/j.amc.2013.03.033. URL: https://doi.org/10.1016/j.amc.2013.03.033.