Implicit kinetic schemes for the shallow water system

Chourouk El Hassanieh[1], Mathieu Rigal[2], Jacques Sainte-Marie[1]

[1]Laboratoire Jacques-Louis Lions, Sorbonne Université

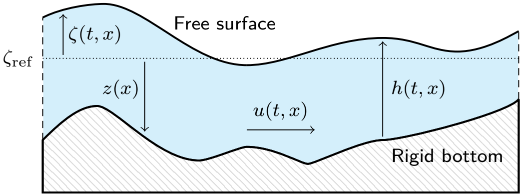
[2]Institut de Mathématiques de Bordeaux

Amiens, November 17, 2023

Why are we interested in geophysical flows?

- water management
- natural disasters
- energy production

→ numerical schemes, simulations



Quantities of interest:

h	$\rightarrow$	water height
u	$\rightarrow$	horizontal velocity
hu	$\rightarrow$	horizontal discharge
z	$\rightarrow$	bathymetry

Free surface flows $\Rightarrow$ evolving fluid geometry

Shallow water equations: vertically averaged model (reduced complexity)

Simplifying assumptions

- shallow flow
- velocity has small variations along the vertical
- no plunging wave

Important properties

1D shallow water system:

$$\begin{cases} \partial_t h + \partial_x hu = 0 \\ \partial_t hu + \partial_x (hu^2 + \frac{g}{2} h^2) = -gh\partial_x z \end{cases} \quad \text{in } \mathbb{R} \tag{SV}$$

Convenient vector notation $\partial_t U + \partial_x F(U) = S(U, z)$ with $U = (h, hu)^T$.

Important properties at the continuous level:

- Positivity ($h \geq 0 \ \forall t$)
- Stationary state $h + z \equiv$ Cst, $u \equiv 0$
- Entropy inequality $\partial_t \eta(U, z) + \partial_x G(U, z) \leq 0$

$$\eta(U, z) = \frac{hu^2}{2} + \frac{gh^2}{2} + ghz, \quad G(U, z) = \Big(\eta(U, z) + \frac{gh^2}{2}\Big)u$$

Important properties

1D shallow water system:

$$\begin{cases} \partial_t h + \partial_x hu = 0 \\ \partial_t hu + \partial_x (hu^2 + \frac{g}{2} h^2) = -gh\partial_x z \end{cases} \quad \text{in } \mathbb{R} \tag{SV}$$

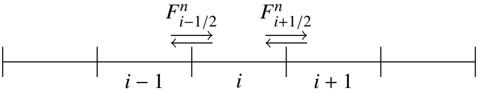
Convenient vector notation $\partial_t U + \partial_x F(U) = S(U, z)$ with $U = (h, hu)^T$.

Important properties at the continuous level:

- Positivity ($h \geq 0 \ \forall t$)
- Stationary state $h + z \equiv$ Cst, $u \equiv 0$
- Entropy inequality $\partial_t \eta(U, z) + \partial_x G(U, z) \leq 0$

$$\eta(U, z) = \frac{hu^2}{2} + \frac{gh^2}{2} + ghz, \quad G(U, z) = \Big(\eta(U, z) + \frac{gh^2}{2}\Big)u$$

Finite volumes: a simple example



Finite volume scheme of the form

$$\begin{cases} \dfrac{U_i^{n+1} - U_i^n}{\Delta t} + \dfrac{1}{\Delta x}(F^n_{i+1/2} - F^n_{i-1/2}) = S_i^n \\ F^n_{i+1/2} = \mathcal{F}(U_i^n, z_i, U^n_{i+1}, z_{i+1}) \end{cases} \qquad (1)$$

For instance, Rusanov flux + centered source

$$F^n_{i+1/2} = \frac{1}{2}\Big(F(U_i^n, z_i) + F(U^n_{i+1}, z_{i+1})\Big) - \frac{a}{2}(U^n_{i+1} - U_i^n), \quad a > 0$$

$$S_i^n = -gh_i^n \frac{z_{i+1} - z_{i-1}}{2\Delta x} \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

Finite volumes: numerical challenges

Problem n°1: if it exists, find $G^n_{i+1/2}$ numerical entropy flux such that

$$\text{Update (1)} \implies \begin{cases} \dfrac{\eta(U_i^{n+1}, z_i) - \eta(U_i^n, z_i)}{\Delta t} + \dfrac{1}{\Delta x}(G^n_{i+1/2} - G^n_{i-1/2}) \le 0 \\ G^n_{i+1/2} = \mathcal{G}(U_i^n, z_i, U_{i+1}^n, z_{i+1}) \end{cases}$$

Problem n°2: preserve lakes at rest ($h + z = 0$, $u = 0$)

Steady state $\partial_t U = 0$ implies $\partial_x F(U) = S(U, z)$, whereas at discrete level:

$$\frac{1}{\Delta x}\left(F^n_{i+1/2} - F^n_{i-1/2}\right) = \begin{pmatrix} \dfrac{a}{2}\,\dfrac{z_{i+1} - 2z_i + z_{i-1}}{\Delta x} \\ \dfrac{g}{2\Delta x}((z_{i+1})^2 - (z_{i-1})^2) \end{pmatrix} \neq \begin{pmatrix} 0 \\ g z_i \dfrac{z_{i+1} - z_{i-1}}{2\Delta x} \end{pmatrix},$$

and therefore $\dfrac{U_i^{n+1} - U_i^n}{\Delta t} \neq 0.$

Finite volumes: numerical challenges

Solution proposed by Audusse et. al (2016)

→ Explicit kinetic scheme preserving lakes at rest...

→ but satisfying a discrete entropy inequality with error term

$$\frac{\eta(U_i^{n+1}, z_i) - \eta(U_i^n, z_i)}{\Delta t} + \frac{1}{\Delta x}(G_{i+1/2}^n - G_{i-1/2}^n) \leq D_i^n,$$

where in some cases $D_i^n > 0$

Our goal is to implicit this scheme to improve its stability

Outline of the talk:

1. Brief recall of the kinetic formalism
2. The case of a flat topography
3. The case of a varying topography

Kinetic representation of the shallow water system

Kinetic equation with BGK collision operator

$$\partial_t f + \xi \partial_x f - g(\partial_x z)\partial_\xi f = \frac{1}{\epsilon}(M(U,\xi) - f) \tag{BGK}$$

- $f(t,x,\xi) \geq 0$ density of particles with velocity ξ
- Moment relations $\int (1,\xi,\xi^2)^T M(U,\xi)\,\mathrm{d}\xi = (h, hu, hu^2 + gh^2/2)^T$
- In the limit $\varepsilon \to 0$, we formally have $f \to M$

Proposition 1 (Audusse, Bouchut, Bristeau, and Sainte-Marie 2016)

If the bathymetry $z(x)$ is Lipschitz continuous, then U is solution of the shallow water system iff $M(U,\xi)$ satisfies the kinetic equation

$$\partial_t M + \xi \partial_x M - g(\partial_x z)\partial_\xi M = Q \tag{2}$$

for some collision term $Q(t,x,\xi)$ that satisfies $\int_{\mathbb{R}} (1,\xi)^T Q\,\mathrm{d}\xi = 0$ for a.e. (t,x).

Definition 1 (Kinetic entropy H)

$H(f,\xi)$ convex in f and satisfying

$$\int_{\mathbb{R}} H(M(U,\xi),\xi)\,\mathrm{d}\xi = \eta(U), \qquad \int_{\mathbb{R}} H(M(U_f,\xi),\xi)\,\mathrm{d}\xi \le \int_{\mathbb{R}} H(f,\xi)\,\mathrm{d}\xi \ \forall f$$

If flat bottom ($z \equiv$ Const), integrate (BGK) against $\partial_1 H(f,\xi)$ to get

$$\underbrace{\partial_t \int_{\mathbb{R}} H(f,\xi)\,\mathrm{d}\xi}_{\xrightarrow{\varepsilon\to 0}\ \partial_t \eta(U_f)} + \underbrace{\partial_x \int_{\mathbb{R}} \xi H(f,\xi)\,\mathrm{d}\xi}_{\xrightarrow{\varepsilon\to 0}\ \partial_x G(U_f)} = \frac{1}{\varepsilon} \underbrace{\int_{\mathbb{R}} \partial_1 H(f,\xi)(M(U_f,\xi) - f)\,\mathrm{d}\xi}_{\le \int_{\mathbb{R}} H(M,\xi) - H(f,\xi)\,\mathrm{d}\xi \le 0}$$

Extends to varying bottoms if

$$\int_{\mathbb{R}} \partial_3 H(f,z,\xi)\,\mathrm{d}\xi = hu,$$

which implies

$$\lim_{\varepsilon\to 0} \int_{\mathbb{R}} \partial_1 H(f,z,\xi)\left(\xi\partial_x f - g(\partial_x z)\partial_\xi f\right)\mathrm{d}\xi = \partial_x G(U,z)$$

Kinetic representation of the shallow water system

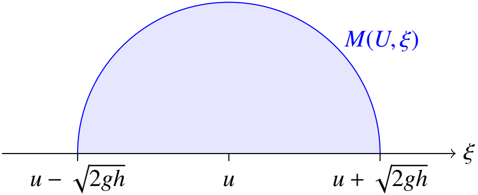
Given a convex H, determine $M(U, \cdot)$ by minimizing

$$f \mapsto \int_{\mathbb{R}} H(f, \xi)\, \mathrm{d}\xi \quad \text{constrained by} \quad \int_{\mathbb{R}} (1, \xi)^T f\, \mathrm{d}\xi = U$$

François Bouchut. "Construction of BGK Models with a Family of Kinetic Entropies for a Given System of Conservation Laws." (1999)

Lemma 1 (Perthame and Simeoni 2001)

$H(f, z, \xi) = \dfrac{\xi^2}{2} f + \dfrac{g^2\pi^2}{6} f^3 + gzf$ *is a kinetic entropy for* $M(U, \xi) = \dfrac{1}{g\pi} \sqrt{(2gh - (\xi - u)^2)_+}$.



Explicit time discretization involving BGK splitting

$$\begin{cases} \dfrac{f^{n+1/2} - f^n}{\Delta t} = \dfrac{1}{\varepsilon}\left(M(U_f^n, \xi) - f^{n+1/2}\right) & \text{collision step} \\ \dfrac{f^{n+1} - f^{n+1/2}}{\Delta t} + \xi \partial_x f^{n+1/2} = 0 & \text{transport step} \end{cases}$$

Explicit first order upwind scheme when $\epsilon \to 0$

$$\frac{f_i^{n+1} - M_i^n}{\Delta t} + \frac{\xi}{\Delta x}\Big(\mathbb{1}_{\xi<0}\big(M_{i+1}^n - M_i^n\big) + \mathbb{1}_{\xi>0}\big(M_i^n - M_{i-1}^n\big)\Big) = 0 \tag{3}$$

Macroscopic rewriting by integrating (3) against $(1, \xi)^T$

$$\frac{U_i^{n+1} - U_i^n}{\Delta t} + \frac{1}{\Delta x}\Big(F(U_i^n, U_{i+1}^n) - F(U_{i-1}^n, U_i^n)\Big) = 0$$

Kinetic numerical flux $F(U_L, U_R) = \int_{\xi<0} \xi \begin{pmatrix} 1 \\ \xi \end{pmatrix} M(U_R, \xi)\,\mathrm{d}\xi + \int_{\xi>0} \xi \begin{pmatrix} 1 \\ \xi \end{pmatrix} M(U_L, \xi)\,\mathrm{d}\xi.$

The case of a flat topography

Explicit time discretization involving BGK splitting

$$\begin{cases} \dfrac{f^{n+1/2} - f^n}{\Delta t} = \dfrac{1}{\varepsilon}(M(U_f^n, \xi) - f^{n+1/2}) & \text{collision step} \\ \dfrac{f^{n+1} - f^{n+1/2}}{\Delta t} + \xi \partial_x f^{n+1/2} = 0 & \text{transport step} \end{cases}$$

Explicit first order upwind scheme when $\epsilon \to 0$

$$\frac{f_i^{n+1} - M_i^n}{\Delta t} + \frac{\xi}{\Delta x}\Big(\mathbb{1}_{\xi<0}\big(M_{i+1}^n - M_i^n\big) + \mathbb{1}_{\xi>0}\big(M_i^n - M_{i-1}^n\big)\Big) = 0 \qquad (3)$$

Macroscopic rewriting by integrating (3) against $(1, \xi)^T$

$$\frac{U_i^{n+1} - U_i^n}{\Delta t} + \frac{1}{\Delta x}\Big(F(U_i^n, U_{i+1}^n) - F(U_{i-1}^n, U_i^n)\Big) = 0$$

Kinetic numerical flux $F(U_L, U_R) = \int_{\xi<0} \xi \begin{pmatrix} 1 \\ \xi \end{pmatrix} M(U_R, \xi)\, \mathrm{d}\xi + \int_{\xi>0} \xi \begin{pmatrix} 1 \\ \xi \end{pmatrix} M(U_L, \xi)\, \mathrm{d}\xi$.

The case of a flat topography

Do we satisfy a discrete counterpart to $\partial_t \eta + \partial_x G \leq 0$?

Proposition 2 (Audusse, Bouchut, Bristeau, and Sainte-Marie 2016)

If the CFL $\frac{\Delta t}{\Delta x}|\xi| \leq 1$ *holds for any* $\xi \in \operatorname{supp} M^n$, *then the explicit kinetic scheme* (3) *satisfies*

$$h_i^{n+1} \geq 0 \quad \textit{together with} \quad \frac{\eta(U_i^{n+1}) - \eta(U_i^n)}{\Delta t} + \frac{1}{\Delta x}(G_{i+1/2}^n - G_{i-1/2}^n) \leq 0$$

Proof: set $\sigma = \frac{\Delta t}{\Delta x}$ and rewrite (3) as

$$f_i^{n+1} = (1 - \sigma|\xi|)M_i^n + \sigma|\xi|M_{i\pm1}^n \geq 0$$

Also $\eta_i^{n+1} = \int_{\mathbb{R}} H(M_i^{n+1}, \xi)\,\mathrm{d}\xi \leq \int_{\mathbb{R}} H(f_i^{n+1}, \xi)\,\mathrm{d}\xi \leq \underbrace{\int_{\mathbb{R}} (1 - \sigma|\xi|)H_i^n + \sigma|\xi|H_{i\pm1}^n\,\mathrm{d}\xi}_{\eta_i^n - \sigma(G_{i+1/2}^n - G_{i-1/2}^n)}$

The case of a flat topography

We study the implicit version of the previous scheme.

$$\frac{f_i^{n+1} - M_i^n}{\Delta t} + \frac{\xi}{\Delta x}\Big(\mathbb{1}_{\xi<0}\big(f_{i+1}^{n+1} - f_i^{n+1}\big) + \mathbb{1}_{\xi>0}\big(f_i^{n+1} - f_{i-1}^{n+1}\big)\Big) = 0 \tag{4}$$

Solve the system $(\mathbf{I} + \sigma\mathbf{L})f^{n+1} = M^n + \sigma B^{n+1}$ with $\sigma = \Delta t/\Delta x$ and

$$\mathbf{L} = |\xi|\begin{pmatrix} 1 & -\mathbb{1}_{\xi<0} & & & 0 \\ -\mathbb{1}_{\xi>0} & 1 & -\mathbb{1}_{\xi<0} & & \\ & \ddots & \ddots & \ddots & \\ & & -\mathbb{1}_{\xi>0} & 1 & -\mathbb{1}_{\xi<0} \\ 0 & & & -\mathbb{1}_{\xi>0} & 1 \end{pmatrix}_{N\times N}, \quad B^{n+1} = |\xi|\begin{pmatrix} M_0^{n+1}\mathbb{1}_{\xi>0} \\ 0 \\ \vdots \\ 0 \\ M_{N+1}^{n+1}\mathbb{1}_{\xi<0} \end{pmatrix}_N$$

In practice, ghost cell contribution B^{n+1} unknown → substitute it by B^n.

Proposition 3 (El Hassanieh, R., Sainte-Marie)

The implicit kinetic scheme (4) is well defined, its update can be computed analytically and it enjoys the same properties as the explicit scheme $\forall \Delta t > 0$.

The case of a flat topography

We study the implicit version of the previous scheme.

$$\frac{f_i^{n+1} - M_i^n}{\Delta t} + \frac{\xi}{\Delta x}\Big(\mathbb{1}_{\xi<0}\big(f_{i+1}^{n+1} - f_i^{n+1}\big) + \mathbb{1}_{\xi>0}\big(f_i^{n+1} - f_{i-1}^{n+1}\big)\Big) = 0 \qquad (4)$$

Solve the system $(\mathbf{I} + \sigma\mathbf{L})f^{n+1} = M^n + \sigma B^{n+1}$ with $\sigma = \Delta t/\Delta x$ and

$$\mathbf{L} = |\xi| \begin{pmatrix} 1 & -\mathbb{1}_{\xi<0} & & & 0 \\ -\mathbb{1}_{\xi>0} & 1 & -\mathbb{1}_{\xi<0} & & \\ & \ddots & \ddots & \ddots & \\ & & -\mathbb{1}_{\xi>0} & 1 & -\mathbb{1}_{\xi<0} \\ 0 & & & -\mathbb{1}_{\xi>0} & 1 \end{pmatrix}_{N\times N}, \quad B^{n+1} = |\xi| \begin{pmatrix} M_0^{n+1}\mathbb{1}_{\xi>0} \\ 0 \\ \vdots \\ 0 \\ M_{N+1}^{n+1}\mathbb{1}_{\xi<0} \end{pmatrix}_N$$

In practice, ghost cell contribution B^{n+1} unknown $\rightarrow$ substitute it by B^n.

Proposition 3 (El Hassanieh, R., Sainte-Marie)

The implicit kinetic scheme (4) *is well defined, its update can be computed analytically and it enjoys the same properties as the explicit scheme* $\forall \Delta t > 0$.

The case of a flat topography

$$\frac{f_i^{n+1} - M_i^n}{\Delta t} + \frac{\xi}{\Delta x}\left(f_{i+1/2}^{n+1} - f_{i-1/2}^{n+1}\right) = 0 \iff (\mathbf{I} + \sigma\mathbf{L})f^{n+1} = M^n + \sigma B^n \tag{4}$$

Sketch of the proof:

Well-defined: The mass matrix has a strictly dominant diagonal $\Rightarrow$ invertible

Positivity: The mass matrix is monotone and RHS is positive

Analytic expression: Decompose $\mathbf{L} = |\xi|\mathbf{I} - \mathbf{N}$ so that

$$(\mathbf{I} + \sigma\mathbf{L})^{-1} = \frac{1}{1+\sigma|\xi|}\left(\mathbf{I} - \frac{\sigma}{1+\sigma|\xi|}\mathbf{N}\right)^{-1} = \frac{1}{1+\sigma|\xi|}\sum_{k=0}^{N}\left(\frac{\sigma}{1+\sigma|\xi|}\mathbf{N}\right)^k$$

Entropy inequality: Multiply (4) by $\partial_1 H(f_i^{n+1}, \xi)$ and use

$$\partial_1 H(b,\xi)(b-a) = H(b,\xi) - H(a,\xi) + \frac{g^2\pi^2}{6}(2b+a)(b-a)^2$$

to obtain

$$\frac{H(f_i^{n+1}) - H(M_i^n)}{\Delta t} + \frac{\xi}{\Delta x}(H_{i+1/2}^{n+1} - H_{i-1/2}^{n+1}) = \widetilde{D}_i(\xi) \leq 0$$

The case of a flat topography

In practice, cannot obtain explicit expression for $\int_{\mathbb{R}} \begin{pmatrix} 1 \\ \xi \end{pmatrix} (\mathbf{I} + \sigma \mathbf{L})^{-1} M \, \mathrm{d}\xi$ with

$$M(U, \xi) = \frac{1}{g\pi} \sqrt{\left(2gh - (\xi - u)^2\right)_+}$$

Substitute M with a simpler Maxwellian satisfying the moment relations

$$\widetilde{M}(U, \xi) = \frac{h}{2\sqrt{3}c} \mathbb{1}_{|\xi - u| \leq \sqrt{3}c}, \quad c = \sqrt{\frac{gh}{2}}$$

- nonlinear implicit update **can be rewritten explicitly**
- counterpart: unlike M, $\widetilde{M}$ doesn't minimize $\int_{\mathbb{R}} H(\cdot, \xi) \, \mathrm{d}\xi$
- as a consequence, no proof of discrete entropy inequality...
- ... but in practice, it seems to dissipate energy (numerical validation)

Explicit writing of the implicit kinetic scheme

Neglecting the ghost cells, the implicit kinetic update writes

$$h^{n+1} = \mathbf{A}(U^n) \cdot \sqrt{h^n}, \qquad hu^{n+1} = \mathbf{B}(U^n) \cdot \sqrt{h^n}$$

$\mathbf{A}(U^n), \mathbf{B}(U^n)$ dense matrices of $\mathbb{R}^{N \times N}$

- matrix vector product has complexity $\mathcal{O}(N^2)$ (cannot do better)
- coefficients of $\mathbf{A}, \mathbf{B}$ involve sums of up to N terms $\Rightarrow$ matrix assembly in $\mathcal{O}(N^3)$

Optimization: assemble matrices in specific order

- each coefficient computed in $\mathcal{O}(1)$ steps from the previous one
- cost of matrix assembly reduced to $\mathcal{O}(N^2)$

$\rightarrow$ Fully vectorized implementation in Python

Explicit writing of the implicit kinetic scheme

Neglecting the ghost cells, the implicit kinetic update writes

$$h^{n+1} = \mathbf{A}(U^n) \cdot \sqrt{h^n}, \qquad hu^{n+1} = \mathbf{B}(U^n) \cdot \sqrt{h^n}$$

$\mathbf{A}(U^n), \mathbf{B}(U^n)$ dense matrices of $\mathbb{R}^{N \times N}$

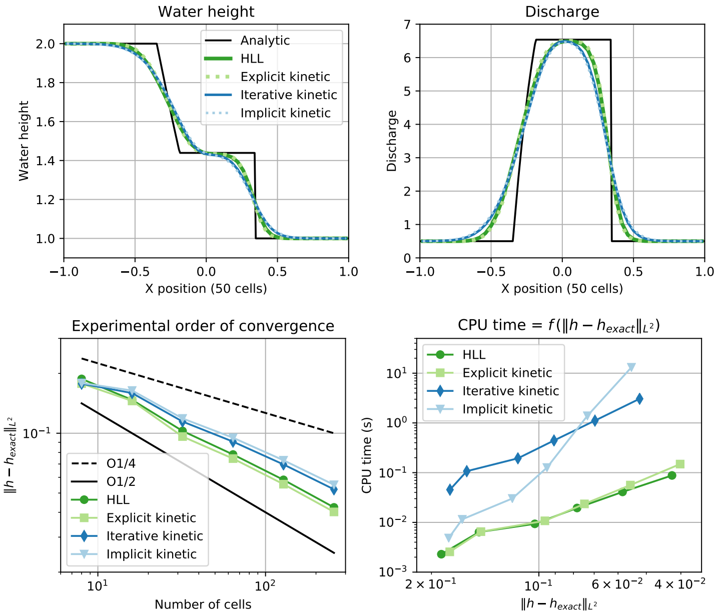
- matrix vector product has complexity $\mathcal{O}(N^2)$ (cannot do better)
- coefficients of $\mathbf{A}, \mathbf{B}$ involve sums of up to N terms $\Rightarrow$ matrix assembly in $\mathcal{O}(N^3)$

Optimization: assemble matrices in specific order

- each coefficient computed in $\mathcal{O}(1)$ steps from the previous one
- cost of matrix assembly reduced to $\mathcal{O}(N^2)$

$\rightarrow$ Fully vectorized implementation in Python

Implicit kinetic scheme: Riemann problem

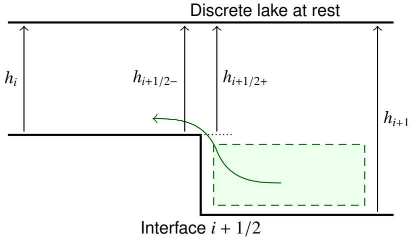


Discretize source term in (SV) $\begin{cases} \partial_t h + \partial_x hu = & 0 \\ \partial_t hu + \partial_x (hu^2 + \frac{g}{2}h^2) = & -gh\partial_x z \end{cases}$

Problem: how to preserve lakes at rest $h + z \equiv$ Cst, $u \equiv 0$?

- Upwinding introduces diffusion on $h \Rightarrow h^{n+1} \neq h^n$
- Pressure variation should balance with source: $\partial_x\left(\frac{g}{2}h^2\right) = -gh\partial_x z$

Hydrostatic reconstruction



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

$$h_{i+1/2-} = (h_i + z_i - z_{i+1/2})_+$$

$$h_{i+1/2+} = (h_{i+1} + z_{i+1} - z_{i+1/2})_+$$

Audusse, Bouchut, Bristeau, Klein, et al. 2004 “A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows.”

The case of a varying topography: hydrostatic reconstruction

Numerical flux and source term using reconstructed values

$$\begin{cases} \widetilde{F}_{i+1/2} = F(U_{i+1/2-}, U_{i+1/2+}) \\ \widetilde{F}_{i-1/2} = F(U_{i-1/2-}, U_{i-1/2+}) \end{cases} , \quad \widetilde{S}_i = \frac{g}{2\Delta x}(h^2_{i+1/2-} - h^2_{i-1/2+}) \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

- $\widetilde{S}_i$ is indeed consistent with the source term

$$\frac{1}{2\Delta x}(h^2_{i+1/2-} - h^2_{i-1/2+}) = \underbrace{\frac{1}{2}(h_{i+1/2-} + h_{i-1/2+})}_{h_i + \mathcal{O}(\Delta z_i)} \times \underbrace{\frac{1}{\Delta x}(h_{i+1/2-} - h_{i-1/2+})}_{-(z_{i+1/2} - z_{i-1/2})/\Delta x} = -h\partial_x z + \mathcal{O}(\Delta x)$$

- If $F(U,U) = F(U)$ (consistency), then over lakes at rest one has

$$U_{i+1/2-} = U_{i+1/2+} \implies \frac{\widetilde{F}_{i+1/2} - \widetilde{F}_{i-1/2}}{\Delta x} = \frac{F(U_{i+1/2-}) - F(U_{i-1/2+})}{\Delta x} = \widetilde{S}_i$$

The case of a varying topography: explicit kinetic scheme

Explicit kinetic scheme with hydrostatic reconstruction:

$$\widetilde{F}_{i+1/2} = \int_{\mathbb{R}} \xi \begin{pmatrix} 1 \\ \xi \end{pmatrix} \Big(\mathbb{1}_{\xi>0} M(U_{i+1/2-}, \xi) + \mathbb{1}_{\xi<0} M(U_{i+1/2+}, \xi) \Big) \, \mathrm{d}\xi$$

$$\frac{U_i^{n+1} - U_i^n}{\Delta t} + \frac{1}{\Delta x} (\widetilde{F}^n_{i+1/2} - \widetilde{F}^n_{i-1/2}) = \widetilde{S}^n_i \qquad (5)$$

Proposition 4 (Audusse, Bouchut, Bristeau, and Sainte-Marie 2016)

Under the CFL condition $\frac{\Delta t}{\Delta x}|\xi| < 1$ *the scheme* (5) *preserves the water height positivity, and admits the discrete entropy inequality*

$$\frac{\eta(U_i^{n+1}, z_i) - \eta(U_i^n, z_i)}{\Delta t} + \frac{1}{\Delta x}(G^n_{i+1/2} - G^n_{i-1/2}) \leq D_i,$$

where D_i *features a quadratic error term, Lipschitz in* $\sigma, \Delta x, \Delta z_i$ *and vanishing when* $u_i^n \to 0$.

⇒ We cannot ensure the dissipation of the total energy $\int_\Omega \eta(U(t,x)) \, \mathrm{d}x$

To solve this issue, implicit the previous scheme

$$\frac{U_i^{n+1} - U_i^n}{\Delta t} + \frac{1}{\Delta x}(\widetilde{F}_{i+1/2}^{n+1} - \widetilde{F}_{i-1/2}^{n+1}) = \widetilde{S}_i^{n+1} \tag{6}$$

Approximate U_i^{n+1} using fixed point iteration $(U_i^{n+1,k})_{k\in\mathbb{N}}$

$$(1+\alpha)U_i^{n+1,k+1} = U_i^n + \alpha U_i^{n+1,k} - \frac{\Delta t}{\Delta x}\left(\widetilde{F}_{i+1/2}^{n+1,k} - \widetilde{F}_{i-1/2}^{n+1,k}\right) + \Delta t\,\widetilde{S}_i^{n+1,k}, \quad \alpha \geq 0 \tag{7}$$

Proposition 5 (El Hassanieh, R., Sainte-Marie)

1. *If $(U_i^{n+1,k})_{k\in\mathbb{N}}$ converges, its limit is the solution U_i^{n+1} of (6).*
2. *We have $h_i^{n+1,k+1} \geq 0$ assuming $\frac{\Delta t}{\Delta x}|\xi| \leq \alpha$ for all $\xi \in \operatorname{supp} M(U_i^{n+1,k},\cdot)$*
3. *The iterative process (7) satisfies a macroscopic entropy inequality of the form*

$$\frac{\eta(U_i^{n+1,k+1}, z_i) - \eta(U_i^n, z_i)}{\Delta t} + \frac{1}{\Delta x}\left(G_{i+1/2}^{n+1,k} - G_{i-1/2}^{n+1,k}\right) \leq D_i^{n+1,k},$$

with $D_i^{n+1,k} \leq 0$ from some rank k assuming (7) converges.

The case of a varying topography: iterative kinetic scheme

To solve this issue, implicit the previous scheme

$$\frac{U_i^{n+1} - U_i^n}{\Delta t} + \frac{1}{\Delta x}(\widetilde{F}_{i+1/2}^{n+1} - \widetilde{F}_{i-1/2}^{n+1}) = \widetilde{S}_i^{n+1} \tag{6}$$

Approximate U_i^{n+1} using fixed point iteration $(U_i^{n+1,k})_{k\in\mathbb{N}}$

$$(1+\alpha)U_i^{n+1,k+1} = U_i^n + \alpha U_i^{n+1,k} - \frac{\Delta t}{\Delta x}\left(\widetilde{F}_{i+1/2}^{n+1,k} - \widetilde{F}_{i-1/2}^{n+1,k}\right) + \Delta t\,\widetilde{S}_i^{n+1,k}, \quad \alpha \geq 0 \tag{7}$$

Proposition 5 (El Hassanieh, R., Sainte-Marie)

1. *If $(U_i^{n+1,k})_{k\in\mathbb{N}}$ converges, its limit is the solution U_i^{n+1} of* (6)*.*
2. *We have $h_i^{n+1,k+1} \geq 0$ assuming $\frac{\Delta t}{\Delta x}|\xi| \leq \alpha$ for all $\xi \in$* supp *$M(U_i^{n+1,k},\cdot)$*
3. *The iterative process* (7) *satisfies a macroscopic entropy inequality of the form*

$$\frac{\eta(U_i^{n+1,k+1}, z_i) - \eta(U_i^n, z_i)}{\Delta t} + \frac{1}{\Delta x}\left(G_{i+1/2}^{n+1,k} - G_{i-1/2}^{n+1,k}\right) \leq D_i^{n+1,k},$$

with $D_i^{n+1,k} \leq 0$ from some rank k assuming (7) *converges.*

The case of a varying topography: iterative kinetic scheme

Sketch of the proof: ① is the consequence of the continuity of $\widetilde{F}_{i+1/2}, \widetilde{S}_i$.

For ② and ③, write iteration (7) at kinetic level

$$\widetilde{F}_{i+1/2} = \int_{\mathbb{R}} \begin{pmatrix} 1 \\ \xi \end{pmatrix} \xi \Big(\underbrace{\mathbb{1}_{\xi>0} M(U_{i+1/2-}, \xi) + \mathbb{1}_{\xi<0} M(U_{i+1/2+}, \xi)}_{M_{i+1/2}} \Big) \, \mathrm{d}\xi$$

Recall that $\widetilde{S}_i = (0, \frac{g}{2\Delta x} h_{i+1/2-}^2 - \frac{g}{2\Delta x} h_{i-1/2+}^2)^T$ so

$$\begin{aligned} \widetilde{S}_i &= \frac{1}{\Delta x} \Big(F(U_{i+1/2-}) - F(U_{i-1/2+}) - u_i \cdot (U_{i+1/2-} - U_{i-1/2+}) \Big) \\ &= \frac{1}{\Delta x} \int_{\mathbb{R}} \begin{pmatrix} 1 \\ \xi \end{pmatrix} (\xi - u_i) \cdot \Big(\underbrace{M(U_{i+1/2-}, \xi)}_{M_{i+1/2-}} - \underbrace{M(U_{i-1/2+}, \xi)}_{M_{i-1/2+}} \Big) \, \mathrm{d}\xi \end{aligned}$$

$\Rightarrow$ At kinetic level, the fixed point iteration can be written

$$(1+\alpha) f_i^{n+1,k+1} = M_i^n + \alpha M_i^{n+1,k} - \sigma \xi \Big(M_{i+1/2}^{n+1,k} - M_{i-1/2}^{n+1,k} \Big) + \sigma (\xi - u_i^{n+1,k}) \Big(M_{i+1/2-}^{n+1,k} - M_{i-1/2+}^{n+1,k} \Big) \quad (8)$$

The case of a varying topography: iterative kinetic scheme

The proof of ❷ and ❸ relies on the kinetic rewriting

$$(1+\alpha) f_i^{n+1,k+1} = M_i^n + \alpha M_i^{n+1,k} - \sigma\xi\left(M_{i+1/2}^{n+1,k} - M_{i-1/2}^{n+1,k}\right) + \sigma(\xi - u_i^{n+1,k})\left(M_{i+1/2-}^{n+1,k} - M_{i-1/2+}^{n+1,k}\right) \quad (8)$$

❷ Positivity: The quantity $(1+\alpha)h_i^{n+1,k}$ equals

$$\int_{\mathbb{R}} \left(M_i^n + \alpha M_i^{n+1,k} - \sigma\xi\left(M_{i+1/2}^{n+1,k} - M_{i-1/2}^{n+1,k}\right)\right) \mathrm{d}\xi \geq \int_{\mathbb{R}} \left(M_i^n + M_i^{n+1,k}\left(\alpha - \sigma|\xi|\right)\right) \mathrm{d}\xi$$

since $h_{i\pm 1/2\mp}^{n+1,k} \leq h_i^{n+1,k} \Rightarrow M_{i\pm 1/2\mp}^{n+1,k} \leq M_i^{n+1,k}$

❸ Entropy inequality: Multiply (8) by $\partial_1 H(M_i^{n+1,k}, z_i, \xi)$ and use convexity of H

$$H(M_i^{n+1,k+1}, z_i) \leq H(M_i^n, z_i) - \sigma\left(\widetilde{G}_{i+1/2}^{n+1,k} - \widetilde{G}_{i-1/2}^{n+1,k}\right) + Q(\xi) + \widetilde{D}_i,$$

with $\int_{\mathbb{R}} Q(\xi)\,\mathrm{d}\xi = 0$ and

$$\widetilde{D}_i = \text{Strictly negative term } + O(M_i^{n+1,k+1} - M_i^{n+1,k})$$

The case of a varying topography: iterative kinetic scheme

Proposition 6 (El Hassanieh, R., Sainte-Marie)

Assume the iterative scheme (7) keeps $U_i^{n+1,k}$ *in* $\{(h,hu)^T,\ \delta \leq h \leq K_1,\ |u| \leq K_2\}$ *for all* k*.*
$\exists C(K_1, K_2, 1/\delta)$ *such that* $\Delta t \leq C\Delta x$ *implies the convergence of* $(U_i^{n+1,k})_{k\in\mathbb{N}}$ *to* U_i^{n+1} *solution of the implicit kinetic scheme with hydrostatic reconstruction.*

$\rightarrow$ In practice, iterative process seems to converge without restriction $\delta \leq h$

Stopping criteria: tolerance + total energy dissipation

$$\|U^{n+1,k+1} - U^{n+1,k}\| \leq \tau \quad \& \quad \frac{1}{\Delta t}\sum_{1\leq i\leq N}\Big(\eta(U_i^{n+1,k+1}, z_i) - \eta(U_i^n, z_i)\Big) + \frac{1}{\Delta x}(G_{N+1/2}^{n+1,k} - G_{1/2}^{n+1,k}) \leq 0$$

Numerical simulations: energy dissipation

Total energy $\int_\Omega \eta \, \mathrm{d}x$ shoud decrease in time due to **entropy inequality**.

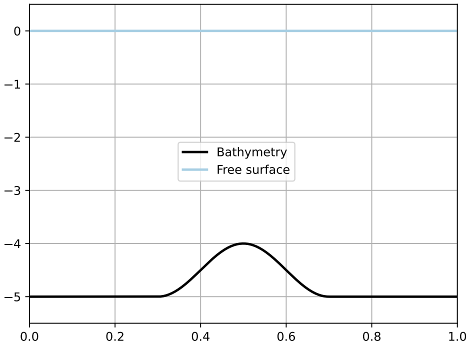
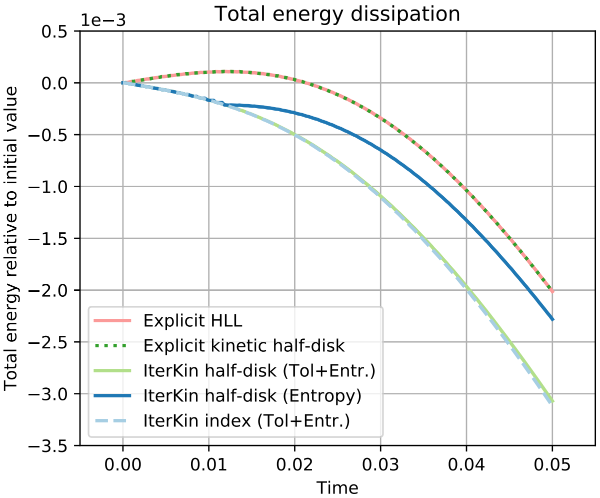


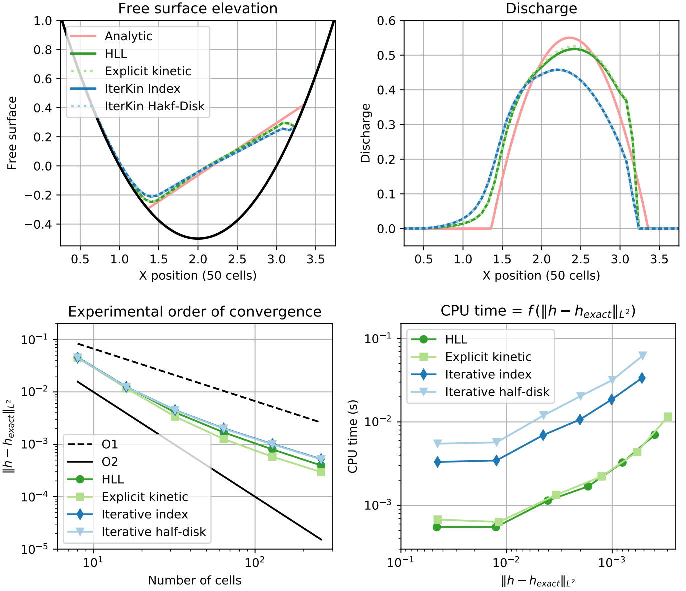
Figure: Initial condition with $h + z \equiv 0$, $u \equiv 1$ and $z \neq$ Cst

→ Compare explicit and implicit kinetic schemes

Numerical simulations: energy dissipation

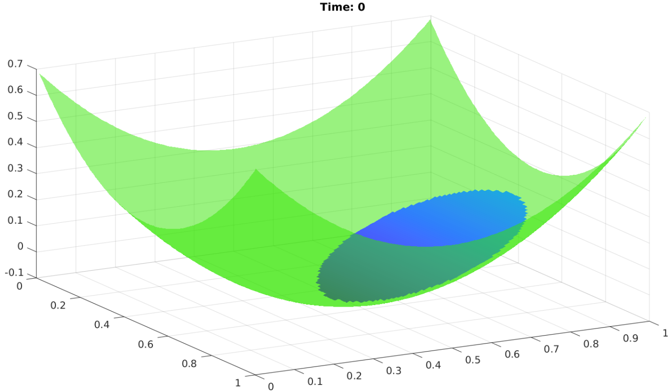


Numerical simulations: 1D parabolic bowl



→ Results still valid in 2D

→ Good approximation of the parabolic bowl (difficult numerical testcase)



Conclusion and perspectives

For a flat topography

- Positivity and entropy inequality obtained unconditionally
- Obtained fully implicit scheme with explicit update for shallow water
- Optimal setting: inversion by hand, no factorization/iterative method
- Computational cost quadratic (cannot be improved further)

With varying bathymetry

- Hydrostatic reconstruction requires iterative strategy
- Positivity and entropy inequality hold under CFL

Advantageous framework for numerical analysis, but costly in practice

Conclusion and perspectives

Perspectives

- Improve convergence proof
- 2D version of implicit scheme
- Increase order of accuracy (iterative only)

Application in oceanography

- Coarse resolution $\Rightarrow$ dissipation D_i very large
- Improve hydrostatic reconstruction by also reconstructing velocity u
- Make $|D_i|$ smaller near Bernoulli equilibrium

Conclusion and perspectives

Perspectives

- Improve convergence proof
- 2D version of implicit scheme
- Increase order of accuracy (iterative only)

Application in oceanography

- Coarse resolution $\Rightarrow$ dissipation D_i very large
- Improve hydrostatic reconstruction by also reconstructing velocity u
- Make $|D_i|$ smaller near Bernoulli equilibrium

Thank you for your attention!