Computers and Fluids 132 (2016) 76–93

Contents lists available at ScienceDirect

Computers and Fluids

journal homepage: www.elsevier.com/locate/compfluid

An ADER-type scheme for a class of equations arising from the water-wave theory

G.I. Montecinos[a,*], J.C. López-Rios[b], R. Lecaros[a], J.H. Ortega[a], E.F. Toro[c]

[a] Center for Matematical Modeling, CMM, Universidad de Chile, Santiago, Chile
[b] Basque Center for Applied Mathematics, BCAM, Basque country, Spain
[c] Laboratory of Applied Mathematics, DICAM, University of Trento, Trento, Italy

ARTICLE INFO

Article history:
Received 5 January 2016
Revised 31 March 2016
Accepted 8 April 2016
Available online 9 April 2016

Keywords:
Finite volume schemes
ADER schemes
Generalized Riemann problems
Water waves equations

ABSTRACT

In this work we propose a numerical strategy to solve a family of partial differential equations arising from the water-wave theory. These problems may contain four terms; a source which is an algebraic function of the solution, a convective part involving first order spatial derivatives of the solution, a diffusive part involving second order spatial derivatives and the transient part. Unlike partial differential equations of hyperbolic or parabolic type, where the transient part is the time derivative of the solution, here the transient part can contain mixed time and space derivatives.

In [Zambra et al. International Journal for Numerical Methods in Engineering 89(2):227-240, 2012], the authors proposed a globally implicit strategy to solve the Richards equation. In that case, transient terms consisted of algebraic expressions of the solution. Motivated by this work, we propose a one-step finite volume method to deal with problems in which transient terms are differential operators. Here, a locally implicit formulation is investigated, which is based on the ADER philosophy. The scheme is divided in three steps: i) a polynomial reconstruction of the data; ii) solutions to Generalized Riemann Problems (GRP); iii) the solution of differential problems. Note that steps i) and ii), are those of conventional ADER schemes for conservation laws. Advantages of the present approach include the possibility to construct high-order approximations in both space and time, for which existing methodologies for hyperbolic problems can be applied. The differential problems associated to the transient term can be non-linear and numerical strategies can be adopted to deal wit it. Convergence of the scheme is proved rigorously and an empirical convergence rates assessment is carried out in order to illustrate the high space and time accuracy of the present scheme.

1. Introduction

The water-wave theory describes the motion of the free surface and the velocity field of an ideal, incompressible and irrotational fluid under the influence of gravity. The force acting on the free surface is expressed in terms of the Bernoulli equation for irrotational flows, in this form the pressure is included in terms of dynamical boundary conditions, [42,43]. The phenomena are governed by a strongly non-linear system of partial differential equations in a time dependent domain formed by a parametrized free surface profile and a bottom surface or topography. However, important simplifications can be achieved by neglecting some amplitude scales associated with the depth, topography and free surface. So, for example for large amplitude, the Green–Naghdi model [34] is derived. Similarly for small topography variations see [6], for medium amplitude topography variations see model in [53]. These problems may contain four terms; a source which is an algebraic function of the solution, a convective part involving first order spatial derivatives of the solution, a diffusive part involving second order spatial derivatives and the transient part. Unlike partial differential equations of hyperbolic or parabolic type, where the transient part is the time derivative of the solution, here the transient part can contain mixed time and space derivatives.

Numerical methods to solve equations derived from the water waves theory, include the splitting methods of Chazel et al. [20] and Bonneton et al. [14], where the Green–Naghdi equations have been solved. Due to the splitting nature these methods are at most second order of accuracy and the stability is governed by a condition on Courant–Friedrich–Levy (CFL) coefficient. The accuracy has been improved by using standard Galerkin-finite element method for spatial discretization and fourth-order explicit Runge–Kutta schemes for marching in time, in [1–3], Boussinesq equations

* Corresponding author. Tel.: +56229784599.
E-mail address: gmontecinos@dim.uchile.cl (G.I. Montecinos).

http://dx.doi.org/10.1016/j.compfluid.2016.04.012
0045-7930/

have been solved with that strategy. Some variants are proposed in [32,52,56], where KdV equations have been solved through finite difference schemes, whose stability is also governed by a CFL type constraint, these schemes are of second-order of accuracy in space. Spline functions for spatial discretization and implicit Runge–Kutta methods for marching in time have been proposed in [10–13]. Regarding accuracy, it is well known that the finite volume method provides accurate, efficient and robust schemes to approximate conservation laws. Due to that, of special interest for us is the work in [30], where the finite volume framework for conservation laws has been extended to approximate solutions of dispersive wave equations, in particular for the classical Boussinesq system. Numerical fluxes are employed to discretize the convective part, whereas dispersive terms are discretized by using centred finite differences, for marching in time a TVD Runge–Kutta method is used; the method is reported to be of third-order of accuracy. To our knowledge, recent contributions for solving this type of problems are [30,31,41,44,54]. A different strategy for the Green–Naghdi model is proposed in [45] and [47], the authors have introduced a new variable, which accounts for terms containing mixed time and spatial derivatives. It allows to reformulate the problem as a coupled system between a parabolic conservation law and an elliptic equation. In [45] a Godunov type scheme is employed to solve the parabolic conservation law, whereas, an ordinary differential equation is constructed to get sought solutions from the elliptic equation. In [47] a central discontinuous Galerkin method is used to solve the conservation laws, whereas, a continuous finite element method solves the parabolic equation. As wee will see later, this strategy can be extended and formulated in a high order finite volume framework.

In this paper we propose a one-step finite volume evolution of differential operators as those in [30], here based on Zambra et al. [70]. Here, a unified framework to discretize convective and dispersive terms will be presented. In [70], a numerical scheme to solve the Richards equation was proposed. An ADER type method was used to evolve the water content as function of the hydraulic pressure in a porous media. The ADER (arbitrary Accuracy DERivative Riemann problems) method was first put forward by Toro et al. [64], in the finite volume framework, for solving linear hyperbolic equations in one and multiple space dimensions on Cartesian meshes; see also Schwartzkopff, Munz and Toro [55]. The extension of ADER finite volume (ADER-FV) to non-linear equations, due to Titarev and Toro [58], is based on a semi-analytical, explicit solution of the generalized Riemann problem put forward by Toro and Titarev [62]. Since then, ADER has also been extended to the discontinuous Galerkin finite element framework by Dumbser [22], giving rise to ADER-DG schemes. For an elementary introduction to ADER schemes and the generalised Riemann problem, the reader is referred to chapters 19 and 20 of the textbook by Toro [65]. Further generalisations were put forward by Dumbser et al. [22], setting ADER-FV and ADER-DG in a generalised framework. The ADER approach has undergone numerous extensions and applications, examples include [4,5,15–17,19,21–24,26–29,38–40,48–51,57–60,62,63].

In general ADER are high-order finite volume methods, which contain two building blocks, a) a reconstruction procedure which is non-linear, and b) the solution of generalized Riemann problems. The ADER finite volume scheme in [70] required a predictor, which was obtained through a globally implicit space-time DG scheme proposed by van der Vegt and van der Ven [66,67]. In [70], the finite volume scheme evolved an algebraic equation of the unknown, hydraulic pressure. Therefore, motivated by this approach we propose a finite volume formulation for the time evolution of differential terms. In the present paper, an extension of the ADER-type scheme of Dumbser et al. [25], is used; predictors are obtained by solving a local weak formulation within cells. The discontinuous Galerkin approach is also used to compute the weak solutions. Then the (Harten–Engquist–Osher–Chakravarthy) HEOC approach, (see [18] for an extension to high order approach of the original Harten et al. method [37]), allows us to interact the predictors, evaluated at both sides of cell interfaces. To solve classical Riemann problems, required for the HEOC approach, a HLL type solver has been developed. This provides a high order approximations to the numerical flux, and a similar procedure is carried out for numerical source terms. Notice that this is a locally implicit method to obtain a prediction, which is distinct to the globally implicit strategy in [70]. Advantages of the present method include; the use of a CFL type condition, which ensures the high-order of accuracy in both space and time; the use of the Dumbser et al. [25] strategy, provides a suitable treatment of balance laws with stiff source terms, reconciling accuracy and stability. Furthermore, in this paper we prove analytically that the present algorithm converges and we carried out a systematic convergence rate assessment in order to illustrate that the present methodology reaches high-order of accuracy.

The paper is organized as follows. In Section 2 the governing equation and the numerical scheme are presented. In Section 3 a theoretical result regarding the convergence of the numerical scheme for the scalar and linear case is presented. In Section 4, suitable tests are solved. Finally, in Section 5 the conclusions are drawn.

2. The model system and the numerical scheme

In this work we are interested in systems of time-dependent partial differential equations in the form

$$\partial_t \mathcal{L}(\mathbf{Q}(x,t)) + \partial_x \mathbf{F}(\mathbf{Q}(x,t)) = \mathbf{S}(\mathbf{Q}(x,t)) \,, x \in [x_L, x_R] \,, t \in [0,T] \,, \tag{1}$$

where $\mathbf{Q} \in \mathbb{R}^m$ is the vector of unknown, $\mathbf{F}(\mathbf{Q})$ is the physical flux, $\mathbf{S}(\mathbf{Q})$ is the source term, $\mathcal{L}(\mathbf{Q})$ is a differential operator which only contains spatial-derivatives, so we can assume

$$\mathcal{L}(\mathbf{Q}) =: \mathcal{G}(\mathbf{Q}) + \mathcal{B}(\mathbf{Q}), \tag{2}$$

with

$$\mathcal{B}(\mathbf{Q}) = \sum_{k=1}^{r} a_k \partial_x^{(k)} \mathbf{Q}, \tag{3}$$

where $\mathcal{G}(\mathbf{Q})$ may be a non-linear algebraic function, $a_k's$ are constant values and r is the maximum order of the spatial derivatives, which in this work is assumed to be $r = 2$. Additionally, we assume that given a regular enough function $\mathbf{W}$, the problem

$$\left.\begin{aligned} \mathcal{L}(\mathbf{Q}) &= \mathbf{W}(x) \,, \\ \mathbf{Q}(x_L) = \mathbf{W}(x_L) \,, \mathbf{Q}(x_R) &= \mathbf{W}(x_R) \,, \end{aligned}\right\} \tag{4}$$

has an exact solution $\mathbf{Q}(x)$, where boundary conditions are given in terms of some prescribed functions $\mathbf{W}(x_L)$ and $\mathbf{W}(x_R)$.

In this work, to solve (1) we propose a numerical scheme based on the finite volume method. So we divide $[x_L, x_R]$ into N uniform cells $I_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$. Then, we integrate (1) on the space-time control volume $I_i^n = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [t^n, t^{n+1}]$, which provides the following one-step finite volume formula:

$$\mathbf{W}_i^{n+1} = \mathbf{W}_i^n - \frac{\Delta t}{\Delta x}\left[\mathbf{F}_{i+\frac{1}{2}} - \mathbf{F}_{i-\frac{1}{2}}\right] + \Delta t \mathbf{S}_i \,, \tag{5}$$

with

$$\left.\begin{aligned} \mathbf{W}_i^n &= \tfrac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \mathcal{L}(\mathbf{Q}(x,t^n))dx\,, \\ \mathbf{F}_{i+\frac{1}{2}} &= \tfrac{1}{\Delta t}\int_{t^n}^{t^{n+1}} \mathbf{F}(\mathbf{Q}(x_{i+\frac{1}{2}},t))dt\,, \\ \mathbf{S}_i &= \tfrac{1}{\Delta t}\tfrac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}}\int_{t^n}^{t^{n+1}} \mathbf{S}(\mathbf{Q}(x,t))dtdx\,. \end{aligned}\right\} \tag{6}$$

These integrals are exact if $\mathbf{Q}(x,\,t)$ is available inside I_i^n. However, due to cell average evolution we are only able to provide approximations to $\mathbf{F}_{i+\frac{1}{2}}$ and $\mathbf{S}_i$. In this paper an ADER type discretization is proposed to approximate these terms.

Once, $\{\mathbf{W}_i^{n+1}\}_{i=1}^N$ is available, $\{\mathbf{Q}_i^{n+1}\}_{i=1}^N$ is obtained from the equation

$$\mathbf{W}(x,t^{n+1}) = \mathcal{L}(\mathbf{Q}(x,t^{n+1}))\,, x\in[x_L,x_R]\,. \tag{7}$$

Here the solution is obtained numerically, so the discretization of Eq. (7) involves the discrete values $\mathbf{W}_i^{n+1}$ and $\mathbf{Q}_i^{n+1}$. Therefore, an algebraic equation for $\{\mathbf{Q}_i^{n+1}\}_{i=1}^N$ has to be solved.

In the following section we provide a succinct review of ADER schemes and GRP solvers.

2.1. ADER finite volume schemes and GRP solvers

The ADER finite volume approach computes high-order approximations to the integral averages (6), to obtain an ADER numerical flux and an ADER numerical source, if present. The ADER methodology is an extension of the second-order method of Ben-Artzi and Falcoviz [7]. The extension concerns the Generalised Riemann Problem (GRP) to evaluate the numerical flux, and is two-fold: (a) the initial condition for the GRP is piece-wise polynomials of any degree, and (b) the equations preserve their source terms, if present originally.

The ADER approach was first put forward by Toro et al. [64], for linear problems on Cartesian meshes, see also [55]. These methods are one-step schemes, fully discrete, containing two main ingredients to determine the numerical flux, namely (i) a high-order, non-linear spatial reconstruction procedure and (ii) solution of a generalised, or high order, Riemann problem (GRP) at each cell interface. If source terms are present, an additional, analogous step is required. Reconstructions should be non-linear, to circumvent Godunov's theorem [33,65].

The first practical solver for the GRP is due to Toro and Titarev [62]. Here, the authors computed numerical fluxes by evaluating the solution of GRP at fixed interface positions. In this solver, the solution is proposed as a Taylor series in time, where the leading term is computed as the solution to a possible non-linear classical Riemann problem and time derivatives are expressed in terms of spatial derivatives by using the well-known Cauchy–Kowalewkaya procedure. A sequence of linearized classical Riemann problems, provides the spatial derivatives. Later, Castro and Toro [18] have re-interpreted the Harten et al. solver [37] in terms of GRP's, which they have called the HEOC solver. Extrapolated solution at the interface position were interacted through a classical, and possible non-linear Riemann problem. Here, Taylor series expansions in time were proposed to obtain these extrapolated values. The Cauchy–Kowalewskaya procedure was again used to obtain extrapolations, and spatial derivatives were obtained from the derivatives of the reconstruction polynomials. Subsequently, Dumbser, Enaux and Toro [25] proposed to use the discontinuous Galerkin method to obtain a local polynomial predictor of the solution within cells. So extrapolation at the interface were computed by evaluating these predictors at both sides of the interface. Notice that this method is an alternative to the Harten et al. re-interpretation provided by Castro and Toro, which avoid the use of the Cauchy–Kowalewskaya procedure. A more recent GRP solver due to Toro and Montecinos, [61], regards an implicit re-interpretations of the existing solvers of Toro-Titarev and the HEOC solver of Castro and Toro.

In particular in this paper, the HEOC solver provided by Castro and Toro [18] is adopted. Therefore, the numerical flux and the source term are computed as

$$\begin{aligned} \mathbf{F}_{i+\frac{1}{2}} &= \tfrac{1}{\Delta t}\int_{t^n}^{t^{n+1}} \mathbf{F}_h(\mathbf{Q}_i(x_{i+\frac{1}{2}},t),\mathbf{Q}_{i+1}(x_{i+\frac{1}{2}},t))dt\,, \\ \mathbf{S}_i &= \tfrac{1}{\Delta t\Delta x}\int_{t^n}^{t^{n+1}}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \mathbf{S}(\mathbf{Q}_i(x,t))dxdt\,, \end{aligned} \tag{8}$$

with $\mathbf{F}_h(\mathbf{Q}_L,\,\mathbf{Q}_R)$ a function of two arguments $\mathbf{Q}_L$ and $\mathbf{Q}_R$, which stands for a numerical flux (classical Riemann solver), as we will see in Section 2.3, it is a HLL type solver. Here, $\mathbf{Q}_i(x,\,t)$ is a prediction of the solution within I_i^n. Therefore, $\mathbf{Q}_i(x_{i+\frac{1}{2}},t)$ and $\mathbf{Q}_{i+1}(x_{i+\frac{1}{2}},t)$ are the boundary extrapolated solutions on the left and right side of the cell interface $x=x_{i+\frac{1}{2}}$ at a given time t. The strategy to compute $\mathbf{Q}_i(x,\,t)$ in I_i^n is presented in the next section. This is motivated by the approach due to Dumbser, Enaux and Toro (DET), see [25].

2.2. The Dumbser–Enaux–Toro (DET) solver to compute the predictor

In this section we present the strategy to solve $\mathbf{Q}_i(x,\,t)$ in I_i^n. This is found by solving

$$\left.\begin{aligned} \partial_t\mathcal{L}(\mathbf{Q}(x,t)) + \partial_x\mathbf{F}(\mathbf{Q}(x,t)) &= \mathbf{S}(\mathbf{Q}(x,t))\,, \\ \mathbf{Q}(x,0) &= \mathbf{P}_i(x)\,, \end{aligned}\right\} \tag{9}$$

here $\mathbf{P}_i(x)$ is the polynomial resulting from the reconstruction procedure. For the sake of simplicity, we may consider problem (9) in terms of variables $0\le\xi\le 1$ and $0\le\tau\le 1$, which are related with x and t by $x(\xi)=x_{i-\frac{1}{2}}+\xi\Delta x$ and $t(\tau)=t^n+\tau\Delta t$. So, (9) takes the form

$$\left.\begin{aligned} \partial_\tau\mathcal{L}^*(\bar{\mathbf{Q}}(\xi,\tau)) + \partial_\xi\mathbf{F}^*(\bar{\mathbf{Q}}(\xi,\tau)) &= \mathbf{S}^*(\mathbf{Q}(x,t))\,, \\ \bar{\mathbf{Q}}(\xi,0) &= \mathbf{P}_i(x(\xi))\,, \end{aligned}\right\} \tag{10}$$

with $\mathbf{F}^*(\bar{\mathbf{Q}}) = \frac{\Delta t}{\Delta x}\mathbf{F}(\bar{\mathbf{Q}})$, $\mathbf{S}^*(\bar{\mathbf{Q}}) = \Delta t\mathbf{S}(\bar{\mathbf{Q}})$ and

$$\mathcal{L}^*(\bar{\mathbf{Q}}) = \Phi(\bar{\mathbf{Q}},\,\Delta x^{-1}\partial_\xi\bar{\mathbf{Q}},\,...,\,\Delta x^{-r}\partial_\xi^{(r)}\bar{\mathbf{Q}})\,.$$

It will be convenient to introduce the differential operator

$$\mathcal{B}^*(\bar{\mathbf{Q}}) := \sum_{k=1}^{r}\frac{a_k}{\Delta x^k}\partial_\xi^{(k)}\bar{\mathbf{Q}}\,. \tag{11}$$

Problem (10) is solved in a weak sense. To this end we set a Lagrangian polynomial basis in [0, 1] × [0, 1], which is denoted by

$$V = span\{\phi_1(\xi,\tau),\,...,\,\phi_{MD}(\xi,\tau)\}\,,$$

where $MD=M^2$ are the degrees of freedom of this polynomial space. Here $M+1$ is the order of accuracy of the approximation. Additionally, let us assume that reconstruction procedure, is supported by a polynomial space spanned by the Legendre polynomials in [0, 1] denoted by $\psi_1(\xi)$, ..., $\psi_M(\xi)$, hence,

$$\mathbf{P}_i^*(\xi) := \mathbf{P}_i(x(\xi)) = \sum_{k=1}^{M}\psi_k(\xi)\hat{\mathbf{w}}_k\,.$$

Therefore, multiplying by a test function, we have

$$\begin{aligned} &\int_0^1\int_0^1 \partial_\tau\bar{\mathcal{L}}\phi_j(\xi,\tau)d\xi d\tau + \int_0^1\int_0^1 \partial_\xi\mathbf{F}^*\phi_j(\xi,\tau)d\xi d\tau \\ &\quad = \int_0^1\int_0^1 \partial_\xi\mathbf{S}_k^*\phi_j(\xi,\tau)d\xi\,. \end{aligned} \tag{12}$$

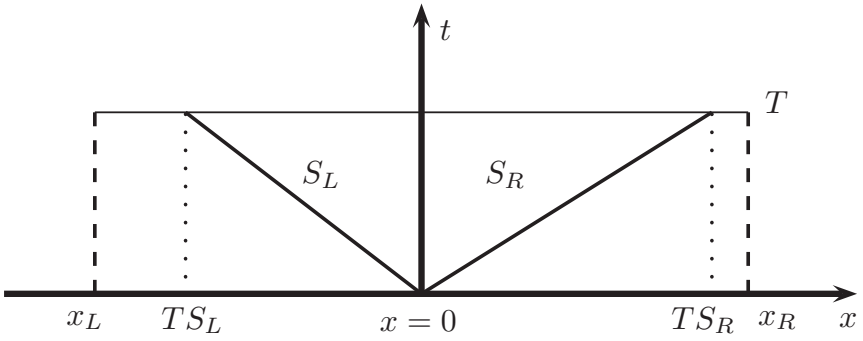

Fig. 1. Sketch of the wave model for HLL type flux.

Now, we propose a solution in V, given by

$$\bar{\mathbf{Q}}_i(\xi,\tau)=\sum_{l=1}^{MD}\phi_l(\xi,\tau)\hat{Q}_l^i\,.$$

As the basis is nodal we consider the following approximations:

$$\begin{aligned}\mathcal{G}(\mathbf{Q})&=\sum_{l=1}^{MD}\phi_l(\xi,\tau)\mathcal{G}(\hat{Q}_l^i)\,,\\ \mathcal{B}^*(\mathbf{Q})&=\sum_{l=1}^{MD}\mathcal{B}^*(\phi_l(\xi,\tau))\hat{Q}_l^i\,,\\ \mathbf{F}^*(\mathbf{Q})&=\sum_{l=1}^{MD}\phi_l(\xi,\tau)\mathbf{F}^*(\hat{Q}_l^i)\,,\ \mathbf{S}^*(\mathbf{Q})=\sum_{l=1}^{MD}\phi_l(\xi,\tau)\mathbf{S}^*(\hat{Q}_l^i)\,.\end{aligned}\tag{13}$$

For a friendly notation, let us define the operators

$$\begin{aligned}\{f,g\}_\tau&=\int_0^1 f(\xi,\tau)g(\xi)d\xi\ ,[f,g]_\tau=\int_0^1 f(\xi,\tau)g(\xi,\tau)d\xi\ ,\langle f,g\rangle\\&=\int_0^1\int_0^1 f(\xi,\tau)g(\xi,\tau)d\xi d\tau\,.\end{aligned}\tag{14}$$

Then the weak formulation for this problem provides

$$\begin{aligned}&[\phi_k,\mathcal{L}^*(\bar{\mathbf{Q}}_i)]_1-\langle\partial_\tau\phi_k,\mathcal{L}^*(\bar{\mathbf{Q}}_i)\rangle+\langle\phi_k,\partial_\xi\mathbf{F}^*(\bar{\mathbf{Q}}_i)\rangle\\&\quad=\langle\phi_k,\mathbf{S}^*(\bar{\mathbf{Q}}_i)\rangle+\{\phi_k,\mathcal{L}^*(\mathbf{P}_i^*)\}_0\,.\end{aligned}\tag{15}$$

We propose an iterative process, which converges linearly to approximations of (15). It is based on a decomposition of the governing equations in two sub-problems

$$\begin{aligned}&\text{Problem A}\ :\partial_\tau(\bar{\mathbf{Q}}(\xi,\tau)+\mathcal{B}^*(\bar{\mathbf{Q}}(\xi,\tau)))+\partial_\xi\mathbf{F}^*(\bar{\mathbf{Q}}(\xi,\tau))\\&\qquad=\mathbf{S}^*(\bar{\mathbf{Q}}(\xi,\tau))\,,\\&\text{Problem B}\ :\partial_\tau(\mathcal{G}(\bar{\mathbf{Q}}(\xi,\tau))-\bar{\mathbf{Q}}(\xi,\tau))=\mathbf{0}\,.\end{aligned}\tag{16}$$

In order to introduce the strategy let us define the following matrices:

$$\begin{aligned}\mathbf{K}^1_{k,l}&=[\phi_k,\phi_l]_1-<\partial_\tau\phi_k,\phi_l>\,,&\mathbf{K}^\tau_{k,l}&=<\phi_k,\partial_\tau\mathcal{B}^*(\phi_l)>\,,\\\mathbf{K}^\xi_{k,l}&=<\phi_k,\partial_\xi\phi_l>\,,&\mathbf{M}_{k,l}&=<\phi_k,\phi_l>\,,\\\mathbf{K}^0_{k,l}&=[\phi_k,\phi_l]_0\,,&\mathbf{F}^0_{k,l}&=\{\phi_k,\psi_l\}_0\,,\\\mathcal{Q}_l&=\hat{Q}_l^i\,,&\mathcal{G}(\mathcal{Q})_l&=\mathcal{G}(\hat{Q}_l^i)\,,\\\mathcal{F}(\mathcal{Q})_l&=\mathbf{F}^*(\hat{Q}_l^i)\,,&\mathcal{S}(\mathcal{Q})_l&=\mathbf{S}^*(\hat{Q}_l^i)\,,\\\mathcal{W}_l&=\hat{w}_l^i\,,&\mathcal{G}(\mathcal{W})_l&=\mathcal{G}(\hat{w}_l^i)\,.\end{aligned}\tag{17}$$

So weak formulations of sub-problems are used to compute the degrees of freedom of the sought solution to (15). It is carried out in terms of the following iterative process:

Step 1: Provide a starting guess $\mathcal{Q}^0$. In this work it is given by $\mathcal{Q}^0_{(k-1)M+l}=\hat{w}_l$, with $l,k=1,\ldots,M$.

Step 2: Solve the weak formulation associated to the Problem A:

$$(\mathbf{K}^\xi+\mathbf{K}^\tau)\mathcal{Q}^{(k+\frac{1}{2})}+\mathbf{K}^\xi\mathcal{F}^*(\mathcal{Q}^{(k)})=\mathbf{M}\mathcal{S}^*(\mathcal{Q}^{(k+\frac{1}{2})})+\mathbf{F}^0\mathcal{W}\tag{18}$$

Step 3 : Solve the weak formulation associated to Problem B, which in matrix form reads

$$\mathbf{K}^1(\mathcal{G}(\mathcal{Q}^{(k+1)})-\mathbf{Q}^{(k+1)})=\mathbf{K}^0(\mathcal{G}(\mathcal{Q}^{(k+\frac{1}{2})})-\mathbf{Q}^{(k+\frac{1}{2})})\,.\tag{19}$$

Step 4 : Stop if $||\mathcal{Q}^{k+1}-\mathcal{Q}^k||_{l_1}<Tol$, otherwise return to Step 2. Here, we set $Tol=1e^{-6}$.

At this point we remark that standard fixed point iteration procedures may be applied to solve (15). Notice that the source term is included in this formulation and solved implicitly through a fixed point iteration procedure. This feature makes this solver able to solve problems in which the source terms are stiff.

2.3. Derivation of an HLL type numerical flux

Let us consider the Riemann problem given by

$$\left.\begin{aligned}&\partial_t\mathcal{L}(\mathbf{Q})+\partial_x\mathbf{F}(\mathbf{Q})=\mathbf{0}\,,\\&\mathbf{Q}(x,0)=\begin{cases}\mathbf{Q}_L\,,x<0\,,\\\mathbf{Q}_R\,,x>0\,,\end{cases}\end{aligned}\right\}\tag{20}$$

where $\mathbf{Q}_L$ and $\mathbf{Q}_R$ are constant states.

In this section we derive a numerical flux of HLL type for the Riemann problem (20). Then let us construct a wave model as depicted in the Fig. 1. Here $x_L\le TS_L$ and $x_R\ge TS_R$, where S_L and S_R are the faster waves speeds and T is a given time. Here, we are interested in the subsonic case, $S_L\le 0\le S_R$. As we will see later, the influence of T disappears. Let us start by approximating the cell average of the operator $\mathcal{L}^*$, that means

$$\mathcal{L}^*:=\tfrac{1}{T(S_R-S_L)}\textstyle\int_{TS_L}^{TS_R}\mathcal{L}(\mathbf{Q}(x,T))dx\,.\tag{21}$$

Notice that this can be achieved by integrating Eq. (20) on $[TS_L, TS_R]\times[0,T]$ and by applying some algebraic manipulations, it yields

$$\mathcal{L}^*=\tfrac{S_R\mathcal{L}_R-S_L\mathcal{L}_L+\mathbf{F}_L-\mathbf{F}_R}{S_R-S_L}\,,\tag{22}$$

where we have introduced the following notation:

$$\begin{aligned}&\mathcal{L}_R:=\mathcal{L}(\mathbf{Q}_R)\,,\mathcal{L}_L:=\mathcal{L}(\mathbf{Q}_L)\,,\\&\mathbf{F}_R:=\mathbf{F}(\mathbf{Q}_R)\,,\mathbf{F}_L:=\mathbf{F}(\mathbf{Q}_L)\,.\end{aligned}\tag{23}$$

On the other hand, if we integrate the governing equation on $[0, TS_R]\times[0,T]$ and after some algebraic manipulations, we obtain

$$\tfrac{1}{T}\textstyle\int_0^{TS_R}\mathcal{L}(\mathbf{Q}(x,T))dx-S_R\mathcal{L}_R+\mathbf{F}_R-\mathbf{F}^*=\mathbf{0}\,,\tag{24}$$

where

$$\mathbf{F}^*:=\tfrac{1}{T}\textstyle\int_0^T\mathbf{F}(\mathbf{Q}(0,t))dt\,.\tag{25}$$

Similarly, by integrating the governing equation on $[TS_L, 0] \times [0, T]$ and after some algebraic manipulations, we obtain

$$\frac{1}{T}\int_{TS_L}^{0} \mathcal{L}(\mathbf{Q}(x,T))dx + S_L\mathcal{L}_L + \mathbf{F}^* - \mathbf{F}_L = \mathbf{0}\,, \tag{26}$$

then by substituting the integrand in (24) and (26) by $\mathcal{L}^*$ in (22) and manipulating, we obtain

$$\begin{aligned} &S_L(\mathcal{L}_L - \mathcal{L}^*) + \mathbf{F}^* - \mathbf{F}_L = \mathbf{0}\,,\\ &S_R(\mathcal{L}^* - \mathcal{L}_R) + \mathbf{F}_R - \mathbf{F}^* = \mathbf{0}\,, \end{aligned} \tag{27}$$

from both equations we can remove $\mathcal{L}^*$ and solve for $\mathbf{F}^*$, providing

$$\mathbf{F}^* = \frac{S_R\mathbf{F}_L - S_L\mathbf{F}_R + S_LS_R(\mathcal{L}_R - \mathcal{L}_L)}{S_R - S_L}\,. \tag{28}$$

So, in virtue of (2), it provides

$$\mathbf{F}^* = \frac{S_R\mathbf{F}_L - S_L\mathbf{F}_R + S_LS_R((\mathcal{G}_R - \mathcal{G}_L) + (\mathcal{B}_R - \mathcal{B}_L))}{S_R - S_L}\,, \tag{29}$$

with $\mathcal{G}(\mathbf{Q}_L) =: \mathcal{G}_L$, $\mathcal{G}(\mathbf{Q}_R) =: \mathcal{G}_R$, $\mathcal{B}(\mathbf{Q}_L) =: \mathcal{B}_L$ and $\mathcal{B}(\mathbf{Q}_R) =: \mathcal{B}_R$.

Notice that, $\mathcal{B}(\mathbf{Q})$ has only a jump discontinuity in $x = 0$, whose contribution is contained in $\mathcal{L}^*$. However, $\mathcal{L}^*$ does not influence the numerical flux. Furthermore, because $\mathbf{Q}(x, 0)$ takes constant values at both sides of the interface $x = 0$, we have $\partial_x^{(k)}\mathbf{Q} = \mathbf{0}$ for $x < 0$ and $x > 0$. Therefore, due to (3), $\mathcal{B}_L = \mathcal{B}_R = \mathbf{0}$. So the HLL flux has a similar form than for the hyperbolic case

$$\mathbf{F}^* = \frac{S_R\mathbf{F}_L - S_L\mathbf{F}_R + S_LS_R(\mathcal{G}_R - \mathcal{G}_L)}{S_R - S_L}\,. \tag{30}$$

We remark that for $-S_L := S_R = \lambda_{max} = max\{|\lambda_m|\}$, where $|\lambda_m|$ plays the role of eigenvalues in hyperbolic systems, we obtain an equivalent Rusanov flux

$$\mathbf{F}^* = \tfrac{1}{2}(\mathbf{F}_L + \mathbf{F}_R) - \tfrac{\lambda_{max}}{2}(\mathcal{G}_R - \mathcal{G}_L)\,. \tag{31}$$

However, in this case λ_{max} contains the eigenvalues of $\mathbf{F}$ with respect to $\mathbf{Q}$ but also a measurement of the variation of $\mathcal{L}(\mathbf{Q})$ with respect to the variation of $\mathbf{Q}$. Further, simplifications may be carried out, if $\mathcal{G}(\mathbf{Q})$ is a differentiable function, in such a case

$$(\mathcal{G}_R - \mathcal{G}_L) = \mathcal{G}'(\theta)(\mathbf{Q}_R - \mathbf{Q}_L)\,,$$

with $\theta_j \in [min(\mathbf{Q}_{R,j}, \mathbf{Q}_{L,j}), max(\mathbf{Q}_{R,j}, \mathbf{Q}_{L,j})]$, $j = 1, ..., m$. So we propose

$$\mathbf{F}^* = \tfrac{1}{2}(\mathbf{F}_L + \mathbf{F}_R) - \tfrac{(\lambda_{max} + \mathcal{G}'_{max})}{2}(\mathbf{Q}_R - \mathbf{Q}_L)\,, \tag{32}$$

where $\mathcal{G}'_{max} = max\{|\mathcal{G}'(\mathbf{Q}_L)|, |\mathcal{G}'(\mathbf{Q}_R)|\}$.

2.4. A strategy to provide $\lambda_{\max}$

In this section we deal the issue of computing λ_{max} to determine the HLL flux and for computing a stable time step. Let us consider a discretization in space of the governing Eq. (1) as follows:

$$\mathbf{L}(\mathbf{Q})\dot{\mathbf{Q}} = \mathbf{B}(\mathbf{Q})\mathbf{Q}\,, \tag{33}$$

where $\mathbf{Q}$ is the vector of N discretized values, the "$\cdot$" symbol represents the derivative in time, $\mathbf{L}(\mathbf{Q})$ is the discretization matrix of the diffusive operator $\mathcal{L}(\mathbf{Q})$ and $\mathbf{B}(\mathbf{Q})$ the discretization matrix for the advective part which also includes the source term, if present. For the sake of simplicity, let us denote a matrix norm as well as a vector norm, by the symbol $||\cdot||$. Additionally, let us assume the following:

- $||\dot{\mathbf{Q}}(t)|| > 0$ for all $t > 0$ or $||\dot{\mathbf{Q}}(t)|| \equiv 0$ for all $t > 0$. That means, if the vector solution is a constant for some time, then it is a constant vector for all times.
- Bounded operators. We assume that there exist constant values K_L and K_B such that, $||\mathbf{L}(\mathbf{Q})|| \leq K_L$ and $||\mathbf{B}(\mathbf{Q})|| \leq K_B$.

On the other hand, we note that a forward Euler integration provides

$$\mathbf{Q}^{n+1} = \mathbf{Q}^n + \Delta t\mathbf{L}^{-1}(\mathbf{Q}^n)\mathbf{B}(\mathbf{Q}^n)\mathbf{Q}^n\,,$$

so by taking norm

$$||\mathbf{Q}^{n+1}|| \leq ||\mathbf{Q}^n|| + \Delta t||\mathbf{L}^{-1}(\mathbf{Q}^n)|| \cdot ||\mathbf{B}(\mathbf{Q}^n)\mathbf{Q}^n||\,. \tag{34}$$

From (33), we have

$$||\mathbf{B}(\mathbf{Q})\mathbf{Q}|| \leq ||\mathbf{L}(\mathbf{Q})|| \cdot ||\dot{\mathbf{Q}}||\,. \tag{35}$$

Thus, we obtain

$$||\mathbf{Q}^{n+1}|| \leq ||\mathbf{Q}^n|| + \Delta t||\mathbf{L}^{-1}(\mathbf{Q}^n)|| \cdot ||\mathbf{L}(\mathbf{Q})|| \cdot ||\dot{\mathbf{Q}}||\,. \tag{36}$$

After some manipulations

$$\frac{||\mathbf{Q}^{n+1}|| - ||\mathbf{Q}^n||}{||\dot{\mathbf{Q}}||} \leq \Delta t||\mathbf{L}^{-1}(\mathbf{Q}^n)|| \cdot ||\mathbf{L}(\mathbf{Q})||\,, \tag{37}$$

so by consistency of the Euler method we can assume that $||\mathbf{L}^{-1}|| \cdot ||\mathbf{L}|| < \infty$, moreover, we may assume the quantity $\Delta t||\mathbf{L}^{-1}|| \cdot ||\mathbf{L}||$ to be small, thus we impose that

$$\Delta t||\mathbf{L}^{-1}|| \cdot ||\mathbf{L}|| < (1 - C_{cfl})\Delta x\,, \tag{38}$$

where C_{cfl} is the Courant–Friedrich–Levy coefficient and $max_{||\mathbf{Q}||=1}\frac{||\mathbf{L}^{-1}(\mathbf{Q})||}{||\mathbf{Q}||} \leq ||\mathbf{L}^{-1}||$. On the other hand, from convection terms we have associated wave speeds, which correspond to the eigenvalues of the Jacobian matrix $\mathbf{A}(\mathbf{Q}) = \frac{\partial \mathbf{F}(\mathbf{Q})}{\partial \mathbf{Q}}$, denoted here by $\lambda_1 \leq ... \leq \lambda_m$. In addition, for hyperbolic problems we know that the stability constraint corresponds to

$$\Delta t\bar{\lambda} = C_{cfl}\Delta x\,. \tag{39}$$

Thus we have

$$\Delta t(\bar{\lambda} + ||\mathbf{L}^{-1}|| \cdot ||\mathbf{L}||) < \Delta x\,, \tag{40}$$

where $\bar{\lambda} = max\{|\lambda_j| j = 1, ..., m\}$. Therefore, the time step is computed as

$$\Delta t = C_{cfl}\frac{\Delta x}{\lambda_{max}}\,, \tag{41}$$

with $\lambda_{max} = \bar{\lambda} + ||\mathbf{L}^{-1}|| \cdot ||\mathbf{L}||$. In the following section we are going to deal the problem of convergence of the algorithm.

3. The L^1_{loc} convergence of the numerical scheme

In this section the convergence of the presented numerical scheme is studied. The main result is presented in this section, whereas, auxiliary lemmas and propositions are available in Appendix A. Here, we consider the scalar case

$$\partial_t\mathcal{L}(q(x,t)) + \partial_x f(q(x,t)) = s(q(x,t)) \tag{42}$$

and we will prove that the numerical scheme converges under regular conditions on $\mathcal{L}(q)$, $f(q)$ and $s(q)$. Let us assume that $w(x,t) := \mathcal{L}(q(x,t)) = q(x,t) + b(q(x,t))$, with b a linear operator.

Proposition 3.1. *Let $\mathcal{L}(q)$ be a linear operator, $f(q)$ and $s(q)$ continuously differentiable functions of q. Then, the numerical scheme with Δt chosen via CFL condition, converges in $L^1_{loc}(\mathbb{R})$ for $t \in [0, T]$.*

Proof. Let us consider

$$\Delta t = h\frac{\Delta x}{\lambda_{max}}, \tag{43}$$

with h a proportionality constant between the wave speed and mesh velocity. It is associated with the CFL coefficient. On the other hand, we define the function $w_h(x, t)$ as $w_h(x,t) = w_i^n$ if $x \in [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$ and $t \in [t^n, t^n + \Delta t]$. Hence

$$||w_h(t)||_{L^1} := \sum_i |w_i^n|\Delta x\,, \tag{44}$$

Table 1
Convergence rates for the linear model equation. We have used a $C_{cfl} = 0.3$, $\lambda = 2$ and $t_{out} = 1$.

Mesh	L_∞ - err	L_∞- ord	L_1 - err	L_1 - ord	L_2 - err	L_2 - ord
			Theoretical order : 2			
64	0.00	$6.69e-01$	0.00	$4.13e-01$	0.00	$4.64e-01$
71	2.53	$5.28e-01$	2.70	$3.21e-01$	2.67	$3.61e-01$
78	2.65	$4.20e-01$	2.79	$2.52e-01$	2.77	$2.84e-01$
85	2.66	$3.40e-01$	2.81	$2.02e-01$	2.79	$2.28e-01$
92	2.69	$2.79e-01$	2.82	$1.64e-01$	2.80	$1.86e-01$
			Theoretical order : 3			
Mesh	L_∞ - err	L_∞- ord	L_1 - err	L_1 - ord	L_2 - err	L_2 - ord
64	0.00	$8.81e-02$	0.00	$5.61e-02$	0.00	$6.23e-02$
71	3.15	$6.55e-02$	3.16	$4.17e-02$	3.16	$4.63e-02$
78	3.15	$5.00e-02$	3.15	$3.18e-02$	3.16	$3.53e-02$
85	3.15	$3.89e-02$	3.16	$2.48e-02$	3.15	$2.75e-02$
92	3.15	$3.09e-02$	3.14	$1.97e-02$	3.14	$2.19e-02$
			Theoretical order : 4			
Mesh	L_∞ - err	L_∞- ord	L_1 - err	L_1 - ord	L_2 - err	L_2 - ord
64	0.00	$1.90e-03$	0.00	$1.13e-03$	0.00	$1.26e-03$
71	4.76	$1.21e-03$	4.73	$7.26e-04$	4.75	$8.09e-04$
78	4.39	$8.32e-04$	4.32	$5.01e-04$	4.34	$5.57e-04$
85	3.83	$6.15e-04$	3.76	$3.72e-04$	3.77	$4.13e-04$
92	2.93	$4.96e-04$	2.90	$3.01e-04$	2.90	$3.34e-04$
			Theoretical order : 5			
Mesh	L_∞ - err	L_∞- ord	L_1 - err	L_1 - ord	L_2 - err	L_2 - ord
64	0.00	$1.70e-03$	0.00	$1.08e-03$	0.00	$1.20e-03$
71	9.33	$7.07e-04$	9.33	$4.50e-04$	9.33	$5.00e-04$
78	10.27	$2.93e-04$	10.27	$1.86e-04$	10.27	$2.07e-04$
85	5.78	$1.85e-04$	5.78	$1.18e-04$	5.78	$1.31e-04$
92	7.88	$1.04e-04$	7.88	$6.62e-05$	7.88	$7.35e-05$

$$TV(w_h(t)) := \sum_i |w_i^n - w_{i-1}^n| \,. \tag{44}$$

From Lemma A.10 and Proposition A.15

$$||w_h(t)|| + TV(w_h(t)) \leq \mathbf{C_2} \,, \tag{45}$$

with $\mathbf{C_2} = max\{\bar{C} + ||\mathbf{w}^0||, (TV(\mathbf{w}^0) + \mathcal{D}\mathcal{A}\delta||\mathbf{w}^0||)\}$.

On the other hand, given t^n and t^p, there exist integers n and p, such that $p\Delta t < t^p < (p+1)\Delta t$ and $n\Delta t < t^n < (n+1)\Delta t$. Then from Lemma A.16

$$||w_h(t^n) - w_h(t^p)|| \leq \mathbf{C_2}(t^n - t^p) \,. \tag{46}$$

Therefore, from Helly's selection theorem, [8,46], given a set $\{\tau_k\}_k^\infty \subset [0, T]$, which is dense in $[0, T]$, we can extract a sequence of $h_i \to 0$ such that $\{w_{h_i}(x, \tau_k)\}_{i=1}^\infty$ converges for all x in $L^1_{loc}(\mathbb{R})$, to the limiting function $w(x, \tau_k)$.

Let $[-X, X] \subset \mathbb{R}$ be an arbitrary interval. Then, given $\varepsilon > 0$, there exists an index $I = I(\varepsilon)$ such that $|t - \tau_k| < \varepsilon$ for $k > I$. So, for $i, j > I$

$$\begin{aligned}
&\int_{-X}^{X} |w_{h_i}(x,t) - w_{h_j}(x,t)|dx \\
&\leq \int_{-X}^{X} \{|w_{h_i}(x,t) - w_{h_i}(x,\tau_k)| + |w_{h_j}(x,t) - w_{h_j}(x,\tau_k)| \\
&\quad + |w_{h_i}(x,\tau_k) - w_{h_j}(x,\tau_k)|\} dx \,.
\end{aligned} \tag{47}$$

Therefore $\int_{-X}^{X} |w_{h_i}(x,t) - w_{h_j}(x,t)|dx \leq (2\mathbf{C_2} + 1)\varepsilon$ and thus $\{w_{h_i}(x,t)\}_{i=1}^\infty$ is a Cauchy sequence in $[-X, X]$. Therefore it must converge to a limiting function $w(x, t)$ in $L^1_{loc}(\mathbb{R})$. On the other hand

$$\begin{aligned}
&\int_{-X}^{X} |w(x,t) - w(x,\tau_k)|dx \\
&\leq \int_{-X}^{X} \Big\{|w(x,t) - w_{h_i}(x,t)| + |w(x,\tau_k) - w_{h_i}(x,\tau_k)| \\
&\quad + |w_{h_i}(x,t) - w_{h_i}(x,\tau_k)|\Big\} dx \leq (2 + \mathbf{C_2})\varepsilon.
\end{aligned} \tag{48}$$

Hence, $w(x,t) = lim_{\tau_k \to t} w(x, \tau_k)$. Therefore, the scheme converges in $L^1_{loc}(\mathbb{R})$ for all $t \in [0, T]$. □

Remark 1. In the proposition a qualitative analysis more than a quantitative study, is carried out. Therefore, the range of values for CFL coefficients cannot be set. Indeed, the proposition provides a theoretical justification of the convergence of the proposed scheme, specially when h goes to zero. It ensures that a time-step and a mesh space related through a CFL coefficient (the role of parameter h), must converge to the exact solution as expected. For simulations presented in this paper, values for the CFL coefficient, are heuristically chosen.

On the other hand, notice that, the influence of the source term $s(q)$ is included in the coefficient $\mathbf{C_2}$. So, stiff source terms must modify such a constant and so the range of values of h has to be accordingly updated. Therefore, the proposition remains valid also for equations with stiff source terms.

4. Numerical results

In this section three tests are solved, the first two, are scalar problems with exact solutions, whereas, the third test corresponds to a Boussinesq type system. Exact solutions are not available for the system, however a numerical solution of reference is used to compare the results of the present scheme.

4.1. A linear model equation

In this section we deal with the simple model equation

$$\begin{aligned}
&\partial_t \mathcal{L}(q) + \lambda \partial_x q = 0 \,, \\
&q(x,0) = sin(2\pi x) \,,
\end{aligned} \tag{49}$$

with λ a constant value, $\mathcal{L}(q) = q - \alpha \partial_x^{(2)} q$ and $\alpha = \frac{\lambda - 1}{4\pi^2}$, which has the exact solution

$$q(x,t) = sin(2\pi(x - t)) \,. \tag{50}$$

Table 1, shows the result of the empirical convergence rate assessment for $\lambda = 2$, $t_{out} = 1$ and $C_{cfl} = 0.3$. We note that the convergence rates are attained up to fifth-order of accuracy, even on coarse meshes. Fig. 2 depicts a comparison between the exact solution (full line) and numerical solutions of fifth (circles), fourth (squares) and third (stars) orders of accuracy computed on a mesh of 64 cells.

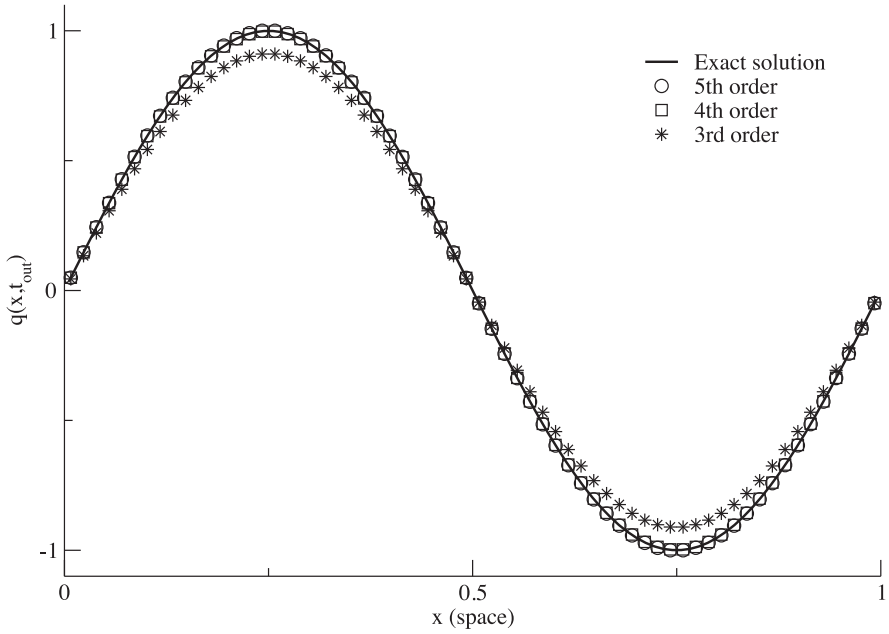

Fig. 2. Linear problem. Exact solution (full line), fifth order solution (circles), fourth order solution (squares) and third order solution (stars). Numerical solutions with 64 cells, $C_{CFL} = 0.3$, $\lambda = 2$ and $t_{out} = 1$.

Table 2
Convergence rates for the scalar model equation. We have used a $C_{cfl} = 0.01$ and $t_{out} = 0.1$.

Mesh	L_∞ - err	L_∞- ord	Theoretical order : 2 L_1 - err	L_1 - ord	L_2 - err	L_2 - ord
80	0.00	4.83e − 01	0.00	2.66e − 01	0.00	3.06e − 01
85	2.26	4.25e − 01	1.70	2.41e − 01	2.08	2.72e − 01
90	2.72	3.67e − 01	1.99	2.17e − 01	2.32	2.40e − 01
95	2.67	3.20e − 01	2.18	1.94e − 01	2.45	2.11e − 01
100	2.88	2.78e − 01	2.35	1.73e − 01	2.58	1.86e − 01
			Theoretical order : 3			
Mesh	L_∞ - err	L_∞- ord	L_1 - err	L_1 - ord	L_2 - err	L_2 - ord
70	0.00	5.78e − 02	0.00	5.26e − 02	0.00	5.27e − 02
75	3.20	4.70e − 02	2.96	4.34e − 02	2.97	4.35e − 02
80	3.23	3.87e − 02	3.02	3.62e − 02	3.03	3.62e − 02
85	3.19	3.22e − 02	3.02	3.04e − 02	3.02	3.05e − 02
90	3.20	2.71e − 02	3.05	2.58e − 02	3.06	2.58e − 02
			Theoretical order : 4			
Mesh	L_∞ - err	L_∞- ord	L_1 - err	L_1 - ord	L_2 - err	L_2 - ord
70	0.00	4.70e − 03	0.00	3.59e − 03	0.00	3.64e − 03
75	4.61	3.49e − 03	4.81	2.63e − 03	4.80	2.67e − 03
80	4.72	2.62e − 03	4.87	1.96e − 03	4.86	1.99e − 03
85	4.54	2.02e − 03	4.92	1.48e − 03	4.89	1.50e − 03
90	4.73	1.57e − 03	5.02	1.13e − 03	4.98	1.15e − 03
			Theoretical order : 5			
Mesh	L_∞ - err	L_∞- ord	L_1 - err	L_1 - ord	L_2 - err	L_2 - ord
22	0.00	1.39e − 01	0.00	1.14e − 01	0.00	1.15e − 01
27	5.83	5.16e − 02	5.21	4.69e − 02	5.28	4.70e − 02
32	5.83	2.21e − 02	5.60	2.08e − 02	5.61	2.08e − 02
37	5.74	1.07e − 02	5.78	9.99e − 03	5.77	1.00e − 02
42	5.74	5.60e − 03	5.98	5.10e − 03	5.97	5.11e − 03
47	5.64	3.17e − 03	6.19	2.73e − 03	6.15	2.74e − 03

4.2. A non-linear scalar model equation

Let us consider the equation

$$\begin{aligned} \partial_t \mathcal{L}(q(x,t)) + \partial_x f(q(x,t)) = 0 \,, \\ q(x,0) = sin(2\pi x)cos(2\pi x) \,, \end{aligned} \tag{51}$$

with

$$\begin{aligned} \mathcal{L}(q) = q^2 - \tfrac{1}{16\pi^2}\partial_x^{(2)} q \,, \\ f(q) = q(q+1) \,. \end{aligned} \tag{52}$$

Periodic boundary conditions are considered. This problem has the exact solution

$$q(x,t) = sin(2\pi\,(x-t))cos(2\pi\,(x-t)) \,. \tag{53}$$

Due to the non-linearity of $\mathcal{L}(q)$, in this test we use only a 1% of the theoretical CFL condition, so we take $C_{cfl} = 0.01$. Table 2 shows the results of the convergence rate assessment, up to the output time $t_{out} = 0.1$ for schemes of 2nd, 3rd, 4th and 5th orders of accuracy. We observe that the orders of accuracy are attained even if the differential operator is non-linear.

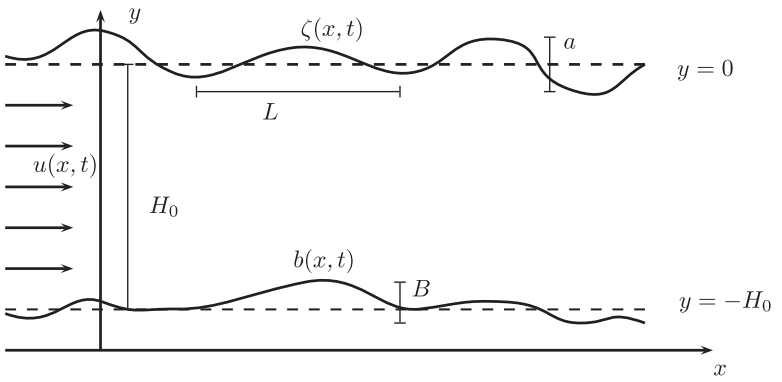

Fig. 3. Schematic representation of the physical phenomenon.

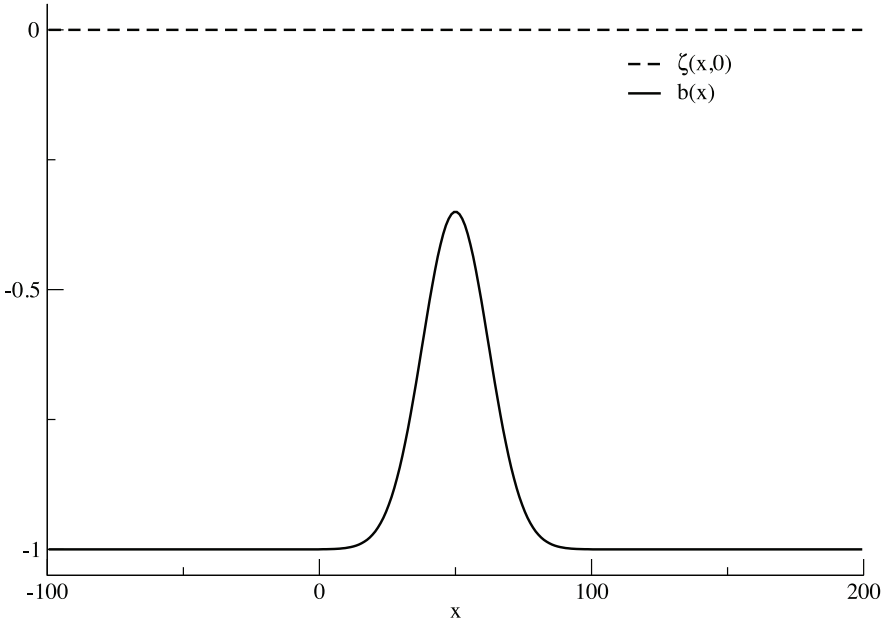

Fig. 4. Initial configuration.

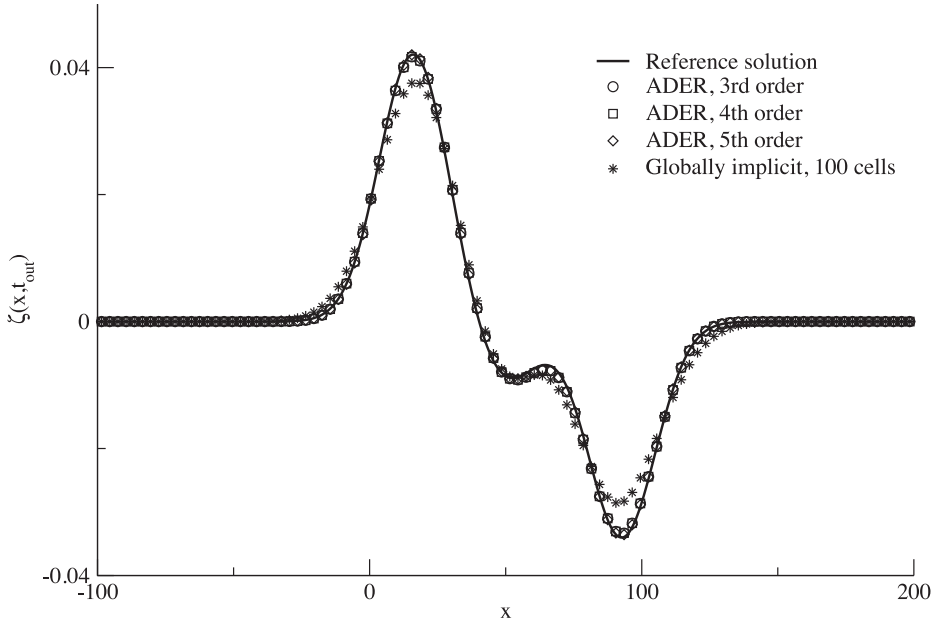

Fig. 5. Profile of the free surface $\zeta(x, t_{out})$ for $b = b(x)$ and $t_{out} = 37$. Comparison of schemes of 3rd (circles), 4th (squares) and 5th (diamonds) orders of accuracy with 100 cells against a reference solution and the solution of the globally implicit method with 100 cells (stars).

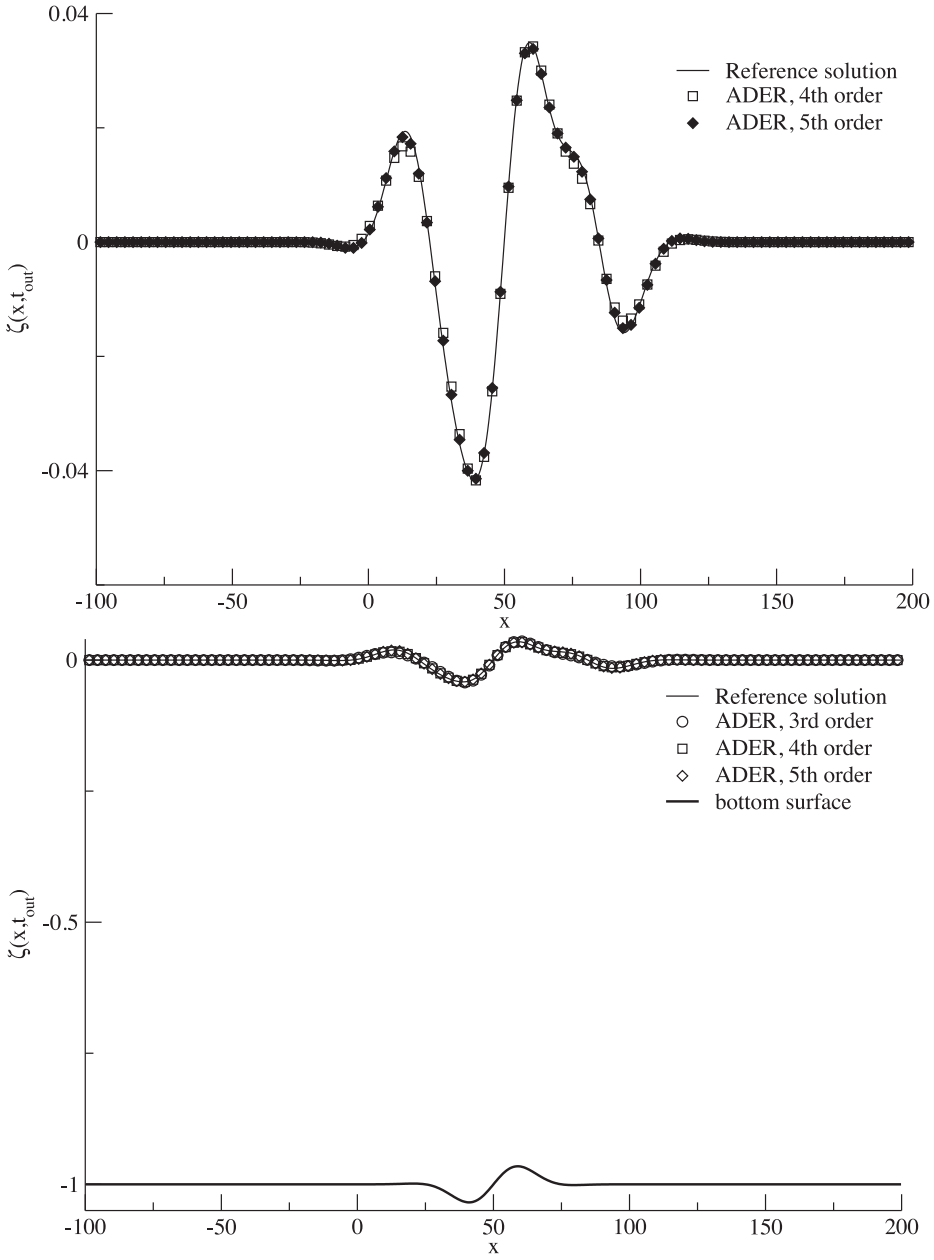

Fig. 6. Profile of the free surface $\zeta(x, t_{out})$ and bottom surface $b = b(x, t)$ for $t_{out} = 30$. **Top**: zoom of free surface. **Bottom**: general profile. Comparison of schemes of 3rd (circles), 4th (squares) and 5th (diamonds) orders of accuracy with 100 cells against a reference solution.

4.3. An asymptotic model in water-waves equations

Here, we are interested in the simulation of the motion of a layer of water in a domain delimited below by a solid moving bottom and above by a free surface, more precisely, we are considering an approximation model within the shallow water regime, in which only the length and depth are relevant. Fig. 3 shows a representation of the phenomenon. The length and depth are parametrized by the variables x and y, respectively. Model variables are; $\zeta(x, t)$ the free surface, $b(x, t)$ the bottom surface and $u(x, t)$ the velocity of a column of water in x at a given time t. On the other hand, model parameters are; $H_0 > 0$ a constant reference depth (note that $y = 0$ corresponds to the still water level), L the characteristic horizontal scale in the longitudinal direction, a the order of the free surface amplitude, B the order of bottom topography variation. The following dimensionless parameters are formed from these four scales; $\mu := H_0^2/L^2$ the shallowness parameter, $\epsilon := \frac{a}{H_0}$ the wave amplitude parameter and $\beta := \frac{B}{H_0}$ the topography parameter. The behaviour of water is governed by a dimensionless Boussinesq type model, [53], which has the form

$$\begin{aligned}
&\partial_t \zeta(x,t) + \partial_x(h(x,t)u(x,t)) = \frac{\beta}{\varepsilon}\partial_t b(x,t)\ ,\\
&\partial_t u(x,t) + \partial_x \zeta(x,t) + \varepsilon u(x,t)\partial_x u(x,t) - \frac{\mu}{3}\partial_{txx}u(x,t)\\
&= -\frac{\mu\beta}{2\varepsilon}\partial_{ttx} b(x,t)\ ,\}
\end{aligned} \tag{54}$$

with $h(x,t) = \varepsilon\zeta(x,t) + 1 - \beta b(x,t)$. We define the differential operator

$$\mathcal{L}(u) := u - \frac{\mu}{3}\partial_{xx}u\ . \tag{55}$$

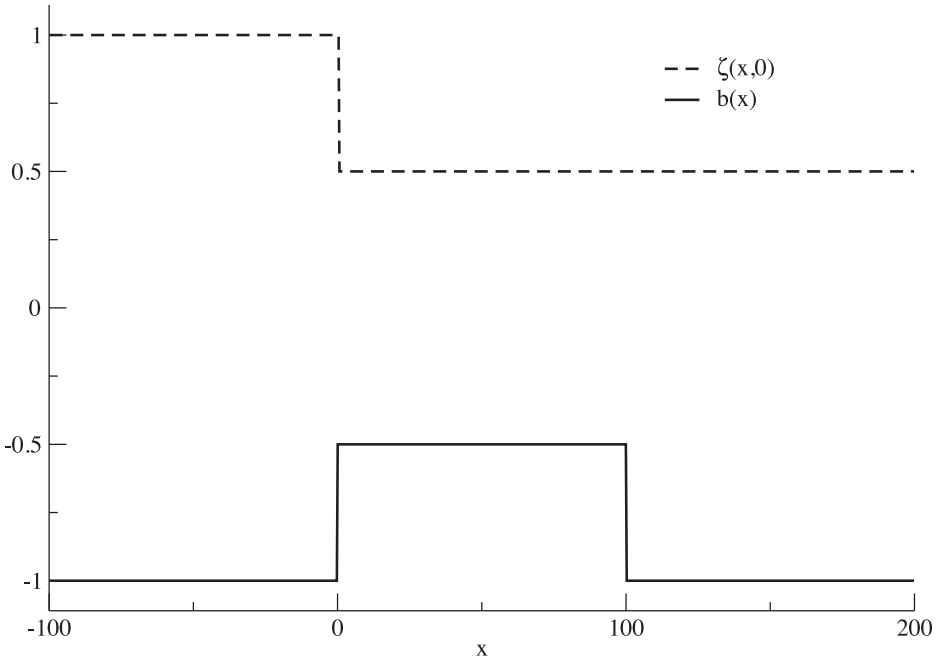

Fig. 7. Initial configuration for the dam break type problem.

So, we observe that (54) can be written as

$$\left.\begin{array}{c}\partial_t \zeta\,(x,t)+\partial_x(h(x,t)u(x,t))=\frac{\beta}{\varepsilon}\partial_t b(x,t)\,,\\ \partial_t \mathcal{L}(u(x,t))+\partial_x\zeta\,(x,t)+\varepsilon u(x,t)\partial_x u(x,t)\\ =-\frac{\mu\beta}{2\varepsilon}\partial_{ttx}b(x,t)\,.\end{array}\right\} \tag{56}$$

So, in a matrix form (56) takes the form (5) with

$$\begin{aligned}\mathbf{Q}(x,t)&=\begin{bmatrix}\zeta\,(x,t)\\ u(x,t)\end{bmatrix},\\ \mathcal{L}(\mathbf{Q}(x,t))&=\begin{bmatrix}\zeta\,(x,t)\\ u(x,t)-\frac{\mu}{3}\partial_{xx}u(x,t)\end{bmatrix},\\ \mathbf{F}(\mathbf{Q}(x,t))&=\begin{bmatrix}h(x,t)\zeta\,(x,t)\\ \zeta\,(x,t)+\varepsilon\frac{u(x,t)^2}{2}\end{bmatrix},\\ \mathbf{S}(\mathbf{Q}(x,t),x,t)&=\begin{bmatrix}\frac{\beta}{\varepsilon}\partial_t b(x,t)\\ -\frac{\mu\beta}{2\varepsilon}\partial_{ttx}b(x,t)\end{bmatrix}.\end{aligned} \tag{57}$$

The model requires $\mu \ll 1$ and $O(\varepsilon)=O(\mu)=O(\beta)$. In order to assess the performance of the present method, we compute reference solutions through a globally implicit finite difference scheme, which is reported in Appendix B. A fine mesh of 500 cells has been used.

4.3.1. Bottom surface varying in space

Here, we simulate the case in which $b(x,t)=b(x)$ on the computational domain $[-100,200]$. In particular we take

$$b(x)=\begin{cases}-1\,, & x<0\,,\\ -1+0.65e^{-33.75(x/100.-.5)^2}\,, & 0\le x\le 100\,,\\ -1\,, & x>100\,.\end{cases} \tag{58}$$

We consider the parameters $\mu=\varepsilon=\beta=0.1$, initial condition $\zeta\,(x,0)=0$ and $u(x,0)=1.2$, which are shown in Fig. 4. Additionally, we apply transmissive boundary conditions. Fig. 5, shows a comparison of the reference solution against the solution obtained with the present scheme for 3rd, 4th and 5th orders of accuracy and the scheme presented in Appendix B. The numerical solutions are compared at the output time $t_{out}=37$ and using 100 cells. We observe a very good agreement of the high-order schemes with respect to a conventional globally implicit method. The linearity of $\mathcal{L}(\mathbf{Q})$ allows the use of $C_{CFL}=0.1$.

4.4. Bottom surface varying in space and time

Now, we simulate the case in which the bottom surface is given by

$$b(x,t)=\begin{cases}-1\,, & x<0\,,\\ -1+0.1sin(4\pi\,(x/100-t))e^{-33.75(x/100.-.5)^2}\,, & 0\le x\le 100\,,\\ -1\,, & x>100\,.\end{cases} \tag{59}$$

An important feature of this test, is that a variation in time, could cause stiffness in the source term. However, the ability of the present scheme to solve problems with stiff source term, should overcome this difficulty. Here, we use the same parameters seen in the previous test. Fig. 6 shows the results up to a time $t_{out}=30$. We observe a very good agreement of the high-order schemes with respect to a conventional globally implicit method.

4.5. A dam-break type problem with a discontinuous bottom surface

In this test we explore the phenomenon consisting of a dam-break in a horizontal channel. The water is assumed to contain two different water contents at both sides of the breakpoint. The evolution of the free surface and velocity profile can be modelled by system (56). These problems are well known for hyperbolic problems, where the ADER schemes have proved to provide very good results, see [9,19,68,69] to mention but a few.

As we will see below, the present methodology allows to solve this problem with discontinuous bottom profiles. So, in this subsection, we solve the system (56) on a channel parametrized in the computational domain $[-100,200]$, the breakpoint originally is assumed to be at $x=\frac{1}{2}$ and the bottom surface is given by

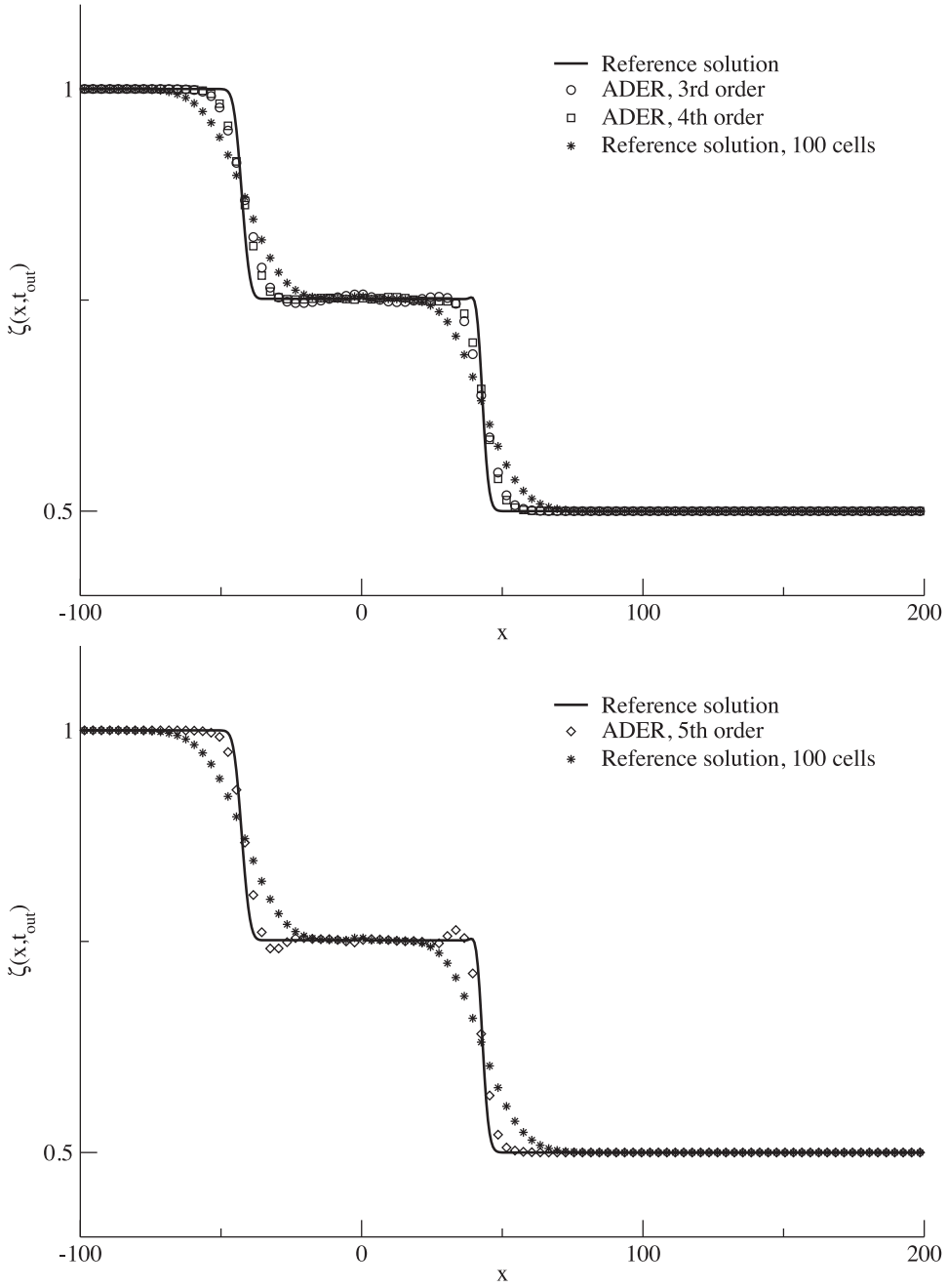

Fig. 8. Dam-break type problem. Profile of the free surface $\zeta(x, t_{out})$ for $t_{out} = 40$, $C_{CFL} = 0.1$. Comparison of novel ADER schemes with 100 cells against reference solution (continuous line). **Top:** ADER of 3rd (circles) and 4th (squares) orders of accuracy against the reference solution with 100 cells (stars). **Bottom:** ADER of 5th (diamond) order of accuracy against the reference solution with 100 cells (stars).

$$b(x) = \begin{cases} -1\,, & x < 0\,, \\ -\frac{1}{2}\,, & 0 \le x \le 100\,, \\ -1\,, & x > 100\,. \end{cases} \tag{60}$$

As we have assumed that the break point is located at $x = \frac{1}{2}$, the initial condition is given by

$$\zeta(x,0) = \begin{cases} 1\,, & x < \frac{1}{2}\,, \\ \frac{1}{2}\,, & x \ge \frac{1}{2}\,, \end{cases} \tag{61}$$

additionally, the fluid is assumed to be at rest so the initial velocity is given by $u(x,0) = 0$. Fig. 7 shows the initial configuration. Fig. 8 shows the result for the free surface $\zeta(x, t_{out})$, at $t_{out} = 40$, parameters in this test are $C_{CFL} = 0.1$ and $\mu = \varepsilon = \beta = 0.1$.

In Fig. 8 is shown a comparison of the reference solution, presented in Appendix B, against the solutions obtained with the present scheme for 3rd, 4th and 5th orders of accuracy, all of these solutions have been computed using 100 cells. We note that fifth order of accuracy shows small oscillations (see Fig. 8, bottom). We observe in general a very good agreement between the computed solutions and the reference solution with a very fine mesh. Similarly, we observe that there is an improvement in the degree of resolution by using high order schemes, which is evident when compared with the reference solution with the same number of cells.

5. Conclusions

Here we have extended the well known ADER finite volume schemes for solving a class of partial differential equations arising from the water wave equations. These equations are character-

ized by transient terms, which may contain mixed space and time derivatives. The present numerical scheme is based on the HEOC solution strategy of generalized Riemann problems at the interfaces of computational cells, similar to ADER methods for hyperbolic problems. To this end, we have derived an HLL type Riemann solver, which is used in the solution of the GRP and we have provided a strategy to estimate the maximum wave speeds in order to reconcile stability and accuracy. Stability constraint depends on the wave speeds associated to convective parts and the discretization of differential transient terms. We have proved for a linear and scalar case that the present scheme converge in L^1_{loc} as usual for Godunov-type schemes applied to hyperbolic problems. We have solved four tests and empirical convergence rate assessment has been carried out for two of them, illustrating in this form that the expected theoretical order of accuracy is attained. Additionally, when an exact solution was not available we have compared the numerical solution of the present schemes with reference solutions obtained with a globally implicit scheme on a fine mesh. We have observed that the degree of resolution with respect to the reference solution is very satisfactory, specially because high order computations were carried out on coarse meshes. Weakness of the present scheme are; i) the differential problems have been solved by using centred finite differences in a context of fixed-point iteration procedures, ii) the influence of the differential operator on the stability constraint is not well understood. Improvements regarding the solution of the differential part and how it affects the stability, will be the aim of future works.

Acknowledgements

G.I. Montecinos thanks FONDECYT in the frame of the research project FONDECYT Postdoctorado 2016, number 3160743. The first four authors were partially supported by Basal-CMM project, PFB 03. J.C. López thanks to BCAM - Basque Center for Applied Mathematics, Mazarredo 14, E48009 Bilbao, Basque Country-Spain.

Appendix A. Auxiliary results

In this appendix we deal with the proof of Proposition 3.1. The aim is to provide technical results to ensure the applicability of the Helly selection theorem [8]. In order to prove the convergence of the present scheme, the results will be presented in the following order; Lemma A.1 and Proposition A.2 will ensure the convergence of the discontinuous Galerkin method, which allows to compute local predictors at each cell of the computational domain. Lemma A.4 will relate the degrees of freedom of polynomials obtained through the discontinuous Galerkin method with the approximate solutions at time level t^n, $\{q_i^n\}$. Lemma A.5 to Lemma A.7 will provide the relationship between the discrete solution of the differential operator $\{w_i^n\}$ and the approximate solutions at time level t^n, $\{q_i^n\}$, it will prove the continuity of w with respect to q. Lemma A.9 to Lemma A.11 will provide inequalities which take into account numerical approximations of the differential operator for successive time steps, $\{w_i^{n+1}\}$ and $\{w_i^n\}$. Lemma A.12 to Lemma A.16 will provide the inequalities for the Total-Variation-Diminishing (TVD) operator and L_1 norm of $\{w_i^n\}$. It is required for Helly's selection theorem, see (44) for a definition of TVD and L_1 norm.

For the sake of simplicity we are going to assume that the reconstruction polynomial p_i can be defined as

$$p_i(x) = \bar{p}(x_{i-\frac{1}{2}} + \xi\,\Delta x) = \sum_{m=1}^{M} \psi_m(\xi)\gamma_m\,, \tag{A.1}$$

where $\{\psi_m(\xi)\}$ is a reconstruction polynomial basis. From the conservation property of reconstruction polynomials, [35], we have

$$q^n_{i+j} = \sum_{m=1}^{M} \int_{i+j}^{i+j+1} \psi_m(\xi)\gamma_m\,, \tag{A.2}$$

with $j = -k_L, \ldots, k_R$. Therefore, coefficients γ_m can be found by solving the linear system

$$\tilde{\mathbf{q}}^i = \mathbf{B}\tilde{\mathbf{\Gamma}}^i\,, \tag{A.3}$$

where

$$\mathbf{B}_{j,m} = \int_{i+j}^{i+j+1} \psi_m(\xi)\,,\ \tilde{\mathbf{q}}^i_j = q^n_{i+j}\,,\ \tilde{\mathbf{\Gamma}}^i_m = \gamma^i_m\,. \tag{A.4}$$

We define the matrices

$$\begin{aligned}
\mathbf{K}^1_{k,l} &= [\theta_k, \theta_l]_\tau - \langle \theta_k, \partial_\tau \theta_l \rangle\,,\\
\mathbf{M}_{k,l} &= \langle \theta_k, \theta_l \rangle\,,\\
\mathbf{K}^\xi_{k,l} &= \langle \partial_\xi \theta_k, \theta_l \rangle\,,\\
\mathbf{K}^b_{k,l} &= \langle \partial_\tau b(\theta_k), \theta_l \rangle\,,\\
(\mathbf{F}_0)_{k,m} &= \int_0^1 \theta_k(\xi, 0)\psi_m(\xi)d\xi\,,\\
\mathcal{F}(\mathcal{Q})_k &= f(q^i_k)\,,\\
\mathcal{S}(\mathcal{Q})_k &= s(q^i_k)\,.
\end{aligned} \tag{A.5}$$

Additionally, in this section we consider the L^1-norm, which for simplicity is denoted by $\|\cdot\|$.

Lemma A.1. *If $f(q)$ and $s(q)$ are differentiable functions of q. The operator*

$$\mathbf{T}(\mathcal{Q}) := \mathbf{F}_0\tilde{\mathbf{\Gamma}}^i - \mathbf{R}(\mathcal{Q})\,, \tag{A.6}$$

with

$$\mathbf{R}(\mathcal{Q}) := \Delta t(\mathbf{K}^1 + \mathbf{K}^{\mathbf{b}})^{-1}\Big(\frac{1}{\Delta x}\mathbf{K}^\xi \mathcal{F}(\mathcal{Q}) - \mathbf{M}\mathcal{S}(\mathcal{Q})\Big)\,, \tag{A.7}$$

is Lipschitz, with constant

$$K := \Delta t||(\mathbf{K}^1 + \mathbf{K}^{\mathbf{b}})||\Big(\frac{1}{\Delta x}||\mathbf{K}^\xi|| \cdot ||\mathcal{A}_F|| + ||\mathbf{M}|| \cdot ||\mathcal{A}_S||\Big)\,. \tag{A.8}$$

where $\mathcal{A}_F$ is the Jacobian of $\mathcal{F}(\mathcal{Q})$ with respect to $\mathcal{Q}$ and $\mathcal{A}_S$ is the Jacobian of $\mathcal{S}(\mathcal{Q})$ with respect to $\mathcal{Q}$.

Proof.

$$\begin{aligned}
\mathbf{T}(\mathcal{Q}_1) - \mathbf{T}(\mathcal{Q}_2) = (\mathbf{K}^1 + \mathbf{K}^{\mathbf{b}})^{-1}\Big(&\frac{\Delta t}{\Delta x}\mathbf{K}^\xi\Big(\mathcal{F}(\mathcal{Q}_1) - \mathcal{F}(\mathcal{Q}_2)\Big)\Big)\\
&-\Delta t\mathbf{M}\Big(\mathcal{S}(\mathcal{Q}_1) - \mathcal{S}(\mathcal{Q}_2)\Big)\Big)\,.
\end{aligned} \tag{A.9}$$

We note that $\mathcal{Q}_1 = [(q_1)_1, \ldots, (q_1)_{MD}]^T$ and $\mathcal{Q}_2 = [(q_2)_1, \ldots, (q_2)_{MD}]^T$. As $f(q)$ and $s(q)$ are differentiable functions of q. Then, there exist constants θ_k and η_k such that

$$\begin{aligned}
f((q_1)_k) - f((q_2)_k) &= \lambda(\theta_k)((q_1)_k - (q_2)_k)\,,\\
s((q_1)_k) - s((q_2)_k) &= \beta(\theta_k)((q_1)_k - (q_2)_k)\,,
\end{aligned} \tag{A.10}$$

with $\lambda = df(q)/dq$ and $\beta = ds(q)/dq$. Therefore,

$$\begin{aligned}
\mathcal{F}(\mathcal{Q}_1) - \mathcal{F}(\mathcal{Q}_2) &= \mathcal{A}_F(\theta)(\mathcal{Q}_1 - \mathcal{Q}_2)\,,\\
\mathcal{S}(\mathcal{Q}_1) - \mathcal{S}(\mathcal{Q}_2) &= \mathcal{A}_S(\eta)(\mathcal{Q}_1 - \mathcal{Q}_2)\,,
\end{aligned} \tag{A.11}$$

where $\mathcal{A}_F(\theta)$ and $\mathcal{A}_S(\beta)$ are diagonal matrices defined by $\mathcal{A}_F(\theta)_{k,k} = \lambda(\theta_k)$ and $\mathcal{A}_S(\eta)_{k,k} = \beta(\eta_k)$, respectively. Taking the norm of (A.9) and after some simple manipulations, the sought result is obtained. □

Proposition A.2. *If $\mathcal{L}(q)$ is a linear operator, $f(q)$ and $s(q)$ continuously differentiable functions. Then the problem (15) contains a local solution.*

Proof. We note that a solution of (15), $\mathcal{Q}^*$, satisfies

$$\mathcal{Q}^* = \mathbf{T}(\mathcal{Q}^*) \,. \tag{A.12}$$

From the Lemma A.1 $\mathbf{T}(\mathcal{Q})$ is a Lipschitzian operator, now given $r > 0$ and $\mathcal{B}_r = \{\mathcal{Q} : ||\mathbf{F}_0\tilde{\mathbf{\Gamma}}^i - \mathcal{Q}|| < r\}$. Then there exists a $\delta > 0$ such that $\Delta t < \delta$ and $\mathbf{T} : \mathcal{B}_r \to \mathcal{B}_r$. Indeed, given a $\mathcal{Q} \in \mathcal{B}_r$

$$sup||\mathbf{T}(\mathcal{Q}) - \mathbf{F}_0\tilde{\mathbf{\Gamma}}^i|| = K \,.$$

Therefore, if we take $\Delta t < \delta$

$$||\mathbf{T}(\mathcal{Q}) - \mathbf{F}_0\tilde{\mathbf{\Gamma}}^i|| < \delta\mathcal{M} \,,$$

with $\mathcal{M} = (\mathbf{K}^1 + \mathbf{K}^{\mathbf{b}})^{-1}\Big(\frac{1}{\Delta x}\mathbf{K}^{\xi}\mathcal{F}(\mathcal{Q}) - \mathbf{M}\mathcal{S}(\mathcal{Q})\Big)$. Hence, δ can be chosen in order to obtain $\delta\mathcal{M} < r$. Finally, if we take $\mathcal{Q}_1, \mathcal{Q}_2 \in \mathcal{B}_r$, then

$$||\mathbf{T}(\mathcal{Q}_1) - \mathbf{T}(\mathcal{Q}_2)|| < K(\mathcal{Q}_1 - \mathcal{Q}_2) < \delta\mathcal{M}(\mathcal{Q}_1 - \mathcal{Q}_2) \,,$$

so, δ can be chosen in order to obtain $\delta < min(r, 1)$. Therefore, $\mathbf{T}$ is a contraction in $\mathcal{B}_r$ and thus $\mathbf{T}$ has a unique fixed point. Therefore, the result holds. □

Definition A.3. A scheme is said to be of Total-Variation-Diminishing (TVD) if

$$TV(q^{n+1}) \leq TV(q^n) \,, \forall n \,, \tag{A.13}$$

where

$$TV(q^n) = \sum_{i=-\infty}^{\infty} |q_i^n - q_{i-1}^n| \,. \tag{A.14}$$

On the other hand, the finite volume schemes have the form

$$w_i^{n+1} = w_i^n - \frac{\Delta t}{\Delta x}[f_{i+\frac{1}{2}} - f_{i-\frac{1}{2}}] + \Delta t s_i \,, \tag{A.15}$$

where $f_{i+\frac{1}{2}}$ is the flux defined by

$$\begin{aligned} f_{i+\frac{1}{2}} &= \frac{1}{\Delta t}\int_0^{\Delta t} f(q(x_{i+\frac{1}{2}}, t))dt \,, \\ s_i &= \frac{1}{\Delta t \Delta x}\int_0^{\Delta t}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} s(q(x,t))dxdt \,. \end{aligned} \tag{A.16}$$

We use a quadrature rule in [0, 1], which is defined by n_G pairs (η_l, ω_l) and thus

$$\begin{aligned} f_{i+\frac{1}{2}} &= \sum_{l=0}^{n_G} f(q(x_{i+\frac{1}{2}}, \eta_l \Delta t))\omega_l \,, \\ s_i &= \sum_{l=0}^{n_G}\sum_{k=0}^{n_G} s(q(x_{i-\frac{1}{2}} + \eta_l\Delta x, \eta_k \Delta t))\omega_l\omega_k \,. \end{aligned} \tag{A.17}$$

We will adopt here the Harten philosophy, see [18,36,37] and references therein for further details. So the numerical flux (A.17) is computed, without loss of generality, using the Rusanov flux

$$\begin{aligned} f(q(x_{i+\frac{1}{2}}, \tau\Delta t)) \quad = \quad & \tfrac{1}{2}(f(q_i(1_-, \tau) + f(q_{i+1}(0_+, \tau)) \\ & - \tfrac{\lambda_{i+\frac{1}{2}}}{2}(q_{i+1}(0_+, \tau) - q_i(1_-, \tau)) \,, \end{aligned} \tag{A.18}$$

where $\lambda_{i+\frac{1}{2}}$ is the maximum wave speed inside the interval

$$[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \cup [x_{i+\frac{1}{2}}, x_{i+\frac{3}{2}}] \,.$$

However, we can choose these values as $\lambda_{i+\frac{1}{2}} = \lambda_\infty = c\frac{\Delta x}{\Delta t}$, where c is the Courant number. Therefore, we have

$$\begin{aligned} q_i^{n+1} = q_i^n &- \frac{\Delta t}{\Delta x}\sum_{k=0}^{n_G}\frac{1}{2}\Big(f(q_i(1_-, \eta_k)) - f(q_{i-1}(1_-, \eta_k))\Big)\omega_k \\ &- \frac{\Delta t}{\Delta x}\sum_{k=0}^{n_G}\frac{1}{2}\Big(f(q_{i+1}(0_+, \eta_k)) - f(q_i(0_+, \eta_k))\Big)\omega_k \\ &+ \frac{\Delta t}{\Delta x}\sum_{k=0}^{n_G}\frac{\lambda_\infty}{2}\Big(q_{i+1}(0_+, \eta_k) - q_i(0_+, \eta_k)\Big)\omega_k \\ &- \frac{\Delta t}{\Delta x}\sum_{k=0}^{n_G}\frac{\lambda_\infty}{2}\Big(q_i(1_-, \eta_k) - q_{i-1}(1_-, \eta_k)\Big)\omega_k \\ &+ \Delta t\sum_{l=0}^{n_G}\sum_{k=0}^{n_G} s(q_i(\eta_l, \eta_k))\omega_l\omega_k \,. \end{aligned} \tag{A.19}$$

On the other hand, $q_k(\xi, \tau)$, with $k = i-1, i, i+1$ are polynomials defined in terms of local variables (ξ, τ) corresponding to intervals $[x_{i-\frac{3}{2}}, x_{i-\frac{1}{2}}]$, $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$ and $[x_{i+\frac{1}{2}}, x_{i+\frac{3}{2}}]$, respectively. From Section 2.2, these polynomials have the form

$$q_r(\xi, \tau) = \sum_{k=0}^{MD}\phi_k(\xi, \tau)\hat{q}_k^r \,. \tag{A.20}$$

As seen in previous sections, coefficients $\hat{q}_k^r$ are obtained from the solution to the algebraic problem

$$(\mathbf{K}^1 + \mathbf{K}^b)\mathcal{Q}^r + \frac{\Delta t}{\Delta x}\mathbf{K}^{\xi}\mathcal{F}(\mathcal{Q}^r) - \Delta t\mathbf{M}\mathcal{S}(\mathcal{Q}^r) = \mathbf{F}_0\tilde{\Gamma}^r \,. \tag{A.21}$$

Here, $\mathbf{K^1}$, $\mathbf{K}^{\xi}$ and $\mathbf{F}_0$ are *finite element* matrices and; $\mathcal{Q}_j^r = \hat{q}_j^r$ and $\tilde{\Gamma}_j^r = \gamma_j^r$. With these observations we note that

$$\begin{aligned} q_{i+1}(0_+, \tau) - q_i(0_+, \tau) &= \sum_{k=0}^{MD}\phi_k(0_+, \tau)(\hat{q}_k^{i+1} - \hat{q}_k^i) \,, \\ q_i(1_-, \tau) - q_{i-1}(1_-, \tau) &= \sum_{k=0}^{MD}\phi_k(1_-, \tau)(\hat{q}_k^i - \hat{q}_k^{i-1}) \,. \end{aligned} \tag{A.22}$$

Lemma A.4. *If $\mathcal{L}(q)$ is a linear operator, $f(q)$ and $s(q)$ are continuously differentiable functions of q. Then*

(a) *there exist constant vectors $\tilde{\lambda} = [\tilde{\lambda}_1^r, ..., \tilde{\lambda}_{M^2}^r]^T$ and $\tilde{\beta} = [\tilde{\beta}_1^r, ..., \tilde{\beta}_{M^2}^r]^T$ such that $\tilde{\beta}_k^r, \tilde{\lambda}_k^r \in [min\{q_k^r, q_k^{r+1}\}, max\{q_k^r, q_k^{r+1}\}]$.*

$$\begin{aligned} \mathcal{F}(\mathcal{Q}^{r+1}) - \mathcal{F}(\mathcal{Q}^r) &= \mathcal{A}_F(\tilde{\lambda}^r)(\mathcal{Q}^{r+1} - \mathcal{Q}^r) \,, \\ \mathcal{S}(\mathcal{Q}^{r+1}) - \mathcal{S}(\mathcal{Q}^r) &= \mathcal{A}_S(\tilde{\beta}^r)(\mathcal{Q}^{r+1} - \mathcal{Q}^r) \,, \end{aligned} \tag{A.23}$$

with $\mathcal{A}_F$ and $\mathcal{A}_S$ the jacobians of $\mathcal{F}(\mathcal{Q})$ and $\mathcal{S}(\mathcal{Q})$ with respect to $\mathcal{Q}$, respectively.

(b) *The following identity is satisfied:*

$$\mathcal{Q}^{r+1} - \mathcal{Q}^r = \mathbf{C}^r(\tilde{\mathbf{q}}^{r+1} - \tilde{\mathbf{q}}^r) \,, \tag{A.24}$$

with

$$\mathbf{C}^r = (\mathbf{K}^1 + \mathbf{K}^b + \frac{\Delta t}{\Delta x}\mathbf{K}^{\xi}\mathbf{A}_F(\tilde{\lambda}^r) - \Delta t\mathbf{M}\mathbf{A}_s(\tilde{\beta}^r))^{-1}(\mathbf{F}_0\mathbf{B}^{-1}). \tag{A.25}$$

Proof. Point a) is a consequence of the mean value theorem applied to $f(q)$ and $s(q)$. To prove point b), we note that, from (A.3) we have

$$\begin{aligned} \mathbf{K}^1(\mathcal{Q}^{r+1} - \mathcal{Q}^r) + \mathbf{K}^{\xi}(\mathcal{F}(\mathcal{Q}^{r+1}) - \mathcal{F}(\mathcal{Q}^r)) - \Delta t\mathbf{M}(\mathcal{S}(\mathcal{Q}^{r+1}) \\ - \mathcal{S}(\mathcal{Q}^r)) = \mathbf{F}_0\mathbf{B}^{-1}(\tilde{q}^{r+1} - \tilde{q}^r) \,. \end{aligned} \tag{A.26}$$

So, from point a) we obtain

$$\begin{aligned} (\mathbf{K}^1 + \mathbf{K}^b + \frac{\Delta t}{\Delta x}\mathbf{K}^{\xi}\mathbf{A}_F(\tilde{\lambda}^r) - \Delta t\mathbf{M}\mathbf{A}_s(\tilde{\beta}^r))(\mathcal{Q}^{r+1} - \mathcal{Q}^r) \\ = \mathbf{F}_0\mathbf{B}^{-1}(\tilde{\mathbf{q}}^{r+1} - \tilde{\mathbf{q}}^r) \,. \end{aligned} \tag{A.27}$$

After some algebraic manipulation, the result holds. □

Let us assume that the problem (4) is solved numerically. As $\mathcal{L}$ is assumed to be linear, then there exists a matrix $\mathbf{A}_L$ such that

$$\mathbf{w}^r = \mathbf{A}_L\mathbf{q}^r \,. \tag{A.28}$$

Lemma A.5. *If $\mathcal{L}(q)$ is a linear operator. Then, the following is satisfied:*

$$TV(w^n) \leq \mathcal{D}_L ||q^n|| \,. \tag{A.29}$$

Proof.

$$w_i^n - w_{i-1}^n = \sum_j ((\mathbf{A}_L)_{i,j} - (\mathbf{A}_L)_{i-1,j}) q_j^n \,. \tag{A.30}$$

Then, we have

$$|w_i^n - w_{i-1}^n| \leq \alpha_i \sum_j |q_j^n| \Delta x \,, \tag{A.31}$$

with $max_j \frac{|(\mathbf{A}_L)_{i,j} - (\mathbf{A}_L)_{i-1,j}|}{\Delta x} = \alpha_i$. Therefore the result holds with

$$\mathcal{D}_L = \sum_i \alpha_i \,. \tag{A.32}$$

□

Lemma A.6. *If $\mathcal{L}(q)$ is a linear operator, $f(q)$ and $s(q)$ are continuously differentiable functions of q. The following identity is satisfied:*

$$q_i(\xi,\tau) - q_{i-1}(\xi,\tau) = \sum_{j=-k_L}^{k_R} \psi_j^i(\xi,\tau)(q_{i+j}^n - q_{i+j-1}^n) \,, \tag{A.33}$$

with

$$\psi_j^i(\xi,\tau) = \sum_{k=0}^{M_D} \phi_k(\xi,\eta) \mathbf{C}_{k,j}^i \,. \tag{A.34}$$

Proof. Note that

$$q_i(\xi,\tau) - q_{i-1}(\xi,\tau) = \sum_{k=0}^{M_D} \phi_k(\xi,\eta)(\hat{q}_k^i - \hat{q}_k^{i-1}) \,. \tag{A.35}$$

So, by using the point b) of Lemma A.4, the result holds. □

Lemma A.7. *If $\mathcal{L}(q)$ is a linear operator, $f(q)$ and $s(q)$ are continuously differentiable functions of q. Then, the numerical scheme can be written as*

$$\begin{aligned} w_i^{n+1} = \; & w_i^n \\ & - \frac{\Delta t}{\Delta x} \sum_{j=-k_L}^{k_R} (-D_j^{i+1,+}) q_{i+j+1}^n \\ & - \frac{\Delta t}{\Delta x} \sum_{j=-k_L}^{k_R} (D_j^{i,-} - D_j^{i+1,+}) q_{i+j}^n \\ & - \frac{\Delta t}{\Delta x} \sum_{j=-k_L}^{k_R} (-D_j^{i,-}) q_{i+j-1}^n \\ & + \Delta t \sum_{l=0}^{n_G} \sum_{k=0}^{n_G} s(q_i(\eta_l,\eta_k)) \omega_l \omega_k \,, \end{aligned} \tag{A.36}$$

with

$$D_j^{i+1,+} = \sum_{k=0}^{n_G} \left(\frac{\lambda_\infty + \lambda(\tilde{\lambda}_k^{i+1})}{2}\right) \psi_j^{i+1}(0_+,\eta_k)\omega_k$$

and

$$D_j^{i,-} = \sum_{k=0}^{n_G} \left(\frac{\lambda_\infty + \lambda(\tilde{\lambda}_k^{i})}{2}\right) \psi_j^{i}(1_-,\eta_k)\omega_k \,,$$

where $\psi_j^i(\xi,\tau)$ is defined in Lemma A.6.

Proof. The numerical scheme has the form

$$\begin{aligned} w_i^{n+1} = \; & w_i^n - \tfrac{\Delta t}{\Delta x} \textstyle\sum_{k=0}^{n_G} \tfrac{1}{2} \Big(f(q_i(1_-,\eta_k)) - f(q_{i-1}(1_-,\eta_k)) \Big) \omega_k \\ & + \tfrac{\Delta t}{\Delta x} \textstyle\sum_{k=0}^{n_G} \tfrac{1}{2} \Big(f(q_{i+1}(0_+,\eta_k)) - f(q_i(0_+,\eta_k)) \Big) \omega_k \\ & + \tfrac{\Delta t}{\Delta x} \textstyle\sum_{k=0}^{n_G} \tfrac{\lambda_\infty}{2} \Big(q_{i+1}(0_+,\eta_k) - q_i(0_+,\eta_k) \Big) \omega_k \\ & - \tfrac{\Delta t}{\Delta x} \textstyle\sum_{k=0}^{n_G} \tfrac{\lambda_\infty}{2} \Big(q_i(1_-,\eta_k) - q_{i-1}(1_-,\eta_k) \Big) \omega_k \\ & + \Delta t \textstyle\sum_{l=0}^{n_G} \sum_{k=0}^{n_G} s(q_i(\eta_l,\eta_k)) \omega_l \omega_k \,. \end{aligned} \tag{A.37}$$

Then, if $f(q)$ is a continuously differentiable function, from Lemma A.4 there exist values $\tilde{\lambda}_k^i$ such that

$$\begin{aligned} w_i^{n+1} = \; & w_i^n - \frac{\Delta t}{\Delta x} \sum_{k=0}^{n_G} \left(\frac{\lambda_\infty + \lambda(\tilde{\lambda}_k^{i})}{2}\right) \Big(q_i(1_-,\eta_k) - q_{i-1}(1_-,\eta_k) \Big) \omega_k \\ & + \frac{\Delta t}{\Delta x} \sum_{k=0}^{n_G} \left(\frac{\lambda_\infty + \lambda(\tilde{\lambda}_k^{i+1})}{2}\right) \Big(q_{i+1}(0_+,\eta_k) - q_i(0_+,\eta_k) \Big) \omega_k \\ & + \Delta t \sum_{l=0}^{n_G} \sum_{k=0}^{n_G} s(q_i(\eta_l,\eta_k)) \omega_l \omega_k \,. \end{aligned} \tag{A.38}$$

Additionally, from Lemma A.6

$$q_i(\xi,\tau) - q_{i-1}(\xi,\tau) = \sum_{j=-k_L}^{k_R} \psi_j^i(\xi,\tau)(q_{i+j}^n - q_{i+j-1}^n) \tag{A.39}$$

and thus, we obtain

$$\begin{aligned} w_i^{n+1} = \; & w_i^n \\ & - \frac{\Delta t}{\Delta x} \sum_{k=0}^{n_G} \left(\frac{\lambda_\infty + \lambda(\tilde{\lambda}_k^{i})}{2}\right) \sum_{j=-k_L}^{k_R} \psi_j^i(1_-,\eta_k) \Big(q_{i+j}^n - q_{i+j-1}^n \Big) \omega_k \\ & + \frac{\Delta t}{\Delta x} \sum_{k=0}^{n_G} \left(\frac{\lambda_\infty + \lambda(\tilde{\lambda}_k^{i+1})}{2}\right) \sum_{j=-k_L}^{k_R} \psi_j^{i+1} \\ & (0_+,\eta_k) \Big(q_{i+j+1}^n - q_{i+j}^n \Big) \omega_k \\ & + \Delta t \sum_{l=0}^{n_G} \sum_{k=0}^{n_G} s(q_i(\eta_l,\eta_k)) \omega_l \omega_k \,. \end{aligned} \tag{A.40}$$

Manipulating the last expression, we obtain

$$\begin{aligned} w_i^{n+1} = \; & w_i^n - \frac{\Delta t}{\Delta x} \sum_{j=-k_L}^{k_R} D_j^{i,-} \Big(q_{i+j}^n - q_{i+j-1}^n \Big) \\ & + \frac{\Delta t}{\Delta x} \sum_{j=-k_L}^{k_R} D_j^{i+1,+} \Big(q_{i+j+1}^n - q_{i+j}^n \Big) \\ & + \Delta t \sum_{l=0}^{n_G} \sum_{k=0}^{n_G} s(q_i(\eta_l,\eta_k)) \omega_l \omega_k \end{aligned} \tag{A.41}$$

and so the result follows. □

Lemma A.8. *If $\mathcal{L}(q)$ is a linear operator, $f(q)$ and $s(q)$ are continuously differentiable functions of q. The following identity is obtained:*

$$
\begin{aligned}
w_i^{n+1} - w_{i-1}^{n+1} &= w_i^n - w_{i-1}^n \\
&- \frac{\Delta t}{\Delta x} \sum_{j=-k_L}^{k_R} D_j^{i,-} \left(q_{i+j}^n - q_{i+j-1}^n \right) \\
&+ \frac{\Delta t}{\Delta x} \sum_{j=-k_L}^{k_R} D_j^{i-1,-} \left(q_{i+j-1}^n - q_{i+j-2}^n \right) \\
&+ \frac{\Delta t}{\Delta x} \sum_{j=-k_L}^{k_R} D_j^{i+1,+} \left(q_{i+j+1}^n - q_{i+j}^n \right) \\
&- \frac{\Delta t}{\Delta x} \sum_{j=-k_L}^{k_R} D_j^{i,+} \left(q_{i+j}^n - q_{i+j-1}^n \right) \\
&+ \Delta t \sum_{j=-k_L}^{k_R} \bar{D}_j^i (q_{i+j}^n - q_{i+j-1}^n) \, , \qquad \text{(A.42)}
\end{aligned}
$$

with

$$
\bar{D}_j^i = \sum_{l=0}^{n_G} \sum_{k=0}^{n_G} \beta(\tilde{\beta}_k^i) \psi_j^i(\eta_l, \eta_k) \omega_l \omega_k \, , \qquad \text{(A.43)}
$$

where $\psi_j^i(\xi, \tau)$ is given in Lemma A.6. $D_j^{i,-}$ and $D_j^{i+1,+}$ are given in Lemma A.7.

Proof. From Lemma A.6, we have

$$
\begin{aligned}
w_i^{n+1} - w_{i-1}^{n+1} &= w_i^n - w_{i-1}^n \\
&- \frac{\Delta t}{\Delta x} \sum_{j=-k_L}^{k_R} D_j^{i,-} \left(q_{i+j}^n - q_{i+j-1}^n \right) \\
&+ \frac{\Delta t}{\Delta x} \sum_{j=-k_L}^{k_R} D_j^{i-1,-} \left(q_{i+j-1}^n - q_{i+j-2}^n \right) \\
&+ \frac{\Delta t}{\Delta x} \sum_{j=-k_L}^{k_R} D_j^{i+1,+} \left(q_{i+j+1}^n - q_{i+j}^n \right) \\
&- \frac{\Delta t}{\Delta x} \sum_{j=-k_L}^{k_R} D_j^{i,+} \left(q_{i+j}^n - q_{i+j-1}^n \right) \\
&+ \Delta t \sum_{l=0}^{n_G} \sum_{k=0}^{n_G} (s(q_i(\eta_l, \eta_k)) \\
&- s(q_{i-1}(\eta_l, \eta_k))) \omega_l \omega_k \, . \qquad \text{(A.44)}
\end{aligned}
$$

From Lemma A.4 there exist coefficients

$$
\tilde{\beta}_{l,k}^i \in [min(q_i(\eta_l, \eta_k), q_{i-1}(\eta_l, \eta_k)), max(q_i(\eta_l, \eta_k), q_{i-1}(\eta_l, \eta_k))] \, ,
$$

such that

$$
\begin{aligned}
&s(q_i(\eta_l, \eta_k)) - s(q_{i-1}(\eta_l, \eta_k)) \\
&\quad = \beta(\tilde{\beta}_{l,k}^i)(q_i(\eta_l, \eta_k) - q_{i-1}(\eta_l, \eta_k)) \, , \qquad \text{(A.45)}
\end{aligned}
$$

where $\beta(q) = \frac{ds(q)}{dq}$. Thus the result holds. □

Lemma A.9. *If $\mathcal{L}(q)$ is a linear operator, $f(q)$ and $s(q)$ are continuously differentiable functions of q. Then, there exist matrices of size N × N; $\mathcal{E}$, $\mathcal{M}$ and $\mathbf{A}_b$ such that*

$$
\mathbf{w}^{n+1} = (\mathbb{I} - \Delta t(\mathcal{E} - \mathcal{M})\mathbf{A}_b)\mathbf{w}^n \, , \qquad \text{(A.46)}
$$

where

$$
\mathbf{w}^n = [w_1^n, ..., w_N^n] \, , \qquad \text{(A.47)}
$$

with N the number of cells.

Proof. From Lemma A.7, we have

$$
\mathbf{w}^{n+1} = \mathbf{w}^n - \Delta t \mathcal{E} \mathbf{q}^n + \Delta t \mathbf{s}^n \, , \qquad \text{(A.48)}
$$

with

$$
\begin{aligned}
\mathcal{E}_{i,k_L-1} &= \frac{D_{k_L}^{i,-}}{\Delta x} \, , \\
\mathcal{E}_{i,j} &= \frac{(-D_{j-1}^{i+1,+}) + (D_j^{i,-} - D_j^{i+1,+}) + (-D_{j+1}^{i,-})}{\Delta x} \, , \; j = -k_L, ..., k_R \, , \\
\mathcal{E}_{i,k_R+1} &= \frac{D_{k_R}^{i+1,+}}{\Delta x} \qquad \text{(A.49)}
\end{aligned}
$$

and $\mathbf{s}^n = [s_1^n, ..., s_N^n]^T$, where $s_i^n = \sum_{l=0}^{n_G} \sum_{k=0}^{n_G} s(q_i(\eta_l, \eta_k)) \omega_l \omega_k$.

As $s(q)$ is continuously differentiable, there exist constant values $\theta_{l,k}$ such that $s(q_i(\eta_l, \eta_k)) = \beta(\theta_{l,k}) q_i(\eta_l, \eta_k)$. Therefore, after some manipulations we obtain

$$
s_i = \sum_{j=-K_L}^{k_R} \mathcal{M}_{i,j} q_{i+j}^n \, , \qquad \text{(A.50)}
$$

with

$$
\mathcal{M}_{i,j} = \sum_{l=0}^{n_G} \sum_{k=0}^{n_G} \beta(\theta_{l,k}) \psi_j^i(\eta_l, \eta_k) \omega_l \omega_k \, . \qquad \text{(A.51)}
$$

Therefore, if we take $\mathbf{A}_b = \mathbf{A}_L^{-1}$, with $\mathbf{A}_L^{-1}$ in (A.28) the result holds. □

Remark 2. The matrices $\mathcal{E}$, $\mathcal{M}$ and $\mathbf{A}_b$ are square matrices. It is assumed that boundary conditions modify the entries of these matrices, but the structure of a $k_R + k_L + 1$-diagonal matrix is assumed to be held.

Lemma A.10. *If $\mathcal{L}(q)$ is a linear operator, $f(q)$ and $s(q)$ are continuously differentiable functions of q. Then*

$$
||\mathbf{w}^{n+1}|| \leq ||\mathbf{w}^n|| \kappa(\Delta t) \, , \qquad \text{(A.52)}
$$

with $\kappa(\Delta t) = ||\mathbf{I} - \Delta t(\mathcal{E} - \mathcal{M})\mathbf{A}_b||$.

Proof. It is a consequence of Lemma A.9. □

We note that there exists a range of values of Δt such that $\kappa(\Delta t) < 1$. It is proved in the next.

Lemma A.11. *If $\mathcal{L}(q) = a_0 q + a_1 b(q)$, with $a_0 > 0$ and a_1 constant values, $b(q)$ a linear operator containing only spatial derivatives, f and s are continuously differentiable functions with $\frac{ds(q)}{dq} = \beta(q) \leq 0$. Then, there exists a range of values of Δt such that*

$$
\kappa(\Delta t) < 1 \, .
$$

Proof. As the result is independent of the initial condition, let us consider an initial condition given by $q(x, 0) = k_0$ a constant. On the other hand, by the mean value theorem, there exists a constant θ_0 such that $s^*(q) = \beta_0 q$, with $\Delta t \beta(\theta_0) = \beta_0$. Additionally, let us consider that the solution at time t^n is given by $w_i^n = k_0 e^{\frac{\beta_0}{a_0} t^n}$. Thus, the weak formulation yields

$$
< \theta_k, \partial_\tau \mathcal{L}(\bar{q}) > + < \theta_k, \partial_\xi f^*(\bar{q}) > = < \theta_k, s^*(\bar{q}) > \, ,
$$

which results into

$$
< \theta_k, \partial_\tau \theta_l > a_0 w_i^n \hat{q}_l = < \theta_k, \theta_l > \beta_0 w_i^n \hat{q}_l \, .
$$

Note that the solution of this problem is $\bar{q}_i(\xi, \tau) \equiv \frac{\beta_0}{a_0} w_i^n$. The exact solution of the generalized Riemann problem provides

$$
q_i(\tau) = w_i^n e^{\frac{\beta_0}{a_0} \tau} \, . \qquad \text{(A.53)}
$$

Therefore, the numerical flux and the numerical source are

$$f_{i+\frac{1}{2}} = \sum_{k=0}^{n_{GP}} \frac{w_i^n}{\Delta t}(e^{\frac{\beta_0}{a_0}\tau_k \Delta t} - 1)\omega_k \,, \qquad s_i = \sum_{k=0}^{n_{GP}} \frac{w_i^n}{\Delta t}(e^{\frac{\beta_0}{a_0}\tau_k \Delta t} - 1)\omega_k \,. \tag{A.54}$$

So, the numerical scheme produces the result

$$w_i^{n+1} = w_i^n \sum_{k=0}^{n_{GP}} e^{\frac{\beta_0}{a_0}\tau_k \Delta t}\omega_k \,, \tag{A.55}$$

which finally gives

$$||w^{n+1}|| \leq ||w^0|| \kappa(\Delta t) \,, \tag{A.56}$$

with $||w^0|| = k_0$ and $\kappa(\Delta t) = \sum_{k=0}^{n_{GP}} e^{\frac{\beta_0}{a_0}((n+\tau_k)\Delta t)}\omega_k$. Thus, the result holds. □

Lemma A.12. *If $\mathcal{L}(q)$ is a linear operator. Then, there exists a constant $\mathcal{A}$ such that*

$$TV(\mathbf{q}^n) \leq \mathcal{A}||\mathbf{w}^n|| \,. \tag{A.57}$$

Proof. By taking $\mathcal{A} = ||\mathbf{A}_L^{-1}||$, with $\mathbf{A}_L$ in (A.28) the result holds. □

Lemma A.13. *If $\mathcal{L}(q)$ is a linear operator, $f(q)$ and $s(q)$ are continuously differentiable functions of q. Then the following identity is satisfied:*

$$TV(\mathbf{w}^{n+1}) \leq TV(\mathbf{w}^n) + \Delta t \mathcal{D} TV(\mathbf{q}^n) \,, \tag{A.58}$$

with

$$\mathcal{D} = \sum_i \sum_{j=-k_L}^{k_R} \left\{ \frac{|D_j^{i,-}|}{\Delta x} + \frac{|D_j^{i-1,-}|}{\Delta x} + \frac{|D_j^{i+1,+}|}{\Delta x} + \frac{|D_j^{i,+}|}{\Delta x} + |\bar{D}_j^i| \right\} \tag{A.59}$$

Proof. The result follows directly from Lemma A.8. After applying the absolute value and summation on index i. □

Lemma A.14. *There exists a constant $\mathcal{A}$ such that*

$$TV(\mathbf{w}^{n+1}) \leq TV(\mathbf{w}^n) + \Delta t \mathcal{D}\mathcal{A}||\mathbf{w}^n|| \,. \tag{A.60}$$

Proof. The proof follows directly from Lemmas A.12 and A.13. □

Proposition A.15. *If $\mathcal{L}(q)$ is a linear operator, $f(q)$ and $s(q)$ are continuously differentiable functions of q. There exists a constant $\bar{C}$ such that*

$$TV(\mathbf{w}^{n+1}) \leq \bar{C} \,, \tag{A.61}$$

with

$$\bar{C} = TV(\mathbf{w}^0) + \Delta t \mathcal{D}\mathcal{A}\delta||\mathbf{w}^0|| \tag{A.62}$$

and

$$\delta = \frac{1}{1 - \kappa(\Delta t)} \,. \tag{A.63}$$

Proof. A successive application of Lemma A.14, provides

$$TV(\mathbf{w}^{n+1}) \leq TV(\mathbf{w}^0) + \Delta t \mathcal{D}\mathcal{A} \sum_{r=0}^{n} ||\mathbf{w}^r|| \,. \tag{A.64}$$

From Lemma A.10, we have

$$\sum_{r=0}^{n} ||\mathbf{w}^r|| \leq ||\mathbf{w}|| \sum_{r=0}^{n} \kappa(\Delta t)^r \,. \tag{A.65}$$

From Lemma A.11, there exists a range of values Δt such that $\kappa(\Delta t) < 1$ and thus the result holds. □

Lemma A.16. *Let $\mathcal{L}(q)$ be a linear operator, $f(q)$ and $s(q)$ continuously differentiable functions of q. Then, given n and p two integers, the following identity is satisfied:*

$$|w_i^n - w_i^p| \leq C_1 |t^n - t^p| \,, \tag{A.66}$$

with $C_1 = (TV(\mathbf{w}^0) + \mathcal{D}\mathcal{A}\delta||\mathbf{w}^0||)$, $t^n = n\Delta t$ and $t^p = p\Delta t$.

Proof. From Lemmas A.6 and A.10,

$$|w_i^{n+1} - w_i^n| \leq \Delta t C_0 \,,$$

with

$$C_0 = \sum_i \sum_{j=-k_L}^{k_R} \left\{ \frac{|D_j^{i,-}|}{\Delta x} + \frac{|D_j^{i+1,+}|}{\Delta x} + |\bar{D}_j^i| \right\} \mathcal{A}\delta||\mathbf{w}^0|| \leq \Delta t \mathcal{D}\mathcal{A}\delta||\mathbf{w}^0|| \,. \tag{A.67}$$

On the other hand, without loss of generality, if we assume $n > p$, then

$$|w_i^n - w_i^p| \leq |w_i^n - w_i^{n-1}| + |w_i^{n-1} - w_i^{n-2}| + \ldots + |w_i^{p+1} - w_i^p| \,.$$

Thus, multiplying by Δx and taking summation on i, we obtain

$$||\mathbf{w}^n - \mathbf{w}^p|| \leq \Delta t C_0 (n-p) \leq (TV(\mathbf{w}^0) + \mathcal{D}\mathcal{A}\delta||\mathbf{w}^0||)(t^n - t^p) \,.$$

□

Appendix B. A globally implicit method. A numerical solution of reference

In order to assess the proposed methodology, we compare its numerical solutions against numerical solutions obtained through a *conventional* method on a very fine mesh. In this appendix we present the conventional scheme, which corresponds to a globally implicit finite difference scheme, given by

$$\left.\begin{array}{l} \frac{\zeta_i^{n+1} - \zeta_i^n}{\Delta t} + \mathbf{D}_i^{(1)}(h^{n+1}u^{n+1}) = \frac{\beta}{\varepsilon}\partial_t b(x_i, t^{n+\frac{1}{2}}) \,, \\ \frac{u_i^{n+1} - u_i^n}{\Delta t} - \frac{\mu}{3}\frac{\mathbf{D}_i^{(2)}(u^{n+1}) - \mathbf{D}_i^{(2)}(u^n)}{\Delta t} u_i^{n+1}\mathbf{D}_i^{(1)}(u^{n+1}) \\ = -\frac{\mu\beta}{2\varepsilon}\partial_{ttx} b(x_i, t^{n+\frac{1}{2}}) \,, \\ + \mathbf{D}_i^{(1)}(\zeta^{n+1}) + \varepsilon u_i^{n+1}\mathbf{D}_i^{(1)}(u^{n+1}) \end{array}\right\} \tag{B.1}$$

where

$$\left.\begin{array}{rcl} \mathbf{D}_i^{(2)}(u^{n+1}) & = & \frac{u_{i+1}^{n+1} - 2u_i^{n+1} + u_{i-1}^{n+1}}{\Delta x^2} \,, \\ \mathbf{D}_i^{(1)}(u^{n+1}) & = & \frac{u_{i+1}^{n+1} - u_{i-1}^{n+1}}{2\Delta x} \,, \\ \mathbf{D}_i^{(1)}(h^{n+1}u^{n+1}) & = & \frac{h_{i+1}^{n+1}u_{i+1}^{n+1} - h_{i-1}^{n+1}u_{i-1}^{n+1}}{2\Delta x} \,. \end{array}\right\} \tag{B.2}$$

System (B.2) is an algebraic system for variables $\{\zeta_i^{n+1}\}_{i=1}^N$ and $\{u_i^{n+1}\}_{i=1}^N$.

References

[1] Antonopoulos DC, Dougalis VA. Numerical solution of the classical Boussinesq system. Math Comput Simul 2012;82(6):984–1007. http://dx.doi.org/10.1016/j.matcom.2011.09.006.

[2] Antonopoulos DC, Dougalis VA. Error estimates for Galerkin approximations of the "classical" Boussinesq system. Math Comput 2013;82(282):689–717. http://www.scopus.com/inward/record.url?eid=2-s2.0-84873277964&partnerID=40&md5=d3c6fc3c0b0c67eee80e6bf1cff24690

[3] Antonopoulos DC, Dougalis VA, Mitsotakis DE. Numerical solution of Boussinesq systems of the Bona-Smith family. Appl Numer Math 2010;60(4):314–36. http://dx.doi.org/10.1016/j.apnum.2009.03.002. http://www.sciencedirect.com/science/article/pii/S0168927409000506

[4] Balsara DS, Rumpf T, Dumbser M, Munz CD. Efficient, high accuracy ADER-WENO schemes for hydrodynamics and divergence-free magnetohydrodynamics. J Comput Phys 2009;228(7):2480–516.

[5] Balsara DS, Meyer C, Dumbser M, Du H, Xu Z. Efficient implementation of ADER schemes for Euler and magnetohydrodynamical flows on structured meshes - speed comparisons with Runge–Kutta methods. J Comput Phys 2013;235(0):934–69.

[6] Barthélemy E. Nonlinear shallow water theories for coastal waves. Surv Geophys 2004;25(3–4):315–37. doi:10.1007/s10712-003-1281-7.

[7] Ben-Artzi M, Falcovitz J. A second order Godunov-type scheme for compressible fluid dynamics. J Comput Phys 1984;55(1):1–32.

[8] Ben-Artzi M, Falcovitz J. Generalized Riemann problems in computational fluid dynamics. Cambridge monographs on applied and computational mathematics. London: Cambridge University Press; 2003. ISBN 9781139439473. http://books.google.ca/books?id=zONALWXYiIEC

[9] Bernetti R, Titarev VA, Toro EF. Exact solution of the riemann problem for the shallow water equations with discontinuous bottom geometry. J Comput Phys 2008;227(6):3212–43. http://dx.doi.org/10.1016/j.jcp.2007.11.033.

[10] Bona JL, Dougalis VA, Karakashian OA. Fully discrete galerkin methods for the Korteweg-de Vries equation. Comput Math Appl 1986;12(7, Part A):859–84. http://dx.doi.org/10.1016/0898-1221(86)90031-3.

[11] Bona JL, Dougalis VA, Karakashian OA, McKinney WR. Conservative, high-order numerical schemes for the generalized Korteweg-de Vries equation. Philos Trans R Soc Lond A 1995;351:107–64.

[12] Bona JL, Dougalis VA, Mitsotakis DE. Numerical solution of KdV-KdV systems of Boussinesq equations. I: the numerical scheme and generalized solitary waves. Math Comput Simul 2007;74(2-3):214–28. http://www.scopus.com/inward/record.url?eid=2-s2.0-33846964163&partnerID=40&md5=f92c8c2c0ad86300f15758879b7e4113

[13] Bona JL, Dougalis VA, Mitsotakis DE. Numerical solution of Boussinesq systems of KdV-KdV type II: evolution of radiating solitary waves. Nonlinearity 2008;21(12):2825–48. http://www.scopus.com/inward/record.url?eid=2-s2.0-58149337147&partnerID=40&md5=dd4af893b728eb50f5299ca11a060600

[14] Bonneton P, Chazel F, Lannes D, Marche F, Tissier M. A splitting approach for the fully nonlinear and weakly dispersive Green–Naghdi model. J Comput Phys 2011;230(4):1479–98. http://dx.doi.org/10.1016/j.jcp.2010.11.015.

[15] Borsche R, Kall J. ADER schemes and high order coupling on networks of hyperbolic conservation laws. J Comput Phys 2014;273(0):658–70. http://dx.doi.org/10.1016/j.jcp.2014.05.042.

[16] Boscheri W, Dumbser M, Balsara DS. High-order ADER-WENO ALE schemes on unstructured triangular meshes-application of several node solvers to hydrodynamics and magnetohydrodynamics. Int J Numer Methods Fluids 2014;76(10):737–78. doi:10.1002/fld.3947.

[17] Boscheri W, Balsara DS, Dumbser M. Lagrangian ADER-WENO finite volume schemes on unstructured triangular meshes based on genuinely multidimensional HLL Riemann solvers. J Comput Phys 2014;267:112–38. doi:10.1016/j.jcp.2014.02.023.

[18] Castro CE, Toro EF. Solvers for the high-order Riemann problem for hyperbolic balance laws. J Comput Phys 2008;227:2481–513.

[19] Castro CE. High-order ADER FV/DG numerical methods for hyperbolic equations. Department of Civil and Environmental Engineering, University of Trento, Italy; 2007. Ph.D. thesis.

[20] Chazel F, Lannes D, Marche F. Numerical simulation of strongly nonlinear and dispersive waves using a Green–Naghdi model. J Sci Comput 2011;48(1–3):105–16. doi:10.1007/s10915-010-9395-9.

[21] Dumbser M, Käser M. Arbitrary high order non-oscillatory finite volume schemes on unstructured meshes for linear hyperbolic systems. J Comput Phys 2007;221(2):693–723.

[22] Dumbser M, Munz CD. ADER discontinuous Galerkin schemes for aeroacoustics. Comptes Rendus Mécanique 2005;333:683–7.

[23] Dumbser M, Munz CD. Building blocks for arbitrary high order discontinuous Galerkin schemes. J Sci Comput 2006;27:215–30.

[24] Dumbser M, Käser M, Toro EF. An arbitrary high order discontinuous Galerkin method for elastic waves on unstructured meshes V: local time stepping and *p*–adaptivity. Geophys J Int 2007;171:695–717.

[25] Dumbser M, Enaux C, Toro EF. Finite volume schemes of very high order of accuracy for stiff hyperbolic balance laws. J Comput Phys 2008;227(8):3971–4001.

[26] Dumbser M, Zanotti O, Hidalgo A, Balsara DS. ADER-WENO finite volume schemes with space-time adaptive mesh refinement. Commun Comput Phys 2013;248:257–86.

[27] Dumbser M, Zanotti O, Loubere R, Diot S. A posteriori subcell limiting of the discontinuous Galerkin finite element method for hyperbolic conservation laws. J Comput Phys 2014;278:47–75.

[28] Dumbser M, Hidalgo A, Zanotti O. High order space-time adaptive ADER-WENO finite volume schemes for non-conservative hyperbolic systems. Comput Methods Appl Mech Eng 2014;268(0):359–87. http://dx.doi.org/10.1016/j.cma.2013.09.022.

[29] Dumbser M. Arbitrary high order schemes for the solution of hyperbolic conservation laws in complex domains. Institut für Aero- un Gasdynamik, Universität Stuttgart, Germany; 2005. Ph.D. thesis.

[30] Dutykh D, Katsaounis T, Mitsotakis D. Finite volume schemes for dispersive wave propagation and runup. J Comput Phys 2011;230(8):3035–61. http://dx.doi.org/10.1016/j.jcp.2011.01.003.

[31] Fang KZ, Zhang Z, Zou ZL, Liu ZB, Sun JW. Modelling of 2-D extended Boussinesq equations using a hybrid numerical scheme. J Hydrodynam B 2014;26(2):187–98. http://dx.doi.org/10.1016/S1001-6058(14)60021-4.

[32] Garcia Alvarado MG, Omel'yanov GA. Interaction of solitary waves for the generalized KdV equation. Commun Nonlinear Sci Numer Simul 2012;17(8):3204–18. http://dx.doi.org/10.1016/j.cnsns.2011.12.001.

[33] Godunov SK. A finite difference method for the computation of discontinuous solutions of the equations of fluid dynamics. Matematicheskii Sbornik 1959;47:357–93.

[34] Green AE, Naghdi PM. A derivation of equations for wave propagation in water of variable depth. J Fluid Mech 1976;78:237–46. doi:10.1017/S0022112076002425.

[35] Harten A, Osher S. Uniformly high-order accurate nonoscillatory schemes. I. SIAM J Numer Anal 1987;24(2):279–309. http://www.jstor.org/stable/2157557

[36] Harten A, Engquist B, Osher S, Chakravarthy SR. Some results on high-order accurate essentially non-oscillatory schemes. Appl Numer Math 1986;2:347–77.

[37] Harten A, Engquist B, Osher S, Chakravarthy SR. Uniformly high order accuracy essentially non–oscillatory schemes III. J Comput Phys 1987;71:231–303.

[38] Käser M, Iske A. Adaptive ADER schemes for the solution of scalar non-linear hyperbolic problems. J Comput Phys 2005;205:489–508.

[39] Käser M. Adaptive methods for the numerical simulation of transport processes. Institute of Numerical Mathematics and Scientific Computing, University of Munich, Germany; 2003. Ph.D. thesis.

[40] Käser M. ADER schemes for the solution of conservation laws on adaptive triangulations. Mathematical methods and modelling in hydrocarbon exploration and production, Vol. 7. Springer–Verlag; 2004.

[41] Kazolea M, Delis AI, Synolakis CE. Numerical treatment of wave breaking on unstructured finite volume approximations for extended Boussinesq-type equations. J Comput Phys 2014;271:281–305. doi:10.1016/j.jcp.2014.01.030.

[42] Lannes D. Well-posedness of the water-waves equations. J Am Math Soc 2005;18(3):605–54.

[43] Lannes D. The water waves problem: mathematical analysis and asymptotics. Mathematical surveys and monographs. Providence: American Mathematical Society; 2013. ISBN 9780821894705. https://books.google.cl/books?id=CbwxAAAAQBAJ

[44] Lannes D, Marche F. A new class of fully nonlinear and weakly dispersive Green–Naghdi models for efficient 2D simulations. J Comput Phys 2015;282:238–68. doi:10.1016/j.jcp.2014.11.016.

[45] Le Métayer O, Gavrilyuk S, Hank S. A numerical scheme for the Green–Naghdi model. J Comput Phys 2010;229(6):2034–45. http://dx.doi.org/10.1016/j.jcp.2009.11.021.

[46] LeFloch PG. Hyperbolic systems of conservation laws: the Theory of classical and nonclassical shock waves. Lectures in Mathematics. *ETH* Zürich: Birkhäuser Basel; 2002. ISBN 9783764366872. http://books.google.cl/books?id=ElBVmKuJH1AC

[47] Li M, Guyenne P, Li F, Xu L. High order well-balanced CDG-FE methods for shallow water waves by a Green–Naghdi model. J Comput Phys 2014;257, Part A(0):169–92. http://dx.doi.org/10.1016/j.jcp.2013.09.050.

[48] Loubere R, Dumbser M, Diot S. A new family of high order unstructured MOOD and ADER finite volume schemes for multidimensional systems of hyperbolic conservation laws. Commun Comput Phys 2014;16(3):718–63. doi:10.4208/cicp.181113.140314a.

[49] Montecinos GI, Toro EF. Reformulations for general advection - diffusion - reaction equations and locally implicit ADER schemes. J Comput Phys 2014;275:415–42.

[50] Montecinos GI, Müller LO, Toro EF. Hyperbolic reformulation of a 1D viscoelastic blood flow model and ADER finite volume schemes. J Comput Phys 2014;266:101–23.

[51] Müller LO, Toro EF. A global multiscale mathematical model for the human circulation with emphasis on the venous system. Int J Numer Methods Biomed Eng 2014;30(7):681–725. doi:10.1002/cnm.2622.

[52] Nwogu O. Alternative form of Boussinesq equations for nearshore wave propagation. J Waterwa Port C Ocean Eng 1993;119(6):618–38.

[53] Peregrine DH. Long waves on a beach. J Fluid Mech 1967;27(2):815–27.

[54] Ricchiuto M, Filippini AG. Upwind residual discretization of enhanced Boussinesq equations for wave propagation over complex bathymetries. J Comput Phys 2014;271(0):306–41.Frontiers in Computational Physics Modeling the Earth System http://dx.doi.org/10.1016/j.jcp.2013.12.048.

[55] Schwartzkopff T, Munz CD, Toro EF. ADER: high–order approach for linear hyperbolic systems in 2D. J Sci Comput 2002;17:231–40.

[56] Skogestad JO, Kalisch H. A boundary value problem for the KdV equation: Comparison of finite-difference and Chebyshev methods. Math Comput Simul 2009;80(1):151–63. http://www.scopus.com/inward/record.url?eid=2-s2.0-70349186536&partnerID=40&md5=b4ed4c4e5d7cb1666003c354438dad56

[57] Takakura Y, Toro EF. Arbitrarily accurate non-oscillatory schemes for a non-linear conservation law. Comput Fluid Dynam 2002;11:7–18.

[58] Titarev VA, Toro EF. ADER: arbitrary high order Godunov approach. J Sci Comput 2002;17:609–18.

[59] Titarev VA, Toro EF. ADER schemes for three–dimensional hyperbolic systems. J Comput Phys 2005;204:715–36.

[60] Toro EF, Montecinos GI. Advection-diffusion-reaction equations: hyperbolisation and high-order ADER discretizations. SIAM J Sci Comput 2014;36(5):A2423–57.

[61] Toro EF, Montecinos GI. Implicit, semi-analytical solution of the generalized riemann problem for stiff hyperbolic balance laws. J Comput Phys 2015;303:146–72. http://dx.doi.org/10.1016/j.jcp.2015.09.039.

[62] Toro EF, Titarev VA. Solution of the generalised Riemann problem for advection–reaction equations. Proceedings of the royal society of London A 2002;458:271–81.

[63] Toro EF, Titarev VA. ADER schemes for scalar non-linear hyperbolic conservation laws with source terms in three-space dimensions. J Comput Phys 2005;202(1):196–215.

[64] Toro EF, Millington RC, Nejad LAM. Towards very high-order Godunov schemes. In: Godunov methods: theory and applications. Edited review, E. F. Toro (Editor). New York: Kluwer Academic/Plenum Publishers; 2001. p. 905–37.

[65] Toro EF. Riemann solvers and numerical methods for fluid dynamics: A practical Introduction. third. Berlin: Springer-Verlag; 2009. ISBN 9783642064388.ISBN 978-3-540-25202-3

[66] van der Vegt JJW, van der Ven H. Space-time discontinuous Galerkin finite element method with dynamic grid motion for inviscid compressible flows: i. general formulation. J Comput Phys 2002;182(2):546–85. http://dx.doi.org/10.1006/jcph.2002.7185.

[67] van der Ven H, van der Vegt JJW. Space-time discontinuous Galerkin finite element method with dynamic grid motion for inviscid compressible flows: II. Efficient flux quadrature. Comput Methods Applied Mech Eng 2002;191(41–42):4747–80. http://dx.doi.org/10.1016/S0045-7825(02)00403-6.

[68] Vázquez-Cendón ME, Toro EF. Exact solution of some hyperbolic systems with source terms. Proceedings of the royal society of London A: mathematical, physical and engineering sciences 2003;459(2029):263–71. doi:10.1098/rspa.2002.0987.

[69] Vignoli G, Titarev VA, Toro EF. ADER schemes for the shallow water equations in channel with irregular bottom elevation. J Comput Phys 2008;227(4):2463–80. http://dx.doi.org/10.1016/j.jcp.2007.11.006.

[70] Zambra CE, Dumbser M, Toro EF, Moraga NO. A novel numerical method of high-order accuracy for flow in unsaturated porous media. Int J Numer Methods Eng 2012;89(2):227–40. doi:10.1002/nme.3241.