XL CILAMCE
IBERO-LATIN AMERICAN CONGRESS ON COMPUTATIONAL METHODS IN ENGINEERING
NOVEMBER 11-14, 2019
Praiamar Natal Hotel & Convention
Natal, RN-BRAZIL

A COMPARATIVE STUDY ON THE USE OF STABILIZED AND FLUX-CORRECTED TRANSPORT FINITE ELEMENT METHODS TO SOLVE THE SHALLOW WATER EQUATIONS

Túlio L. Santos
tulio.santos@inpetu.org.br
Instituto de Pesquisa Aplicada Alan Turing (INPETU)
Department of Civil Engineering, COPPE/Federal University of Rio de Janeiro
Alvaro L. G. A. Coutinho
alvaro@nacad.ufrj.br
Department of Civil Engineering, COPPE/Federal University of Rio de Janeiro

Abstract. A study is carried out to assess the use of stabilized and flux-corrected transport (FCT) finite element approaches to solve the shallow water equations. The adopted stabilized formulation employs a Streamline Upwind Petrov-Galerkin term supplemented by a shock-capturing operator. In the examined FCT technique, the low-order equation considers an artificial viscosity term based on a Rusanov-like scalar dissipation. Anti-diffusive fluxes are linearized around the low-order solution and limited with the Zalesak's algorithm. Both approaches use semi-discrete finite elements with implicit time integration over time steps constrained by a CFL condition. Numerical results are presented for standard test problems available in the literature, and the accuracy and stability of the solutions are discussed.

Keywords: Stabilized finite element, SUPG, Shock capturing, Flux-corrected transport, Shallow water.

1 Introduction

Shallow water models have been substantially studied and applied over the years, having applications in atmospheric (Giraldo [1]) and oceanographic (Sármány and Hubbard [2]) modelling, gas flow dynamics (Karel'skii et al. [3]), magnetohydrodynamics (Klimachkov and Petrosyan [4]), floodings (Creed et al. [5]), and turbidity currents (de Luna et al. [6]). All these models assume that the shallow water hypothesis holds. This approximation presupposes the horizontal extent of the studied physical phenomenon is significantly higher than its vertical scale. Thus, standard governing equations can be simplified, producing a so-called depth-averaged or shallow water model. These modifications are fundamental to reduce computational costs, allowing, for example, the simulation of larger areas. In contrast, depth-resolving strategies can provide more accurate results, although they require additional computational effort. Meiburg et al. [7] discuss the strengths and challenges associated with the different approaches in the context of gravity-driven flow simulations.

To solve the shallow water equations, often finite difference (Groenenberg et al. [8]) and finite volume (Hou et al. [9]) methods are employed. Also, several discontinuous (Ambati and Bokhove [10]) and continuous (Hervouet [11], Castro et al. [12]) finite element techniques have been applied. Within the finite element group, stabilized formulations, formed by adding consistent and numerically stabilizing terms to the Galerkin method, have achieved considerable success (Behzadi [13], Castro [14], Takase et al. [15]). Usually, they use a Streamline Upwind Petrov-Galerkin (SUPG) (Hughes and Mallet [16]) term combined with a shock-capturing operator, such as the Consistent Approximate Upwind (CAU) (Galeão and do Carmo [17]) operator. Santos and Coutinho [18] evaluate the use of different SUPG and shock-capturing techniques to solve the shallow water equations. Similarly, regarding continuous finite elements, flux-corrected transport (FCT) (Kuzmin et al. [19], Sheu and Fang [20], Ortiz et al. [21]) methods constitute a relevant subgroup. Classical FCT techniques use a diffusive low-order formulation that is numerically stable and suppresses undershoots and overshoots. Besides, the built-in numerical diffusion can be defined to enforce the positivity constraint. Then, the obtained solution is corrected by anti-diffusive fluxes limited to avoid the creation or growth of extrema. Alternatively, it is possible to blend low- and high-order equations in a high-resolution scheme (Kuzmin [22]).

In this work, we compare a stabilized finite element formulation with a flux-corrected transport scheme. The examined stabilized approach comes from the work of Santos and Coutinho [18]. From the operators they tested, we use the SUPG operator proposed by Tezduyar [23], later adapted by Takase et al. [24] to the shallow water equations. Plus, we employ the $YZ\beta$ (Rispoli et al. [25], Tezduyar and Senga [26]) and δ_{91-MOD} (Rispoli et al. [25], Rispoli and Saavedra [27]) shock-capturing operators. For the FCT scheme, we follow the work of Santos et al. [28], in which the low-order formulation is formed by adding a Rusanov-like scalar dissipation to standard Galerkin equations, and the high-order system is composed by adding limited anti-diffusive fluxes to the low-order equations. Both approaches use semi-discrete finite elements with suitable implicit time integration techniques. Also, time steps are adaptively adjusted throughout the simulations under the same CFL condition.

Altogether, we seek to evaluate whether one of the examined numerical methods is more suitable to solve the shallow water equations. To this end, we address some test problems and assess the accuracy and stability of the obtained solutions. The remainder of this paper is organized as follows: in Section (2), the adopted shallow water model is introduced, while Section (3) presents the numerical methods used to solve its governing equations. Then, Section (4) shows and compares the numerical results obtained with the stabilized and FCT approaches. At last, concluding remarks and suggestions for future works are displayed at Section (5).

2 Physical model

A schematic representation of the adopted shallow water model is presented in Fig. 1. It depicts a fluid with density ρ and height (or depth) $h(x, y, t)$ that flows over an irregular terrain whose elevation from a datum fixed at its lowest point is $z_b(x, y)$. Thus, this model could easily represent the flow of

sea-water across the seabed.

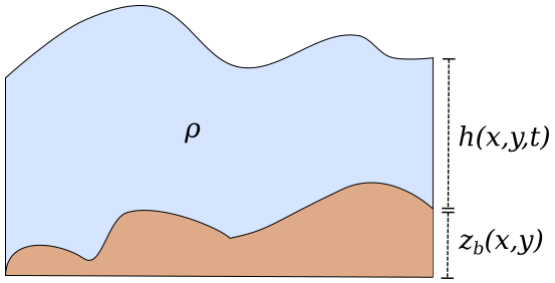


Figure 1. Schematic representation of the adopted shallow water model. It shows a fluid with density ρ and height/thickness $h(x, y, t)$ that flows over a terrain whose elevation is $z_b(x, y)$.

The model's unknown variables are $\boldsymbol{U}^T = [h,\, hu,\, hv] = [h,\, q_x,\, q_y]$. Here, $\boldsymbol{q}^T = [q_x,\, q_y]$ is the vector of the fluid's specific discharges (or discharges per unit width) and $\boldsymbol{u}^T = [u(x, y, t),\, v(x, y, t)]$ is its depth-averaged velocity vector.

The system of equations that governs the flow can be written as a generalized convection-diffusion equation:

$$\frac{\partial \boldsymbol{U}}{\partial t} + \boldsymbol{A}\nabla\boldsymbol{U} - \nabla\cdot[\boldsymbol{K}\nabla\boldsymbol{U}] = \boldsymbol{S}, \tag{1}$$

where:

$$\boldsymbol{A} = \begin{bmatrix}\boldsymbol{A_1}\, \boldsymbol{A_2}\end{bmatrix}, \qquad \nabla\boldsymbol{U} = \begin{bmatrix}\boldsymbol{I_3}\,\partial/\partial x \\ \boldsymbol{I_3}\,\partial/\partial y\end{bmatrix}\boldsymbol{U}, \qquad (\nabla\cdot) = \begin{bmatrix}\boldsymbol{I_3}\dfrac{\partial}{\partial x} & \boldsymbol{I_3}\dfrac{\partial}{\partial y}\end{bmatrix}, \tag{2}$$

$$\boldsymbol{A_1} = \begin{bmatrix} 0 & 1 & 0 \\ gh - u^2 & 2u & 0 \\ -uv & v & u \end{bmatrix}, \qquad \boldsymbol{A_2} = \begin{bmatrix} 0 & 0 & 1 \\ -uv & v & u \\ gh - v^2 & 0 & 2v \end{bmatrix}, \qquad \boldsymbol{S} = \begin{bmatrix} 0 \\ -g\dfrac{\partial z_b}{\partial x}h - \gamma q_x \\ -g\dfrac{\partial z_b}{\partial y}h - \gamma q_y \end{bmatrix}, \tag{3}$$

$$\boldsymbol{K} = \begin{bmatrix}\boldsymbol{K_{11}} & \boldsymbol{K_{12}} \\ \boldsymbol{K_{21}} & \boldsymbol{K_{22}}\end{bmatrix}, \qquad \boldsymbol{K_{11}} = \boldsymbol{K_{22}} = \frac{\mu}{\rho}\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad \boldsymbol{K_{12}} = \boldsymbol{K_{21}} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \tag{4}$$

and $\boldsymbol{I_3}$ is the third-order identity matrix. Also, $\mu = 10^{-3}\mathrm{Pa\,s}$ is the fluid's dynamic viscosity, $\rho = 10^3\mathrm{kg\,m^{-3}}$ is its density, $g = 9.81\,\mathrm{m\,s^{-2}}$ is the gravitational acceleration and $\gamma = C_f||\boldsymbol{q}||$. Here, C_f is the bed friction coefficient, defined by the Manning's equation $C_f = gn^2h^{-7/3}$, where $n = 0.018\,\mathrm{sm}^{-1/3}$ is the Manning coefficient.

Therefore, we consider the following initial value problem: given the closed domain $\bar{\Omega} \in \mathbb{R}^2 \times [t_0, t_1]$, of interior region Ω and boundary $\Gamma = \Gamma_e \cup \Gamma_n$, with $\Gamma_e \cap \Gamma_n = \emptyset$, we solve (1) for $\boldsymbol{U}(\boldsymbol{x}, t)$, subject to the initial condition:

$$\boldsymbol{U}(\boldsymbol{x}, t_0) = \boldsymbol{U}_0(\boldsymbol{x}), \tag{5}$$

and to the essential and natural boundary conditions:

$$\boldsymbol{U} = \boldsymbol{G} \text{ on } \Gamma_e \times]t_0, t_1], \tag{6}$$

$$\boldsymbol{K}\frac{\partial \boldsymbol{U}}{\partial x_k}n_k = (\boldsymbol{K}\nabla\boldsymbol{U})\cdot\boldsymbol{n} = \boldsymbol{0} \text{ on } \Gamma_n \times]t_0, t_1], \tag{7}$$

where $\boldsymbol{n}$ is the outward-pointing unit normal at the boundary. In Eq. (7), the diffusive flux across the boundary Γ_n is null and thus all the flow across the boundary is advective. This is a common approach in convective-diffusive physical models. Another possible boundary condition is the non-penetration condition $\boldsymbol{q}\cdot\boldsymbol{n} = 0$ on $\Gamma_e \times]t_0, t_1]$.

3 Numerical models

We present the finite element approaches employed to solve Eq. (1) numerically. Initially, we obtain the variational formulation of the problem. In sequence, we introduce the employed stabilized and flux-correct transport methods. Afterwards, we show how the time steps vary according to a CFL condition and discuss how to avoid the numerical instabilities that might arise near the transition between wet and dry regions as a flooding front advances on a terrain. Here all finite element related implementations were assisted by the deal.II library (Bangerth et al. [29]) and its module for parallel computing.

3.1 Variational formulation

To determine the variational formulation of the problem, let the sets of finite-dimensional trial and test functions be respectively defined as $\mathcal{S}^h = \{\boldsymbol{U}^h \in (H^{1h}(\bar{\Omega}))^3 \mid \boldsymbol{U}^h = \boldsymbol{G} \text{ on } \Gamma_e\}$ and $\mathcal{V}^h = \{\boldsymbol{W}^h \in (H^{1h}(\bar{\Omega}))^3 \mid \boldsymbol{W}^h = \boldsymbol{0} \text{ on } \Gamma_e\}$, where $H^{1h}(\bar{\Omega})$ is the finite-dimensional first-order Hilbert space, specified in the closed domain $\bar{\Omega}$. Then, employing the continuous Galerkin method, we consider $\boldsymbol{U}^h = \boldsymbol{V}\boldsymbol{N}$ and $\boldsymbol{W}^h = \boldsymbol{C}\boldsymbol{N}$, where $\boldsymbol{V}$ contains the nodal values of the solution, $\boldsymbol{C}$ defines arbitrary constants and $\boldsymbol{N}$ holds the finite element basis (or interpolation) functions. Next, we adopt the following weak formulation:

$$\int_\Omega \left(\boldsymbol{W}^h\right)^T \left(\frac{\partial \boldsymbol{U}^h}{\partial t} + \boldsymbol{A}^h \nabla \boldsymbol{U}^h - \boldsymbol{S}^h\right) d\Omega + \int_\Omega \left(\nabla \boldsymbol{W}^h\right)^T \left(\boldsymbol{K} \nabla \boldsymbol{U}^h\right) d\Omega = 0, \tag{8}$$

where the forms $\boldsymbol{A}^h$ and $\boldsymbol{S}^h$ indicate that their respective matrices should be computed based on $\boldsymbol{U}^h$. Then, with a suitable choice of the constants $\boldsymbol{C}$, the semi-discrete system of governing equations can be defined:

$$\boldsymbol{M}\frac{\partial \boldsymbol{V}}{\partial t} = -\boldsymbol{D}\boldsymbol{V} + \boldsymbol{F}, \tag{9}$$

where $\boldsymbol{M}$ and $\boldsymbol{D}$ are generalized mass and stiffness matrices, and $\boldsymbol{F}$ is the source term. In this case:

$$\boldsymbol{M} = \sum_{e=1}^{n_{el}} \int_{\Omega^e} \boldsymbol{N}^T \boldsymbol{N}\, d\Omega^e, \qquad \boldsymbol{F} = \sum_{e=1}^{n_{el}} \int_{\Omega^e} \boldsymbol{N}^T \boldsymbol{S}^h\, d\Omega^e, \tag{10}$$

$$\boldsymbol{D} = \sum_{e=1}^{n_{el}} \int_{\Omega^e} \left(\boldsymbol{N}^T \boldsymbol{A}^h \nabla \boldsymbol{N} + (\nabla \boldsymbol{N})^T \boldsymbol{K} \nabla \boldsymbol{N}\right) d\Omega^e, \tag{11}$$

where Ω^e, $e = 1, ..., n_{el}$ are the n_{el} elements that comprise the domain Ω, with $\Omega^i \cap \Omega^j = \emptyset$, $\forall\,(i, j)$. It should be noted that $\boldsymbol{D}$ and $\boldsymbol{F}$ depend on $\boldsymbol{V}$, which confers nonlinearity to the system. The discrete problem specification is completed by considering discrete versions of the initial and boundary conditions defined in Eqs. (5)-(7).

3.2 Stabilized method

The stabilized formulation is obtained by adding, to their respective counterparts in Eqs. (10) and (11), the matrices:

$$\boldsymbol{M}_{\text{SUPG}} = \sum_{e=1}^{n_{el}} \int_{\Omega^e} \left(\tau \boldsymbol{A}^h \nabla \boldsymbol{N}\right)^T \boldsymbol{N}\, d\Omega^e, \qquad \boldsymbol{F}_{\text{SUPG}} = \sum_{e=1}^{n_{el}} \int_{\Omega^e} \left(\tau \boldsymbol{A}^h \nabla \boldsymbol{N}\right)^T \boldsymbol{S}^h\, d\Omega^e, \tag{12}$$

$$\boldsymbol{D}_{\text{SUPG}} = \sum_{e=1}^{n_{el}} \int_{\Omega^e} \left(\tau \boldsymbol{A}^h \nabla \boldsymbol{N}\right)^T \left(\boldsymbol{A}^h \nabla \boldsymbol{N} - \boldsymbol{K} \nabla^2 \boldsymbol{N}\right) d\Omega^e, \tag{13}$$

$$\boldsymbol{D}_{\text{Shock}} = \sum_{e=1}^{n_{el}} \int_{\Omega^e} \delta (\nabla \boldsymbol{N})^T \nabla \boldsymbol{N}\, d\Omega^e. \tag{14}$$

In this case, τ and δ are, respectively, the SUPG and shock-capturing operators' coefficients.

The adopted SUPG technique was proposed by Tezduyar [23] and adapted by Takase et al. [24] to the shallow water equations, being defined as:

$$\tau = \left(\frac{1}{(\tau_{SUGN1})^2} + \frac{1}{(\tau_{SUGN2})^2} \right)^{-1/2}, \tag{15}$$

with:

$$\tau_{SUGN1} = \left(\sum_{i=1}^{n_{npe}} c|\boldsymbol{j} \cdot \nabla N_i| + |\boldsymbol{u} \cdot \nabla N_i| \right)^{-1}, \qquad \tau_{SUGN2} = \frac{\Delta t}{2}. \tag{16}$$

Here, Δt is the current time step, $n_{npe} = 4$ is the number of nodes per element, $c = \sqrt{gh}$ is the propagation speed of a perturbation on the surface, and $\boldsymbol{j} = \nabla\eta / ||\nabla\eta||$ is the normalized gradient of the free surface elevation $\eta = h + z_b$.

Moreover, we employ the shock-capturing operator proposed by Rispoli and Saavedra [27], but, instead of just using the advective term, we take into account the residual of Eq. (1) without the transient term. Its coefficient is defined as:

$$\delta = \delta_{91-MOD} = max(0, \delta_{91} - \delta_\tau), \tag{17}$$

with:

$$\delta_{91} = \frac{\left\| \boldsymbol{R}^h \right\|_{\widetilde{\boldsymbol{A}_0}^{-1}}}{\left(\sum_{j=1}^{2} \left\| \frac{\partial \xi_j}{\partial x_k} \frac{\partial \boldsymbol{U}^h}{\partial x_k} \right\|^2_{\widetilde{\boldsymbol{A}_0}^{-1}} \right)^{1/2}}, \qquad \delta_\tau = \frac{\left\| \boldsymbol{R}^h \right\|_{\widetilde{\boldsymbol{A}_0}^{-1}\boldsymbol{\tau}}}{\left\| \frac{\partial \boldsymbol{U}^h}{\partial x_k} \right\|^2_{\widetilde{\boldsymbol{A}_0}^{-1}}}, \tag{18}$$

$$\boldsymbol{R}^h = \boldsymbol{A}^h \nabla \boldsymbol{U}^h - \boldsymbol{K} \nabla^2 \boldsymbol{U}^h - \boldsymbol{S}^h, \tag{19}$$

where ξ_i, with $i \in \{1, 2\}$, are the canonical reference coordinates. For a vector $\boldsymbol{v}$ and a matrix $\boldsymbol{M}$, the norm $||\boldsymbol{v}||_{\boldsymbol{M}} = (\boldsymbol{v}^T \boldsymbol{M} \boldsymbol{v})^{1/2}$ is employed. Also, $\boldsymbol{\tau} = \tau \boldsymbol{I_3}$ and $\widetilde{\boldsymbol{A}_0}^{-1}$ is the inverse Jacobian of the transformation from entropy to conservation variables. Hence, we consider the energy functional:

$$E = \frac{gh^2 + u^2 h + v^2 h}{2}, \tag{20}$$

from which the entropy variables $\boldsymbol{V}$ and the matrix $\widetilde{\boldsymbol{A}_0}^{-1}$ can be defined (Tadmor and Zhong [30]):

$$\boldsymbol{V} = \frac{\partial E}{\partial \boldsymbol{U}} = \begin{bmatrix} gh - \dfrac{u^2 + v^2}{2} \\ u \\ v \end{bmatrix}, \qquad \widetilde{\boldsymbol{A}_0}^{-1} = \frac{\partial \boldsymbol{V}}{\partial \boldsymbol{U}} = \frac{1}{h} \begin{bmatrix} gh + u^2 + v^2 & -u & -v \\ -u & 1 & 0 \\ -v & 0 & 1 \end{bmatrix}. \tag{21}$$

From a simple analysis, it can be seen that $\widetilde{\boldsymbol{A}_0}^{-1}$ is symmetric positive definite if $h > 0$, which should be true by definition.

Another shock-capturing operator adopted is the $YZ\beta$ (Rispoli et al. [25]), whose coefficient is:

$$\delta = \frac{\delta_{\beta=1} + \delta_{\beta=2}}{2}, \tag{22}$$

with:

$$\delta_\beta = ||\boldsymbol{Y}^{-1}\boldsymbol{Z}|| \left(\sum_{i=1}^{2} \left\| \boldsymbol{Y}^{-1} \frac{\partial \boldsymbol{U}^h}{\partial x_i} \right\|^2 \right)^{\beta/2-1} \|\boldsymbol{Y}^{-1}\boldsymbol{U}^h\|^{1-\beta} \left(\frac{h_{\text{shoc}}}{2} \right)^\beta, \tag{23}$$

$$\boldsymbol{Y} = \begin{bmatrix} (h)_{\text{ref}} & 0 & 0 \\ 0 & (q_x)_{\text{ref}} & 0 \\ 0 & 0 & (q_y)_{\text{ref}} \end{bmatrix}, \qquad \boldsymbol{Z} = \boldsymbol{A_i}^h \frac{\partial \boldsymbol{U}^h}{\partial x_i}, \qquad h_{\text{shoc}} = 2 \left(\sum_{a=1}^{n_{\text{npe}}} |\boldsymbol{j} \cdot N_a| \right)^{-1}. \tag{24}$$

In this case, $(h)_{\text{ref}}$, $(q_x)_{\text{ref}}$ and $(q_y)_{\text{ref}}$ are reference values for the variables h, q_x and q_y. Alternatively, it could be used $\delta = \delta_{\beta=1}$ or $\delta = \delta_{\beta=2}$, for smoother or sharper shocks, respectively.

Following a semi-discrete approach, the nodal values are integrated over time using the predictor multi-corrector algorithm introduced by Aliabadi and Tezduyar [31], which is summarized in Algorithm 1. The nonlinear correction iterations are stopped if the l^2-norm of linear system's residual is less than TOL_1 times its value at the first iteration, or the relative difference between its current and previous values is less than TOL_2. Here we set $TOL_1 = 10^{-4}$ and $TOL_2 = 10^{-6}$. Also, the number of iterations is limited to 100. The linear system is solved with the iterative GMRes (Generalized Minimal Residual) algorithm (Saad and Schultz [32]), using a Krylov space of dimension 35 and the ILU(0) preconditioner.

Algorithm 1 Predictor multi-corrector algorithm employed for the time integration in the stabilized method. Here, $\mathbf{A} = \partial \boldsymbol{V}/\partial t$, $\theta = 0.5$ is a parameter that controls the stability and precision of the method, and Δt is the time step.

- Prediction phase:

$\mathbf{A}^{(0)} = \mathbf{0}$.
$\boldsymbol{V}^{(0)} = \boldsymbol{V}^n + (1-\theta)\Delta t \mathbf{A}^n$.

- Correction phase:

for $m = 0, 1, 2, ...$ until the convergence criteria is met, **do**:
  $\boldsymbol{R}^{(m)} = \boldsymbol{F}^{(m)} - (\boldsymbol{M}\mathbf{A}^{(m)} + \boldsymbol{D}^{(m)}\boldsymbol{V}^{(m)})$.
  $\boldsymbol{M}^* = \boldsymbol{M} + \theta \Delta t \boldsymbol{D}^{(m)}$.
  Solve $\boldsymbol{M}^*(\Delta \mathbf{A}^{(m)}) = \boldsymbol{R}^{(m)}$.
  $\mathbf{A}^{(m+1)} = \mathbf{A}^{(m)} + \Delta \mathbf{A}^{(m)}$.
  $\boldsymbol{V}^{(m+1)} = \boldsymbol{V}^{(m)} + \theta \Delta t \Delta \mathbf{A}^{(m)}$.
end for.

3.3 FCT method

To define the FCT method, first, we use the generalized trapezoidal method to discretize Eq. (9) in time, obtaining an initial high-order equation:

$$\left(\boldsymbol{M} + \theta \Delta t \boldsymbol{D}^{n+1}\right) \boldsymbol{V}^{n+1} = \left(\boldsymbol{M} - (1-\theta)\Delta t \boldsymbol{D}^n\right) \boldsymbol{V}^n + \Delta t \boldsymbol{F}^{n+\theta}, \tag{25}$$

where n and $n+1$ denote the current and next states, $\Delta t = t^{n+1} - t^n$ is the time step, and $\theta = 0.5$ is a parameter that controls the stability and precision of the method.

Then, to establish the low-order method, we initially note that it should respect the positivity constraint, i.e., it should not produce nonphysical negative values. Here, the fluid height should never be negative. Thus, we start by replacing the consistent mass matrix $\boldsymbol{M}$ in Eq. (9) by its diagonally lumped version $\boldsymbol{M_L}$, computed as:

$$\boldsymbol{M_L} = \text{diag}\{\boldsymbol{M_L}_i\}, \qquad \boldsymbol{M_L}_i = \sum_j \boldsymbol{M}_{ij}, \qquad \forall i, \tag{26}$$

so we have:

$$\boldsymbol{M_L}_i \frac{\partial \boldsymbol{V}_i}{\partial t} = -\boldsymbol{D}_{ij}\boldsymbol{V}_j + \boldsymbol{F}_i\,, \ \forall i. \tag{27}$$

To obey the positivity criterion that requires that $\boldsymbol{V}_i \geq 0, \forall i, \forall t > 0$ if $\boldsymbol{V}_{0i} \geq 0$, $\forall i$; it is sufficient to ascertain that:

$$\boldsymbol{M_L}_i > 0, \qquad \boldsymbol{D}_{ij} \leq 0, \qquad \boldsymbol{F}_i > 0, \qquad \forall i, \forall j \neq i. \tag{28}$$

Going back to Eq. (25), it can be written in the form:

$$\boldsymbol{A}\boldsymbol{V}^{n+1} = \boldsymbol{B}\boldsymbol{V}^{n}. \tag{29}$$

If we ensure that $\boldsymbol{A}$ is monotonic, then, by definition, $\boldsymbol{V}^{n+1} \geq 0$ if $\boldsymbol{A}\boldsymbol{V}^{n+1} \geq 0$. So, we restrict $\boldsymbol{A}$ to being a M-matrix, which is the subset of monotone matrices that satisfies:

$$\boldsymbol{A}_{ii} > 0,\ \forall i, \tag{30}$$

$$\boldsymbol{A}_{ij} \leq 0,\ \forall\, i \neq j, \tag{31}$$

$$\sum_j \boldsymbol{A}_{ij} \geq 0,\ \forall i. \tag{32}$$

Therefore, assuming the conditions (28) hold, the scheme is positivity preserving if $\boldsymbol{A}$ meets the requirements (30)-(32) and $\boldsymbol{B}$ does not have negative values. Next, we adjust the system matrices to meet such conditions.

To start, we add artificial diffusion to $\boldsymbol{D}$ to remove its positive off-diagonal entries. Following Guermond et al. [33] and Audusse et al. [34], let the hydrostatic reconstruction of the fluid's height at the i-th node in respect to the j-th one be:

$$h_{ij} = \max(0,\, h_i + z_{b_i} - \max(z_{b_i}, z_{b_j})). \tag{33}$$

Then, the associated reconstructed nodal specific discharge vector is $\boldsymbol{q}_{ij} = h_{ij}\boldsymbol{u}_i$.

Now, we define $\boldsymbol{L}(\boldsymbol{V}) = \boldsymbol{D}(\boldsymbol{V}) + \boldsymbol{C}(\boldsymbol{V})$, where $\boldsymbol{C}(\boldsymbol{V})$ is the Rusanov-like scalar dissipation matrix proposed by Santos et al. [28]:

$$\begin{aligned} \boldsymbol{C}_{ij}(\boldsymbol{V}) &= -\max(c_{ij}, c_{ji})\boldsymbol{I_3}\,,\ \forall\, i \neq j, \\ \boldsymbol{C}_{ii}(\boldsymbol{V}) &= -\sum_{j\neq i} \boldsymbol{C}_{ij}, \end{aligned} \tag{34}$$

with:

$$c_{ij} = |\boldsymbol{e}_{ij}\cdot\boldsymbol{u}_j| + |\boldsymbol{e}_{ij}\cdot\boldsymbol{q}_{ji}|\sqrt{g h_{ji}}\,, \qquad \boldsymbol{e}_{ij} = \int_{\Omega^e} N_i \nabla N_j d\Omega^e, \tag{35}$$

where N_i is the basis function of the i-th node. Formed this way, $\boldsymbol{C}$ has zero block-wise row and column sums, and thus, being a generalized diffusion operator, conserves mass. Also, note that $\boldsymbol{L}$ should obey the condition $\boldsymbol{L}_{ij} \leq 0\ \forall i,\ \forall j \neq i$ of the positivity criteria in Eq. (28).

In addition, following Santos et al. [28], a shock detector is used to add more diffusion near discontinuities in detriment to where the solution is smooth. Consider $\Delta h_{ij} = h_i - h_j$ and the function $\mathrm{sign}(x)$ that returns 1 if $x \geq 0$ and -1, otherwise. Then, we define:

$$\psi_i = \begin{cases} \dfrac{|\sum_{j\in S(i)} \mathrm{sign}(\Delta h_{ij})(\Delta h_{ij} - c)^4|}{\sum_{j\in S(i)}(\Delta h_{ij} - c)^4}, & \text{if } \sum_{j\in S(i)}(\Delta h_{ij} - c)^4 \neq 0, \\ 0, & \text{otherwise,} \end{cases} \tag{36}$$

where $c = 0.001$ and $S(i)$ is the set of the neighbouring nodes of i, which includes all those belonging to the elements adjacent to i. Then, the matrices $\boldsymbol{C}_{ij}$ are scaled:

$$\boldsymbol{C}_{ij}(\boldsymbol{V}) = -\max(\psi_i, \psi_j)\max(c_{ij}, c_{ji})\boldsymbol{I_3},\ \forall\, i \neq j. \tag{37}$$

As a result, the equation for the low-order scheme becomes:

$$\left(\boldsymbol{M_L} + \theta\Delta t \boldsymbol{L}^{(m)}\right)\boldsymbol{V}^{(m+1)} = \left(\boldsymbol{M_L} - (1-\theta)\Delta t\boldsymbol{L}^n\right)\boldsymbol{V}^n + \Delta t \boldsymbol{F}^{n+\theta}, \tag{38}$$

which is solved iteratively using the same stopping criteria and linear system solver used with the stabilized approach. After convergence, we assign $\boldsymbol{V}^L = \boldsymbol{V}^{(m+1)}$.

In sequence, as suggested by Hsu [35] and Kuzmin et al. [19], we linearize the source term about $\boldsymbol{V}^{n+1}$ to prevent the positivity condition of being violated by a negative source term. We set:

$$\boldsymbol{F} = \boldsymbol{F_C} + \boldsymbol{F_P}\boldsymbol{V}^{n+1}, \tag{39}$$

with $\boldsymbol{F_C}_i > 0$ and $\boldsymbol{F_P}_i \leq 0, \forall i$. This splitting can be done with:

$$\boldsymbol{F} = \boldsymbol{F}^+ + \left(\frac{\boldsymbol{F}^-}{\boldsymbol{V}^{(m)}}\right)\boldsymbol{V}^{(m+1)}, \tag{40}$$

where $\boldsymbol{F}^+$ and $\boldsymbol{F}^-$ are, respectively, the positive and negative parts of the source term, i.e., $\boldsymbol{F_C} = \boldsymbol{F}^+$ and $\boldsymbol{F_P} = \boldsymbol{F}^-/\boldsymbol{V}^{(m)}$. This separation is computed for each quadrature point contribution independently.

Therefore, the final low-order equation is:

$$\begin{aligned}\left[\boldsymbol{M_L} + \theta\Delta t\left(\boldsymbol{L}^{(m)} - \boldsymbol{F_P}^{(m)}\boldsymbol{I}\right) - (1-\theta)\Delta t\boldsymbol{F_P}^n\boldsymbol{I}\right]\boldsymbol{V}^{(m+1)} = \\ \left[\boldsymbol{M_L} - (1-\theta)\Delta t\boldsymbol{L}^n\right]\boldsymbol{V}^n + \theta\Delta t\boldsymbol{F_C}^{(m)} + (1-\theta)\Delta t\boldsymbol{F_C}^n,\end{aligned} \tag{41}$$

where $\boldsymbol{I}$ is the identity matrix. We remark that this linearization reinforces the requirements (30)-(32) needed for positivity preservation since the terms added to the diagonal of the left-hand side matrix are positive ($-\boldsymbol{F_P}_i \geq 0$). Besides, the contribuition of the source term to the right-hand side is compelled to be positive ($\boldsymbol{F_C}_i > 0$), meeting the criteria (28).

At last, to guarantee that the low-order scheme preserves positivity, the time steps still have to be constrained. While the left-hand side of Eq. (41) meets the necessary criteria by construction, its right-hand side requires that:

$$\boldsymbol{M_L} - (1-\theta)\Delta t\boldsymbol{L}^n \geq 0. \tag{42}$$

Since all off-diagonal entries of $\boldsymbol{L}^n$ should already be non-positive, the actual time step constraint depends on its diagonal:

$$\Delta t \leq \frac{1}{1-\theta}\min_i\left\{\frac{\boldsymbol{M_L}_i}{\boldsymbol{L}^n_{ii}}\,\middle|\,\boldsymbol{L}^n_{ii} > 0\right\}. \tag{43}$$

Although the high-order system is defined by Eq. (25), the high-order method is built by adding the anti-diffusive fluxes $\boldsymbol{\mathcal{F}}$ to the right-hand side of the low-order system (41), recovering the initial high-order equation:

$$\begin{aligned}\left[\boldsymbol{M_L} + \theta\Delta t\left(\boldsymbol{L}^{(m)} - \boldsymbol{F_P}^{(m)}\boldsymbol{I}\right) - (1-\theta)\Delta t\boldsymbol{F_P}^n\boldsymbol{I}\right]\boldsymbol{V}^{(m+1)} = \\ \left[\boldsymbol{M_L} - (1-\theta)\Delta t\boldsymbol{L}^n\right]\boldsymbol{V}^n + \theta\Delta t\boldsymbol{F_C}^{(m)} + (1-\theta)\Delta t\boldsymbol{F_C}^n + \boldsymbol{\mathcal{F}}^{(m)}.\end{aligned} \tag{44}$$

Consequently, the anti-diffusive term should correspond to the difference between the high-order Eq. (25) and the low-order Eq. (41), which results in:

$$\boldsymbol{\mathcal{F}}(\boldsymbol{V}^{(m)}, \boldsymbol{V}^n) = (\boldsymbol{M_L} - \boldsymbol{M})\left(\boldsymbol{V}^{(m)} - \boldsymbol{V}^n\right) + \theta\Delta t\boldsymbol{C}^{(m)}\boldsymbol{V}^{(m)} + (1-\theta)\Delta t\boldsymbol{C}^n\boldsymbol{V}^n. \tag{45}$$

Because we use the diagonal lumped mass matrix and $\boldsymbol{C}$ has zero row sums, the anti-diffusive flux can be written in terms of edge contributions:

$$\boldsymbol{\mathcal{F}}_i = \sum_{j\in S(i)}\boldsymbol{\mathcal{F}}_{ij}\,, \tag{46}$$

with:

$$\boldsymbol{\mathcal{F}}_{ij} = \left(\boldsymbol{M}_{ij} - \theta\Delta t\boldsymbol{C}^{(m)}_{ij}\right)(\boldsymbol{V}^{(m)}_i - \boldsymbol{V}^{(m)}_j) - \left(\boldsymbol{M}_{ij} + (1-\theta)\Delta t\boldsymbol{C}^n_{ij}\right)(\boldsymbol{V}^n_i - \boldsymbol{V}^n_j). \tag{47}$$

Next, we scale the edge contributions to adaptively switch between the low- and high- order methods:

$$\boldsymbol{\mathcal{F}}_i = \sum_{j \in S(i)} \alpha_{ij} \boldsymbol{\mathcal{F}}_{ij}. \tag{48}$$

Here, the correction factors are computed using the Zalesak-type flux limiter used by Santos et al. [28], in which the individual factors of the solution components (α_{ij}^h,$\alpha_{ij}^{q_x}$ and $\alpha_{ij}^{q_y}$) are synchronized into $\alpha_{ij} = \alpha_{ij}^h$. Thus, let $\mathcal{F}_{ij}^h$ be the anti-diffusive flux component related to h. Then, the correction factor is:

$$\alpha_{ij} = \begin{cases} \min\left(R_i^+, R_j^-\right), & \text{if } \mathcal{F}_{ij}^h > 0, \\ \min\left(R_i^-, R_j^+\right), & \text{otherwise}, \end{cases} \tag{49}$$

with:

$$R_i^+ = \min\left(1, \frac{M_{L_i} Q_i^+}{P_i^+}\right), \qquad R_i^- = \min\left(1, \frac{M_{L_i} Q_i^-}{P_i^-}\right), \tag{50}$$

$$P_i^+ = \sum_{j \in S(i)} \max(0, \mathcal{F}_{ij}^h), \qquad P_i^- = \sum_{j \in S(i)} \min(0, \mathcal{F}_{ij}^h), \tag{51}$$

$$Q_i^+ = \max\left[0, \max_{j \in S(i)} (h_{ji} - h_i)\right], \qquad Q_i^- = \min\left[0, \min_{j \in S(i)} (h_{ji} - h_i)\right]. \tag{52}$$

Besides, before the computation of the correction factors, the fluxes are prelimited using a minmod strategy, as suggested by Kuzmin [22]. This is done to avoid the consistent mass matrix of reversing the sign of $\boldsymbol{\mathcal{F}}_{ij}$ or increasing its magnitude:

$$\boldsymbol{\mathcal{F}}_{ij} = \mathrm{minmod}\left[\boldsymbol{\mathcal{F}}_{ij}, -\theta \Delta t \boldsymbol{C}_{ij}^{(m)} \left(\boldsymbol{V}_i^{(m)} - \boldsymbol{V}_j^{(m)}\right) - (1-\theta)\Delta t \boldsymbol{C}_{ij}^n \left(\boldsymbol{V}_i^n - \boldsymbol{V}_j^n\right)\right]. \tag{53}$$

By definition, the minmod function returns zero if the arguments have opposite signs, or the argument with the smallest magnitude, otherwise. This test is performed individually for each flux component.

Lastly, to reduce the computational cost, we use $\boldsymbol{V}^L$ as an approximation to $\boldsymbol{V}^{(m)}$ - and, thus, to $\boldsymbol{V}^{n+1}$ - when computing the fluxes and correction factors. Hence, they have to be computed only once per time step. Nonetheless, the low-order solution must be obtained before solving Eq. (44). We present a summary of the full FCT scheme in Algorithm 2.

Algorithm 2 Summary of the adopted FCT scheme.

1: Choose Δt based on the CFL condition (55).
2: Compute the low-order solution $\boldsymbol{V}^L$.
3: **if** condition (43) is not met **then** choose new Δt.
4: Compute the anti-diffusive fluxes $\boldsymbol{\mathcal{F}}(\boldsymbol{V}^L, \boldsymbol{V}^n)$.
5: Prelimit the fluxes.
6: Compute the α_{ij} factors and limit the fluxes.
7: Compute the high-order solution $\boldsymbol{V}^H$ and set $\boldsymbol{V}^{n+1} = \boldsymbol{V}^H$.

3.4 Adaptive time stepping

Throughout the simulations, we constrain the time step to limit the maximum CFL number of an element e, defined as:

$$CFL_e = \left(|\boldsymbol{u_e}| + \sqrt{g h_e}\right) \frac{\Delta t}{l_e}, \tag{54}$$

where l_e is the element's characteristic length, defined here as the square root of its area. Also, the speeds $|\boldsymbol{u}_e|$ and $\sqrt{gh_e}$ are evaluated at the element's barycenter. Then, the related restriction is given by:

$$\Delta t \leq CFL \min_e \left(\frac{l_e}{|\boldsymbol{u_e}| + \sqrt{gh_e}} \right). \tag{55}$$

For the simulations ran in this work, $CFL = 0.5$. Thus, as we march through time, we adaptively choose time steps that comply with condition (55). Do note that the adopted FCT scheme further constrain the time step to ensure positivity, choosing time steps that also obey the Eq. (43).

3.5 Dry/wet handling

In real-world applications of a shallow water model, it is often required to simulate the fluid flow over an initially dry irregular domain. In these cases, it is fundamental that we properly handle the transitions between dry and wet states as the flooding front advances across the terrain. Otherwise, instabilities and nonphysical behaviours might arise near the dry/wet interface. So, initially, it conveys to distinguish between wet and dry elements. We classify a node as being wet if its height h is greater than the threshold $h_{\text{dry}} = 0.01\,\text{m}$. Then, an element is wet or dry according to whether it has only wet or dry nodes. Otherwise, it is in a dry/wet front. This type of procedure is commonly applied to shallow water models (Brufau and García-Navarro [36], Ricchiuto and Bollermann [37], Kesserwani and Liang [38]).

Another frequent plan is to define a cut-off height value, under which point velocities are considered null. Often this value is the same as the one used for the dry/wet element classification (Kesserwani and Liang [38]). Based on the work of Ricchiuto and Bollermann [37], we adopt:

$$\boldsymbol{u} = \begin{cases} \dfrac{\boldsymbol{q}}{h}, & \text{if } h \geq C_{\boldsymbol{u}}, \\ \boldsymbol{0}, & \text{otherwise}, \end{cases} \tag{56}$$

where $C_{\boldsymbol{u}} = l/L_{\text{ref}}$, with $L_{\text{ref}} = \max_{i,j\in\Omega} (\|\boldsymbol{x}_i - \boldsymbol{x}_j\|)$ and l is the characteristic length of the associated element. In addition, to allow the method to accept negative height values, we assign $h = ||h||$ when assembling the element contributions.

Besides, spurious velocities can arise near dry/wet fronts and violate mass conservation, as a result of trying to simulate a continuous surface elevation on a discrete mesh. This is particularly important at elements with adverse slopes. An element is said to have an adverse slope if it has a wet node i and a dry node j, where $z_{b_j} > z_{b_i}$ (Ricchiuto and Bollermann [37]). For fronts over flat or downward sloping surfaces, the discrete equilibrium correctly induces the flow in the direction of the dry nodes. However, in adverse slopes, the discrete linear finite element approximation causes the momentum balance to produce spurious velocities downslope. This undesired behaviour can be avoided by the solution proposed by Brufau and García-Navarro [36], where the bed gradient is locally redefined to obey the equilibrium condition $\nabla z_b = -\nabla h$. For an element with an adverse slope, the bed elevation of its dry nodes is updated as:

$$z_{b_{\text{dry}}} = \max_{i \text{ is Wet}} (z_{b_i} + h_i). \tag{57}$$

In our implementation, the corrections are made locally during each element's matrix assembly operation.

Therefore, the adopted correction algorithm keeps the same discrete fluid volume and preserves the mass and the steady-state as it avoids the creation of nonphysical velocities. For fronts propagating over adverse slopes, we apply the same procedure. However, after each time step, to avoid some fluid quickly jumping to a dry node, we nullify the solution discharges using the same cut-off condition used for the velocities in Eq. (56).

Another aspect to be considered concerns how the friction term varies when h tends to zero. As we use the Manning formula to compute the drag coefficient, it could become arbitrarily large and bring numerical problems. Nonetheless, based on the approach of Heniche et al. [39] in which the bed friction

is increased when $h \to 0$ as a means to decrease and stabilize the flow, we linearly vary the Manning coefficient at dry nodes: $n_{\text{dry}} = n[1 + \beta(h_{\text{dry}} - h)]$, with $\beta = 10^2$.

At last, in the stabilized formulation, the previously defined SUPG and shock-capturing operators are nullified when evaluating a dry quadrature point's contribution.

4 Numerical results

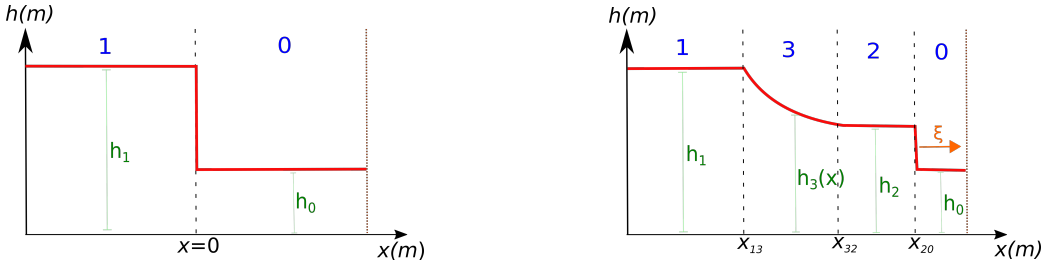
In this section, numerical results of some test cases are presented, and comparisons are made with analytical or literature available solutions. Also, when it is relevant to compute the mass/volume relative error V_{error}, we use the expression:

$$V_{\text{error}}(t) = \frac{V(t) - V(0)}{V(0)}, \tag{58}$$

where $V(t)$ is the volume stored in the domain at the time t.

4.1 1D dam break

The first problem we solve is the 1D dam break configuration where, initially, two reservoirs containing the same fluid are separated by a dam. One reservoir has fluid with depth h_1, while, in the other, the fluid height is h_0, with $h_1 > h_0$. At a given time, the barrier is instantly removed. Thus, we calculate the subsequent fluid flow. For this arrangement, Stoker [40] presents the analytical solution we use as a reference. A schematic representation of the solution and the initial state of the problem can be seen in Fig. 2.



(a) (b)

Figure 2. Schematic representations of the 1D dam break problem (a) and its analytical solution (b). The numbers represent the different regions of the initial state and the analytical solution.

The exact solution can be divided into 4 regions. In the transition between regions 2 and 0, a shock wave propagates to the right with speed ξ. Also, the fluid has zero velocity in regions 1 and 0 ($u_1 = u_0 = 0\,\text{m}\,\text{s}^{-1}$). Let the propagation speed of a perturbation on the surface of each region be defined as $c_i = \sqrt{g h_i}$, with $i = 0, 1, 2, 3$. Then, the shock wave speed is obtained by solving the nonlinear equation:

$$\frac{\xi}{2} + \frac{c_0^2}{8\xi}\left[1 + \sqrt{1 + 8\left(\frac{\xi}{c_0}\right)^2}\right] + \left[\frac{c_0^2}{2}\left(\sqrt{1 + 8\left(\frac{\xi}{c_0}\right)^2} - 1\right)\right]^{1/2} = c_1\,, \tag{59}$$

and the result is used as an input to the exact solution computation.

In sequence, it is defined:

$$c_2 = c_0\left[\frac{1}{2}\sqrt{1 + 8\left(\frac{\xi}{c_0}\right)^2} - \frac{1}{2}\right]^{1/2}, \qquad c_3 = \frac{1}{3}\left(2c_1 - \frac{x}{t}\right), \tag{60}$$

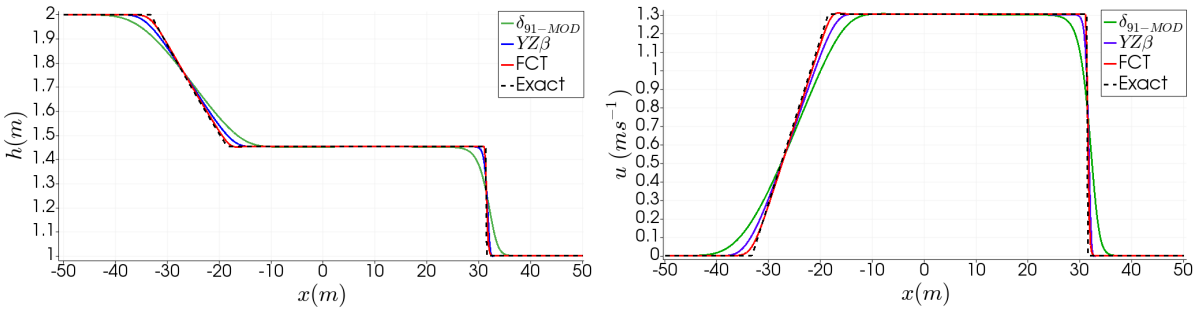
$$u_2 = \xi - \frac{c_0^2}{4\xi}\left(1 + \sqrt{1 + 8\left(\frac{\xi}{c_0}\right)^2}\right), \qquad u_3 = \frac{2}{3}\left(c_1 + \frac{x}{t}\right), \tag{61}$$

where t is the elapsed time since the dam's removal. So, the coordinates of each transition between regions can be defined as:

$$x_{13} = -c_1 t, \qquad x_{32} = (u_2 - c_2)t, \qquad x_{20} = \xi t. \tag{62}$$

Here we consider $x \in [-50, 50]\,\mathrm{m}$, $t \in [0, 10]\,\mathrm{s}$, $h_1 = 2\,\mathrm{m}$ and $h_0 = 1\,\mathrm{m}$, and compute $\xi = 4.183\,128\,\mathrm{m\,s^{-1}}$. Simulation is performed on a 2D regular mesh with 302×4 rectangular elements comprising a $100\,\mathrm{m} \times 1\,\mathrm{m}$ area. Non-penetration boundary conditions are applied at the limits of the numerical domain. Also, we consider a frictionless horizontal bottom. To compute the $YZ\beta$ operator, we set $(h)_{\text{ref}} = h_1$, $(q_x)_{\text{ref}} = h_1 c_1$ and $(q_y)_{\text{ref}} = 10^{10}\,\mathrm{m^2\,s^{-1}}$.

Figure 3 presents the exact solution and the results obtained with the stabilized and FCT methods at $t = 7.5\,\mathrm{s}$. We observe that the results obtained with the FCT scheme are in better consonance with the analytical solution than the ones produced by the stabilized techniques that use the δ_{91-MOD} and $YZ\beta$ shock-capturing operators.



(a) (b)

Figure 3. Exact and simulated solutions at $t = 7.5\,\mathrm{s}$ of the 1D dam break problem. (a) Fluid height; (b) Water velocity.

4.2 Transcritical flow with a shock

We simulate the steady-state solution of a transcritical flow that presents a hydraulic shock. Following Delestre et al. [41], we neglect frictional forces and viscous stresses ($\mu = 0\,\mathrm{Pa\,s}$). The simulation domain is a $25\,\mathrm{m} \times 5\,\mathrm{m}$ area whose bed elevation is:

$$z_b(x) = \begin{cases} 0.2 - 0.05(x - 10)^2, & \text{if } 8\,\mathrm{m} < x < 12\,\mathrm{m}, \\ 0, & \text{otherwise.} \end{cases} \tag{63}$$

The adopted boundary conditions are $q(x = 0) = q_0 = 0.18\,\mathrm{m^2\,s^{-1}}$ and $h(x = L = 25) = h_L = 0.33\,\mathrm{m}$, while the initial state is $q(x) = q_0$ and $h(x) = h_L$. According to Delestre et al. [41], for this configuration, the exact solution at the steady state can be computed by solving:

$$h(x)^3 + \left(z(x) - \frac{q_0^2}{2gh_c} - h_c - z_{\max}\right)h(x)^2 + \frac{q_0^2}{2g} = 0, \ \forall x \in [0, x_{\text{shock}}), \tag{64}$$

$$h(x)^3 + \left(z(x) - \frac{q_0^2}{2gh_L} - h_L\right)h(x)^2 + \frac{q_0^2}{2g} = 0, \ \forall x \in (x_{\text{shock}}, L], \tag{65}$$

$$q_0^2\left(\frac{1}{h_1^2} - \frac{1}{h_2^2}\right) + \frac{g}{2}\left(h_1^2 - h_2^2\right) = 0, \ \text{for } x = x_{\text{shock}}, \tag{66}$$

where $h_c = (q_0^2/g)^{1/3}$ is the critical water level at the subcritical to supercritical transition, $z_{\max} = 0.2\,\mathrm{m}$ is the maximum bed elevation. $h_1 = h^-(x = x_{\text{shock}})$ and $h_2 = h^+(x = x_{\text{shock}})$ are the water height upstream and downstream of the shock. The shock position x_{shock} can be obtained by solving Eq. (66). For the present case, $x_{\text{shock}} \approx 11.665\,615\,\mathrm{m}$ and $h_c \approx 0.148\,922\,\mathrm{m}$. Computation is performed on a mesh with 200×10 quadrangular elements. To compute the $YZ\beta$ operator, we set $(h)_{\text{ref}} = h_L$ and $(q_x)_{\text{ref}} = (q_y)_{\text{ref}} = 10^{10}\,\mathrm{m^2\,s^{-1}}$. We also tested it with $(h)_{\text{ref}} = h_L$, $(q_x)_{\text{ref}} = q_0$ and $(q_y)_{\text{ref}} = 10^{10}\,\mathrm{m^2\,s^{-1}}$, but the obtained solution was too oscillatory.

A comparison between the analytical and simulated solutions can be seen in Fig. 4. We observe that the use of the δ_{91-MOD} operator produced an undesired oscillation to the left of the mound. In contrast, the $YZ\beta$ approach created a nonphysical peak at the downstream side of the shock. At last, the FCT scheme generated the best results, which are in good agreement with the exact solution.

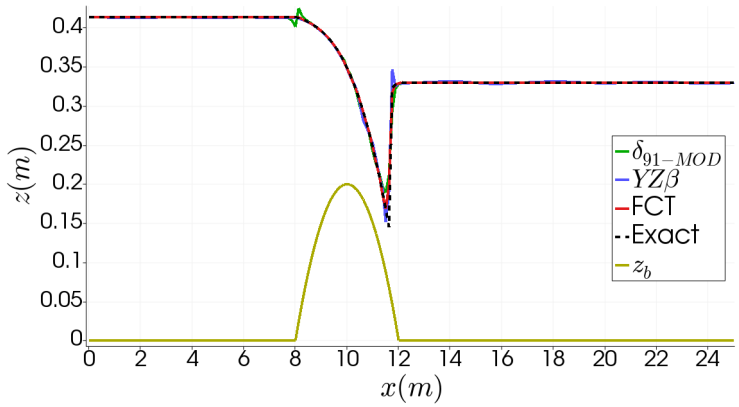


Figure 4. Exact and simulated water elevation ($\eta = h + z_b$) solutions of the transcritical flow with a shock problem.

4.3 Asymmetric dam break

We simulate the frictionless inviscid flow triggered by the instantaneous break of the dam separating two reservoirs connected by a channel. The initial water height at each reservoir is $h_1 = 10\,\mathrm{m}$ and $h_2 = 5\,\mathrm{m}$. Here, we run the simulations up to $t = 7.2\,\mathrm{s}$. Also, we enforce non-penetration boundary conditions. A diagram of the domain's geometry, dam's placement and the initial fluid height distribution can be seen in Fig. 5a. Plus, Fig. 5b depicts a detailed view of the employed 13488-element mesh and the initial water heights near the dam. To compute the $YZ\beta$ operator, let $h_2 = 7.27\,\mathrm{m}$ and $u_2 = 2.92\,\mathrm{m\,s^{-1}}$ be the fluid height and velocity at the region 2 of the 1D dam break problem that has an initial fluid height distribution analogue to the present 2D case. Then, we set $(h)_{\text{ref}} = h_2$, $(q_x)_{\text{ref}} = h_2 u_2$ and $(q_y)_{\text{ref}} = 10^{10}\,\mathrm{m^2\,s^{-1}}$.

Figure 6 shows the final water height distribution with 40 contours between $h = 5\,\mathrm{m}$ and $h = 10\,\mathrm{m}$. We observe that both stabilized formulations produced undesired water height perturbations past the wavefront in the right-most reservoir, where the water surface should be flat, as reproduced by the FCT scheme. Figure 7 compares our results with the solutions obtained by Ricchiuto et al. [42] and Ricchiuto [43]. In general, our results are in sound agreement with these reference solutions, even considering the 3D views and contours presented in their respective papers. Among the tested approaches, the one with the $YZ\beta$ operator produced a smoother profile, and the FCT scheme created the sharpest. Throughout the simulations, the stabilized methods' volume errors remained under 10^{-13}, while the FCT's volume error stayed under 10^{-9}. All in all, we remark that the FCT technique produced better results as it did not create spurious perturbations in regions where the shock wave still had not arrived, and the obtained height profiles are closer to the reference solutions.

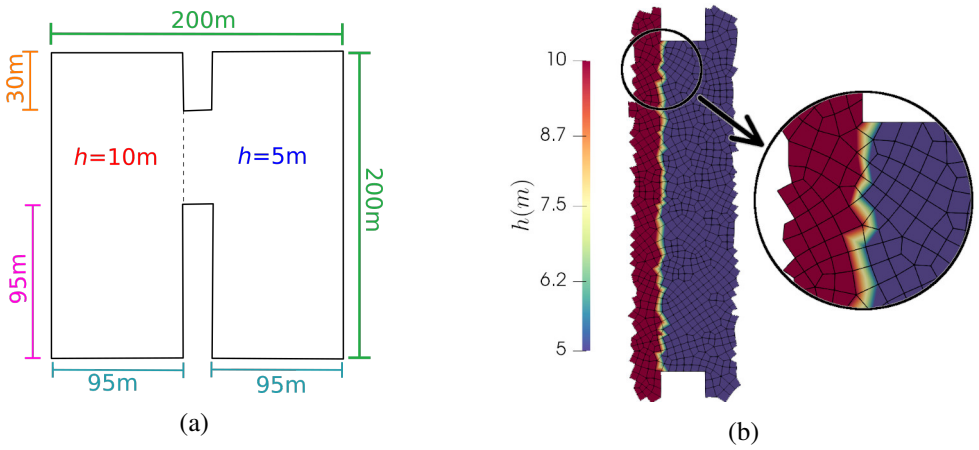


Figure 5. Initial configuration of the asymmetric dam break problem. (a) Domain's geometry and initial fluid height distribution; (b) Detail of the mesh and the initial water heights near the dam.

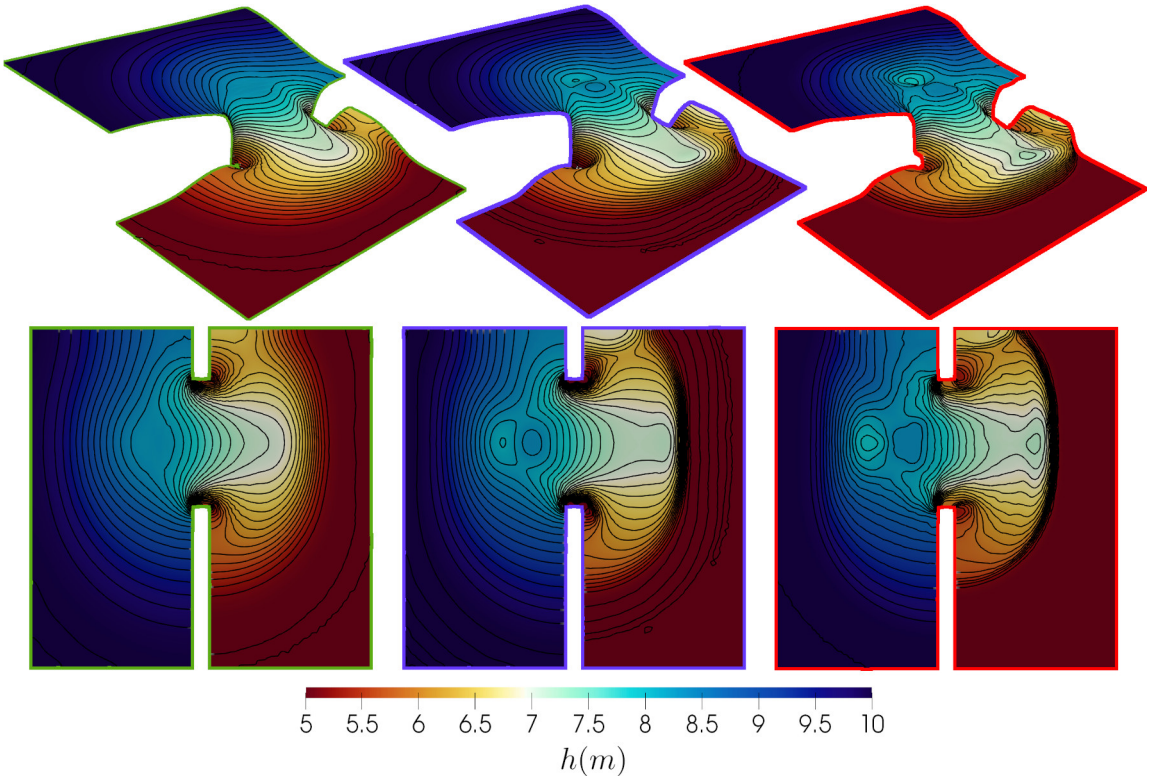
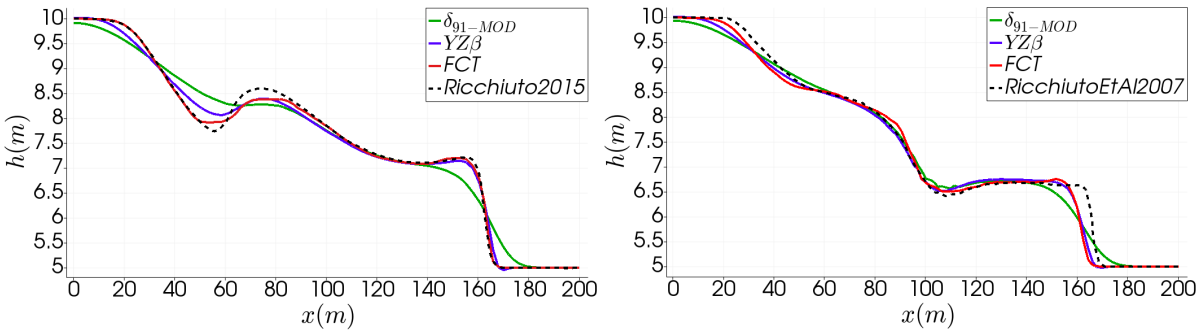


Figure 6. 3D and map views of the water surface with height contours obtained at $t = 7.2\,\mathrm{s}$ of the asymmetric dam break problem. The ones with the green and blue borders were computed with the stabilized approach using, respectively, the δ_{91-MOD} and $YZ\beta$ shock-capturing operators. Those with the red borders are outcomes of the FCT scheme.

4.4 Dam break over a channel with bumps

We simulate the flow generated by a dam break over an initially dry bed that presents three bumps. This problem was introduced by Kawahara and Umetsu [44], being later revisited by Brufau and García-Navarro [36], and Liang and Borthwick [45]. The initial dam encloses a reservoir $16\,\mathrm{m}$ long that contains water $1.875\,\mathrm{m}$ deep. Here, the $75\,\mathrm{m} \times 30\,\mathrm{m}$ domain is discretized with $1\,\mathrm{m}^2$ elements. Also, non-penetration boundary conditions are enforced during the simulations carried out until $t = 300\,\mathrm{s}$. As



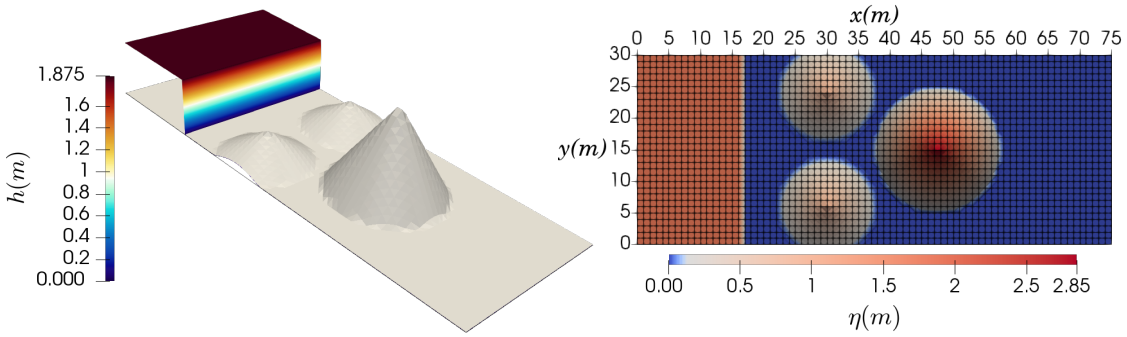
(a) (b)

Figure 7. Computed and reference solutions for the asymmetric dam break problem plotted along two sections at $t = 7.2\,\mathrm{s}$. In (a), the reference is the result of Ricchiuto [43] at $y = 132\,\mathrm{m}$. In (b), the reference is the solution of Ricchiuto et al. [42] at $y = 160\,\mathrm{m}$.

employed by Liang and Borthwick [45], the bed elevation is defined by:

$$z_b(x,y) = \max\left[0,\, 1 - \frac{1}{8}\sqrt{(x-30)^2 + (y-6)^2}\,,\, 1 - \frac{1}{8}\sqrt{(x-30)^2 + (y-24)^2},\right.$$
$$\left. 3 - \frac{3}{10}\sqrt{(x-47.5)^2 + (y-15)^2}\right]. \quad (67)$$

To compute the $YZ\beta$ operator, we set $(h)_{\mathrm{ref}} = 1.875\,\mathrm{m}$ and $(q_x)_{\mathrm{ref}} = (q_y)_{\mathrm{ref}} = 10^{10}\,\mathrm{m^2\,s^{-1}}$. A 3D view of the problem's initial state can be seen in Fig. 8a. Figure 8b depicts the simulation mesh and the initial water surface elevation ($\eta = z_b + h$).



(a) (b)

Figure 8. 3D view of the initial state (a) and 2D map of the simulation mesh (b) of the dam break over a channel with bumps problem.

Results obtained at key simulation times are presented in Fig. 9, while Fig. 10 shows free surface profiles computed along a section at $y = 15\,\mathrm{m}$ and $t = 300\,\mathrm{s}$. After the dam release, the flooding front advances and covers the small mounds, generating a reflection wave upstream. At $t = 12\,\mathrm{s}$, the initial wave is being reflected by the highest mound, creating another upstream-directed wave. Meanwhile, some of the incoming fluid goes downstream around the mound following the top and bottom walls. By $t = 30\,\mathrm{s}$, the reflection of the first wave on the right wall is climbing the largest bump from its right side. Then, at $t = 300\,\mathrm{s}$, after a series of reflections and dry/wet transitions, all bumps are left partially submerged, and the water surface remains flat.

In general, the results achieved with the $YZ\beta$ and FCT techniques are in better agreement with the works of Liang and Borthwick [45] and Guermond et al. [33]. At $t = 12\,\mathrm{s}$, among the tested approaches, the FCT scheme produced the sharpest water surface, as can be seen by the larger maximum fluid height and the more detailed fringes that spread downstream along the top and bottom walls. At $t = 30\,\mathrm{s}$, we

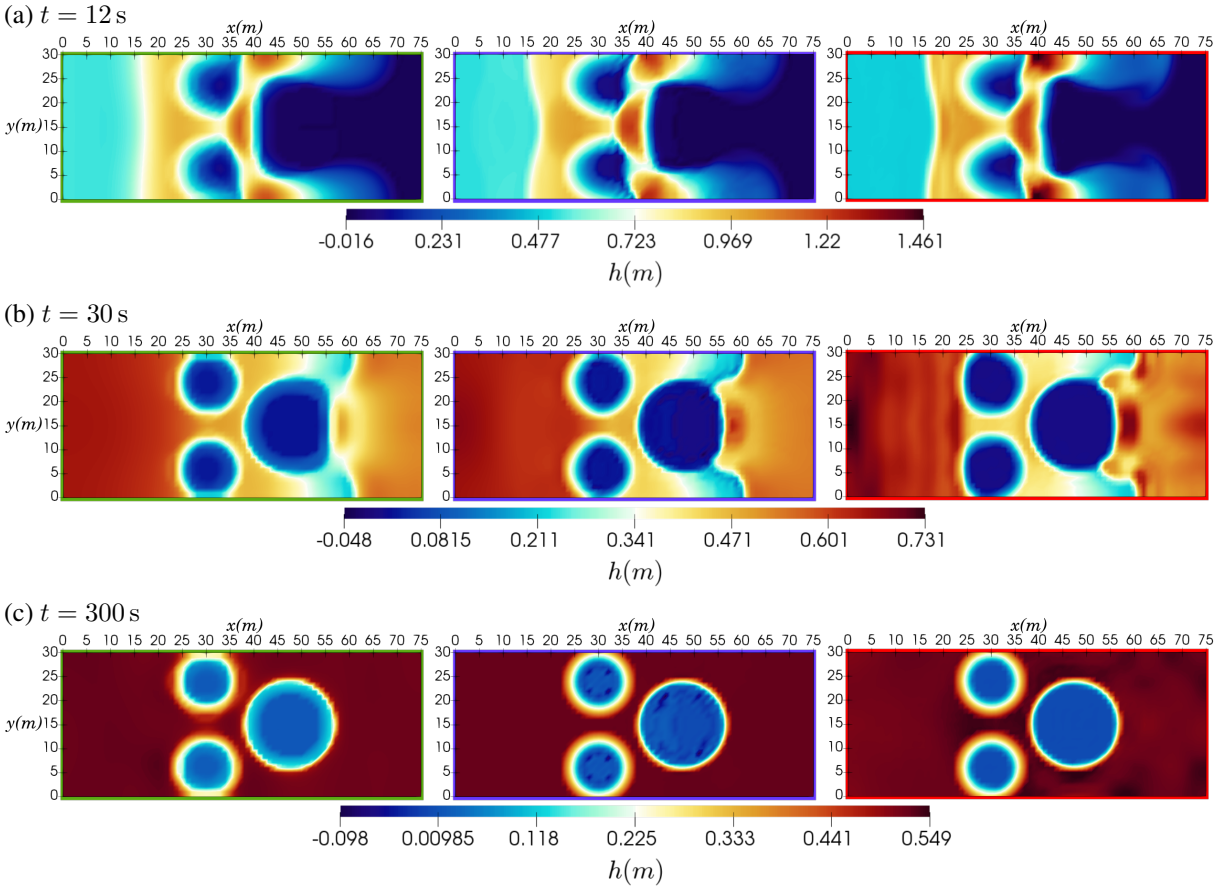


Figure 9. Map views of the obtained water heights at key simulation times of the dam break over a channel with bumps problem. The green and blue borders indicate results computed with the stabilized approach using, respectively, the δ_{91-MOD} and $YZ\beta$ shock-capturing operators. The maps with red borders are outcomes of the FCT scheme.

observe that the FCT scheme presented some ripples to the left of the mounds, while the other methods' surfaces are flat. By the end of the simulation, the solution of the δ_{91-MOD} technique has fluid in all the domain, and the $YZ\beta$ formulation has made some fluid go up the highest bump, creating negative fluid heights at a few points. In this case, the FCT method best represented the dry and wet regions and their transitions. This can be seen in Fig. 9c by the water height distribution near and at the bumps' regions, and the water surface profile in Fig. 10. In terms of the volume error, the δ_{91-MOD} and FCT approaches presented similar errors of the order of 10^{-10}, while the $YZ\beta$ technique showed an error of the order of 10^{-2}.

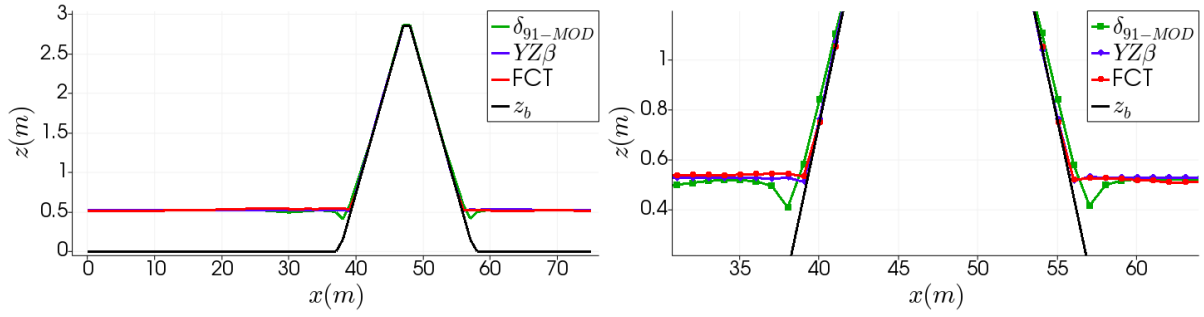


Figure 10. Water surface elevation along a section at $y = 15\,\mathrm{m}$ and $t = 300\,\mathrm{s}$ (a) and detail of the free surface near the highest bump (b) of the dam break over a channel with bumps problem.

5 Conclusion

We examined the use of two finite element approaches to solve the shallow water equations. The first one is a stabilized method built by adding Streamline Upwind Petrov-Galerkin and shock-capturing terms to the Galerkin formulation. From the operators studied by Santos and Coutinho [18], we adopted the SUPG operator presented by Takase et al. [24] and the $YZ\beta$ (Rispoli et al. [25]) and δ_{91-MOD} (Rispoli and Saavedra [27]) shock-capturing operators. In this case, time integration is performed with a predictor multi-corrector algorithm. The second technique is the flux-corrected transport scheme introduced by Santos et al. [28]. Here, the low-order formulation is created by adding a Rusanov-like scalar dissipation scaled by a shock-capturing operator to standard Galerkin equations. Meanwhile, its high-order system is composed by summing, to the low-order one, limited anti-diffusive fluxes linearized around the low-order solution. Limiting is performed with a Zalesak flux limiter that considers the hydrostatic reconstruction of the fluid's height, together with a minmod prelimiter. Here, an iterative nonlinear implicit time integration scheme is employed.

With both techniques, as the fluid height tends to zero, velocities are desingularized using a cut-off value based on the local ratio between element and mesh sizes. Also, the bed elevation at dry nodes is corrected to avoid unnatural dynamics due to the discretization of the fluid and bed surfaces, and we linearly vary the bed friction near the bed to help stabilize the flow. Besides, time steps are adaptively updated throughout the simulation to enforce a maximum CFL constraint. Also, in this work, all implementations related to these finite element techniques were aided by the deal.II library (Bangerth et al. [29]).

We evaluated their performance in several test problems and found that the FCT scheme is more robust, presenting good results in all the cases tested. Regarding the stabilized method, the $YZ\beta$ technique also produced plausible results. However, its usage requires some tweaking with the reference values for the variables h, q_x and q_y. We remark that the perturbations it created past the advancing wavefront in Section 4.3 might have a greater impact on simulations with dry and wet cells. For the example in Section 4.4, it made some fluid go up the larger bump and produced negative height values at some dry points. Thus, this method might be unsuitable to regions with more irregular terrains and more complex flow dynamics.

As future work, we suggest the evaluation of other stabilized and flux-corrected transport approaches, such as those presented by Ortiz [46] and Castro [14], possibly coupled with different dry/wet handling procedures, like the one proposed by Barros et al. [47]. Besides, other finite element techniques could be evaluated, such as residual distribution schemes (Ricchiuto [43]), a characteristic-based split (CBS) method (Ortiz et al. [48]), or even discontinuous Galerkin techniques (Gandham et al. [49]). Plus, we can perform an analysis regarding the effect of different mesh sizes on the simulation, which could also include adaptive mesh refinement and coarsening.

At last, the studied methods should be applied to real-world scenarios, where the terrain is uneven and the flow may pass from subcritical to supercritical (and vice-versa) in different portions of the domain. In these cases, the implementation of absorbing boundary conditions (Paz et al. [50]) might be necessary, especially if different parts of the boundary switch between inlet/outlet or subcritical/supercritical states during the simulation.

References

[1] Giraldo, F., 2014. Continuous and discontinuous galerkin methods for atmospheric modeling. In *Seminar on Recent Developments in Numerical Methods for Atmosphere and Ocean Modelling, 2-5 September 2013*, pp. 167–181, Shinfield Park, Reading. ECMWF.

[2] Sármány, D. & Hubbard, M., 2013. Upwind residual distribution for shallow-water ocean modelling. *Ocean Modelling*, vol. 64, pp. 1–11.

[3] Karel'skii, K. V., Petrosyan, A. S., & Chernyak, A. V., 2013. Nonlinear theory of the compressible

gas flows over a nonuniform boundary in the gravitational field in the shallow-water approximation. *Journal of Experimental and Theoretical Physics*, vol. 116, n. 4, pp. 680–697.

[4] Klimachkov, D. A. & Petrosyan, A. S., 2016. Nonlinear theory of magnetohydrodynamic flows of a compressible fluid in the shallow water approximation. *Journal of experimental and theoretical physics*, vol. 123, n. 3, pp. 520–539.

[5] Creed, M., Apostolidou, I.-G., Taylor, P., & Borthwick, A., 2016. A finite volume shock-capturing solver of the fully coupled shallow water-sediment equations. *International Journal for Numerical Methods in Fluids*. Fld.4359.

[6] de Luna, T. M., Castro-Díaz, M. J., Madroñal, C. P., & Fernández-Nieto, E. D., 2009. On a shallow water model for the simulation of turbidity currents. *Communications in Computational Physics*, pp. 848–882.

[7] Meiburg, E., Radhakrishnan, S., & Nasr-Azadani, M., 2015. Modeling gravity and turbidity currents: computational approaches and challenges. *Applied Mechanics Reviews*, vol. 67, n. 4, pp. 040802.

[8] Groenenberg, R. M., Sloff, K., & Weltje, G. J., 2009. A high-resolution 2-DH numerical scheme for process-based modeling of 3-D turbidite fan stratigraphy. *Computers & Geosciences*, vol. 35, n. 8, pp. 1686–1700.

[9] Hou, J., Liang, Q., Zhang, H., & Hinkelmann, R., 2015. An efficient unstructured MUSCL scheme for solving the 2D shallow water equations. *Environ. Model. Softw.*, vol. 66, n. C, pp. 131–152.

[10] Ambati, V. & Bokhove, O., 2007. Space–time discontinuous Galerkin finite element method for shallow water flows. *Journal of Computational and Applied Mathematics*, vol. 204, n. 2, pp. 452–462. Special Issue: The Seventh International Conference on Mathematical and Numerical Aspects of Waves (WAVES'05).

[11] Hervouet, J., 2007. *Solving transport equations*, chapter 6, pp. 177–194. John Wiley & Sons, Ltd.

[12] Castro, M. J., García-Rodríguez, J. A., González-Vida, J. M., Macías, J., & Parés, C., 2007. Improved fvm for two-layer shallow-water models: Application to the strait of gibraltar. *Advances in Engineering Software*, vol. 38, n. 6, pp. 386–398. Advances in Numerical Methods for Environmental Engineering.

[13] Behzadi, F., 2016. *Solution of fully-coupled shallow water equations and contaminant transport using a primitive variable Riemann solver and a semi-discrete SUPG method*. PhD thesis, University of Tennessee.

[14] Castro, R. S., 2014. Space-time finite element formulation for shallow water equations with shock-capturing operator. *Pesquimat*, vol. 3, n. 1.

[15] Takase, S., Kashiyama, K., Tanaka, S., & Tezduyar, T., 2011. Space-time SUPG finite element computation of shallow-water flows with moving shorelines. *Computational Mechanics*, vol. 48, n. 3, pp. 293–306.

[16] Hughes, T. J. R. & Mallet, M., 1986. A new finite element formulation for computational fluid dynamics: III. The generalized streamline operator for multidimensional advective-diffusive systems. *Computer Methods in Applied Mechanics and Engineering*, vol. 58, n. 3, pp. 305–328.

[17] Galeão, A. C. & do Carmo, E. G. D., 1988. A consistent approximate upwind Petrov-Galerkin method for convection-dominated problems. *Computer Methods in Applied Mechanics and Engineering*, vol. 68, n. 1, pp. 83 – 95.

[18] Santos, T. L. & Coutinho, A. L. G. A., 2017. A continuous finite element approach to the well-balanced shallow water equations. *CILAMCE 2017 – XXXVIII Ibero-Latin American Congress on Computational Methods in Engineering*.

[19] Kuzmin, D., Möller, M., & Turek, S., 2003. Multidimensional FEM-FCT schemes for arbitrary time stepping. *International Journal for Numerical Methods in Fluids*, vol. 42, n. 3, pp. 265–295.

[20] Sheu, T. W. H. & Fang, C. C., 2001. High resolution finite-element analysis of shallow water equations in two dimensions. *Computer Methods in Applied Mechanics and Engineering*, vol. 190, n. 20–21, pp. 2581–2601.

[21] Ortiz, P., Anguita, J., & Riveiro, M., 2015. Free surface flows over partially erodible beds by a continuous finite element method. *Environmental Earth Sciences*, vol. 74, n. 11, pp. 7357–7370.

[22] Kuzmin, D., 2009. Explicit and implicit FEM-FCT algorithms with flux linearization. *Journal of Computational Physics*, vol. 228, n. 7, pp. 2517 – 2534.

[23] Tezduyar, T. E., 2004. *Finite element methods for fluid dynamics with moving boundaries and interfaces*, chapter 17. John Wiley & Sons, Ltd.

[24] Takase, S., Kashiyama, K., Tanaka, S., & Tezduyar, T. E., 2010. Space-time SUPG formulation of the shallow-water equations. *International Journal for Numerical Methods in Fluids*, vol. 64, n. 10-12, pp. 1379–1394.

[25] Rispoli, F., Saavedra, R., Corsini, A., & Tezduyar, T. E., 2007. Computation of inviscid compressible flows with the V-SGS stabilization and YZβ shock-capturing. *International Journal for Numerical Methods in Fluids*, vol. 54, n. 6-8, pp. 695–706.

[26] Tezduyar, T. E. & Senga, M., 2007. Supg finite element computation of inviscid supersonic flows with yzβ shock-capturing. *Computers & Fluids*, vol. 36, n. 1, pp. 147 – 159. Challenges and Advances in Flow Simulation and Modeling.

[27] Rispoli, F. & Saavedra, G. Z. R., 2006. A stabilized finite element method based on sgs models for compressible flows. *Computer Methods in Applied Mechanics and Engineering*, vol. 196, n. 1, pp. 652 – 664.

[28] Santos, T. L., Lopes, A. A. O., & Coutinho, A. L. G. A., 2019. A shallow water event-driven approach to simulate turbidity currents at stratigraphic scale. Manuscript submitted for publication.

[29] Bangerth, W., Hartmann, R., & Kanschat, G., 2007. deal.II – a general-purpose object-oriented finite element library. *ACM Transactions on Mathematical Software*, vol. 33, n. 4.

[30] Tadmor, E. & Zhong, W., 2008. *Energy-Preserving and Stable Approximations for the Two-Dimensional Shallow Water Equations*, pp. 67–94. Springer Berlin Heidelberg, Berlin, Heidelberg.

[31] Aliabadi, S. K. & Tezduyar, T. E., 1995. Parallel fluid dynamics computations in aerospace applications. *International Journal for Numerical Methods in Fluids*, vol. 21, n. 10, pp. 783–805.

[32] Saad, Y. & Schultz, M. H., 1986. GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Stat. Comput.*, vol. 7, n. 3, pp. 856–869.

[33] Guermond, J.-L., de Luna, M. Q., Popov, B., Kees, C. E., & Farthing, M. W., 2018. Well-balanced second-order finite element approximation of the shallow water equations with friction. *SIAM Journal on Scientific Computing*, vol. 40, n. 6, pp. A3873–A3901.

[34] Audusse, E., Bouchut, F., Bristeau, M.-O., Klein, R., & Perthame, B., 2004. A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows. *SIAM Journal on Scientific Computing*, vol. 25, n. 6, pp. 2050–2065.

[35] Hsu, C.-J., 1981. Numerical heat transfer and fluid flow. *Nuclear Science and Engineering*, vol. 78, n. 2, pp. 196–197.

[36] Brufau, P. & García-Navarro, P., 2003. Unsteady free surface flow simulation over complex topography with a multidimensional upwind technique. *Journal of Computational Physics*, vol. 186, n. 2, pp. 503 – 526.

[37] Ricchiuto, M. & Bollermann, A., 2009. Stabilized residual distribution for shallow water simulations. *Journal of Computational Physics*, vol. 228, n. 4, pp. 1071–1115.

[38] Kesserwani, G. & Liang, Q., 2010. Well-balanced RKDG2 solutions to the shallow water equations over irregular domains with wetting and drying. *Computers & Fluids*, vol. 39, n. 10, pp. 2040–2050.

[39] Heniche, M., Secretan, Y., Boudreau, P., & Leclerc, M., 2000. A two-dimensional finite element drying-wetting shallow water model for rivers and estuaries. *Advances in Water Resources*, vol. 23, n. 4, pp. 359 – 372.

[40] Stoker, J. J., 1992. *Water waves: The mathematical theory with applications*, volume 36. John Wiley & Sons, Inc.

[41] Delestre, O., Lucas, C., Ksinant, P.-A., Darboux, F., Laguerre, C., Vo, T.-N.-T., James, F., & Cordier, S., 2013. SWASHES: a compilation of shallow water analytic solutions for hydraulic and environmental studies. *International Journal for Numerical Methods in Fluids*, vol. 72, pp. 269–300.

[42] Ricchiuto, M., Abgrall, R., & Deconinck, H., 2007. Application of conservative residual distribution schemes to the solution of the shallow water equations on unstructured meshes. *Journal of Computational Physics*, vol. 222, n. 1, pp. 287–331.

[43] Ricchiuto, M., 2015. An explicit residual based approach for shallow water flows. *Journal of Computational Physics*, vol. 280, pp. 306–344.

[44] Kawahara, M. & Umetsu, T., 1986. Finite element method for moving boundary problems in river flow. *International Journal for Numerical Methods in Fluids*, vol. 6, n. 6, pp. 365–386.

[45] Liang, Q. & Borthwick, A. G. L., 2009. Adaptive quadtree simulation of shallow flows with wet–dry fronts over complex topography. *Computers & Fluids*, vol. 38, n. 2, pp. 221–234.

[46] Ortiz, P., 2014. Shallow water flows over flooding areas by a flux-corrected finite element method. *Journal of Hydraulic Research*, vol. 52, n. 2, pp. 241–252.

[47] Barros, M. L. C., Rosman, P. C. C., & Telles, J. C. F., 2015. An effective wetting and drying algorithm for numerical shallow water flow models. *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, vol. 37, n. 3, pp. 803–819.

[48] Ortiz, P., Zienkiewicz, O. C., & Szmelter, J., 2006. Hydrodynamics and transport in estuaries and rivers by the CBS finite element method. *International Journal for Numerical Methods in Engineering*, vol. 66, n. 10, pp. 1569–1586.

[49] Gandham, R., Medina, D., & Warburton, T., 2015. GPU accelerated discontinuous Galerkin methods for shallow water equations. *Communications in Computational Physics*, vol. 18, n. 1, pp. 37–64.

[50] Paz, R. R., Storti, M. A., & Garelli, L., 2009. Absorbing boundary condition for nonlinear hyperbolic partial differential equations with unknown Riemann invariants. *Mecánica Computacional*, vol. 28, pp. 1593–1620.