Well-Balanced Positivity Preserving Central-Upwind Scheme for the Shallow Water System with Friction Term

Alexander Kurganov

Tulane University
Mathematics Department

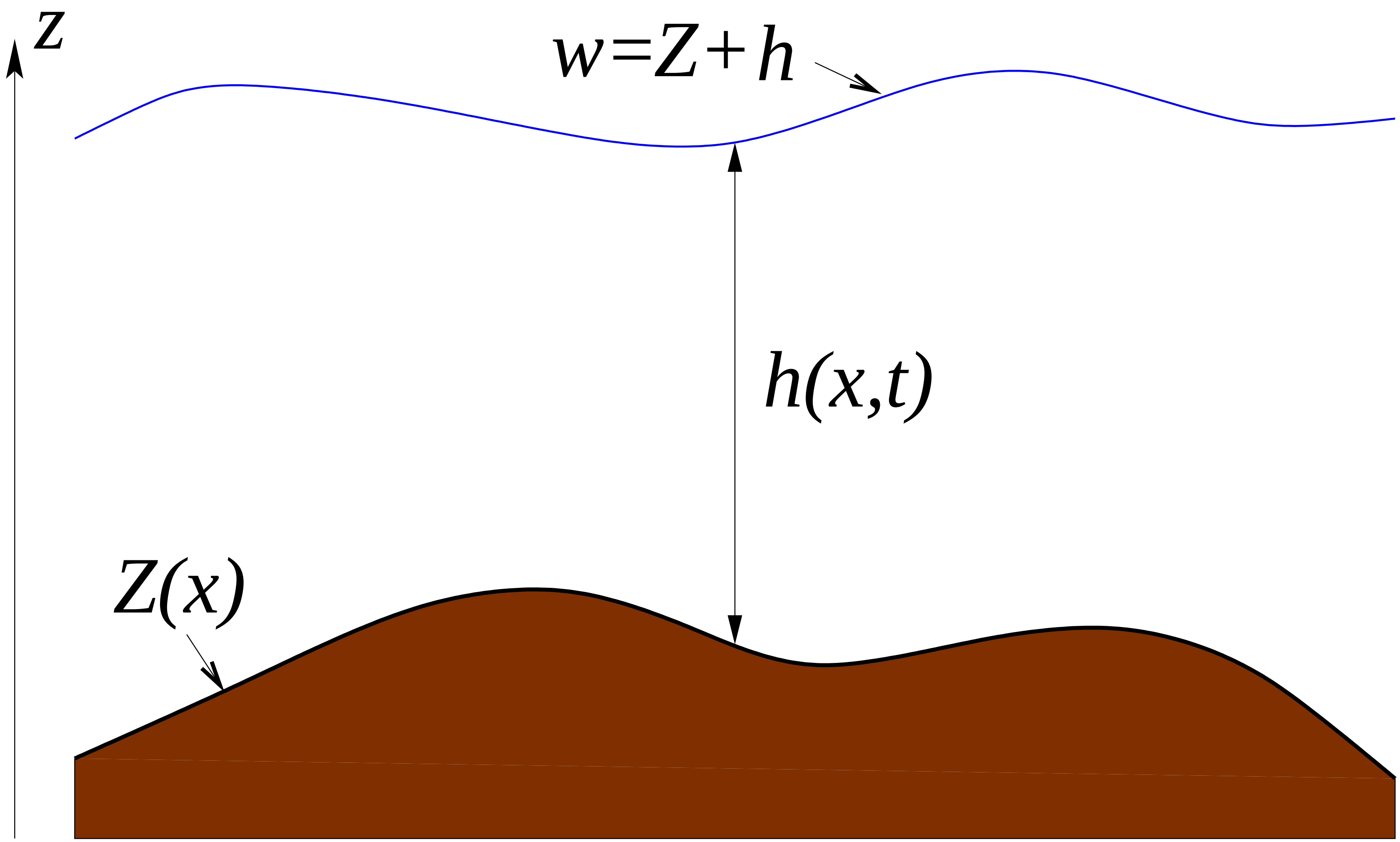
www.math.tulane.edu/∼kurganov

Alina Chertock, North Carolina State University, USA

Shumo Cui, Temple University, USA

Tong Wu, Tulane University, USA

Supported by NSF and ONR


z
w=Z+h
h(x,t)
Z(x)

1-D Saint-Venant System

$$\begin{cases} h_t + q_x = 0 \\ q_t + \left(hu^2 + \frac{g}{2}h^2\right)_x = -ghZ_x \end{cases}$$

This is a system of hyperbolic balance laws

$$\boldsymbol{U}_t + \boldsymbol{F}(\boldsymbol{U}, Z)_x = \boldsymbol{S}(\boldsymbol{U}, Z), \quad \boldsymbol{U} := (h, q)$$

h: depth

u: velocity

$q := hu$: discharge

Z: bottom topography

g: gravitational constant

Saint-Venant System — Numerical Challenges

$$\begin{cases} h_t + q_x = 0 \\ q_t + \left(hu^2 + \frac{g}{2}h^2\right)_x = -ghZ_x \end{cases}$$

- Steady-state solutions:

$$q = \text{Const}, \quad \frac{u^2}{2} + g(h + Z) = \text{Const}$$

- "Lake at rest" steady-state solutions:

$$u = 0, \quad h + Z = \text{Const}$$

- Dry ($h = 0$) or near dry ($h \sim 0$) states

Finite-Volume Methods

1-D System: $\boldsymbol{U}_t + \boldsymbol{F}(\boldsymbol{U})_x = \mathbf{0}$

$$\overline{\boldsymbol{U}}_j(t) \approx \frac{1}{\Delta x} \int\limits_{C_j} \boldsymbol{U}(x,t)\,dx$$

: cell averages over $C_j := (x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})$

This solution is approximated by a piecewise polynomial (conservative, high-order accurate, non-oscillatory) reconstruction:

$$\widetilde{\boldsymbol{U}}(x) = \boldsymbol{P}_j(x) \quad \text{for } x \in C_j$$

Second-order schemes employ piecewise linear reconstructions:

$$\widetilde{\boldsymbol{U}}(x) = \overline{\boldsymbol{U}}_j + (\boldsymbol{U}_x)_j(x - x_j) \quad \text{for } x \in C_j$$

For example,

$$\boxed{(\boldsymbol{U}_x)_j = \text{minmod}\left(\theta\frac{\overline{\boldsymbol{U}}_j - \overline{\boldsymbol{U}}_{j-1}}{\Delta x}, \frac{\overline{\boldsymbol{U}}_{j+1} - \overline{\boldsymbol{U}}_{j-1}}{2\Delta x}, \theta\frac{\overline{\boldsymbol{U}}_{j+1} - \overline{\boldsymbol{U}}_j}{\Delta x}\right)} \quad \theta \in [1, 2]$$

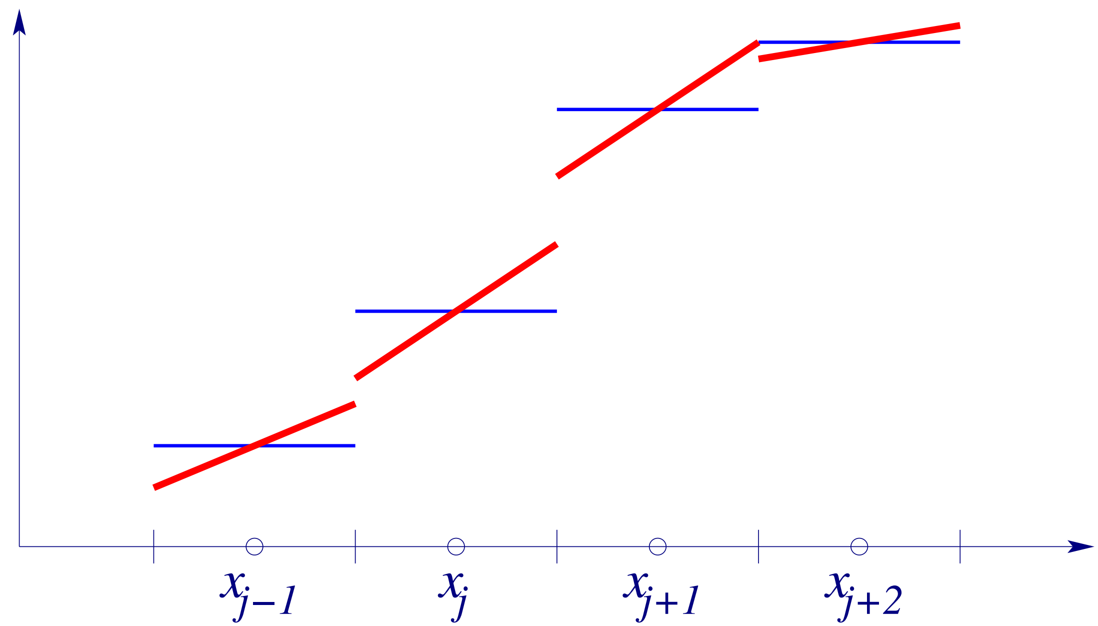
where the minmod function is defined as:

$$\boxed{\text{minmod}(z_1, z_2, ...) := \begin{cases} \min_j\{z_j\}, & \text{if} \quad z_j > 0 \quad \forall j, \\ \max_j\{z_j\}, & \text{if} \quad z_j < 0 \quad \forall j, \\ 0, & \quad \text{otherwise.} \end{cases}}$$

The reconstructed point values at cell interfaces are:

$$\boldsymbol{U}^-_{j+\frac{1}{2}} := \boldsymbol{P}_j(x_{j+\frac{1}{2}}) = \overline{\boldsymbol{U}}_j + \frac{\Delta x}{2}(\boldsymbol{U}_x)_j$$

$$\boldsymbol{U}^+_{j+\frac{1}{2}} := \boldsymbol{P}_{j+1}(x_{j+\frac{1}{2}}) = \overline{\boldsymbol{U}}_{j+1} - \frac{\Delta x}{2}(\boldsymbol{U}_x)_{j+1}$$



The discontinuities appearing at the reconstruction step at the interface points $\{x_{j+\frac{1}{2}}\}$ propagate at finite speeds estimated by:

$$a^{+}_{j+\frac{1}{2}} := \max\left\{\lambda_N\Big(A(\boldsymbol{U}^{-}_{j+\frac{1}{2}})\Big), \lambda_N\Big(A(\boldsymbol{U}^{+}_{j+\frac{1}{2}})\Big), 0\right\}$$

$$a^{-}_{j+\frac{1}{2}} := \min\left\{\lambda_1\Big(A(\boldsymbol{U}^{-}_{j+\frac{1}{2}})\Big), \lambda_1\Big(A(\boldsymbol{U}^{+}_{j+\frac{1}{2}})\Big), 0\right\}$$

$\lambda_1 < \lambda_2 < \ldots < \lambda_N$: N eigenvalues of the Jacobian $A(\boldsymbol{U}) := \dfrac{\partial \boldsymbol{F}}{\partial \boldsymbol{U}}$

Central-Upwind Schemes

Godunov-type central schemes with a built-in upwind nature

[Kurganov, Tadmor; 2000]

[Kurganov, Petrova; 2000, 2001]

[Kurganov, Noelle, Petrova; 2001]

[Kurganov, Lin; 2007]

1-D Semi-Discrete Central-Upwind Scheme

$$\frac{d}{dt}\overline{\boldsymbol{U}}_j(t) = -\frac{\boldsymbol{H}_{j+\frac{1}{2}}(t) - \boldsymbol{H}_{j-\frac{1}{2}}(t)}{\Delta x}$$

The central-upwind numerical flux is:

$$\boldsymbol{H}_{j+\frac{1}{2}} = \frac{a^+_{j+\frac{1}{2}}\boldsymbol{F}(\boldsymbol{U}^-_{j+\frac{1}{2}}) - a^-_{j+\frac{1}{2}}\boldsymbol{F}(\boldsymbol{U}^+_{j+\frac{1}{2}})}{a^+_{j+\frac{1}{2}} - a^-_{j+\frac{1}{2}}} + a^+_{j+\frac{1}{2}}a^-_{j+\frac{1}{2}}\left[\frac{\boldsymbol{U}^+_{j+\frac{1}{2}} - \boldsymbol{U}^-_{j+\frac{1}{2}}}{a^+_{j+\frac{1}{2}} - a^-_{j+\frac{1}{2}}} - \boxed{\boxed{\boldsymbol{d}_{j+\frac{1}{2}}}}\right]$$

The built-in "anti-diffusion" term is:

$$\boldsymbol{d}_{j+\frac{1}{2}} = \text{minmod}\left(\frac{\boldsymbol{U}^+_{j+\frac{1}{2}} - \boldsymbol{U}^*_{j+\frac{1}{2}}}{a^+_{j+\frac{1}{2}} - a^-_{j+\frac{1}{2}}}, \frac{\boldsymbol{U}^*_{j+\frac{1}{2}} - \boldsymbol{U}^-_{j+\frac{1}{2}}}{a^+_{j+\frac{1}{2}} - a^-_{j+\frac{1}{2}}}\right)$$

The intermediate values $\boldsymbol{U}^*_{j+\frac{1}{2}}$ are:

$$\boldsymbol{U}^*_{j+\frac{1}{2}} = \frac{a^+_{j+\frac{1}{2}}\boldsymbol{U}^+_{j+\frac{1}{2}} - a^-_{j+\frac{1}{2}}\boldsymbol{U}^-_{j+\frac{1}{2}} - \left\{\boldsymbol{F}(\boldsymbol{U}^+_{j+\frac{1}{2}}) - \boldsymbol{F}(\boldsymbol{U}^-_{j+\frac{1}{2}})\right\}}{a^+_{j+\frac{1}{2}} - a^-_{j+\frac{1}{2}}}$$

Remarks

1. $\boldsymbol{d}_{j+\frac{1}{2}} \equiv 0$ corresponds to the original central-upwind scheme from [Kurganov, Noelle, Petrova; 2001]

$\boldsymbol{d}_{j+\frac{1}{2}} \equiv 0$ and $a^{+}_{j+\frac{1}{2}} \equiv -a^{-}_{j+\frac{1}{2}}$ correspond to the scheme from [Kurganov, Tadmor; 2000]

2. For the system of balance laws

$$\boldsymbol{U}_t + \boldsymbol{F}(\boldsymbol{U})_x = \boldsymbol{S}$$

the central-upwind scheme is:

$$\frac{d}{dt}\overline{\boldsymbol{U}}_j(t) = -\frac{\boldsymbol{H}_{j+\frac{1}{2}}(t) - \boldsymbol{H}_{j-\frac{1}{2}}(t)}{\Delta x} + \boxed{\overline{\boldsymbol{S}}_j(t)}$$

where

$$\overline{\boldsymbol{S}}_j(t) \approx \frac{1}{\Delta x}\int\limits_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \boldsymbol{S}(x,t)\,dx$$

2-D Semi-Discrete Central-Upwind Scheme

Rectangular Grid

[Kurganov, Petrova; 2001]

[Kurganov, Noelle, Petrova; 2001]

[Kurganov, Tadmor; 2002]

[Kurganov, Lin; 2007]

Triangular Grid

[Kurganov, Petrova; 2005]

Well-Balanced Positivity Preserving Central-Upwind Scheme

[Kurganov, Petrova; 2007]

- $w = h + Z$: water surface $\Longrightarrow$ "Lake at rest" states: $\boxed{q \equiv 0, \ \ w \equiv \text{Const}}$

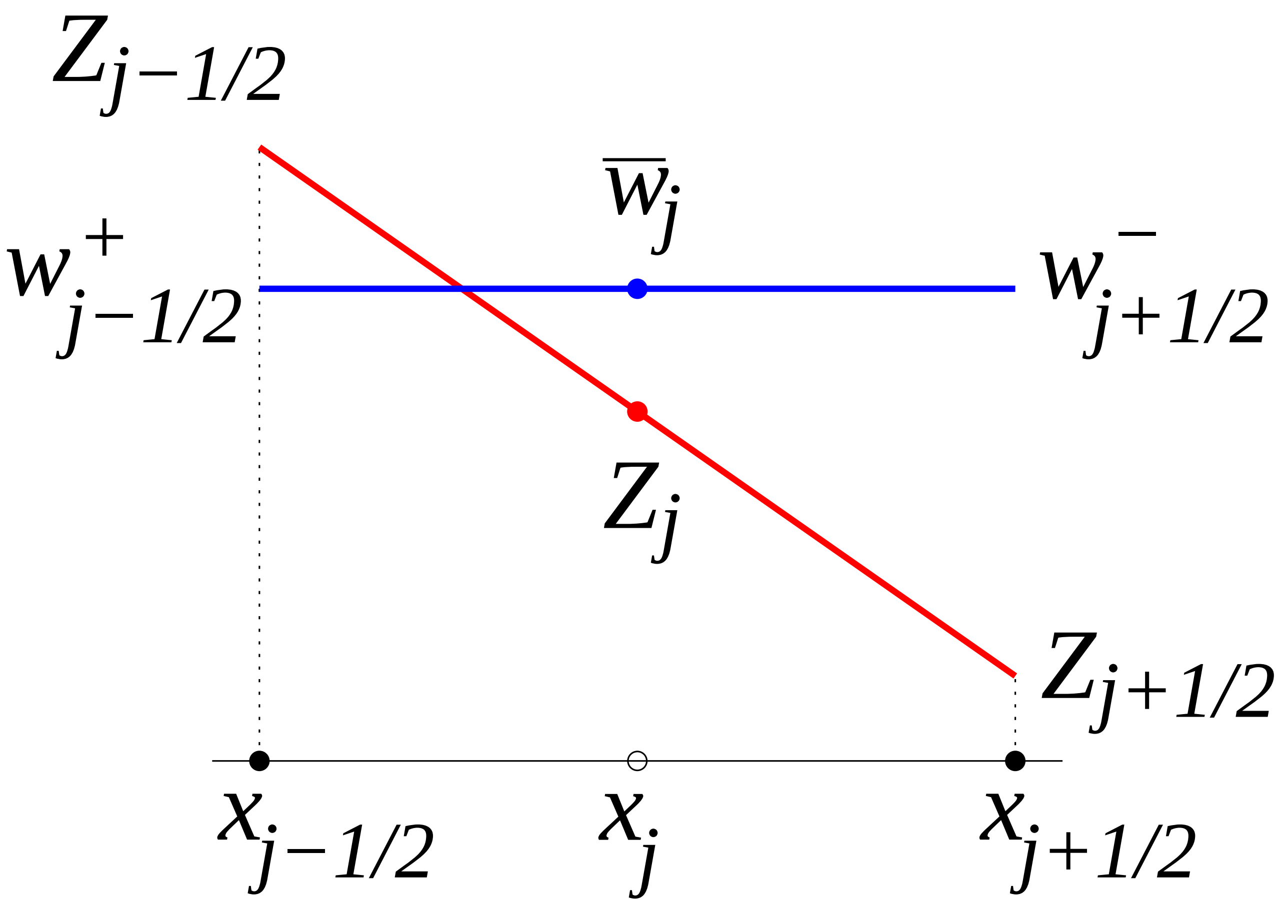
$\Longrightarrow$ Reconstruct the equilibrium variables w and q rather than h and q

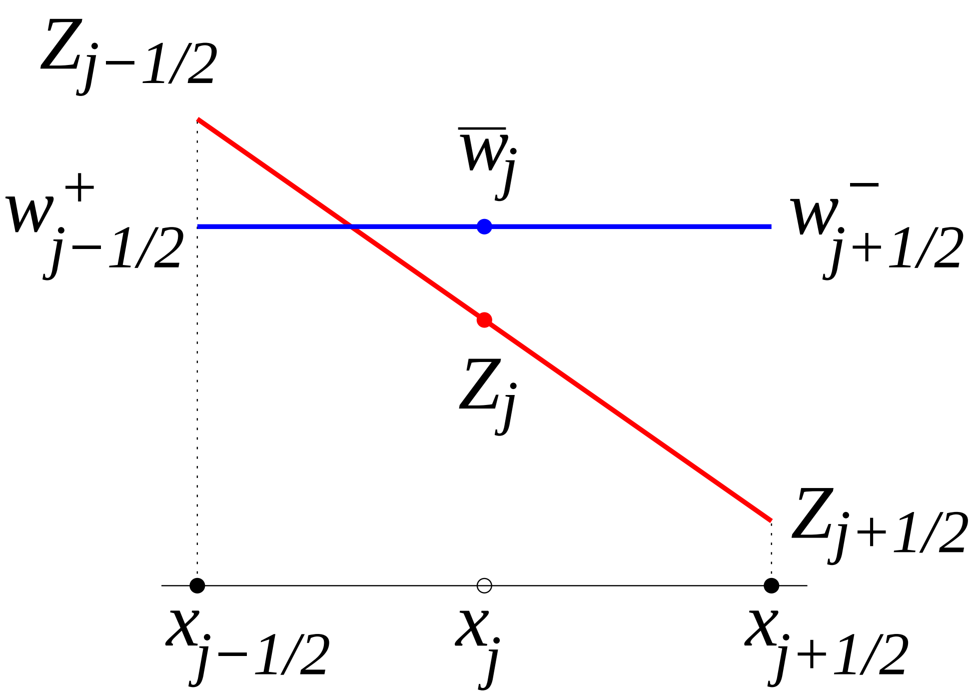
- Use the well-balanced quadrature

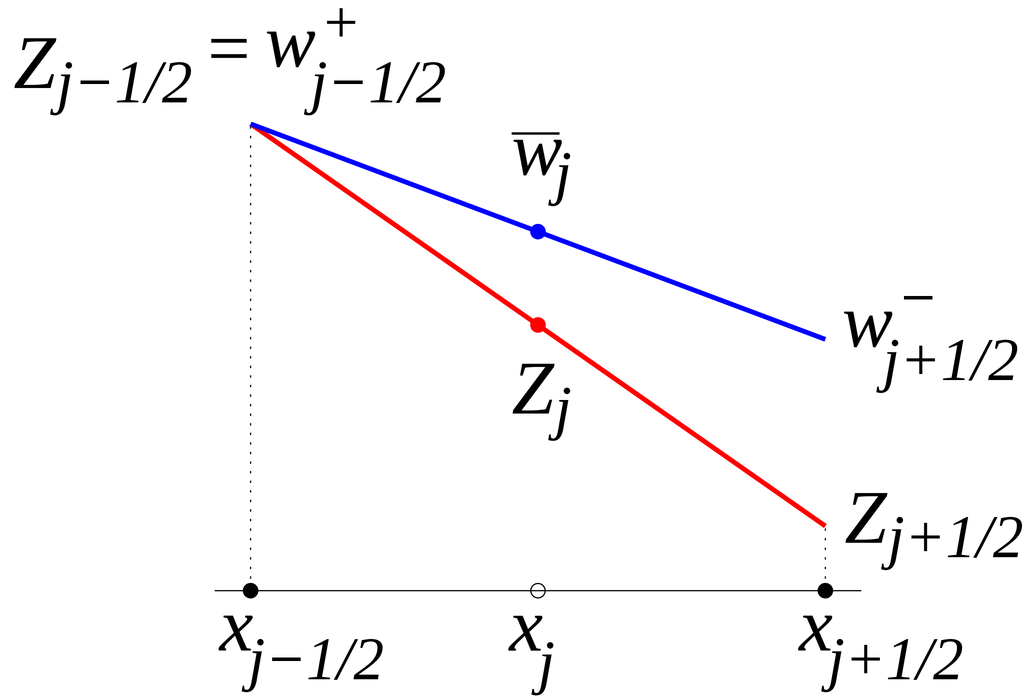
$$\int\limits_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} hZ_x\,dx = \left(\overline{w}_j - \frac{Z(x_{j+\frac{1}{2}}) + Z(x_{j-\frac{1}{2}})}{2}\right) \cdot \Big(Z(x_{j+\frac{1}{2}}) - Z(x_{j-\frac{1}{2}})\Big)$$

- Make positivity preserving correction of the reconstruction of w

$$h^{\pm}_{j+\frac{1}{2}} = \boxed{w^{\pm}_{j+\frac{1}{2}} - Z_{j+\frac{1}{2}}}$$




$Z_{j-1/2}$
$\overline{w}_j$
$w^{+}_{j-1/2}$
$w^{-}_{j+1/2}$
Z_j
$Z_{j+1/2}$
$x_{j-1/2}$
x_j
$x_{j+1/2}$


$Z_{j-1/2} = w^{+}_{j-1/2}$
$\overline{w}_j$
$w^{-}_{j+1/2}$
Z_j
$Z_{j+1/2}$
$x_{j-1/2}$
x_j
$x_{j+1/2}$

- Desingularize the computation of $u = \dfrac{q}{h}$ for small $h < \varepsilon$

– Simplest:

$$u = \begin{cases} \dfrac{q}{h}, & \text{if } h \geq \varepsilon \\ 0, & \text{if } h < \varepsilon \end{cases}$$

– More sophisticated (smoother transition for small h):

$$u = \frac{2\,h\,q}{h^2 + \max(h^2, \varepsilon^2)} \quad \text{or} \quad u = \frac{\sqrt{2}\,h\,q}{\sqrt{h^4 + \max(h^4, \varepsilon^4)}}$$

Remark: For consistency, one has to recompute the discharge:

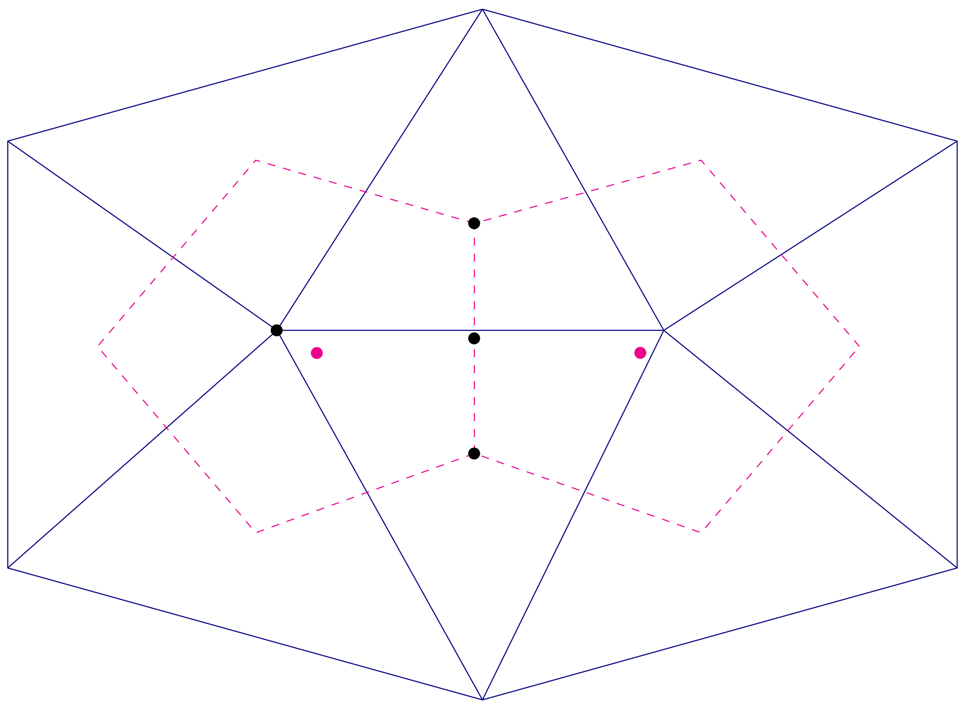
$$q = h \cdot u$$

Central-Upwind Schemes for the 2-D Saint-Venant System

Cartesian Grid: [Kurganov, Levy, 2002], [Kurganov, Petrova; 2007]

Triangular Grid: [Bryson, Epshteyn, Kurganov, Petrova; 2011]

Polygonal Cell-Vertex Mesh:
[Beljadid, Mohammadian, Kurganov; preprint]



Shallow Water System with Friction Terms

[Chertock, Cui, Kurganov, Wu; 2015]

$$\begin{cases} h_t + q_x = 0 \\ q_t + \left(hu^2 + \frac{g}{2}h^2\right)_x = -ghZ_x - g\frac{n^2}{h^{1/3}}|u|u \end{cases}$$

n: Manning coefficient

Special Steady-State Solutions

$$q \equiv \text{Const}, \quad h \equiv \text{Const}, \quad Z_x \equiv \text{Const}$$

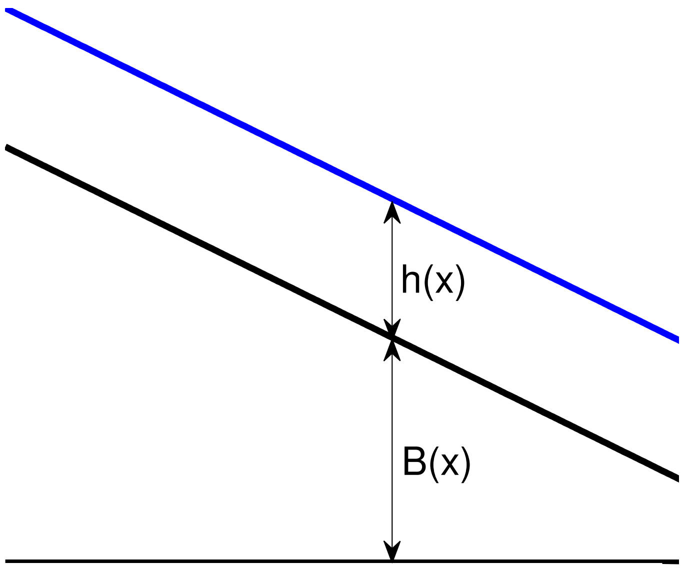
correspond to the situation when the water flows over a slanted infinitely long surface with a constant slope.

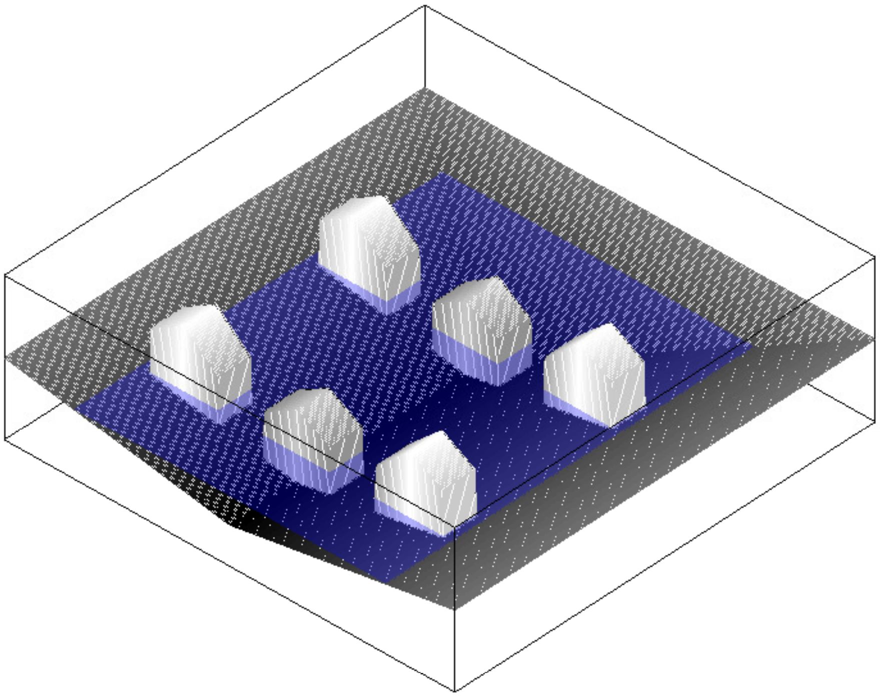
A straightforward midpoint discretization of the friction term leads to the well-balanced positivity preserving **semi-discrete** central-upwind scheme

The bottom setting in numerical examples:

Left: 1-D steady state

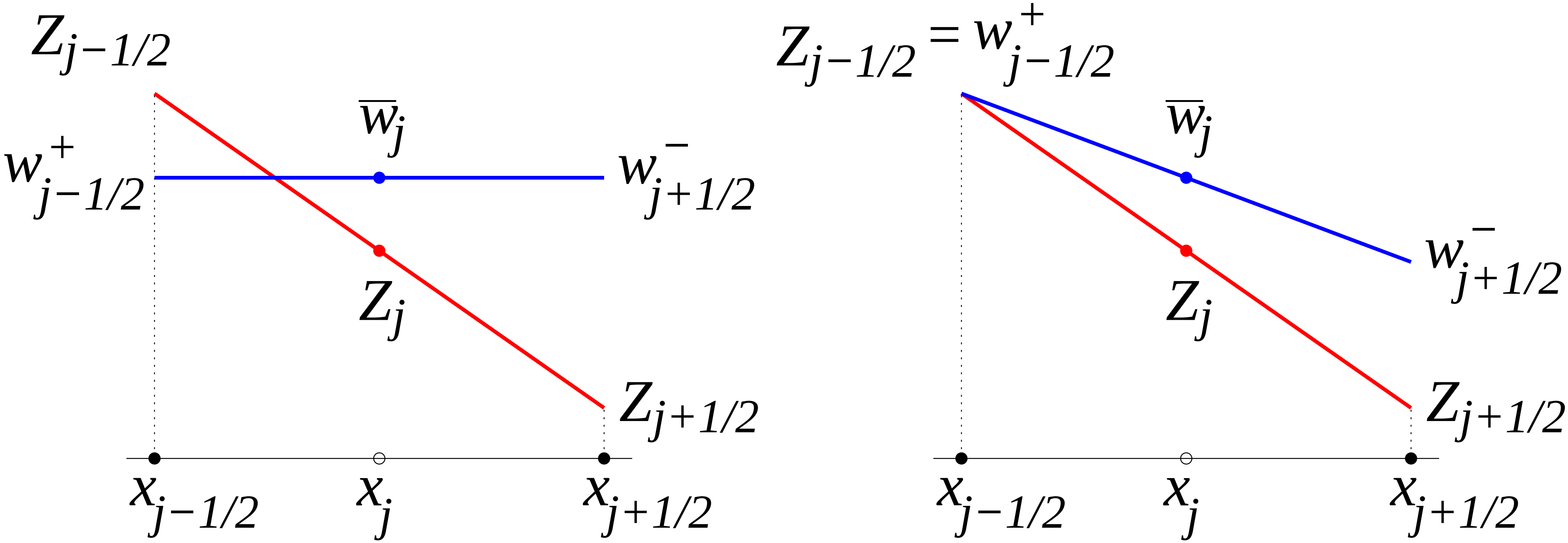
Right: a case of urban draining with obstacles like houses



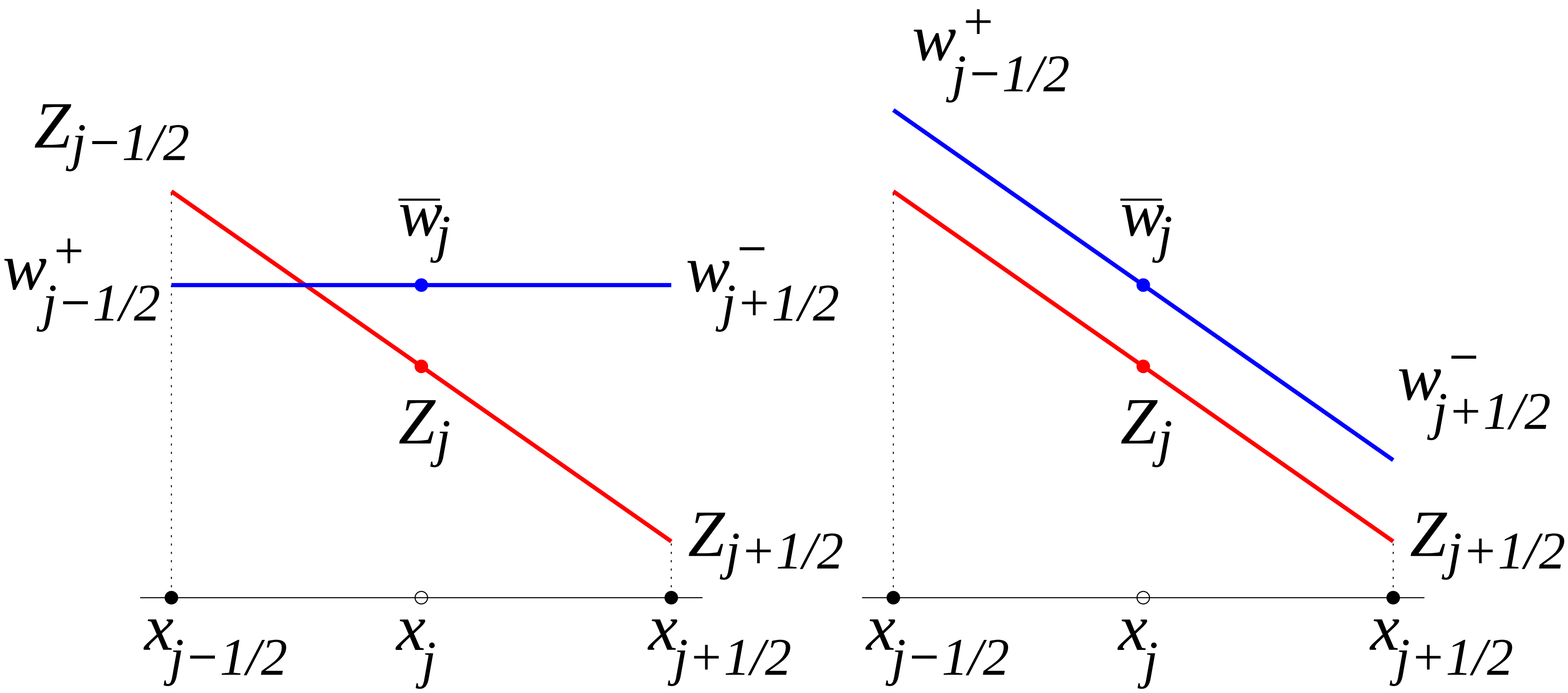


Modified Positivity Correction

Instead of

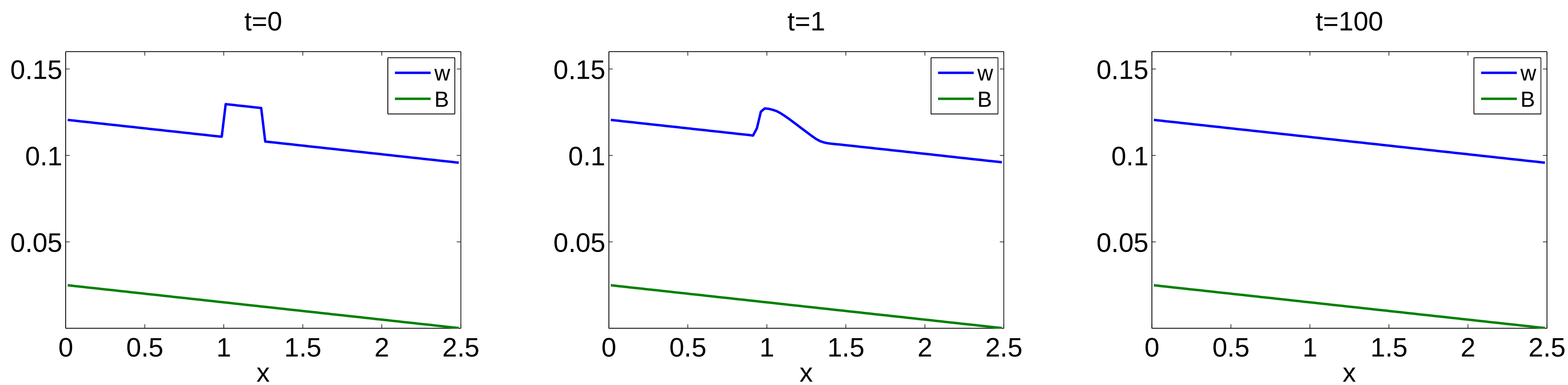


we do

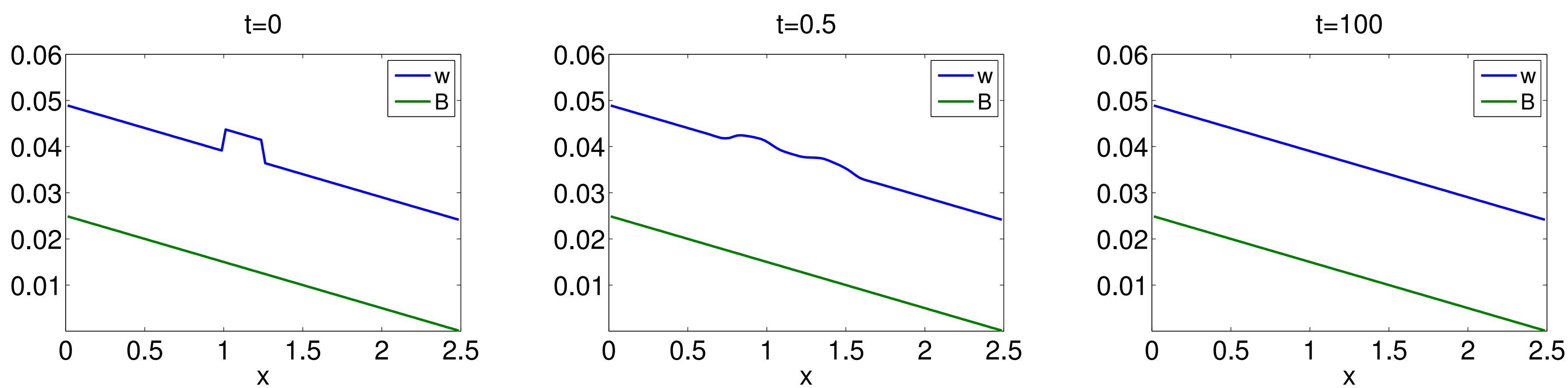

$Z_{j-1/2}$
$w^+_{j-1/2}$
$\overline{w}_j$
$w^-_{j+1/2}$
Z_j
$Z_{j+1/2}$
$x_{j-1/2}$
x_j
$x_{j+1/2}$
$w^+_{j-1/2}$
$\overline{w}_j$
$w^-_{j+1/2}$
Z_j
$Z_{j+1/2}$
$x_{j-1/2}$
x_j
$x_{j+1/2}$

Example — Small Perturbation of a Steady Flow Over a Slanted Surface ($R \equiv 0$)

$$h(x,0) = h_n + \begin{cases} 0.2h_n, & 1 \le x \le 1.25, \\ 0, & \text{otherwise}, \end{cases} \qquad q(x,0) \equiv q_0$$



Supercritical case

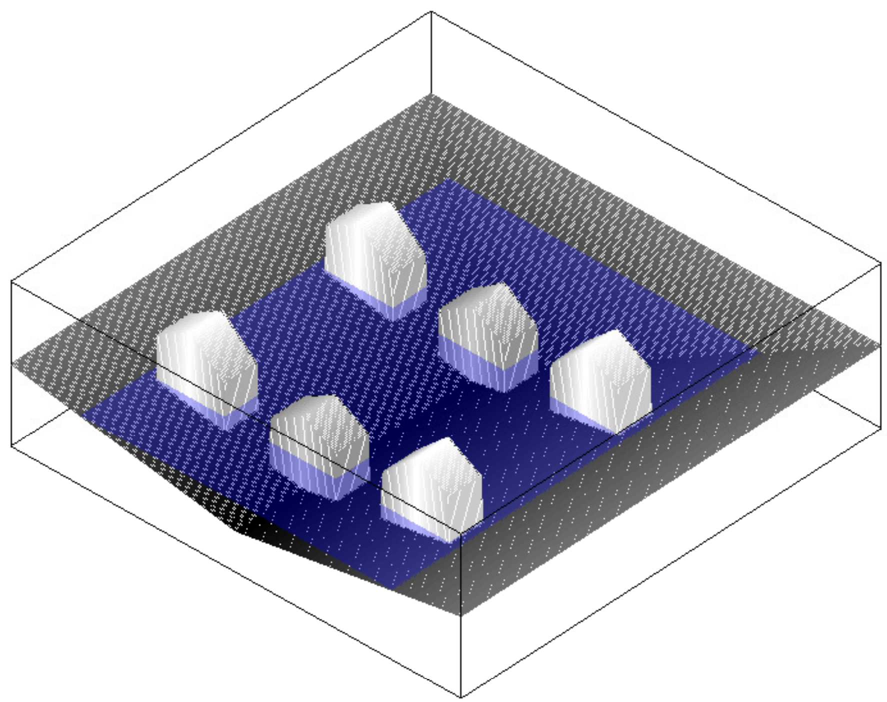


Subcritical case

Two-Dimensional Examples

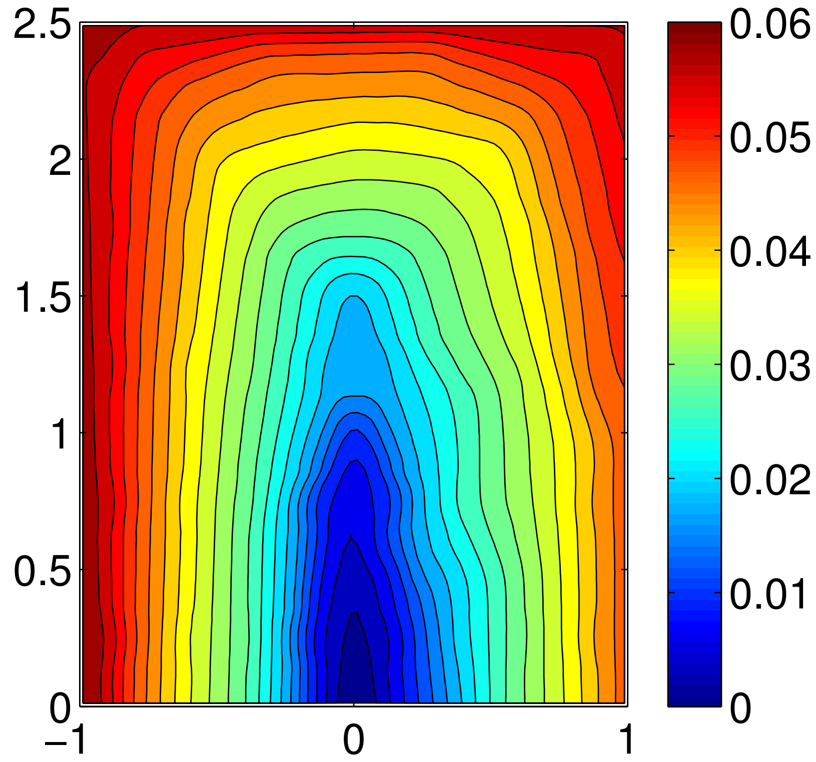
Example — Rainfall-Runoff Over An Urban Area

We consider a rainfall-runoff situation, which occurs over a 2-D surface containing houses as outlined in



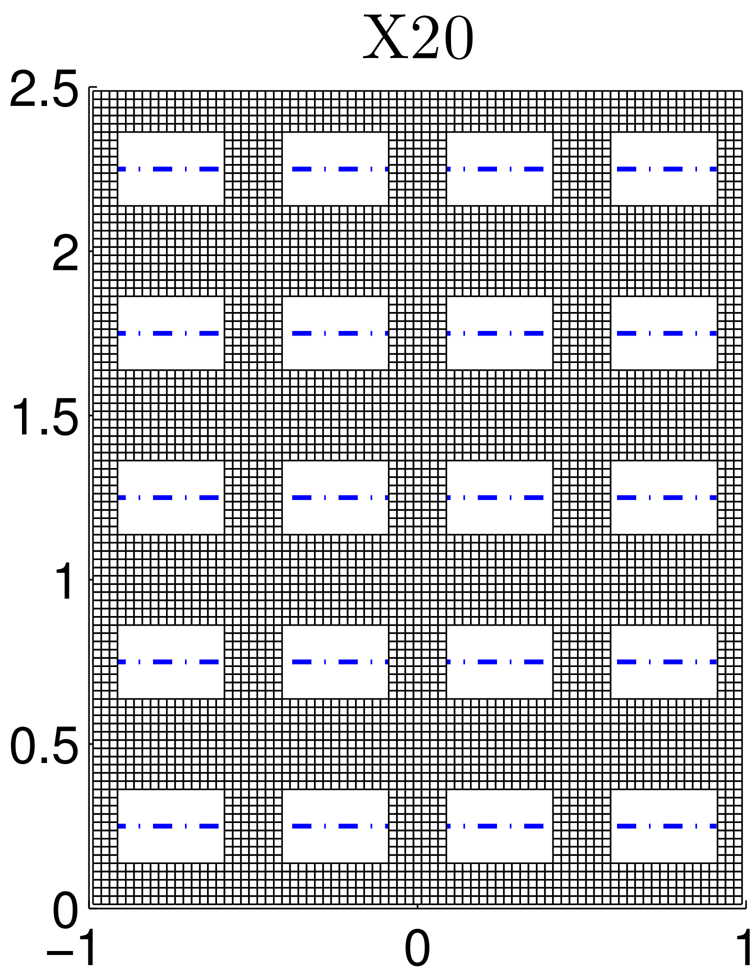
The setting corresponds to the laboratory experiment reported in [Cea, Garrido, Puertas; 2010].

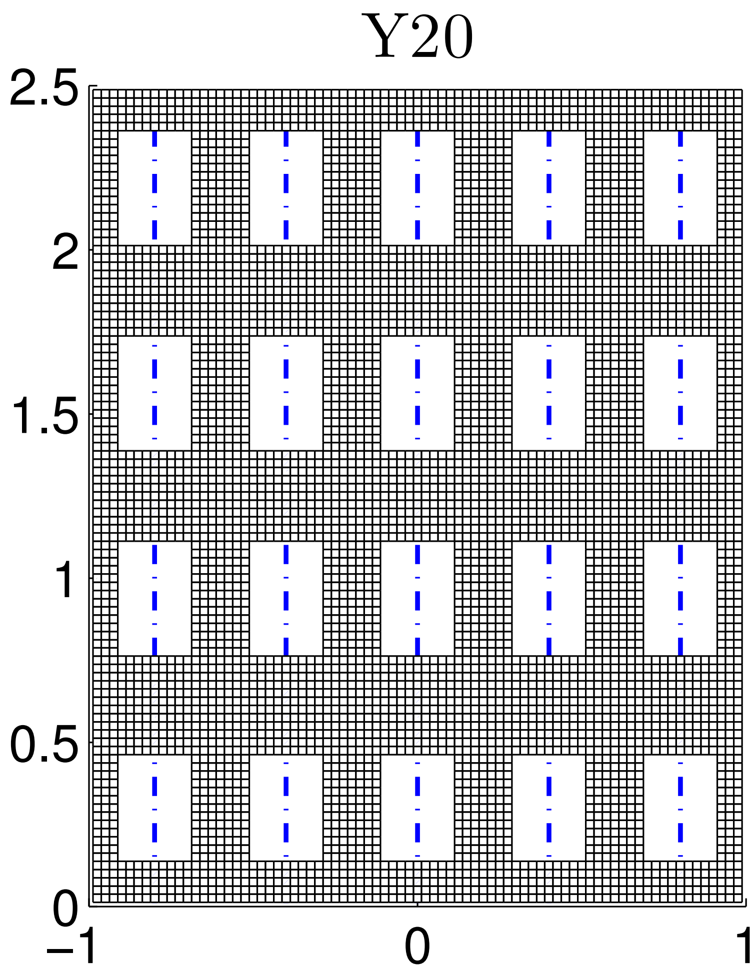
The surface structure is shown in

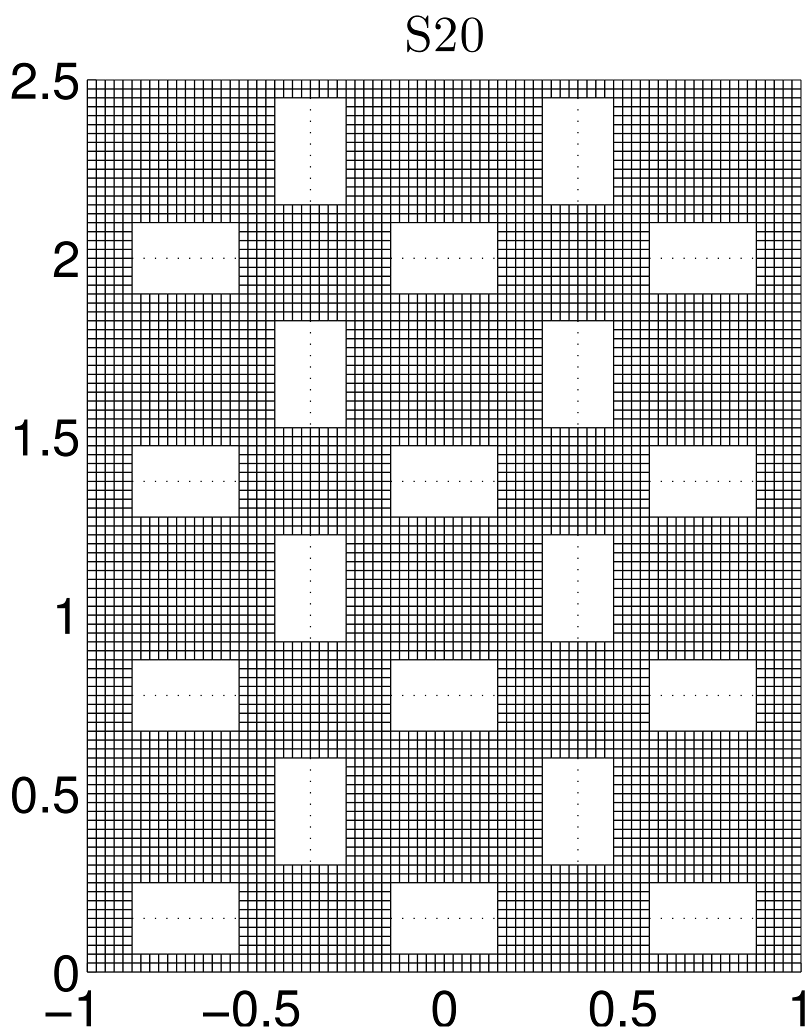


The precise data was provided by Dr. Luis Cea

The experiment was built to mimic an urban area within the laboratory simulator of size $2m \times 2.5m$. To model urban buildings, several blocks were placed onto the surface according to three different geometries:




Y20
2.5
2
1.5
1
0.5
0
−1
0
1



Notice that across the walls of the houses the bottom topography is discontinuous and thus the bilinear interpolant $\tilde{Z}$ has very sharp gradients there

We set the almost dry initial conditions:

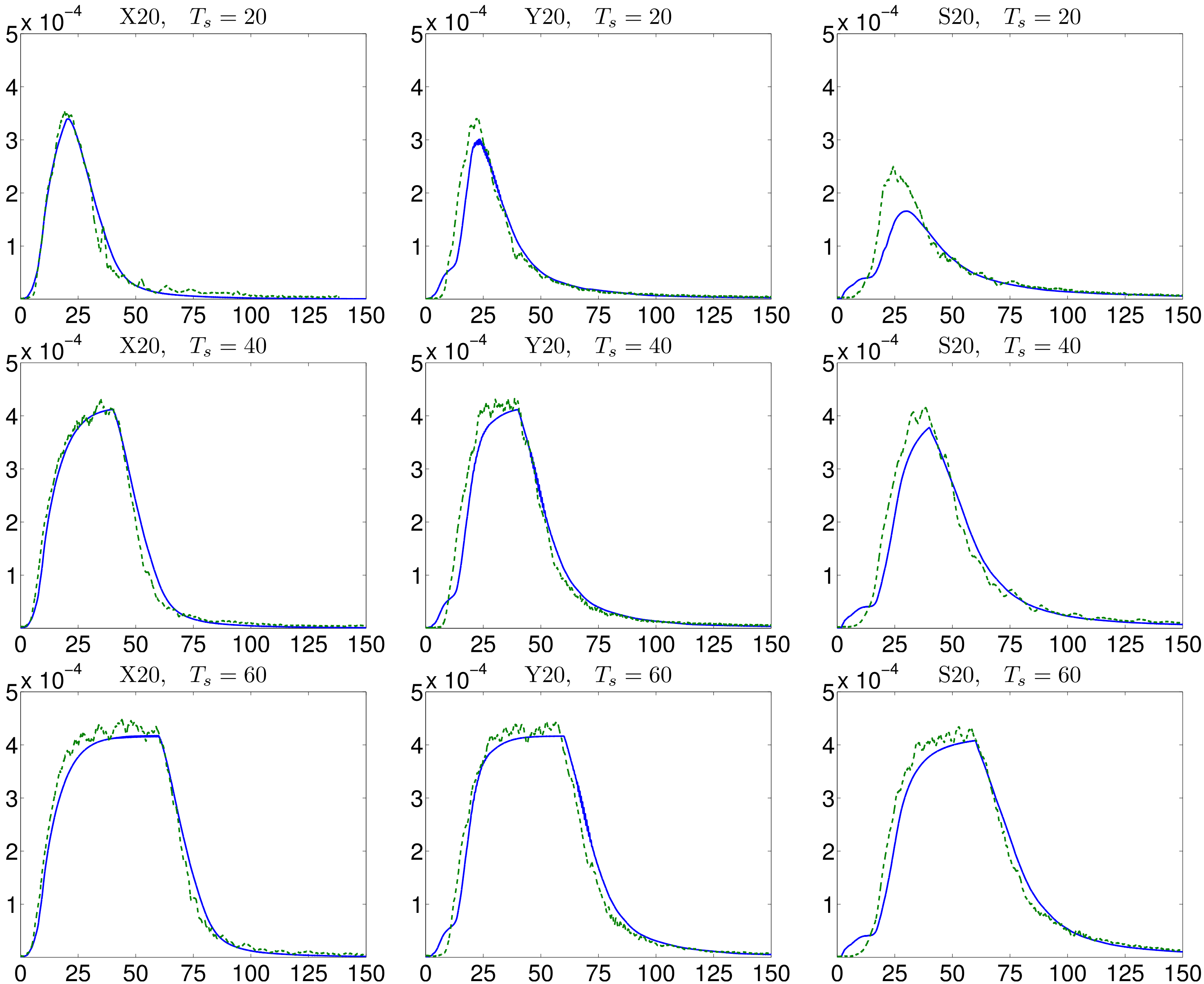
$$h(x,y,0) \equiv 10^{-8}, \quad u(x,y,0) \equiv v(x,y,0) \equiv 0$$

The rain of a constant intensity starts falling at time $t = 0$ and stops at $t = T_s$:

$$R(x,y,t) = \begin{cases} \dfrac{1}{12000}, & 0 \le t \le T_s \\ 0, & \text{otherwise} \end{cases}$$

We take $T_s = 20$, 40 or 60

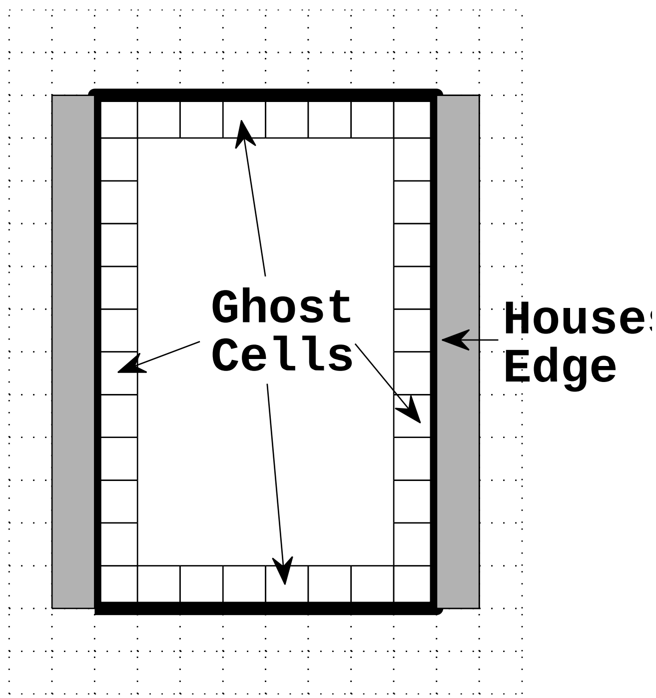
At the lower part of boundary, the total outlet discharge is recorded in the laboratory experiments at different time moments and then it is compared with the computed values of $\sum\limits_{j=1}^{N_x} \left(H^y_{j,1/2}\right)^{(1)}$

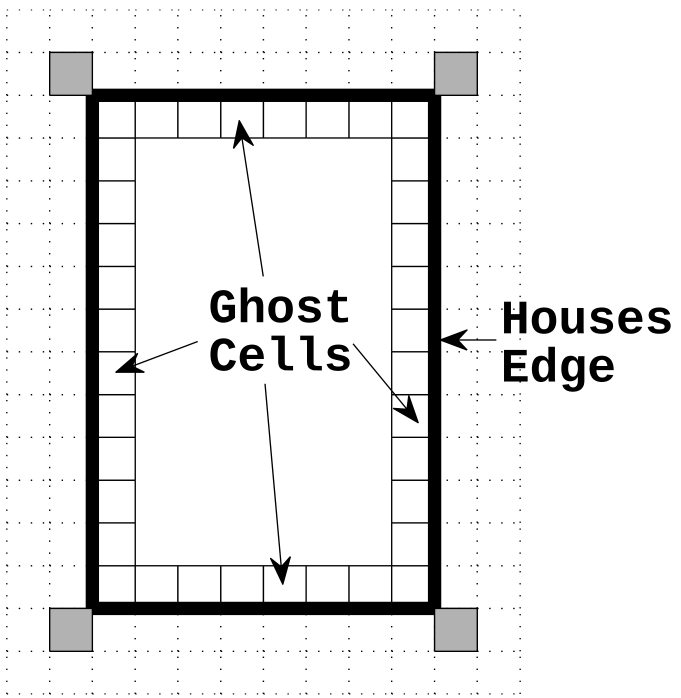


X20, $T_s = 20$
Y20, $T_s = 20$
S20, $T_s = 20$
X20, $T_s = 40$
Y20, $T_s = 40$
S20, $T_s = 40$
X20, $T_s = 60$
Y20, $T_s = 60$
S20, $T_s = 60$
x 10^{-4}
0 25 50 75 100 125 150
1 2 3 4 5

Example — Rainfall Runoff Over An Urban Area (REVISED)

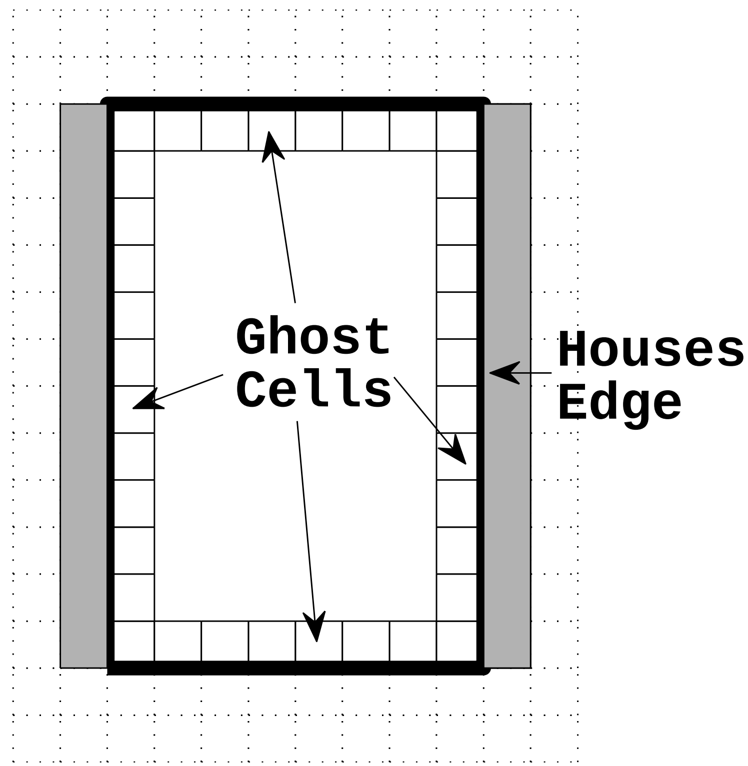
First, we remove the houses from the computational domain which becomes a punctured rectangle. Each of the holes is depicted in



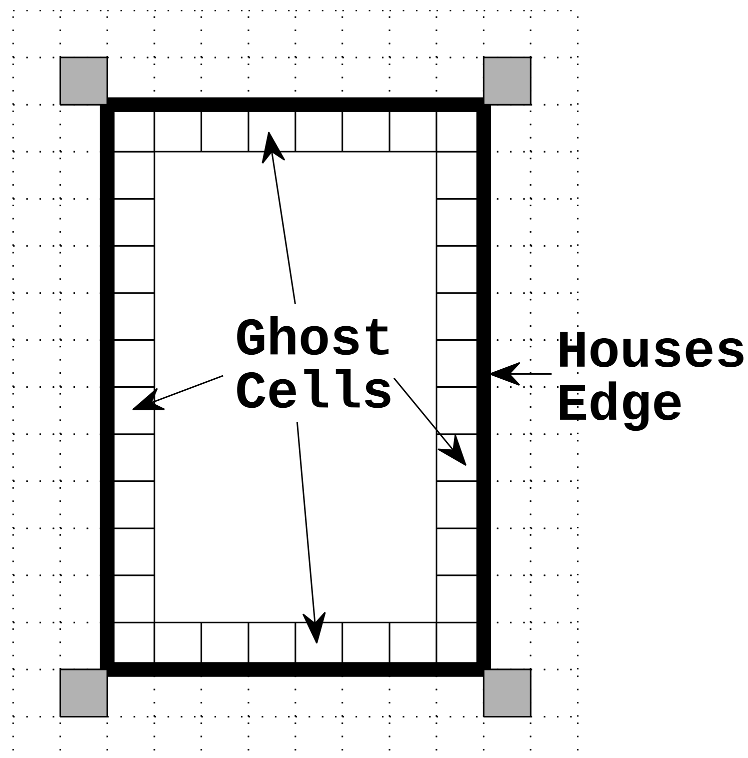


The house walls become the internal boundary, which is numerically treated using a solid wall ghost cell technique.

Second, we need to redistribute the rain water falling onto the roof so that it is placed inside the modified computational domain. In the laboratory experiment, the water falling on the house blocks streams down from the long (lower) edges and finally joins the surface water flow:



In reality, the gutter system is commonly used and the rain water streams down from the rain pipes typically located at the house corners:



In both cases, the building-roof rainfall is uniformly distributed on the shaded cells near and outside the building edges.

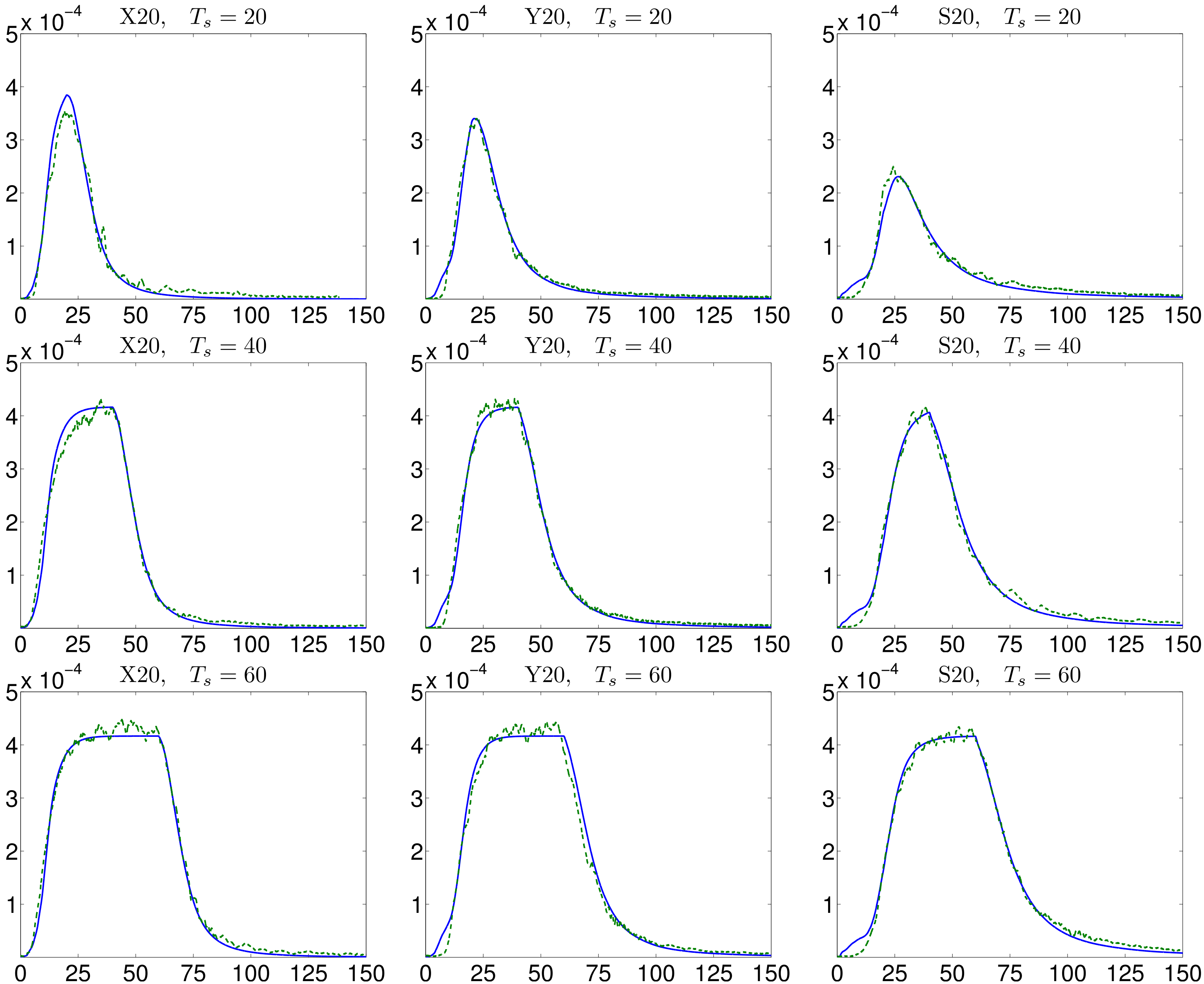
The modified rain source can then be written as follows:

$$\widehat{R}(x,y,t) = \begin{cases} \dfrac{1}{12000}\left(1+\dfrac{A_h}{A_s}\right), & \text{in the shaded cells} \\ \dfrac{1}{12000}, & \text{otherwise} \end{cases}$$

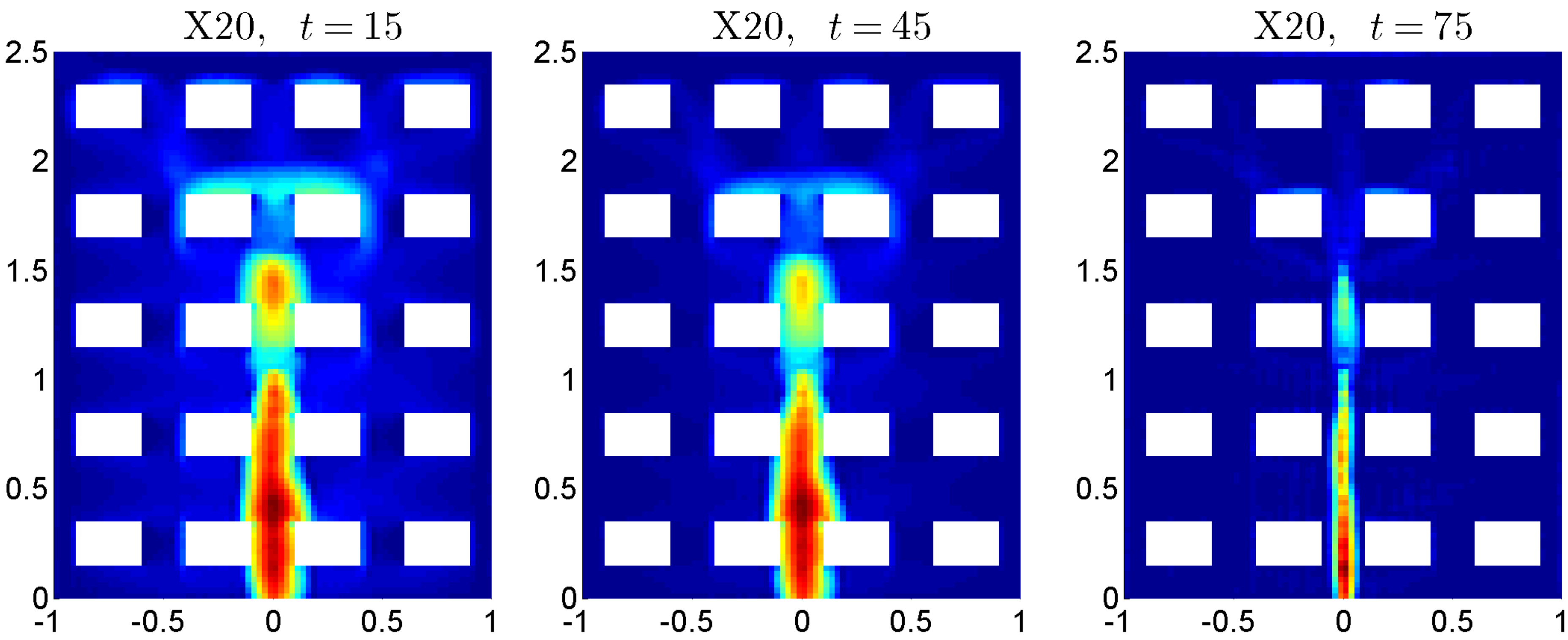
which is as before switched on only for $t \in [0, T_s]$

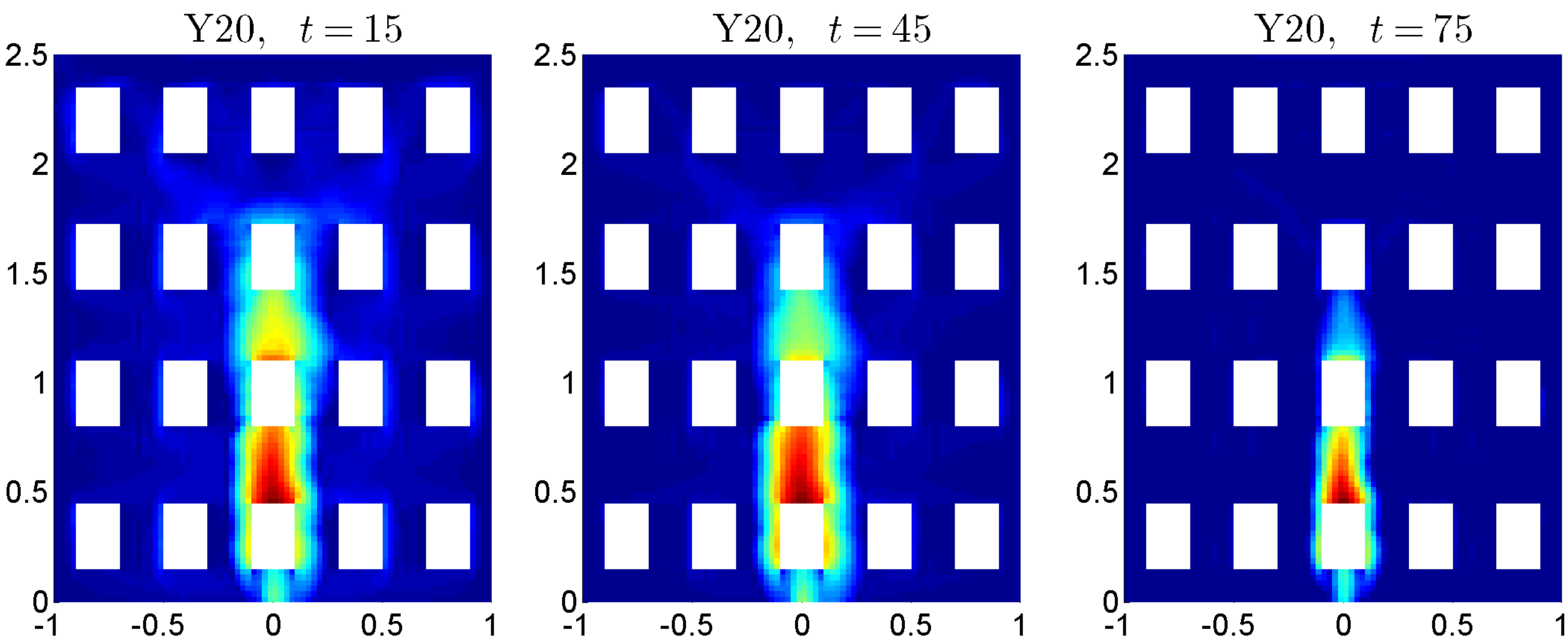
A_h: the area of the house

A_s: the area of the shaded region near that house

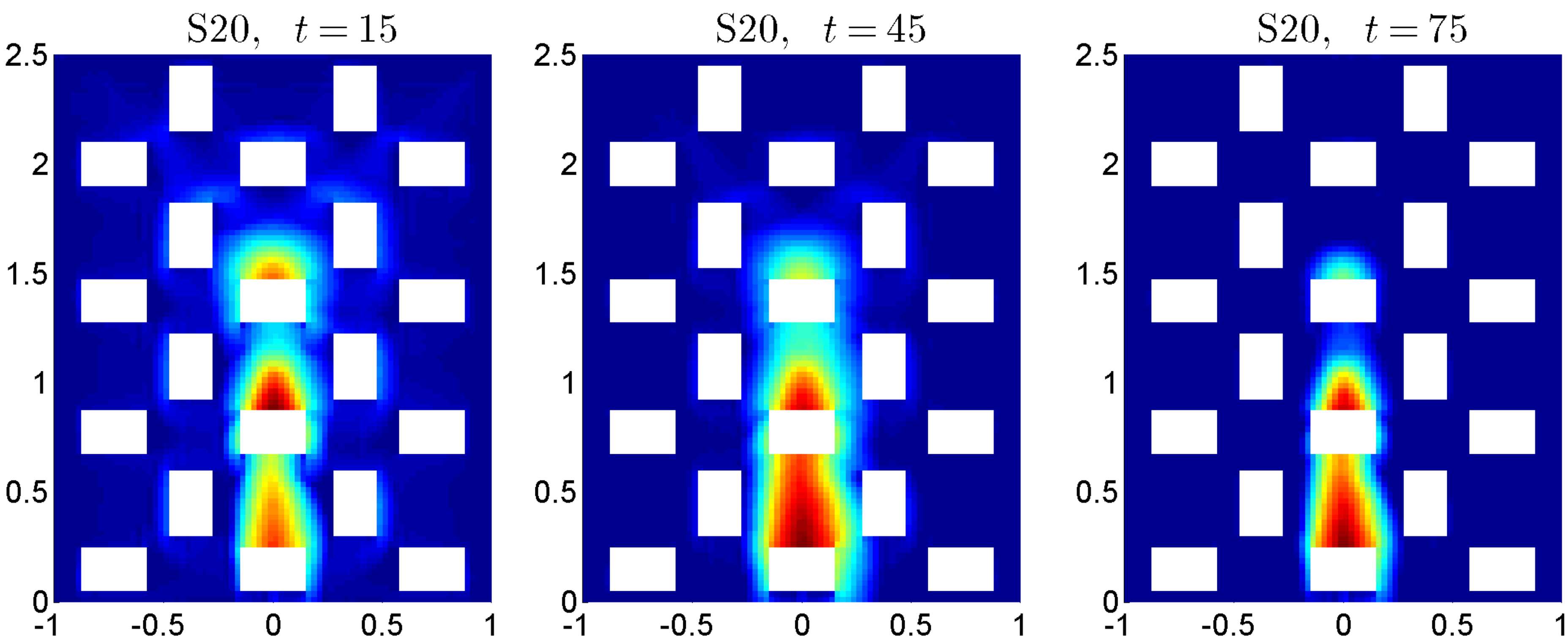

X20, $T_s = 20$
Y20, $T_s = 20$
S20, $T_s = 20$
X20, $T_s = 40$
Y20, $T_s = 40$
S20, $T_s = 40$
X20, $T_s = 60$
Y20, $T_s = 60$
S20, $T_s = 60$
x 10^{-4}
0 25 50 75 100 125 150
1 2 3 4 5

Water height snapshots for $T_s = 40$





Y20, $t = 15$
Y20, $t = 45$
Y20, $t = 75$
2.5
2
1.5
1
0.5
0
-1
-0.5
0
0.5
1


S20, $t = 15$
S20, $t = 45$
S20, $t = 75$
2.5
2
1.5
1
0.5
0
-1
-0.5
0
0.5
1

Steady State and Sign Preserving Semi-Implicit Runge-Kutta Methods for ODEs with Stiff Damping Term

[Chertock, Cui, Kurganov, Wu; 2015]

Consider

$$\boldsymbol{u}' = \boldsymbol{f}(\boldsymbol{u},t) + G(\boldsymbol{u},t)\boldsymbol{u}$$

$\boldsymbol{u}(t) \in \mathbb{R}^N$: unknown vector function

$\boldsymbol{f} : \mathbb{R}^N \to \mathbb{R}^N$: given vector field

$G : \mathbb{R}^{N\times N} \to \mathbb{R}^{N\times N}$: diagonal non-positive definite matrix representing a (stiff) damping term

Steady States: $\boldsymbol{u}(t) \equiv \widehat{\boldsymbol{u}}$ s.t. $\boldsymbol{f}(\widehat{\boldsymbol{u}},t) \equiv -G(\widehat{\boldsymbol{u}},t)\widehat{\boldsymbol{u}}$

Sign Preservation provided $\{\boldsymbol{u}(0) \ge 0, \quad \boldsymbol{f} \ge 0\}$ or $\{\boldsymbol{u}(0) \le 0, \quad \boldsymbol{f} \le 0\}$

Explicit vs. Implicit vs. Semi-Implicit Methods

For simplicity, consider a scalar ODE

$$u' = f(u,t) + g(u,t)u, \quad g(u,t) \leq 0$$

Example: **First-Order Explicit (Forward Euler) Method**

$$u^{n+1} = u^n + \Delta t\Big[f(u^n,t^n) + g(u^n,t^n)u^n\Big]$$

Example: **First-Order Implicit (Backward Euler) Method**

$$u^{n+1} = u^n + \Delta t\Big[f(u^{n+1},t^{n+1}) + g(u^{n+1},t^{n+1})u^{n+1}\Big]$$

Example: **First-Order Semi-Implicit Method**

$$u^{n+1} = u^n + \Delta t\Big[f(u^n,t^n) + g(u^n,t^n)u^{n+1}\Big]$$

Explicit m-stage SSP (TVD) RK Methods

[Shu; 1988] [Shu, Osher; 1988] [Gottlieb, Shu, Tadmor; 2001]

For simplicity, consider a scalar ODE

$$u' = f(u,t) + g(u,t)u, \quad g(u,t) \le 0$$

$f(u,t)$: nonstiff term, $g(u,t)u$: stiff damping term

A general explicit m-stage RK method is

$$\begin{aligned} u^{(0)} &= u^n \\ u^{(i)} &= \sum_{k=0}^{i-1} \alpha_{i,k} \left[u^{(k)} + \beta_{i,k} \Delta t (f^{(k)} + g^{(k)} u^{(k)}) \right], \quad i = 1, \ldots, m \\ u^{n+1} &= u^{(m)} \end{aligned}$$

where $f^{(k)} := f(u^{(k)}, t^{(k)})$, $g^{(k)} := g(u^{(k)}, t^{(k)})$, $t^{(k)} := t^n + D_k \Delta t$, $t^{n+1} := t^n + \Delta t$ and D_k are given by

$$D_0 = 0, \quad D_i = \sum_{k=0}^{i-1} \alpha_{i,k} (D_k + \beta_{i,k})$$

The RK method is fully determined by its coefficients $\{\alpha_{i,k}, \beta_{i,k}\}$

Consistency requirements: $\sum_{k=0}^{i-1} \alpha_{i,k} = 1, \quad i = 1, \ldots, m, \quad D_m = 1$

The RK method is a linear combination of the first-order FE steps:

$$u^{(i)} = \sum_{k=0}^{i-1} \alpha_{i,k} u_{i,k}^{\mathsf{FE}}$$

where

$$u_{i,k}^{\mathsf{FE}} := u^{(k)} + \beta_{i,k} \Delta t (f^{(k)} + g^{(k)} u^{(k)})$$

According to [Gottlieb, Shu, Tadmor; 2001], the RK method is SSP provided

$$\alpha_{i,k} \geq 0 \quad \text{for all} \quad i, k$$

and an appropriate time step restriction is imposed.

Negative time increments are avoided if $\beta_{i,k} \geq 0 \quad \text{for all} \quad i, k$

New Semi-Implicit Methods

We first replace the FE evolution steps by the semi-implicit (SI) ones:

$$u_{i,k}^{\text{SI}} := u^{(k)} + \beta_{i,k}\Delta t(f^{(k)} + g^{(k)}u_{i,k}^{\text{SI}}) \quad \Longleftrightarrow \quad u_{i,k}^{\text{SI}} = \frac{u^{(k)} + \beta_{i,k}\Delta t f^{(k)}}{1 - \beta_{i,k}\Delta t g^{(k)}}$$

This leads to the following SI scheme:

$$\begin{aligned}
&u^{(0)} = u^n \\
&u^{(i)} = \sum_{k=0}^{i-1} \alpha_{i,k}\left(\frac{u^{(k)} + \beta_{i,k}\Delta t f^{(k)}}{1 - \beta_{i,k}\Delta t g^{(k)}}\right), \quad i = 1, \ldots, m \\
&u^{n+1} = u^{(m)}
\end{aligned}$$

Unfortunately, this scheme is at most first-order accurate

We, therefore, propose an order correction step:

$$u^{n+1} = \frac{u^{(m)} - C_m(\Delta t)^2 f^{(m)} g^{(m)}}{1 + C_m(\Delta t g^{(m)})^2}$$

where

$$C_0 = 0, \quad C_i = \sum_{k=0}^{i-1} \alpha_{i,k}(C_k + \beta_{i,k}^2), \quad i = 1, \ldots, m$$

New class of second-order semi-implicit Runge-Kutta (SI-RK) methods:

$$
\begin{aligned}
&u^{(0)} = u^n \\
&u^{(i)} = \sum_{k=0}^{i-1} \alpha_{i,k} \left(\frac{u^{(k)} + \beta_{i,k}\Delta t f^{(k)}}{1 - \beta_{i,k}\Delta t g^{(k)}} \right), \quad i = 1, \ldots, m \\
&u^{n+1} = \frac{u^{(m)} - C_m(\Delta t)^2 f^{(m)} g^{(m)}}{1 + C_m(\Delta t g^{(m)})^2}
\end{aligned}
$$

The set of coefficients $\{\alpha_{i,k}, \beta_{i,k}\}$ is taken directly from the explicit SSP-RK method of an appropriate order.

Remark. Note that in the degenerate case of $g \equiv 0$, the SI-RK methods are identical to the corresponding explicit RK methods

Theorem (Second-Order Accuracy) If the SSP-RK method is at least second-order accurate, then the corresponding SI-RK method with the same set of coefficients $\alpha_{i,k}, \beta_{i,k} \geq 0$ is second-order.

Theorem ($A(\alpha)$-Stability and Stiff Decay) Let us assume that the SI-RK methods are applied to the test equation $u' = \lambda u$, where $\lambda \in \mathbb{C}$ is a constant with $\mathrm{Re}\,\lambda < 0$. Then, the resulting methods, which can be written as

$$u^{n+1} = R(z)u^n, \quad z = \lambda \Delta t$$

satisfy the following two requirements:

$$|R(z)| \leq 1, \ \forall z \in \mathbb{C} \text{ s.t. } \mathrm{Re}\,z \leq -|\mathrm{Im}\,z| \quad \left(A(\alpha)\text{-stability with } \alpha = \frac{\pi}{4}\right)$$

and

$$R(z) \to 0 \text{ as } \mathrm{Re}\,z \to -\infty$$

provided $\alpha_{i,k} \geq 0$ and $\beta_{i,k} \geq 0$ for all i, k.

Theorem (Steady State Preserving Property) Let $\beta_{i,k} \geq 0\ \forall i,k$. Then, if the computed solution is at a steady state at time t^n, i.e., $u^n = \widehat{u}$ such that

$$f(\widehat{u},t) \equiv -g(\widehat{u},t)\widehat{u}$$

it will remain at the same steady state, namely,

$$u^{n+1} = \widehat{u}$$

Theorem (Sign Preserving Property) Let the initial condition u^0 and function f satisfy

$$\{u^0 \geq 0, \quad f \geq 0\} \quad \text{or} \quad \{u^0 \leq 0, \quad f \leq 0\}$$

Then,

$$\text{sgn}(u^n) \equiv \text{sgn}(u^0)$$

for all n provided $\alpha_{i,k} \geq 0$ and $\beta_{i,k} \geq 0$ for all i,k.

Absolute Stability of Two SSP-Based SI-RK Methods

The first SI-RK2 method is based on the 2-order SSP-RK solver:

$$
\begin{aligned}
u^{(1)} &= \frac{u^n + \Delta t f^n}{1 - \Delta t g^n} \\
u^{(2)} &= \frac{1}{2} u^n + \frac{1}{2} \cdot \frac{u^{(1)} + \Delta t f^{(1)}}{1 - \Delta t g^{(1)}} \\
u^{n+1} &= \frac{u^{(2)} - (\Delta t)^2 f^{(2)} g^{(2)}}{1 + (\Delta t g^{(2)})^2}
\end{aligned}
$$

The second SI-RK3 method is based on the 3-order SSP-RK solver:

$$
\begin{aligned}
u^{(1)} &= \frac{u^n + \Delta t f^n}{1 - \Delta t g^n} \\
u^{(2)} &= \frac{3}{4} u^n + \frac{1}{4} \cdot \frac{u^{(1)} + \Delta t f^{(1)}}{1 - \Delta t g^{(1)}} \\
u^{(3)} &= \frac{1}{3} u^n + \frac{2}{3} \cdot \frac{u^{(2)} + \Delta t f^{(2)}}{1 - \Delta t g^{(2)}} \\
u^{n+1} &= \frac{u^{(3)} - (\Delta t)^2 f^{(3)} g^{(3)}}{1 + (\Delta t g^{(3)})^2}
\end{aligned}
$$

To analyze the absolute stability, we consider the following test problem:

$$u' = \lambda_1 u + \lambda_2 u, \quad \lambda_1 \in \mathbb{C}, \ \mathrm{Re}(\lambda_1) \leq 0, \quad \lambda_2 \in \mathbb{R}, \ \lambda_2 \leq 0$$

$\lambda_1 u$: nonstiff part, $\lambda_2 u$: stiff part

We denote $z_1 := \lambda_1 \Delta t$ and $z_2 := \lambda_2 \Delta t$.

We denote the stability regions of the second- and third-order SSP-RK methods by $\mathcal{D}_{\mathrm{SSP2}}$ and $\mathcal{D}_{\mathrm{SSP3}}$, respectively.

We denote the corresponding time step restrictions by $\Delta t \leq \Delta t_{\mathrm{SSP2}}$ and $\Delta t \leq \Delta t_{\mathrm{SSP3}}$.

Theorem (Absolute Stability of the SI-RK2 Method) The region of absolute stability of the SI-RK2 method contains $\mathcal{D}_{\mathrm{SSP2}}$, i.e., for any $z_2 \le 0$, the solution of

$$
\begin{aligned}
u^{(1)} &= \frac{1+z_1}{1-z_2}\, u^n \\
u^{(2)} &= \frac{1}{2} u^n + \frac{1}{2} \cdot \frac{1+z_1}{1-z_2}\, u^{(1)} \\
u^{n+1} &= \frac{1-z_1 z_2}{1+z_2^2}\, u^{(2)}
\end{aligned}
$$

satisfies $|u^{n+1}| \le |u^n|$ provided $\Delta t \le \Delta t_{\mathrm{SSP2}}$.

Conjecture (Absolute Stability of the SI-RK3 Method) The region of absolute stability of the SI-RK3 method contains $\mathcal{D}_{\mathsf{SSP3}}$, i.e., for any $z_2 \le 0$, the solution of

$$
\begin{aligned}
u^{(1)} &= \frac{1+z_1}{1-z_2}\, u^n \\
u^{(2)} &= \frac{3}{4}u^n + \frac{1}{4}\cdot\frac{1+z_1}{1-z_2}\, u^{(1)} \\
u^{(3)} &= \frac{1}{3}u^n + \frac{2}{3}\cdot\frac{1+z_1}{1-z_2}\, u^{(2)} \\
u^{n+1} &= \frac{1-z_1 z_2}{1+z_2^2}\, u^{(3)}
\end{aligned}
$$

satisfies $|u^{n+1}| \le |u^n|$ provided $\Delta t \le \Delta t_{\mathsf{SSP3}}$.

Numerical Examples

We test the second-order SI-RK3 method and compare the results with the ones obtained using the second-order IMEX-SSP3(3,3,2) method of Pareschi and Russo.

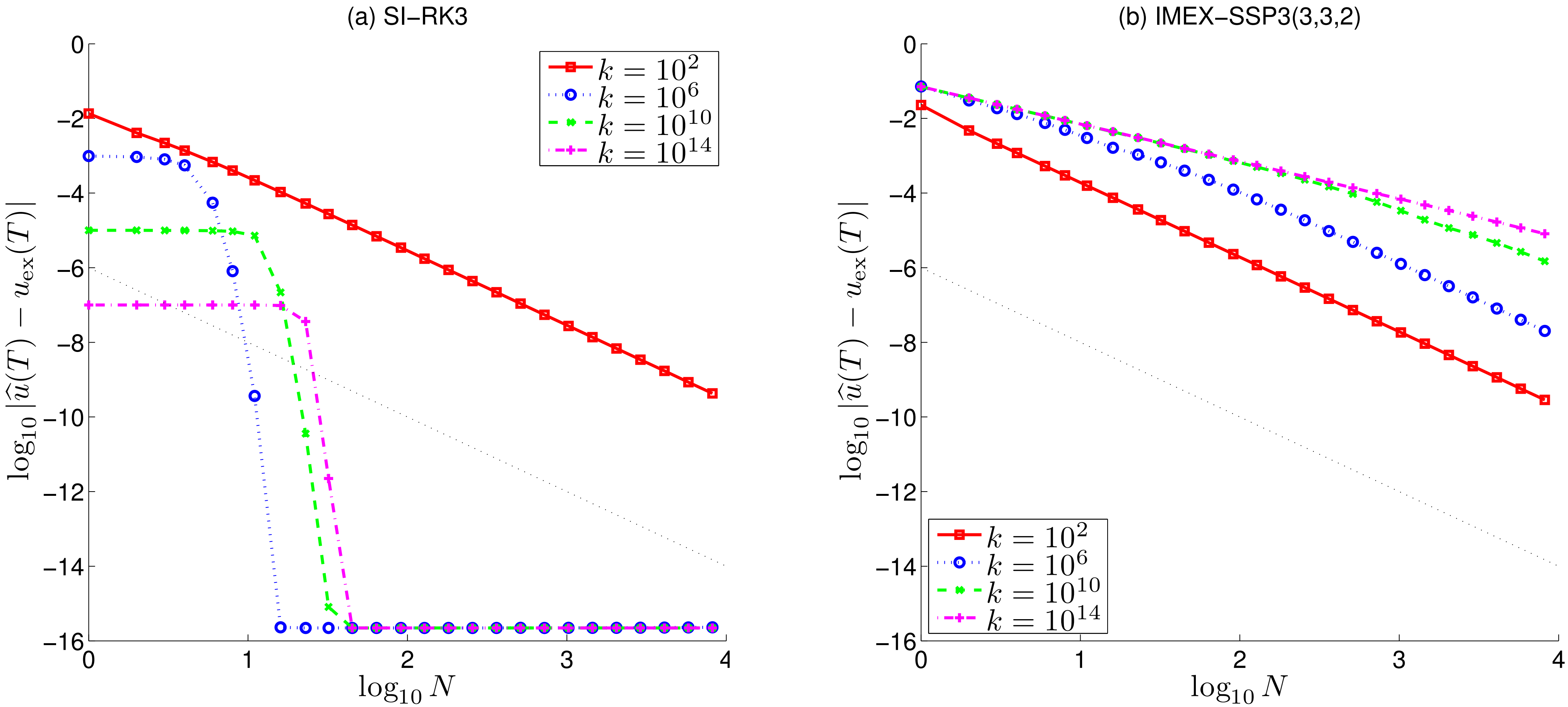
The obtained results clearly demonstrate that the new SI-RK3 method outperforms the IMEX-SSP3(3,3,2) when a large time step and/or coarse grid are used.

Example — Scalar ODE

$$u' = 1 - k|u|u, \quad k > 0$$

It has one equilibrium point $u^* = 1/\sqrt{k}$

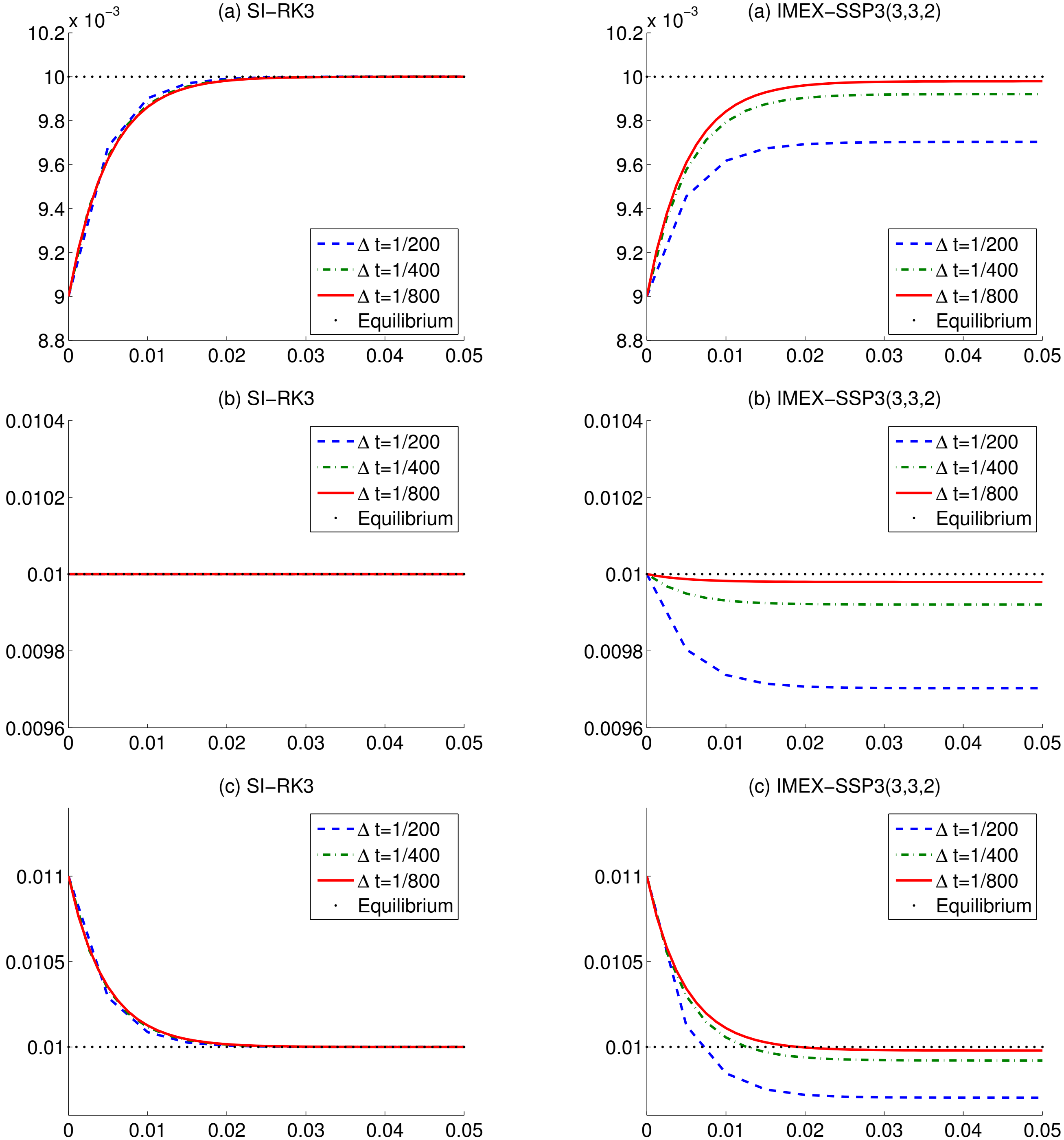
Accuracy Test



Steady State Preserving Test

We take $k = 10000$ with the corresponding equilibrium point $u^* = 0.01$. We consider three different initial values:

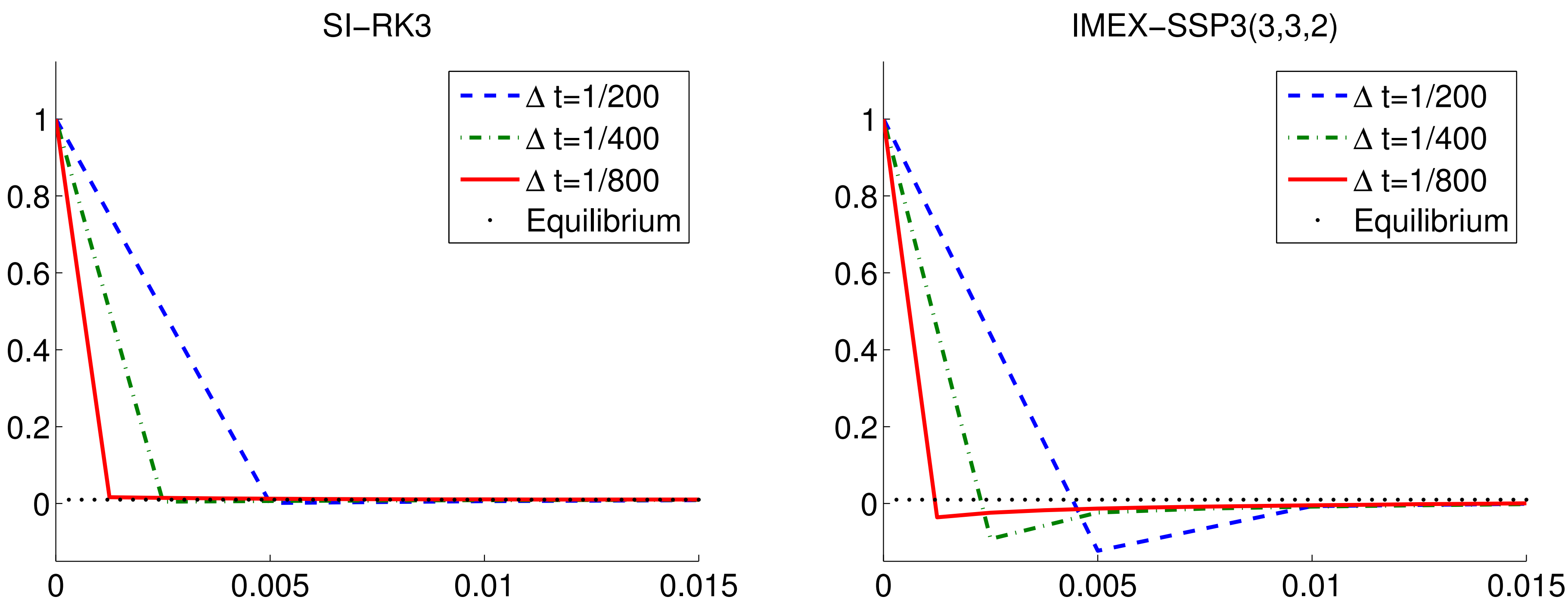
$$\text{(a) } u(0) = 0.9u^*, \quad \text{(b) } u(0) = u^*, \quad \text{(c) } u(0) = 1.1u^*$$


(a) SI−RK3
(a) IMEX−SSP3(3,3,2)
(b) SI−RK3
(b) IMEX−SSP3(3,3,2)
(c) SI−RK3
(c) IMEX−SSP3(3,3,2)
x 10−3
Δ t=1/200
Δ t=1/400
Δ t=1/800
Equilibrium

Sign Preserving Test

We take $k = 10000$ with the corresponding equilibrium point $u^* = 0.01$. We consider large initial value:

$$u(0) = 1$$



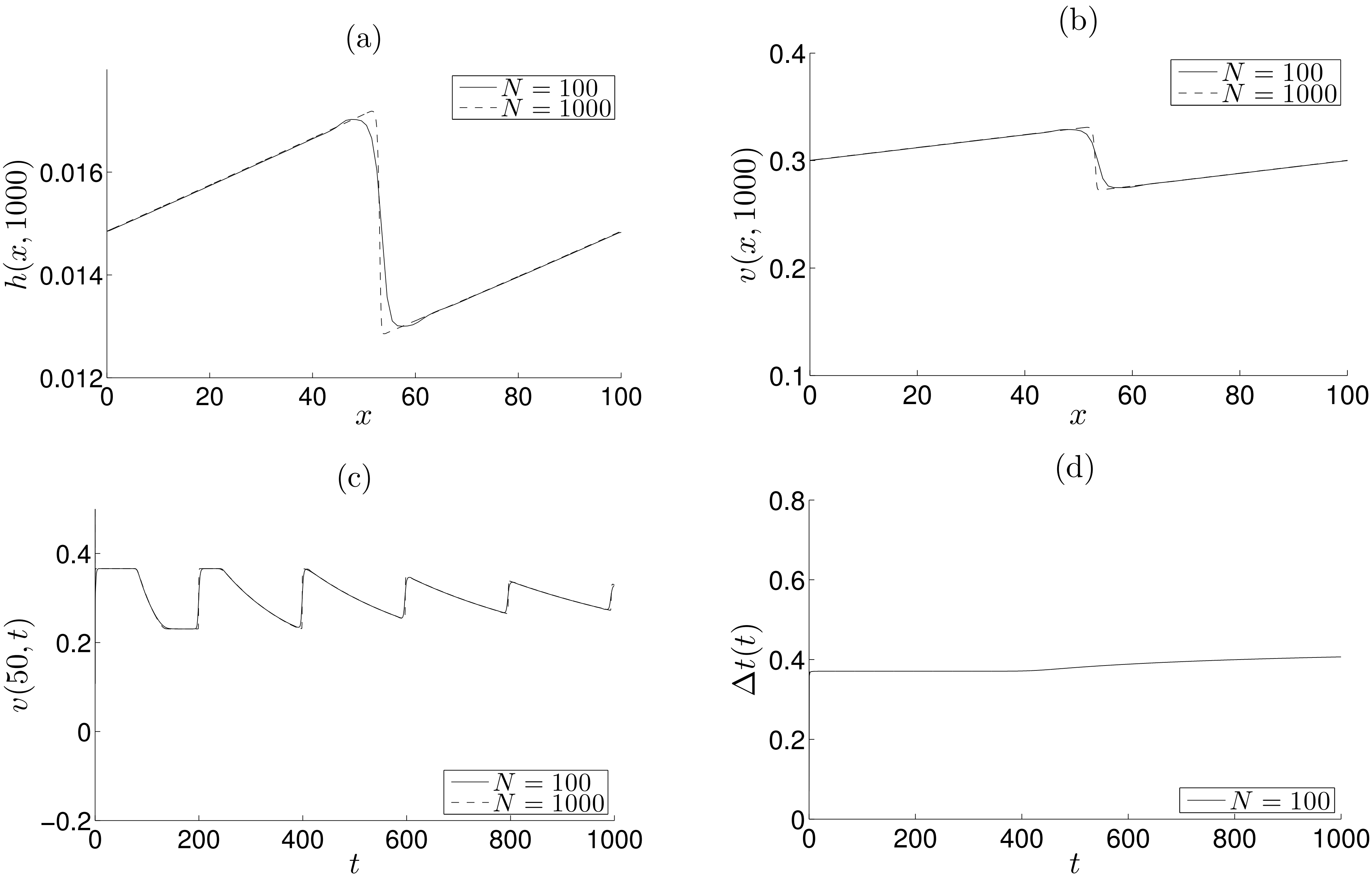
Example — Shallow Water System with Friction

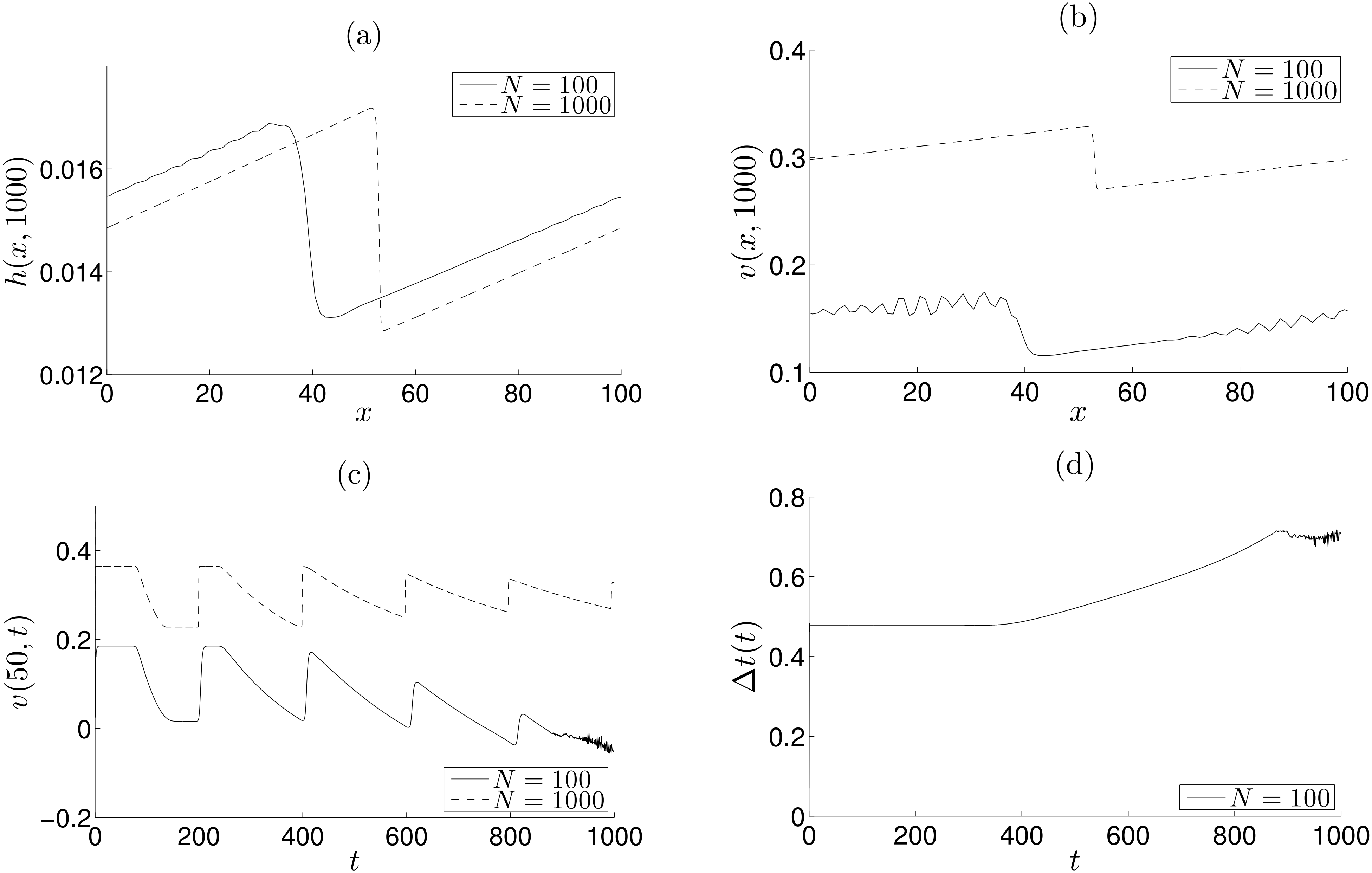
We take $Z_x \equiv -0.2$, $n = 0.09$ and the following initial conditions:

$$h(x,0) = \begin{cases} 0.02, & x < 50 \\ 0.01, & x > 50 \end{cases} \qquad q(x,0) = \begin{cases} 0, & x < 50 \\ 0.04, & x > 50 \end{cases}$$

We restrict the computational domain to $[0, 100]$, which is divided into N uniform cells, and impose the periodic boundary conditions

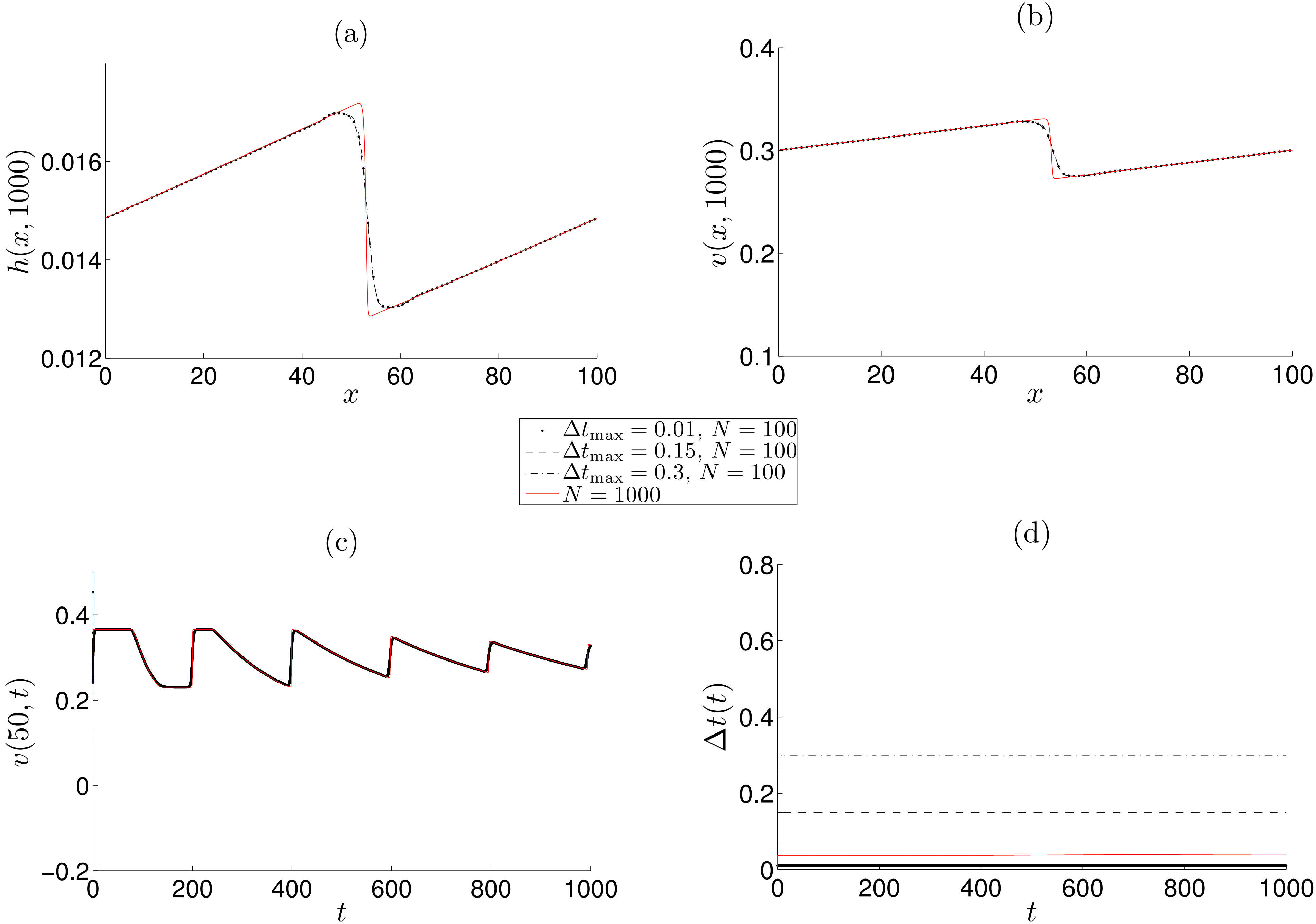
Time Steps Restricted by the CFL Condition (the CFL number is 0.3)

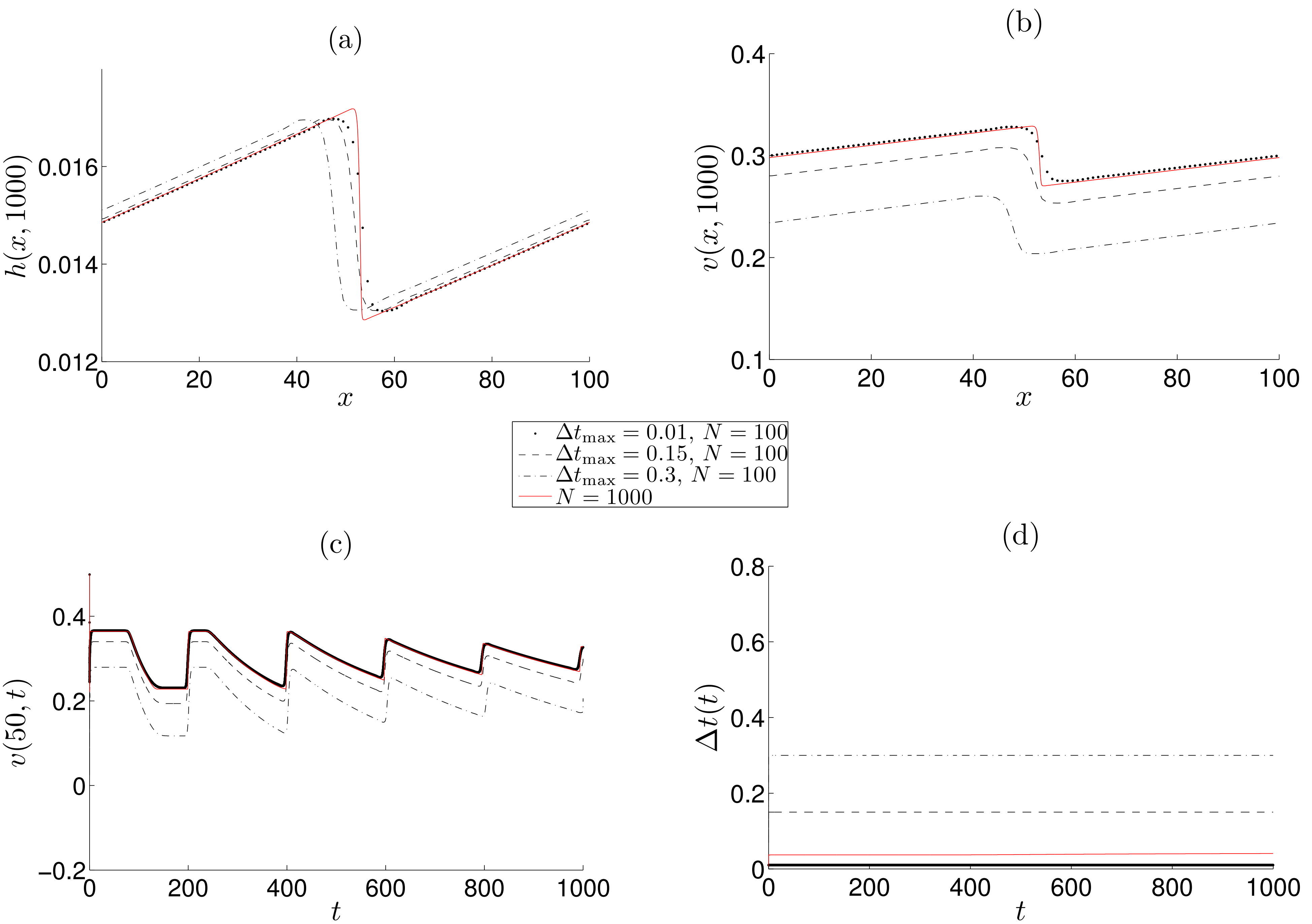




(a)
$h(x,1000)$
0.012
0.014
0.016
0 20 40 60 80 100
x
$N=100$
$N=1000$
(b)
$v(x,1000)$
0.1
0.2
0.3
0.4
0 20 40 60 80 100
x
$N=100$
$N=1000$
(c)
$v(50,t)$
−0.2
0
0.2
0.4
0 200 400 600 800 1000
t
$N=100$
$N=1000$
(d)
$\Delta t(t)$
0
0.2
0.4
0.6
0.8
0 200 400 600 800 1000
t
$N=100$

Fixed Time Step Restriction ($\Delta t = \min\{\Delta t_{CFL}, \Delta t_{max}\}$)




(a)
(b)
(c)
(d)
$h(x, 1000)$
$v(x, 1000)$
$v(50, t)$
$\Delta t(t)$
x
t
$\Delta t_{\max} = 0.01,\ N = 100$
$\Delta t_{\max} = 0.15,\ N = 100$
$\Delta t_{\max} = 0.3,\ N = 100$
$N = 1000$