Article submitted to IMA Journal of Numerical Analysis

Galerkin-finite element methods for the Shallow Water equations with characteristic boundary conditions †

D. C. ANTONOPOULOS‡ AND V. A. DOUGALIS§
Department of Mathematics, University of Athens, 15784 Zographou, Greece, and
Institute of Applied and Computational Mathematics, FORTH, 70013 Heraklion, Greece

[Submitted on 14 February 2016]

We consider the shallow water equations in the supercritical and subcritical cases in one space variable, posed in a finite spatial interval with characteristic boundary conditions at the endpoints, which, as is well known, are transparent, i.e. allow outgoing waves to exit without generating spurious reflected waves. Assuming that the resulting initial-boundary-value problems have smooth solutions, we approximate them in space using standard Galerkin-finite element methods and prove L^2-error estimates for the semidiscrete problems on quasiuniform meshes. We discretize the problems in the temporal variable using an explicit, fourth-order accurate Runge-Kutta scheme and check, by means of numerical experiment, that the resulting fully discrete schemes have excellent absorption properties.

Keywords: Shallow water equations, characteristic boundary conditions,Galerkin methods, error estimates.

Mathematical subject classification 2000: 65M60,35L60

1. Introduction

In this paper we consider the system of *shallow water equations*

$$\begin{aligned} &\eta_t + u_x + (\eta u)_x = 0\,, \\ &u_t + \eta_x + u u_x = 0\,, \end{aligned} \tag{1.1}$$

a well known approximation of the two-dimensional Euler equations of water-wave theory, modelling two-way propagation of long surface waves of finite amplitude in a uniform horizontal channel of finite depth, Whitham (1974). The variables in (1.1) are non-dimensional and unscaled; $x \in \mathbb{R}$ and $t \geqslant 0$ are proportional to position along the channel and time, respectively, while $\eta = \eta(x,t)$ and $u = u(x,t)$ are proportional to the elevation of the free surface above a level of rest and to the horizontal velocity of the fluid, respectively. The latter is depth-independent to the order of approximation represented by the scaled analog of (1.1). In the variables of (1.1) the bottom of the channel lies at a depth equal to -1.

It is well known that the initial-value problem for (1.1), posed with smooth initial data $\eta(x,0)$ and $u(x,0)$ for $x \in \mathbb{R}$, has, in general, smooth solutions only locally in t, cf. e.g. Majda (1984), Ch. 2. In this paper we will pose (1.1) in the finite 'channel' $[0,L]$ with given initial values at $t = 0$,

$$\eta(x,0) = \eta^0(x), \quad u(x,0) = u^0(x), \quad 0 \leqslant x \leqslant L, \tag{1.2}$$

†Work supported by the project PEFYKA of the action KRIPIS of GSRT at IACM/FORTH. The project is funded by Greece and the European Development Fund of EU under the NSRF and the O.P. Competitiveness and Entepreneurship

‡Email: antonod@math.uoa.gr

§Corresponding author. Email : doug@math.uoa.gr

and consider *transparent boundary conditions* at $x = 0$ and $x = L$, i.e. conditions that permit the waves to exit the 'computational' domain $[0,L]$ without generating spurious reflected waves that pollute the solution inside $[0,L]$. The transparent boundary conditions that we will use are *nonlinear characteristic boundary conditions* for subcritical and supercritical flows governed by (1.1). Such conditions were first used, to our knowledge, by Nycander *et al.* (2008), in numerical experiments with finite difference discretizations of shallow water models. In the paper at hand we will analyze Galerkin-finite element approximations for smooth solutions of the initial-boundary-value problems (ibvp's) resulting from the application of characteristic boundary conditions to (1.1). The well-posedness of these ibvp's was analyzed in Petcu & Temam (2011) and Huang *et al.* (2011).

The characteristic boundary conditions may be derived as follows. We write the system (1.1) as

$$\begin{pmatrix} \eta_t \\ u_t \end{pmatrix} + A \begin{pmatrix} \eta_x \\ u_x \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

The matrix $A = \begin{pmatrix} u & 1+\eta \\ 1 & u \end{pmatrix}$ has eigenvalues $\lambda_1 = u + \sqrt{1+\eta}$, $\lambda_2 = u - \sqrt{1+\eta}$. Assuming always that $\eta > -1$, we consider two types of flows:

$$\begin{aligned} \textit{Supercritical}: &\quad u > \sqrt{1+\eta}\,, \quad 0 < \lambda_2 < \lambda_1\,, \\ \textit{Subcritical}: &\quad u < \sqrt{1+\eta}\,, \quad \lambda_2 < 0 < \lambda_1\,. \end{aligned}$$

It is well known , cf. e.g. Whitham (1974), that along the family of *characteristic curves* with $\dot{x}(t) = \lambda_1 = u + \sqrt{1+\eta}$, the quantity $r_1 := u + 2\sqrt{1+\eta}$ is constant, while along the curves with $\dot{x}(t) = \lambda_2 = u - \sqrt{1+\eta}$, $r_2 := u - 2\sqrt{1+\eta}$ is preserved. If η and u are expressed in terms of the *Riemann invariants* r_1, r_2 one obtains the system of ordinary differential equations (ode's)

$$\begin{aligned} \frac{dr_1}{dt} = 0 \quad \text{on} \quad x(t): \quad \frac{dx}{dt} = \lambda_1(r_1,r_2) = \frac{3r_1 + r_2}{4}, \\ \frac{dr_2}{dt} = 0 \quad \text{on} \quad x(t): \quad \frac{dx}{dt} = \lambda_2(r_1,r_2) = \frac{r_1 + 3r_2}{4}, \end{aligned}$$

which is equivalent to the original pde system (1.1) and whose discretization yields the classical *method of characteristics* for solving (1.1). If we pose (1.1) in the spatial interval $[0,L]$ it is straightforword to see, cf. Whitham (1974), Section 5.4, that the temporal integration of the ode system requires in the supercritical case that $r_1(0,t)$ and $r_2(0,t)$ be given for $t \geqslant 0$, as both families of characteristics are incoming at $x = 0$; this is equivalent to prescribing $u(0,t)$ and $\eta(0,t)$ for $t \geqslant 0$. In the subcritical case $r_1(0,t)$ and $r_2(L,t)$ should be given for $t \geqslant 0$, since they correspond to the incoming characteristics at $x = 0$ and at $x = L$ respectively.

Following Nycander *et al.* (2008) we assume that outside the interval $[0,L]$ the flow is uniform and is given by $\eta(x,t) = \eta_0$, $u(x,t) = u_0$, where η_0, u_0 are known constants. Therefore, in the *supercritical case* the characteristic boundary conditions are simply

$$\eta(0,t) = \eta_0\,, \quad u(0,t) = u_0\,, \tag{1.3}$$

with $u_0 > \sqrt{1+\eta_0}$, while, in the *subcritical case*, they are of the form

$$\begin{aligned} u(0,t) + 2\sqrt{1+\eta(0,t)} = u_0 + 2\sqrt{1+\eta_0}, \\ u(L,t) - 2\sqrt{1+\eta(L,t)} = u_0 - 2\sqrt{1+\eta_0}, \end{aligned} \tag{1.4}$$

where now it is assumed that $u_0^2 < 1 + \eta_0$. In both cases we may view the solution (η, u) of (1.1) for $0 \leqslant x \leqslant L$, $t \geqslant 0$, generated by the initial conditions (1.2), as a perturbation of the uniform flow (η_0, u_0) to which the solution inside the computational domain $[0, L]$ will revert once the waves generated by the initial conditions exit this interval. It is straightforward to check, using the definitions of characteristics and Riemann invariants and considering e.g. initial conditions that differ from η_0, u_0 in a subinterval of $[0, L]$, that the boundary conditions (1.3) and (1.4) are transparent.

As previously mentioned, the characteristic boundary conditions (1.3) and (1.4) were used by Nycander *et al.* (2008), in finite difference simulations of the shallow water equations, in one space dimension, actually in more complicated instances of hydraulic and geophysical interest, including single- and two-layer flows in channels of variable width and variable bottom topography, time-dependent forcing in the boundary conditions, examples where transcritical flows develop, *et al.* (In the case of two-layer flows an approximate SW system was used in which the barotropic and baroclinic modes are decoupled; this allows using the analogs of the (local) characteristic boundary conditions in this case too. Also, for reasons of numerical stability, a diffusive term with a small viscosity coefficient was added in the momentum equations.) The finite difference spatial discretization was effected on a staggered grid and the leap-frog scheme was used for time stepping. Similar model equations and characteristic boundary conditions were applied in simulations of two-layer hydraulic exchange flows in Frankcombe & Hogg (2007).

The characteristic boundary conditions (1.3) and (1.4) were also used by Shiue *et al.* (2011), in the case of the one-dimensional, single-layer shallow water equations in channels of variable bottom topography in the presence of Coriolis terms and with the addition of a cross-velocity variable that depends only on x. The system, written in balance law form, was discretized in space using midpoint quadrature for the source cell integral and a 'central-upwind', Kurganov *et al.* (2001), Kurganov & Petrova (2000), Godunov-type approximation of the flux term; a second-order, explicit Runge-Kutta method was used for time stepping. Many numerical experiments performed with this scheme are reported in Shiue *et al.* (2011); they simulate interesting cases of subcritical, transcritical and supercritical flows over variable-bottom topographies. The characteristic boundary conditions and the same numerical scheme were subsequently used in Bousquet *et al.* (2013) in the case of two-layer problems in one dimension under the decoupling assumptions of Nycander *et al.* (2008). (The local well-posedness of this two-layer problem was studied by Petcu & Temam (2013).)

In case the elevation of the free surface η is a small perturbation of the steady state η_0 one may derive *linearized* approximations to the characteristic boundary conditions (1.4) in the subcritical case. These linearized conditions are also considered in Nycander *et al.* (2008) and in Shiue *et al.* (2011), where they are compared to the nonlinear exact conditions and found in general to cause spurious reflections that enter the computational domain. (The linearized boundary conditions are easily seem to be (exactly) transparent for the *linearized* shallow water equations obtained by linearizing (1.1) about the steady state (η_0, u_0). In Shiue *et al.* (2011) it is shown that the ibvp for the linearized system supplemented by the linearized boundary conditions is well posed.) As pointed out in Nycander *et al.* (2008), Shiue *et al.* (2011), and in Nycander & Döös (2003), the linearized characteristic boundary conditions have been extensively used in the computational fluid dynamics literature; the last reference contains a review of several other absorbing boundary conditions for the shallow water equations at artificial boundaries, including absorbing ('sponge') layer conditions *et al.*.

Of particular interest for our purposes is the rigorous analysis of the well-posedness of the ibvp's (locally in time) for the shallow water equations with characteristic boundary conditions carried out in Huang *et al.* (2011) and Petcu & Temam (2011). The ibvp in the supercritical case, was studied by Huang *et al.* (2011), in fact in the more general setting of shallow water supercritical flows over a

variable bottom in the presence of Coriolis terms and a lateral component of the horizontal velocity depending on x, and also with nonhomogeneous boundary conditions satisfying appropriate compatibility conditions. The hypotheses of Huang *et al.* (2011) on u_0, η_0 and the initial data, briefly reviewed in section 2 in the sequel, guarantee the existence and uniqueness, locally in time, of a smooth solution of the ibvp (1.1)-(1.3) with positive $1+\eta$, satisfying the strong supercriticality property $u^2-(1+\eta)\geqslant c_0^2$ for some positive constant c_0. The well-posedness of the ibvp in the subcritical case, i.e. of the ibvp (1.1), (1.2), (1.4), was studied by Petcu & Temam (2011). The assumptions of Petcu & Temam (2011) (reviewed in section 3 below) imply the existence and uniqueness, locally in time, of a smooth solution of the ibvp with positive $1+\eta$ and satisfying the strong subcriticality condition $u^2-(1+\eta)\leqslant -c_0^2$, where c_0 is a positive constant.

In this paper we will analyze standard Galerkin-finite element spatial discretizations of the ibvp's (1.1)-(1.3) and (1.1), (1.2), (1.4), under the hypothesis that they have smooth solutions. In both cases the basic approximation will be effected by C^{r-2} functions which are piecewise polynomials of degree $r-1$, $r\geqslant 2$, on quasiuniform partitions of $[0,L]$. In section 2 we consider the supercritical case and prove an L^2-error estimate of $O(h^{r-1})$ accuracy for the Galerkin approximations of η and u. (It is well known that this is the expected best order of convergence in L^2 for standard Galerkin semidiscretizations of first-order hyperbolic problems on general quasiuniform meshes. For *uniform meshes*, better results hold, cf. Dupont (1973) for the analysis in the case of a linear model problem. In Antonopoulos & Dougalis (to appear) it was proved that the order of convergence in L^2 for piecewise linear continuous elements on a uniform mesh is equal to 2 in the case of an ibvp for (1.1) with the homogeneous boundary conditions $u(0,t)=u(L,t)=0$. This superaccuracy result is expected to hold for the ibvp's under consideration as well and this is indeed what the numerical experiments of section 4 indicate.) For the proof of the error estimate we assume that a strengthened supercriticality condition holds for the solution of (1.1)-(1.3); cf. (H1)-(H3) in section 2. The proof also requires that $r\geqslant 3$ so that a certain bootstrap argument, based on the boundedness of the $\|\cdot\|_{1,\infty}$ norm of an error term, goes through as in Dupont (1974), Antonopoulos & Dougalis (to appear).

In section 3 we turn to the subcritical case. We write the ibvp (1.1), (1.2), (1.4) in its classical diagonal form in which the new unknowns are analogs of the two Riemann invariants in the context of the ibvp at hand and satisfy homogeneous Dirichlet boundary conditions one at $x=0$ and the other at $x=L$. The diagonal system is discretized in space on a quasiuniform mesh by the same type of standard Galerkin method as before, and a L^2 error estimate of $O(h^{r-1})$ is proved for both components of the solution. A change of variables of this semidiscrete approximation yields approximations of the original unknowns η and u of $O(h^{r-1})$ accuracy. The proof requires that a strengthened form of the subcriticality property holds for the solution of the ibvp (1.1), (1.2), (1.4) (cf. (Y1), (Y2) in section 3), and the technical assumption that $r\geqslant 3$.

Section 4 is a report of various numerical experiments that we performed with the Galerkin-finite element methods of sections 2 and 3 and some of their variants. We use spatial discretizations with piecewise linear continuous functions on uniform meshes and discretize them in the temporal variable by the 'classical', explicit, four-stage, fourth-order Runge-Kutta scheme. The resulting fully discrete methods are stable under a Courant number restriction. (Stability and convergence of high order explicit Runge-Kutta methods was established for closely related pde systems in Antonopoulos & Dougalis (2013) and Antonopoulos & Dougalis (to appear).) Our main purpose in the numerical experiments is to check the stability and the numerical order of convergence of the fully discrete Galerkin methods and study by computational means their absorption properties. Although the full discretizations of the characteristic boundary conditions are not exactly transparent of course, the numerical experiments show that they are practically transparent. In the subcritical case we also implement the analogous fully

discrete Galerkin method for the original, nondiagonal form (1.1), (1.2), (1.4) of the system and check that it gives results close but somewhat inferior to those of the analogous discretization of the diagonal form of the system.

In summary, the main contribution of the paper is the derivation of error estimates for the Galerkin-finite element spatial discretization of smooth solutions of the initial-boundary-value problems with characteristic boundary conditions for the 1D shallow water equations, that are appropriate to supercritical and subcritical flows in a channel of finite length, and the numerical study of the stability, order of convergence, and absorption properties of these methods, when coupled with a high-order explicit time-stepping scheme.

In Section 4 of an extended version of the present paper, Antonopoulos & Dougalis (2015), interested readers may also find a detailed comparison of the nonlinear characteristic boundary conditions (1.4) in the subcritical case with their linearized analogs previously mentioned, when both sets of boundary conditions are discretized in the context of the fully discrete Galerkin-finite element schemes considered in this paper. The Galerkin method with the linearized boundary conditions is less accurate as the discretized linearized conditions are absorbing but in general allow spurious reflections to form and enter the computational domain as in the numerical experiments of Nycander *et al.* (2008) and Shiue *et al.* (2011).

In the error estimates in the sequel, we let the spatial interval be $[0,1]$ for simplicity. We let $C^k = C^k[0,1]$, $k = 0,1,2,\ldots$, be the space of k times continuously differentiable functions on $[0,1]$. The norm and inner product on $L^2 = L^2(0,1)$ are denoted by $\|\cdot\|$, $(\cdot,\cdot)$, respectively. For integer $k \geqslant 0$, H^k, $\|\cdot\|_k$ will denote the usual, L^2-based Sobolev spaces of classes of functions and the associated norms. The norms on $L^\infty = L^\infty(0,1)$ and on the L^∞-based Sobolev spaces $W^k_\infty = W^k_\infty(0,1)$ will be denoted by $\|\cdot\|_\infty$, $\|\cdot\|_{k,\infty}$, respectively. Finally, we let $\mathbb{P}_r$ be the polynomials of degree $\leqslant r$.

2. Semidiscretization of the supercritical shallow water equations

In this section we consider the shallow water equations with characteristic boundary conditions in the *supercritical* case. Specifically, for $(x,t) \in [0,1]\times[0,T]$ we seek $\eta = \eta(x,t)$ and $u = u(x,t)$ satisfying the ibvp

$$\begin{aligned} &\left.\begin{aligned} &\eta_t + u_x + (\eta u)_x = 0,\\ &u_t + \eta_x + uu_x = 0, \end{aligned}\right\} \quad 0 \leqslant x \leqslant 1,\ 0 \leqslant t \leqslant T,\\ &\eta(x,0) = \eta^0(x), \quad u(x,0) = u^0(x), \quad 0 \leqslant x \leqslant 1,\\ &\eta(0,t) = \eta_0, \quad u(0,t) = u_0, \quad 0 \leqslant t \leqslant T, \end{aligned} \tag{SW1}$$

where η^0, u^0 are given functions on $[0,1]$ and η_0, u_0 constants such that $1+\eta_0 > 0$, $u_0 > 0$, $u_0 > \sqrt{1+\eta_0}$.

As mentioned in the Introduction, the ibvp (SW1) was studied by Huang *et al.* (2011), in fact in the more general case of a shallow water supercritical flow with nonhomogeneous boundary conditions over a variable bottom for a nonzero Coriolis parameter and also in the presence of a lateral component of the horizontal velocity depending on x only. In the simpler case of (SW1), the proof of the main result of Huang *et al.* (2011) amounts to the selection of a suitable constant solution (η_0, u_0) of (SW1) and of sufficiently smooth initial conditions close to the constant solution and satisfying appropriate compatibility relations at $x = 0$. Under these hypotheses the conclusion of Huang *et al.* (2011) is that given positive constants c_0, α_0, $\underline{\zeta}_0$, and $\overline{\zeta}_0$, there exists a $T > 0$ such that a sufficiently smooth solution

of (SW1) exists satisfying for $(x,t) \in [0,1] \times [0,T]$ the strong supercriticality properties

$$u^2 - (1+\eta) \geqslant c_0^2, \tag{P1}$$

$$u \geqslant \alpha_0, \tag{P2}$$

$$\underline{\zeta}_0 \leqslant (1+\eta) \leqslant \overline{\zeta}_0. \tag{P3}$$

For the purposes of the error estimation to follow we will assume that (SW1) has a sufficiently smooth solution (η,u) that satisfies a strengthened supercriticality condition of the following form: There exist positive constants α and β, such that for $(x,t) \in [0,1] \times [0,T]$ it holds that

$$1+\eta \geqslant \beta, \tag{H1}$$

$$u \geqslant 2\alpha, \tag{H2}$$

$$1+\eta \leqslant (u-\alpha)(u-\tfrac{2\alpha}{3}). \tag{H3}$$

Obviously (H1) and (H3) imply that $u > \sqrt{1+\eta}$. It is not hard to see that (H3) follows from (P1)-(P3) if e.g. α_0 is taken sufficiently small and c_0 sufficiently large. We also remark here that in the error estimates to follow (H3) will be needed only at $x = 1$ for $t \in [0,T]$.

We will approximate the solution of (SW1) in a slightly tranformed form. We let $\widetilde{\eta} = \eta - \eta_0$, $\widetilde{u} = u - u_0$ and rewrite (SW1) as an ibvp for $\widetilde{\eta}$ and $\widetilde{u}$ with homogeneous boundary conditions. Dropping the tildes we obtain the system

$$\begin{aligned} &\eta_t + u_0\eta_x + (1+\eta_0)u_x + (\eta u)_x = 0, \\ &u_t + \eta_x + u_0 u_x + u u_x = 0, \\ &\eta(x,0) = \eta^0(x) - \eta_0, \qquad u(x,0) = u^0(x) - u_0, \quad 0 \leqslant x \leqslant 1, \\ &\eta(0,t) = 0, \quad u(0,t) = 0, \quad 0 \leqslant t \leqslant T. \end{aligned} \qquad 0 \leqslant x \leqslant 1,\ 0 \leqslant t \leqslant T, \tag{SW1a}$$

In terms of the new variables η and u, our hypotheses (H1)-(H3) become

$$1 + \eta + \eta_0 \geqslant \beta, \tag{H1a}$$

$$u + u_0 \geqslant 2\alpha, \tag{H2a}$$

$$1 + \eta + \eta_0 \leqslant (u + u_0 - \alpha)(u + u_0 - \tfrac{2\alpha}{3}). \tag{H3a}$$

In the sequel, for integer $k \geqslant 0$, let $\mathring{C}^k = \{v \in C^k[0,1] : v(0) = 0\}$, and $\mathring{H}^{k+1} = \{v \in H^{k+1}(0,1) : v(0) = 0\}$. For a positive integer N let $0 = x_1 < x_2 < \cdots < x_{N+1} = 1$ be a quasiuniform partition of $[0,1]$ with $h := \max_i (x_{i+1} - x_i)$, and for integer $r \geqslant 2$ define $\mathring{S}_h = \{\phi \in \mathring{C}^{r-2} : \phi\big|_{[x_j, x_{j+1}]} \in \mathbb{P}_{r-1}, 1 \leqslant j \leqslant N\}$. It is well known that if $v \in \mathring{H}^r$, there exists $\chi \in \mathring{S}_h$ such that

$$\|v - \chi\| + h\|v' - \chi'\| \leqslant Ch^r \|v\|_r, \tag{2.1}$$

and, cf. Schreiber (1980), if $r \geqslant 3$,

$$\|v - \chi\|_2 \leqslant Ch^{r-2}\|v\|_r. \tag{2.2}$$

(Here and in the sequel C will denote a generic constant independent of h.) In addition, if P is the L^2-projection operator onto $\mathring{S}_h$, then it follows that, cf. Douglas *et al.* (1975),

$$\|Pv\|_\infty \leqslant C\|v\|_\infty, \quad \text{if} \quad v \in L^\infty, \tag{2.3}$$

$$\|Pv - v\|_\infty \leqslant Ch^r \|v\|_{r,\infty}, \quad \text{if} \quad v \in W^{r,\infty} \cap \mathring{H}^1. \tag{2.4}$$

As a consequence of the quasiuniformity of the mesh the inverse inequalities

$$\|\chi\|_1 \leqslant Ch^{-1}\|\chi\|, \tag{2.5}$$

$$\|\chi\|_{j,\infty} \leqslant Ch^{-(j+1/2)}\|\chi\|, \quad j = 0,1, \tag{2.6}$$

hold for $\chi \in \mathring{S}_h$. (In (2.6) $\|\cdot\|_{0,\infty} = \|\cdot\|_\infty$.)

The standard Galerkin semidiscretization of (SW1a) is defined as follows: We seek η_h, $u_h : [0,T] \to \mathring{S}_h$ such that for $0 \leqslant t \leqslant T$

$$(\eta_{ht},\phi) + (u_0\eta_{hx},\phi) + ((1+\eta_0)u_{hx},\phi) + ((\eta_h u_h)_x,\phi) = 0, \quad \forall \phi \in \mathring{S}_h, \tag{2.7}$$

$$(u_{ht},\phi) + (\eta_{hx},\phi) + (u_0 u_{hx},\phi) + (u_h u_{hx},\phi) = 0, \quad \forall \phi \in \mathring{S}_h, \tag{2.8}$$

with

$$\eta_h(0) = P(\eta^0(\cdot) - \eta_0), \quad u_h(0) = P(u^0(\cdot) - u_0). \tag{2.9}$$

The main result of this section is:

PROPOSITION 2.1 Let (η,u) be the solution of (SW1a), and assume that the hypotheses (H1a), (H2a), (H3a) hold, that $r \geqslant 3$, and h is sufficiently small. Then the semidiscrete ivp (2.7)-(2.9) has a unique solution (η_h,u_h) for $0 \leqslant t \leqslant T$ satisfying

$$\max_{0\leqslant t\leqslant T}(\|\eta(t) - \eta_h(t)\| + \|u(t) - u_h(t)\|) \leqslant Ch^{r-1}. \tag{2.10}$$

Proof. Let $\rho = \eta - P\eta$, $\theta = P\eta - \eta_h$, $\sigma = u - Pu$, $\xi = Pu - u_h$. After choosing a basis for $\mathring{S}_h$, it is straightforward to see that the semidiscrete problem (2.7)-(2.9) represents an ivp for an ode system which has a unique solution locally in time. While this solution exists, it follows from (2.7), (2.8) and the pde's in (SW1a), that

$$(\theta_t,\phi) + (u_0(\rho_x+\theta_x),\phi) + ((1+\eta_0)(\sigma_x+\xi_x),\phi) + ((\eta u - \eta_h u_h)_x,\phi) = 0, \quad \forall \phi \in \mathring{S}_h, \tag{2.11}$$

$$(\xi_t,\phi) + (\rho_x+\theta_x,\phi) + (u_0(\sigma_x+\xi_x),\phi) + (uu_x - u_h u_{hx},\phi) = 0, \quad \forall \phi \in \mathring{S}_h. \tag{2.12}$$

Since $\eta u - \eta_h u_h = \eta(\sigma+\xi) + u(\rho+\theta) - (\rho+\theta)(\sigma+\xi)$, $uu_x - u_h u_{hx} = (u\sigma)_x + (u\xi)_x - (\sigma\xi)_x - \sigma\sigma_x - \xi\xi_x$, it follows that

$$(\eta u - \eta_h u_h)_x = (\eta\xi)_x + (u\theta)_x - (\theta\xi)_x + \widetilde{R}_1, \tag{2.13}$$

$$uu_x - u_h u_{hx} = (u\xi)_x - \xi\xi_x + \widetilde{R}_2, \tag{2.14}$$

where

$$\widetilde{R}_1 = (\eta\sigma)_x + (u\rho)_x - (\rho\sigma)_x - (\rho\xi)_x - (\theta\sigma)_x, \tag{2.15}$$

$$\widetilde{R}_2 = (u\sigma)_x - (\sigma\xi)_x - \sigma\sigma_x. \tag{2.16}$$

Therefore, the equations (2.11), (2.12) may be written as

$$(\theta_t,\phi) + (u_0\theta_x,\phi) + (\gamma_x,\phi) + ((u\theta)_x,\phi) - ((\theta\xi)_x,\phi) = -(R_1,\phi), \quad \forall \phi \in \mathring{S}_h, \tag{2.17}$$

$$(\xi_t,\phi) + (\theta_x,\phi) + (u_0\xi_x,\phi) + ((u\xi)_x,\phi) - (\xi\xi_x,\phi) = -(R_2,\phi), \quad \forall \phi \in \mathring{S}_h. \tag{2.18}$$

where $\gamma = (1+\eta_0+\eta)\xi$ and

$$R_1 = u_0\rho_x + (1+\eta_0)\sigma_x + \widetilde{R}_1, \tag{2.19}$$

$$R_2 = \rho_x + u_0\sigma_x + \widetilde{R}_2. \tag{2.20}$$

Putting $\phi = \theta$ in (2.17), using integration by parts, and suppressing the dependence on t we have

$$\begin{aligned}\tfrac{1}{2}\tfrac{d}{dt}\|\theta\|^2 - (\gamma,\theta_x) + \tfrac{1}{2}(u_0+u(1))\theta^2(1) + (1+\eta_0+\eta(1))\xi(1)\theta(1)\\ - \tfrac{1}{2}\xi(1)\theta^2(1) = -\tfrac{1}{2}(u_x\theta,\theta) + \tfrac{1}{2}(\xi_x\theta,\theta) - (R_1,\theta).\end{aligned} \tag{2.21}$$

Take now $\phi = P\gamma = P[(1+\eta_0+\eta)\xi]$ in (2.18) and get

$$(\xi_t,\gamma) + (\theta_x,\gamma) + (u_0\xi_x,\gamma) + ((u\xi)_x,\gamma) - (\xi\xi_x,\gamma) = -(R_3, P\gamma-\gamma) - (R_2,P\gamma), \tag{2.22}$$

where

$$R_3 = \theta_x + u_0\xi_x + (u\xi)_x - \xi\xi_x. \tag{2.23}$$

Integration by parts in various terms in (2.22) gives

$$\begin{aligned}(u_0\xi_x,\gamma) &= (u_0\xi_x,(1+\eta_0+\eta)\xi) = \tfrac{1}{2}u_0(1+\eta_0+\eta(1))\xi^2(1) - \tfrac{1}{2}(u_0\eta_x\xi,\xi),\\ ((u\xi)_x,\gamma) &= ((u\xi)_x,(1+\eta_0+\eta)\xi) = (u_x\xi,(1+\eta_0+\eta)\xi) + (u\xi_x,(1+\eta_0+\eta)\xi)\\ &= \tfrac{1}{2}u(1)(1+\eta_0+\eta(1))\xi^2(1) + \tfrac{1}{2}(u_x(1+\eta_0+\eta),\xi^2) - \tfrac{1}{2}(u\eta_x\xi,\xi),\\ (\xi\xi_x,\gamma) &= (\xi\xi_x,(1+\eta_0+\eta)\xi) = \tfrac{1}{3}(1+\eta_0+\eta(1))\xi^3(1) - \tfrac{1}{3}(\eta_x\xi^2,\xi).\end{aligned}$$

Hence (2.22) becomes

$$\begin{aligned}(\xi_t,\gamma) + (\theta_x,\gamma) + \tfrac{1}{2}(u_0+u(1))(1+\eta_0+\eta(1))\xi^2(1) - \tfrac{1}{3}(1+\eta_0+\eta(1))\xi^3(1)\\ = (R_4,\xi) - (R_3,P\gamma-\gamma) - (R_2,P\gamma),\end{aligned} \tag{2.24}$$

where

$$R_4 = \tfrac{1}{2}u_0\eta_x\xi - \tfrac{1}{2}u_x(1+\eta_0+\eta)\xi + \tfrac{1}{2}u\eta_x\xi - \tfrac{1}{3}\eta_x\xi^2. \tag{2.25}$$

Adding now (2.21) and (2.24) we obtain

$$\begin{aligned}\tfrac{1}{2}\tfrac{d}{dt}[\|\theta\|^2 + ((1+\eta_0+\eta)\xi,\xi)] + \omega = \tfrac{1}{2}(\eta_t\xi,\xi) - \tfrac{1}{2}(u_x\theta,\theta)\\ + \tfrac{1}{2}(\xi_x\theta,\theta) - (R_1,\theta) + (R_4,\xi) - (R_3,P\gamma-\gamma) - (R_2,P\gamma),\end{aligned} \tag{2.26}$$

where

$$\begin{aligned}\omega =& \tfrac{1}{2}(u_0+u(1))\theta^2(1) + \tfrac{1}{2}(u_0+u(1))(1+\eta_0+\eta(1))\xi^2(1)\\ &+ (1+\eta_0+\eta(1))\xi(1)\theta(1) - \tfrac{1}{2}\xi(1)\theta^2(1) - \tfrac{1}{3}(1+\eta_0+\eta(1))\xi^3(1).\end{aligned} \tag{2.27}$$

In view of (2.9), by continuity we conclude that there exists a maximal temporal instance $t_h > 0$ such that (η_h,u_h) exist and $\|\xi_x\|_\infty \leqslant \alpha$ for $t \leqslant t_h$. Suppose that $t_h < T$. Then, since $\|\xi\|_\infty \leqslant \|\xi_x\|_\infty$, it follows from (2.27) that for $t \in [0,t_h]$

$$\begin{aligned}\omega \geqslant& \tfrac{1}{2}(u_0+u(1)-\alpha)\theta^2(1) + \tfrac{1}{2}(1+\eta_0+\eta(1))(u_0+u(1)-\tfrac{2\alpha}{3})\xi^2(1)\\ &+ (1+\eta_0+\eta(1))\xi(1)\theta(1) = \tfrac{1}{2}(\theta(1),\xi(1))^T\begin{pmatrix}\mu & \lambda\\ \lambda & \lambda\nu\end{pmatrix}\begin{pmatrix}\theta(1)\\ \xi(1)\end{pmatrix},\end{aligned} \tag{2.28}$$

where $\mu = u_0 + u(1) - \alpha$, $\lambda = 1 + \eta_0 + \eta(1)$, $\nu = u_0 + u(1) - \frac{2\alpha}{3}$. The hypotheses (H1a) and (H2a) give that $0 < \mu < \nu$, $\lambda > 0$. It is easy to see then that the matrix in (2.28) will be positive semidefinite precisely when (H3a) holds. We conclude from (2.28) that $\omega \geqslant 0$.
We now estimate the various terms in the right-hand side of (2.26) for $0 \leqslant t \leqslant t_h$. We obviously have

$$|(\eta_t \xi, \xi)| + |(u_x \theta, \theta)| \leqslant C(\|\xi\|^2 + \|\theta\|^2), \tag{2.29}$$

and

$$|(\xi_x \theta, \theta)| \leqslant \alpha \|\theta\|^2. \tag{2.30}$$

In addition, from (2.19), (2.14), and the inverse and approximation properties of $\mathring{S}_h$ and (2.3), (2.4) we have

$$\begin{aligned} |(R_1, \theta)| \leqslant & Ch^{r-1}\|\theta\| + \|\rho\|_\infty \|\xi_x\| \|\theta\| + \|\rho_x\|_\infty \|\xi\| \|\theta\| \\ & + \|\sigma_x\|_\infty \|\theta\|^2 + \|\sigma\|_\infty \|\theta_x\| \|\theta\| \\ \leqslant & Ch^{r-1}\|\theta\| + C(\|\theta\|^2 + \|\xi\|^2). \end{aligned} \tag{2.31}$$

Also, from (2.25)

$$|(R_4, \xi)| \leqslant C\|\xi\|^2 + C\|\xi\|_\infty \|\xi\|^2 \leqslant C(1+\alpha)\|\xi\|^2. \tag{2.32}$$

By (2.20), (2.16), (2.3), (2.4) and the inverse and approximation properties of $\mathring{S}_h$ we have

$$\begin{aligned} |(R_2, P\gamma)| &\leqslant Ch^{r-1}\|\xi\| + C\|\sigma\|_\infty \|\xi_x\| \|\xi\| + C\|\sigma_x\|_\infty \|\xi\|^2 \\ &\leqslant Ch^{r-1}\|\xi\| + C\|\xi\|^2. \end{aligned} \tag{2.33}$$

Finally, using a well-known *superapproximation* property of $\mathring{S}_h$, cf. Douglas *et al.* (1975), Dupont (1974), in order to estimate the term $P\gamma - \gamma$ by

$$\|P\gamma - \gamma\| = \|P[(1+\eta+\eta_0)\xi] - (1+\eta+\eta_0)\xi\| \leqslant Ch\|\xi\|, \tag{2.34}$$

we obtain by (2.23) and the inverse properties of $\mathring{S}_h$ that

$$\begin{aligned} |(R_3, P\gamma - \gamma)| &\leqslant |(\theta_x, P\gamma - \gamma)| + |(u_0 \xi_x, P\gamma - \gamma)| + |((u\xi)_x, P\gamma - \gamma)| + |(\xi \xi_x, P\gamma - \gamma)| \\ &\leqslant Ch\|\theta_x\| \|\xi\| + Ch\|\xi_x\| \|\xi\| + Ch\|\xi\|^2 + Ch\|\xi\|_\infty \|\xi_x\| \|\xi\| \\ &\leqslant C\|\theta\| \|\xi\| + C\|\xi\|^2 + C\alpha\|\xi\|^2 \\ &\leqslant C\|\theta\|^2 + C(1+\alpha)\|\xi\|^2. \end{aligned} \tag{2.35}$$

Therefore, (2.26), the fact that $\omega \geqslant 0$, and the inequalities (2.30)-(2.33), (2.35) give for $0 \leqslant t \leqslant t_h$

$$\tfrac{d}{dt}[\|\theta\|^2 + ((1+\eta_0+\eta)\xi, \xi)] \leqslant Ch^{r-1}(\|\theta\| + \|\xi\|) + C(\|\theta\|^2 + \|\xi\|^2),$$

where C is a constant independent of h and t_h. By (H1a) the norm $((1+\eta_0+\eta)\cdot,\cdot)^{1/2}$ is equivalent to that of L^2 uniformly for $t \in [0,T]$. Hence, Gronwall's inequality and the fact that $\theta(0) = \xi(0) = 0$ yield for a constant $C = C(T)$

$$\|\theta\| + \|\xi\| \leqslant Ch^{r-1} \quad \text{for} \quad 0 \leqslant t \leqslant t_h. \tag{2.36}$$

We conclude from (2.6) that $\|\xi_x\|_\infty \leqslant Ch^{r-5/2}$ for $0 \leqslant t \leqslant t_h$, and, since $r \geqslant 3$, if h is taken sufficiently small, t_h is not maximal. Hence we may take $t_h = T$ and (2.10) follows from (2.36). □

3. Semidiscretization of the subcritical shallow water equations

In this section we consider the shallow water equations with characteristic boundary conditions in the *subcritical* case. Specifically, for $(x,t) \in [0,1] \times [0,T]$ we seek $\eta = \eta(x,t)$ and $u = u(x,t)$ satisfying the ibvp

$$\begin{aligned} &\left.\begin{aligned} &\eta_t + u_x + (\eta u)_x = 0, \\ &u_t + \eta_x + u u_x = 0, \end{aligned}\right\} \quad 0 \leqslant x \leqslant 1,\ 0 \leqslant t \leqslant T, \\ &\eta(x,0) = \eta^0(x), \quad u(x,0) = u^0(x), \quad 0 \leqslant x \leqslant 1, \\ &u(0,t) + 2\sqrt{1+\eta(0,t)} = u_0 + 2\sqrt{1+\eta_0}, \quad 0 \leqslant t \leqslant T, \\ &u(1,t) - 2\sqrt{1+\eta(1,t)} = u_0 - 2\sqrt{1+\eta_0}, \quad 0 \leqslant t \leqslant T, \end{aligned} \tag{SW2}$$

where η^0, u^0 are given functions on $[0,1]$ and η_0, u_0 constants such that $1+\eta_0 > 0$ and $u_0^2 < 1+\eta_0$.

As mentioned in the Introduction, the ibvp (SW2) was studied by Petcu & Temam (2011). They used the hypotheses that there exists a constant $c_0 > 0$ such that $u_0^2 - (1+\eta_0) \leqslant -c_0^2$ and that the initial conditions $\eta^0(x)$ and $u^0(x)$ are sufficiently smooth and satisfy the condition $(u^0(x))^2 - (1+\eta^0(x)) \leqslant -c_0^2$ (with $1+\eta^0(x)$ positive) and suitable compatibility relations at $x=0$ and $x=1$. Under these assumptions one may infer from the theory of Petcu & Temam (2011) that there exists a $T > 0$ such that a sufficiently smooth solution (η,u) of (SW2) exists for $(x,t) \in [0,1] \times [0,T]$ with the properties that $1+\eta$ is positive and the strong subcriticality condition

$$u^2 - (1+\eta) \leqslant -c_0^2, \tag{Π}$$

holds for $(x,t) \in [0,1] \times [0,T]$. For the purposes of the error estimation to follow we will assume that (SW2) has a sufficiently smooth solution (η,u) such that $1+\eta > 0$ and satisfies a stronger subcriticality condition. Specifically we assume that for some constant $c_0 > 0$ it holds that

$$u_0 + \sqrt{1+\eta_0} \geqslant c_0, \quad u_0 - \sqrt{1+\eta_0} \leqslant -c_0, \tag{Y1}$$

and for $(x,t) \in [0,1] \times [0,T]$ that

$$u + \sqrt{1+\eta} \geqslant c_0, \quad u - \sqrt{1+\eta} \leqslant -c_0. \tag{Y2}$$

Obviously (Y1) and (Y2) imply the subcriticality conditions $u_0^2 - (1+\eta_0) \leqslant -c_0^2$ and (Π) of Petcu & Temam (2011), and approximate the latter better as c_0 decreases.

In this section we will approximate the solution of (SW2) with a Galerkin-finite element method after transforming (1.1) in its classical diagonal form. As in the Introduction, we write the system as

$$\begin{pmatrix} \eta_t \\ u_t \end{pmatrix} + A \begin{pmatrix} \eta_x \\ u_x \end{pmatrix} = 0, \tag{3.1}$$

where $A = \begin{pmatrix} u & 1+\eta \\ 1 & u \end{pmatrix}$. The matrix A has the eigenvalues $\lambda_1 = u + \sqrt{1+\eta}$, $\lambda_2 = u - \sqrt{1+\eta}$, (note that by (Y2) $\lambda_1 \geqslant c_0$ and $\lambda_2 \leqslant -c_0$ in $[0,1] \times [0,T]$), with associated eigenvectors $X_1 = (\sqrt{1+\eta},1)^T$, $X_2 = (-\sqrt{1+\eta},1)^T$. If S is the matrix with columns X_1, X_2 it follows from (3.1) that

$$S^{-1} \begin{pmatrix} \eta_t \\ u_t \end{pmatrix} + \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} S^{-1} \begin{pmatrix} \eta_x \\ u_x \end{pmatrix} = 0. \tag{3.2}$$

If we try to define now functions v, w on $[0,1]\times[0,T]$ by the equations $S^{-1}\begin{pmatrix}\eta_t\\ u_t\end{pmatrix}=\begin{pmatrix}v_t\\ w_t\end{pmatrix}$, $S^{-1}\begin{pmatrix}\eta_x\\ u_x\end{pmatrix}=\begin{pmatrix}v_x\\ w_x\end{pmatrix}$, we see that these equations are consistent and their solutions are given by $v=\frac{1}{2}(u+2\sqrt{1+\eta})+c_v$, $w=\frac{1}{2}(u-2\sqrt{1+\eta})+c_w$, for arbitrary constants c_v, c_w. Choosing the constants c_v, c_w so that $v(0,t)=0$, $w(1,t)=0$, and using the boundary conditions in (SW2) we get

$$v=\tfrac{1}{2}[u-u_0+2(\sqrt{1+\eta}-\delta_0)],\quad w=\tfrac{1}{2}[u-u_0-2(\sqrt{1+\eta}-\delta_0)], \tag{3.3}$$

where $\delta_0=\sqrt{1+\eta_0}$. The original variables η, u are given in terms of v and w by the formulas

$$\eta=[\tfrac{1}{2}(v-w)+\delta_0]^2-1,\quad u=v+w+u_0. \tag{3.4}$$

Since

$$\lambda_1=u+\sqrt{1+\eta}=u_0+\delta_0+\frac{3v+w}{2},\quad \lambda_2=u-\sqrt{1+\eta}=u_0-\delta_0+\frac{v+3w}{2}, \tag{3.5}$$

we see that the ibvp (SW2) becomes

$$\begin{aligned}&\begin{pmatrix}v_t\\ w_t\end{pmatrix}+\begin{pmatrix}u_0+\delta_0+\frac{3v+w}{2} & 0\\ 0 & u_0-\delta_0+\frac{v+3w}{2}\end{pmatrix}\begin{pmatrix}v_x\\ w_x\end{pmatrix}=0,\quad 0\leqslant x\leqslant 1,\ 0\leqslant t\leqslant T,\\ &\qquad v(x,0)=v^0(x),\quad w(x,0)=w^0(x),\quad 0\leqslant x\leqslant 1,\\ &\qquad v(0,t)=0,\quad w(1,t)=0,\quad 0\leqslant t\leqslant T,\end{aligned} \tag{SW2a}$$

where $v^0(x)=\frac{1}{2}[u^0(x)-u_0+2(\sqrt{1+\eta^0(x)}-\delta_0)]$, $w^0(x)=\frac{1}{2}[u^0(x)-u_0-2(\sqrt{1+\eta^0(x)}-\delta_0)]$. Under our hypotheses (SW2a) has a unique solution (v,w) on $[0,1]\times[0,T]$ which will be assumed to be smooth enough for the purposes of the error estimation that follows. Of course, v and w represent analogs of the Riemann invariants of the shallow water system in the context of the ibvp at hand; the system of pde's in (SW2a) and (3.4), (3.5) imply that the solution (η,u) of (SW2) may be expressed in terms of two waves v and w that propagate to the right and left, respectively, with speeds $u+\sqrt{1+\eta}$ and $u-\sqrt{1+\eta}$.

Given a quasiuniform partition of $[0,1]$ as in section 2, in addition to the spaces defined there, let for integer $k\geqslant 0$ $\mathring{\mathscr{C}}^k=\{f\in C^k[0,1]: f(1)=0\}$, $\mathring{\mathscr{H}}^{k+1}=\{f\in H^{k+1}(0,1), f(1)=0\}$, and, for integer $r\geqslant 2$, $\mathring{\mathscr{S}}_h=\{\phi\in\mathring{\mathscr{C}}^{r-2}:\phi\big|_{[x_j,x_{j+1}]}\in\mathbb{P}_{r-1}, 1\leqslant j\leqslant N\}$. Note that the analogs of the approximation and inverse properties (2.1), (2.2), (2.5), (2.6) hold for $\mathring{\mathscr{S}}_h$ as well, and that (2.3), (2.4) are also valid for the L^2 projection $\mathscr{P}$ onto $\mathring{\mathscr{S}}_h$, *mutatis mutandis*.

The (standard) Galerkin semidiscretization of (SW2a) is then defined as follows: Seek $v_h:[0,T]\to\mathring{S}_h$, $w_h:[0,T]\to\mathring{\mathscr{S}}_h$, such that for $t\in[0,T]$

$$(v_{ht},\phi)+((u_0+\delta_0)v_{hx},\phi)+\tfrac{3}{2}(v_hv_{hx},\phi)+\tfrac{1}{2}(w_hv_{hx},\phi)=0,\quad\forall\phi\in\mathring{S}_h, \tag{3.6}$$

$$(w_{ht},\chi)+((u_0-\delta_0)w_{hx},\chi)+\tfrac{3}{2}(w_hw_{hx},\chi)+\tfrac{1}{2}(v_hw_{hx},\chi)=0,\quad\forall\chi\in\mathring{\mathscr{S}}_h, \tag{3.7}$$

with

$$v_h(0)=P(v^0),\quad w_h(0)=\mathscr{P}(w^0). \tag{3.8}$$

The main result of this section is

PROPOSITION 3.1 Let (v,w) be the solution of (SW2a) and assume that the hypotheses (Y1) and (Y2) hold, that $r \geqslant 3$, and that h is sufficiently small. Then the semidiscrete ivp (3.6)-(3.8) has a unique solution (v_h, w_h) for $0 \leqslant t \leqslant T$ that satisfies

$$\max_{0\leqslant t\leqslant T}(\|v-v_h\|+\|w-w_h\|) \leqslant Ch^{r-1}. \tag{3.9}$$

If (η,u) is the solution of (SW2) and we define

$$\eta_h = [\tfrac{1}{2}(v_h-w_h)+\delta_0]^2-1, \quad u_h = v_h+w_h+u_0, \tag{3.10}$$

then

$$\max_{0\leqslant t\leqslant T}(\|\eta-\eta_h\|+\|u-u_h\|) \leqslant Ch^{r-1}.$$

Proof. Let $\rho = v-Pv$, $\theta = Pv-v_h$, $\sigma = w-\mathscr{P}w$, $\xi = \mathscr{P}w-w_h$. After choosing bases for $\mathring{S}_h$ and $\mathring{\mathscr{S}}_h$ we see that the ode ivp (3.6)-(3.8) has a unique solution locally in time. From (SW2a) and (3.6), (3.7) we obtain, as long as the solution exists,

$$(\theta_t,\phi)+((u_0+\delta_0)(\theta_x+\rho_x),\phi)+\tfrac{3}{2}(vv_x-v_hv_{hx},\phi)+\tfrac{1}{2}((wv_x-w_hv_{hx},\phi)=0, \quad \forall\phi\in\mathring{S}_h, \tag{3.11}$$

$$(\xi_t,\chi)+((u_0-\delta_0)(\sigma_x+\xi_x),\chi)+\tfrac{3}{2}(ww_x-w_hw_{hx},\chi)+\tfrac{1}{2}(vw_x-v_hw_{hx},\chi)=0, \quad \forall\chi\in\mathring{\mathscr{S}}_h. \tag{3.12}$$

Now, since

$$\begin{aligned}
vv_x-v_hv_{hx} &= (v\rho)_x+(v\theta)_x-(\rho\theta)_x-\rho\rho_x-\theta\theta_x,\\
wv_x-w_hv_{hx} &= w(\rho_x+\theta_x)+v_x(\sigma+\xi)-(\rho_x+\theta_x)(\sigma+\xi),\\
ww_x-w_hw_{hx} &= (w\sigma)_x+(w\xi)_x-(\sigma\xi)_x-\sigma\sigma_x-\xi\xi_x,\\
vw_x-v_hw_{hx} &= v(\sigma_x+\xi_x)+w_x(\rho+\theta)-(\sigma_x+\xi_x)(\rho+\theta),
\end{aligned}$$

it follows that

$$vv_x-v_hv_{hx} = (v\theta)_x-\theta\theta_x+R_{11}, \qquad wv_x-w_hv_{hx} = -\theta_x\xi+R_{12}, \tag{3.13}$$

$$ww_x-w_hw_{hx} = (w\xi)_x-\xi\xi_x+R_{21}, \qquad vw_x-v_hw_{hx} = -\xi_x\theta+R_{22}, \tag{3.14}$$

where

$$R_{11} = (v\rho)_x-(\rho\theta)_x-\rho\rho_x, \qquad R_{12} = w\rho_x+w\theta_x+v_x\sigma+v_x\xi-\rho_x\sigma-\rho_x\xi-\theta_x\sigma, \tag{3.15}$$

$$R_{21} = (w\sigma)_x-(\sigma\xi)_x-\sigma\sigma_x, \qquad R_{22} = v\sigma_x+v\xi_x+w_x\rho+w_x\theta-\sigma_x\rho-\sigma_x\theta-\xi_x\rho. \tag{3.16}$$

Putting now $\phi=\theta$ in (3.11) and $\chi=\xi$ in (3.12) we obtain

$$\begin{aligned}
&\tfrac{1}{2}\tfrac{d}{dt}\|\theta\|^2+((u_0+\delta_0)\theta_x,\theta)+\tfrac{3}{2}((v\theta)_x,\theta)-\tfrac{3}{2}(\theta\theta_x,\theta)\\
&\qquad = -((u_0+\delta_0)\rho_x,\theta)-\tfrac{3}{2}(R_{11},\theta)+\tfrac{1}{2}(\theta_x\xi,\theta)-\tfrac{1}{2}(R_{12},\theta),
\end{aligned} \tag{3.17}$$

$$\begin{aligned}
&\tfrac{1}{2}\tfrac{d}{dt}\|\xi\|^2+((u_0-\delta_0)\xi_x,\xi)+\tfrac{3}{2}((w\xi)_x,\xi)-\tfrac{3}{2}(\xi\xi_x,\xi)\\
&\qquad = -((u_0-\delta_0)\sigma_x,\xi)-\tfrac{3}{2}(R_{21},\xi)+\tfrac{1}{2}(\xi_x\theta,\xi)-\tfrac{1}{2}(R_{22},\theta),
\end{aligned} \tag{3.18}$$

Integration by parts yields (we suppress the t-dependence)

$$((u_0+\delta_0)\theta_x,\theta)=\frac{u_0+\delta_0}{2}\theta^2(1),\quad ((v\theta)_x,\theta)=\frac{1}{2}(v_x\theta,\theta)+\frac{1}{2}v(1)\theta^2(1),$$
$$(\theta\theta_x,\theta)=\frac{1}{3}\theta^3(1),\quad ((u_0-\delta_0)\xi_x,\xi)=-\frac{u_0-\delta_0}{2}\xi^2(0),$$
$$((w\xi)_x,\xi)=\frac{1}{2}(w_x\xi,\xi)-\frac{1}{2}w(0)\xi^2(0),\quad (\xi\xi_x,\xi)=-\frac{1}{3}\xi^3(0).$$

Hence, (3.17) becomes

$$\begin{aligned}&\tfrac{1}{2}\tfrac{d}{dt}\|\theta\|^2+\tfrac{1}{2}(u_0+\delta_0+\tfrac{3}{2}v(1)-\theta(1))\theta^2(1)\\&\quad=-((u_0+\delta_0)\rho_x,\theta)-\tfrac{3}{4}(v_x\theta,\theta)+\tfrac{1}{2}(\theta_x\xi,\theta)-\tfrac{3}{2}(R_{11},\theta)-\tfrac{1}{2}(R_{12},\theta).\end{aligned}$$

By (Y2) and (3.5) we see that $u_0+\delta_0+\frac{3}{2}v(1)\geqslant c_0>0$. Therefore the above equation gives

$$\begin{aligned}\tfrac{1}{2}\tfrac{d}{dt}\|\theta\|^2+\tfrac{1}{2}(c_0-\theta(1))\theta^2(1)&\leqslant-((u_0+\delta_0)\rho_x,\theta)\\&\quad-\tfrac{3}{4}(v_x\theta,\theta)+\tfrac{1}{2}(\theta_x\xi,\theta)-\tfrac{3}{2}(R_{11},\theta)-\tfrac{1}{2}(R_{12},\theta).\end{aligned}\tag{3.19}$$

Similarly, from (3.18) we obtain

$$\begin{aligned}&\tfrac{1}{2}\tfrac{d}{dt}\|\xi\|^2+\tfrac{1}{2}(-(u_0-\delta_0+\tfrac{3}{2}w(0))+\xi(0))\xi^2(0)\\&\quad=-((u_0-\delta_0)\sigma_x,\xi)+\tfrac{1}{2}(\xi_x\theta,\xi)-\tfrac{3}{4}(w_x\xi,\xi)-\tfrac{3}{2}(R_{21},\xi)-\tfrac{1}{2}(R_{22},\xi).\end{aligned}$$

Again, by (Y2) and (3.5) we get $u_0-\delta_0+\frac{3}{2}w(0)\leqslant-c_0<0$. We conclude that

$$\begin{aligned}\tfrac{1}{2}\tfrac{d}{dt}\|\xi\|^2+\tfrac{1}{2}(c_0+\xi(0))\xi^2(0)&\leqslant-((u_0-\alpha)\sigma_x,\xi)\\&\quad+\tfrac{1}{2}(\xi_x\theta,\xi)-\tfrac{3}{4}(w_x\xi,\xi)-\tfrac{3}{2}(R_{21},\xi)-\tfrac{1}{2}(R_{22},\xi).\end{aligned}\tag{3.20}$$

Finally, adding (3.19) and (3.20) we get, as long as the solution of (3.6)-(3.8) exists, that

$$\begin{aligned}&\tfrac{1}{2}\tfrac{d}{dt}(\|\theta\|^2+\|\xi\|^2)+\tfrac{1}{2}(c_0-\theta(1))\theta^2(1)+\tfrac{1}{2}(c_0+\xi(0))\xi^2(0)\\&\quad\leqslant-((u_0+\delta_0)\rho_x,\theta)-((u_0-\delta_0)\sigma_x,\xi)-\tfrac{3}{4}(v_x\theta,\theta)-\tfrac{3}{4}(w_x\xi,\xi)\\&\quad+\tfrac{1}{2}(\theta_x\xi,\theta)+\tfrac{1}{2}(\xi_x\theta,\xi)-\tfrac{3}{2}(R_{11},\theta)-\tfrac{1}{2}(R_{12},\theta)-\tfrac{3}{2}(R_{21},\xi)-\tfrac{1}{2}(R_{22},\xi).\end{aligned}\tag{3.21}$$

In view of (3.8), by continuity we conclude that there exists a maximal temporal instance $t_h>0$ such that v_h, w_h exist for $t\leqslant t_h$ and

$$\|\theta(t)\|_{1,\infty}+\|\xi(t)\|_{1,\infty}\leqslant c_0,\quad t\in[0,t_h].\tag{3.22}$$

Suppose that $t_h<T$. For $t\in[0,t_h]$ we have by (3.22)

$$\tfrac{1}{2}(c_0-\theta(1))\theta^2(1)+\tfrac{1}{2}(c_0+\xi(0))\xi^2(0)\geqslant 0,\tag{3.23}$$

and

$$\tfrac{1}{2}|(\theta_x\xi,\theta)|+\tfrac{1}{2}|(\xi_x\theta,\xi)|\leqslant\frac{c_0}{2}\|\theta\|\|\xi\|.\tag{3.24}$$

We obviously have

$$|(v_x\theta,\theta)|+|(w_x\xi,\xi)|\leqslant C(\|\theta\|^2+\|\xi\|^2). \tag{3.25}$$

Using now the approximation and inverse properties (2.1)-(2.6) for $\mathring{S}_h$ (and also for $\mathring{\mathscr{S}}_h$) we estimate the rest of the terms in the right-hand side of (3.21) as follows. We first clearly have

$$|((u_0+\delta_0)\rho_x,\theta)|+|((u_0-\delta_0)\sigma_x,\xi)|\leqslant Ch^{r-1}(\|\theta\|+\|\xi\|). \tag{3.26}$$

Integrating by parts we see by (3.15) that

$$\begin{aligned}(R_{11},\theta)&=((v\rho)_x,\theta)-((\rho\theta)_x,\theta)-(\rho\rho_x,\theta)\\&=v(1)\rho(1)\theta(1)-(v\rho,\theta_x)-\rho(1)\theta^2(1)+(\rho\theta,\theta_x)-(\rho\rho_x,\theta).\end{aligned}$$

Therefore

$$\begin{aligned}|(R_{11},\theta)|&\leqslant C\|\rho\|_\infty\|\theta\|_\infty+C\|\rho\|_\infty\|\theta_x\|+\|\rho\|_\infty\|\theta\|_\infty^2+\|\rho\|_\infty\|\theta\|\|\theta_x\|+\|\rho\|_\infty\|\rho_x\|\|\theta\|\\&\leqslant Ch^r\|\theta\|_\infty+Ch^r\|\theta_x\|+Ch^r\|\theta\|_\infty^2+Ch^r\|\theta\|\|\theta_x\|+Ch^{2r-1}\|\theta\|\\&\leqslant Ch^{r-1}(\|\theta\|+\|\theta\|^2).\end{aligned} \tag{3.27}$$

Integration by parts and (3.15) yield for the R_{12} term that

$$(R_{12},\theta)=(w\rho_x,\theta)-\tfrac{1}{2}(w_x\theta,\theta)+(v_x\sigma,\theta)+(v_x\xi,\theta)-(\rho_x\sigma,\theta)-(\rho_x\xi,\theta)-(\theta_x\sigma,\theta).$$

Hence, similarly as above

$$\begin{aligned}|(R_{12},\theta)|&\leqslant Ch^{r-1}\|\theta\|+C\|\theta\|^2+Ch^r\|\theta\|+C\|\xi\|\|\theta\|\\&\quad+Ch^{2r-1}\|\theta\|+Ch^{r-1}\|\xi\|_\infty\|\theta\|+Ch^r\|\theta\|_\infty\|\theta_x\|\\&\leqslant Ch^{r-1}\|\theta\|+C\|\theta\|^2+C\|\xi\|\|\theta\|.\end{aligned} \tag{3.28}$$

Again, using integration by parts and (3.15) for the R_{21} term, we obtain

$$(R_{21},\xi)=-(w\sigma,\xi_x)-w(0)\sigma(0)\xi(0)+(\sigma\xi,\xi_x)+\sigma(0)\xi^2(0)-(\sigma\sigma_x,\xi).$$

Therefore

$$\begin{aligned}|(R_{21},\xi)|&\leqslant C\|\sigma\|\|\xi_x\|+C\|\sigma\|_\infty\|\xi\|_\infty+\|\sigma\|_\infty\|\xi\|\|\xi_x\|+\|\sigma\|_\infty\|\xi\|_\infty^2+\|\sigma\|_\infty\|\sigma_x\|\|\xi\|\\&\leqslant Ch^r\|\xi_x\|+Ch^r\|\xi\|_\infty+Ch^r\|\xi\|\|\xi_x\|+Ch^r\|\xi\|_\infty^2+Ch^{2r-1}\|\xi\|\\&\leqslant Ch^{r-1}(\|\xi\|+\|\xi\|^2).\end{aligned} \tag{3.29}$$

Finally, by (3.15) and integration by parts we have for the R_{22} term

$$(R_{22},\xi)=(v\sigma_x,\xi)-\tfrac{1}{2}(v_x\xi,\xi)+(w_x\rho,\xi)+(w_x\theta,\xi)-(\sigma_x\rho,\xi)-(\sigma_x\theta,\xi)-(\rho\xi_x,\xi).$$

Hence,

$$\begin{aligned}|(R_{22},\xi)|&\leqslant C\|\sigma_x\|\|\xi\|+C\|\xi\|^2+C\|\rho\|\|\xi\|+C\|\theta\|\|\xi\|\\&\quad+\|\sigma_x\|\|\rho\|_\infty\|\xi\|+\|\sigma_x\|\|\theta\|_\infty\|\xi\|+\|\rho\|_\infty\|\xi_x\|\|\xi\|\\&\leqslant Ch^{r-1}\|\xi\|+C\|\xi\|^2+Ch^r\|\xi\|+C\|\theta\|\|\xi\|+Ch^{2r-1}\|\xi\|\\&\quad+Ch^{r-1}\|\theta\|_\infty\|\xi\|+Ch^r\|\xi_x\|\|\xi\|\\&\leqslant Ch^{r-1}\|\xi\|+C\|\xi\|^2+C\|\theta\|\|\xi\|.\end{aligned} \tag{3.30}$$

By (3.21), taking into account (3.23)-(3.30) we see that

$$\tfrac{1}{2}\tfrac{d}{dt}(\|\theta\|^2+\|\xi\|^2)\leqslant Ch^{r-1}(\|\theta\|+\|\xi\|)+C(\|\theta\|^2+\|\xi\|^2),\quad t\in[0,t_h].$$

An application of Gronwall's Lemma and (3.8) yield

$$\|\theta(t)\|+\|\xi(t)\|\leqslant Ch^{r-1},\quad t\in[0,t_h], \tag{3.31}$$

from which by inverse assumptions it follows that $\|\theta\|_{1,\infty}+\|\xi\|_{1,\infty}\leqslant Ch^{r-5/2}$ for $t\in[0,t_h]$. Since it was assumed that $r\geqslant 3$ this contradicts the maximality of t_h and (3.31) holds for $0\leqslant t\leqslant T$. The estimate (3.9) follows. Since now $\|v-v_h\|_\infty\leqslant\|\rho\|_\infty+\|\theta\|_\infty\leqslant Ch^{r-3/2}$ and similarly $\|w-w_h\|_\infty\leqslant Ch^{r-3/2}$, and since

$$\eta-\eta_h=[\delta_0+\tfrac{1}{4}((v-w)+(v_h-w_h))][(v-w)-(v_h-w_h)],$$

we conclude that $\|\eta-\eta_h\|\leqslant C(\|v-v_h\|+\|w-w_h\|)\leqslant Ch^{r-1}$. Similarly $\|u-u_h\|\leqslant\|v-v_h\|+\|w-w_h\|\leqslant Ch^{r-1}$, and the proof of Proposition 3.1 is now complete. □

4. Numerical implementation and experiments

4.1 *Supercritical case*

We will implement the standard Galerkin method for the shallow water equations with characteristic boundary conditions in the supercritical case using the space of piecewise linear continuous functions on a uniform mesh in $[0,1]$ in the usual manner that problems with nonhomogeneous Dirichlet boundary conditions are approximated in practice. For this purpose we let $x_i=ih$, $0\leqslant i\leqslant N$, $Nh=1$, define $S_h=\{\phi\in C^0[0,1]:\phi\big|_{[x_j,x_{j+1}]}\in\mathbb{P}_1,0\leqslant j\leqslant N-1\}$, and let $\mathring{S}_h$ consist of the functions in S_h that vanish at $x=0$. We seek η_h, $u_h:[0,T]\to S_h$, the semidiscrete approximation of the solution of (SW1), satisfying for $0\leqslant t\leqslant T$ $\eta_h(0,t)=\eta_0$, $u_h(0,t)=u_0$, and the system of ode's

$$\begin{aligned}&(\eta_{ht},\phi)+(u_{hx},\phi)+((\eta_h u_h)_x,\phi)=0,\quad\forall\phi\in\mathring{S}_h,\\&(u_{ht},\phi)+(\eta_{hx},\phi)+(u_h u_{hx},\phi)=0,\quad\forall\phi\in\mathring{S}_h,\end{aligned} \tag{4.1}$$

with initial values $\eta_h(0)=P\eta^0$, $u_h(0)=Pu^0$, where P is the L^2 projection onto S_h. We discretize this ode ivp in time by the 'classical' explicit 4^{th}-order accurate Runge-Kutta scheme, written in the case of the ode $y'=f(t,y)$, $0\leqslant t\leqslant T$, in the form

$$\begin{aligned}y^{n,1}&=y^n+\tfrac{k}{2}f(t^n+\tfrac{k}{2},y^n),\\y^{n,2}&=y^n+\tfrac{k}{2}f(t^n+\tfrac{k}{2},y^{n,1}),\\y^{n,3}&=y^n+kf(t^n+k,y^{n,2}),\\y^{n+1}&=y^n+k\big(\tfrac{1}{6}f(t^n,y^n)+\tfrac{1}{3}f(t^n+\tfrac{k}{2},y^{n,1})\\&\qquad+\tfrac{1}{3}f(t^n+\tfrac{k}{2},y^{n,2})+\tfrac{1}{6}f(t^n+k,y^{n,3})\big),\end{aligned} \tag{4.2}$$

where k is the time step, $t^n=nk$, $n=0,1,\dots,M-1$, $Mk=T$, and y^n approximates $y(t^n)$. Theoretical and numerical evidence from linear stability theory and previous work by the authors, Antonopoulos & Dougalis (2013), Antonopoulos & Dougalis (to appear), on similar nonlinear systems, suggests that the resulting fully discrete scheme is stable under a Courant-number restriction of the form $k/h\leqslant r_0$.

In our first numerical experiment we check the spatial rate of convergence of this fully discrete scheme. We consider (SW1) with $\eta_0 = 1$ and $u_0 = 3$ and add right-hand sides to the pde's so that the exact solution of the ibvp is $\eta(x,t) = xe^{-xt} + \eta_0$, $u(x,t) = (1 - x - \cos(\pi x))e^{2t} + u_0$. With $h = 1/N$, $k = h/10$ (so that the temporal error is negligible), we obtain the L^2 errors and associated rates of convergence of (essentially) the semidiscrete problem at $T = 1$ shown in Table 1. The experimental

TABLE 1. *L^2 errors and spatial orders of convergence, supercritical case.*

N	η	*order*	u	*order*
40	1.243098(−3)	–	5.623510(−3)	–
80	3.110525(−4)	1.99871	1.405648(−3)	2.00024
160	7.778520(−5)	1.99959	3.513979(−4)	2.00006
320	1.944737(−5)	1.99992	8.784876(−5)	2.00001
480	8.643341(−6)	1.99998	3.904381(−5)	2.00001
520	7.364768(−6)	1.99996	3.326806(−5)	2.00001

rates of convergence for both components of the solution are clearly equal to 2, i.e. superaccurate, as the expected rates for a general quasiuniform mesh would be equal to 1. A numerical study of stability for this example indicates that the errors remain of the same order of magnitude at $T = 1$ up to about $k/h = 0.13$. For larger values of k/h blow-up eventually occurs. It should be noted that for the same test problem the alternative standard Galerkin formulation (2.7)-(2.9) (analyzed in section 2) coupled with the same Runge-Kutta scheme gives L^2 errors that coincide with those of Table 1 to at least 5 significant digits; the stability condition was also the same. (Recall that the result of Proposition 2.1 strictly holds for $r \geqslant 3$ due to the technical requirement in the proof for controlling the $W^{1,\infty}$ norm of the error. The numerical results suggest that the scheme converges for $r = 2$ as well and that the superaccurate order of convergence for a uniform mesh for $r = 2$, proved in Antonopoulos & Dougalis (to appear) for an ibvp for the shallow water equations with homogeneous Dirichlet boundary conditions on u, persists in the case of the ibvp (SW1) too.)

Since the temporal error is much smaller than the spatial one, the experimental estimation of the temporal order of convergence may be done in the following way, used in Bona *et al.* (1995). Let H_h^n be

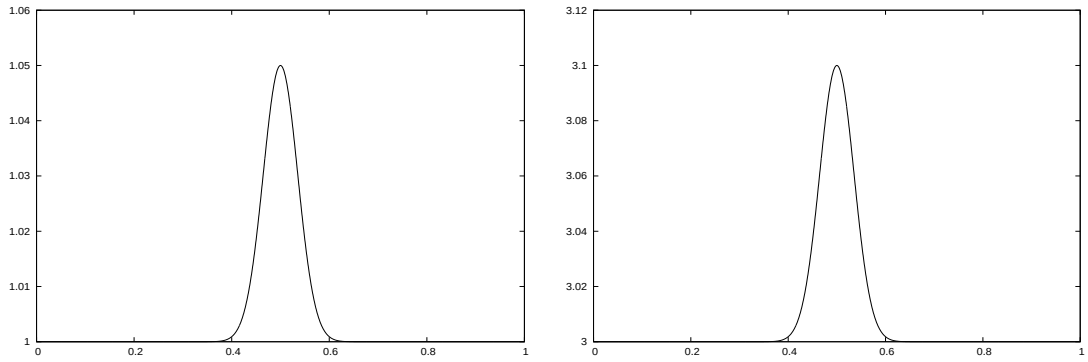
TABLE 2. *Temporal order of convergence, supercritical case, scheme (4.1)-(4.2), $h = 1/100$, $T = 1$, $k_{ref} = h/120$.*

k/h	$E^*(T)$	*order*	$E(T)$
1/35	2.6618459890(−8)	–	1.9910684230(−4)
1/40	1.6020860073(−8)	3.8022	1.9910680992(−4)
1/45	1.0112973048(−8)	3.9061	1.9910679532(−4)
1/50	6.6717108025(−9)	3.9478	1.9910678792(−4)
1/55	4.5726218272(−9)	3.9638	1.9910678370(−4)
1/60	3.2362144361(−9)	3.9728	1.9910678102(−4)
1/64	2.5020819256(−9)	3.9865	1.9910677950(−4)
1/64.5	2.4254282105(−9)	3.9983	1.9910677934(−4)
1/65	2.3516195603(−9)	4.0020	1.9910677918(−4)

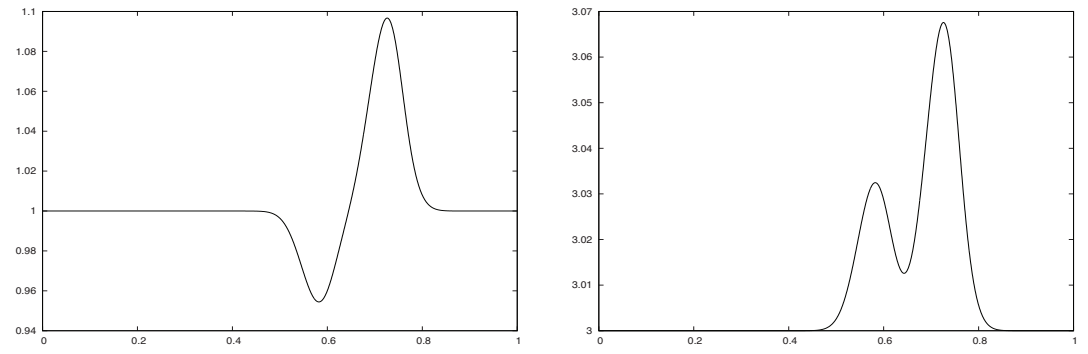
the fully discrete approximation of $\eta(t^n)$. For a fixed value of h we make a reference computation with

a very small value $k = k_{ref}$. The approximate solution $H_h^m = H_h^m(h,k_{ref})$, where $mk_{ref} = T$, differs from the exact solution $\eta(\cdot,T)$ by an amount which is practically the error of the spatial discretization. For the same value of h we define a modified L^2 error for small values of k, that are nevertheless considerably larger than k_{ref}, by the formula $E^*(T) = \|H_h^n(h,k) - H_h^m(h,k_{ref})\|$, where $nk = T$. Since taking the difference $H_h^n(h,k) - H_h^m(h,k_{ref})$ essentially cancels the spatial error of $H_h^n(h,k)$, we expect that $E^*(T)$ will decrease at the temporal order of convergence of the scheme as k decreases. This is illustrated in the case of the test problem under consideration and the fully discrete scheme (4.1)-(4.2) in Table 2, where $h = 1/100$, $T = 1$, $k_{ref} = h/120$, and $E(T)$ denotes the L^2 error $\|H_h^n(h,k) - \eta(t^n)\|$. For this range of k's the expected temporal order of convergence, equal to 4, clearly emerges. The analogous experiment with the Galerkin method (2.7)-(2.9) discretized in time with the same Runge-Kutta scheme yields fourth-order temporal convergence in L^2 as well.

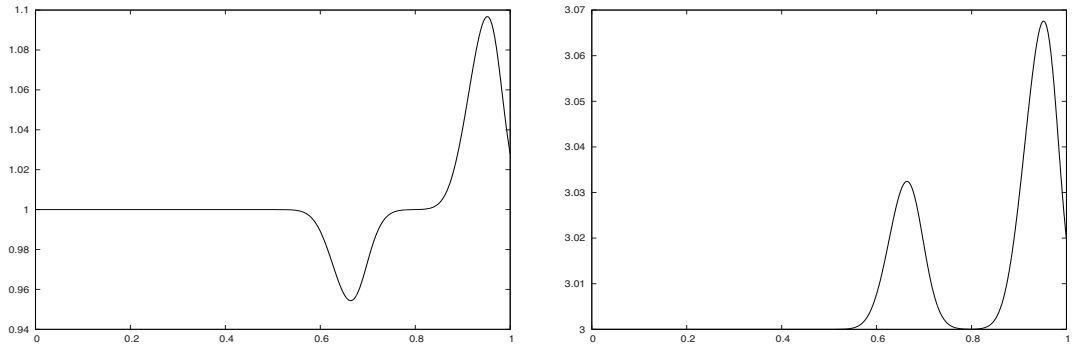
In the next numerical experiment we integrate (SW1) with the fully discrete scheme (4.1)-(4.2), taking $h = 1/N$, $N = 2000$, $k = h/10$, $\eta_0 = 1$, $u_0 = 3$, and initial conditions $\eta^0(x) = 0.05\exp(-400(x-0.5)^2) + \eta_0$, $u^0(x) = 0.1\exp(-400(x-0.5)^2) + u_0$, $0 \leqslant x \leqslant 1$. (Small-amplitude initial conditions were taken to ensure that no discontinuities in the derivatives of the solution develop before the wave profiles



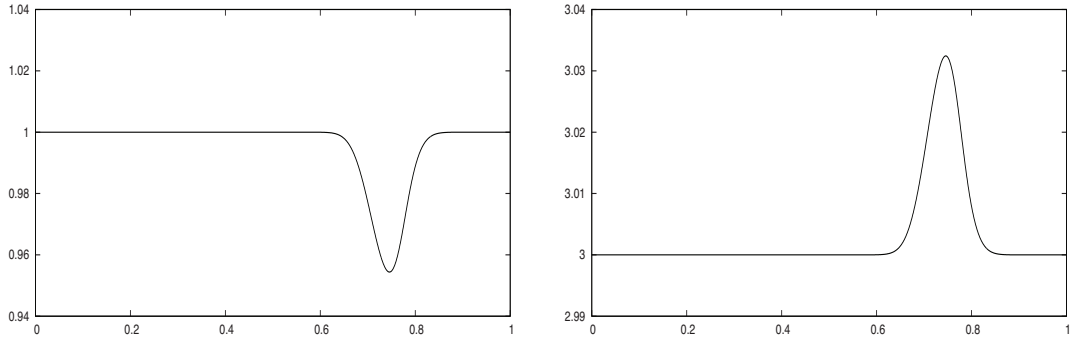
(a) η and u at $t = 0.0$



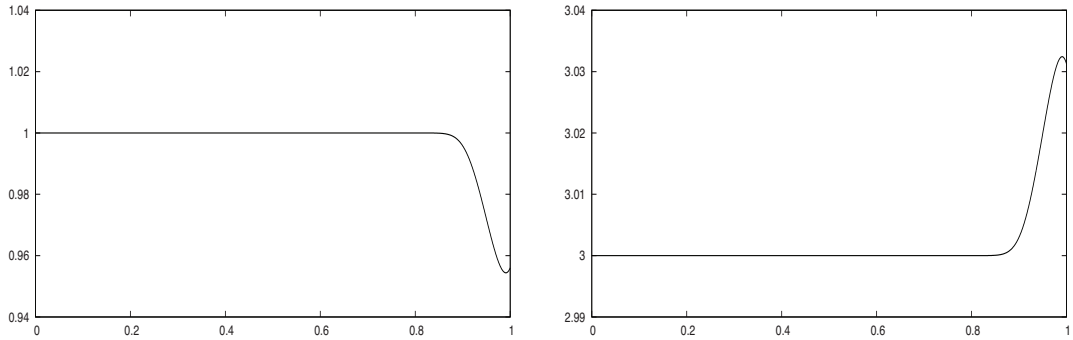
(b) η and u at $t = 0.05$



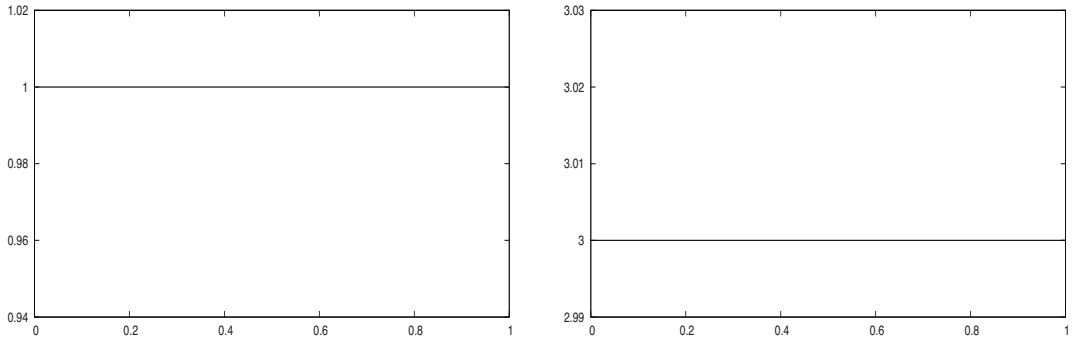
(c) η and u at $t = 0.1$



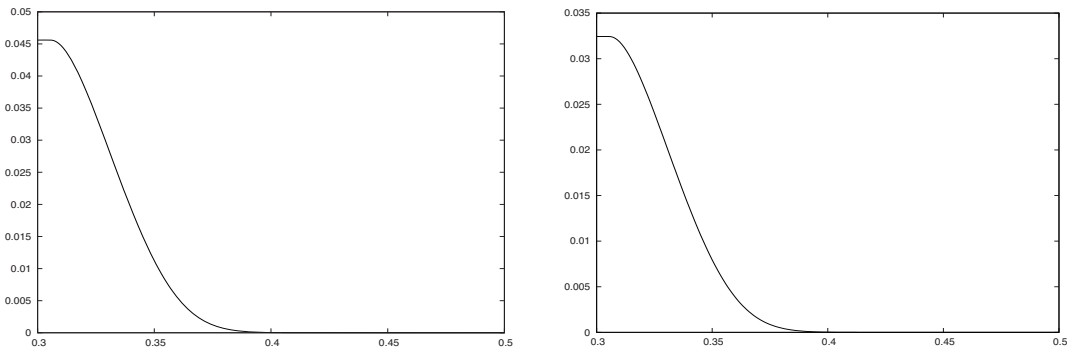
(d) η and u at $t = 0.15$



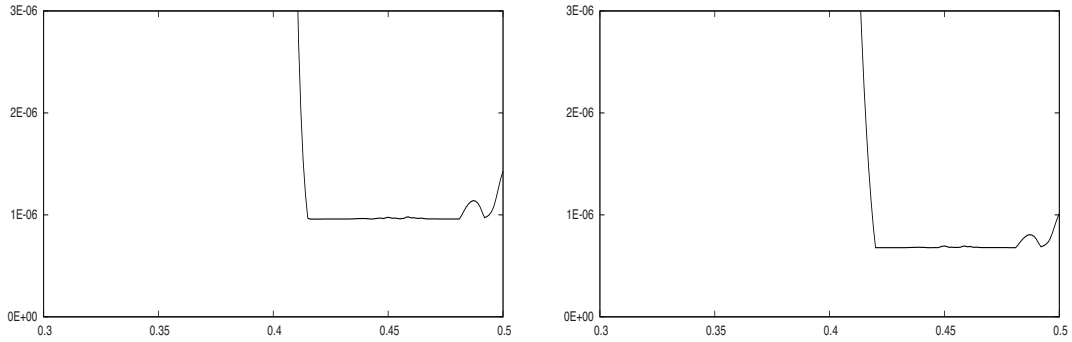
(e) η and u at $t = 0.3$



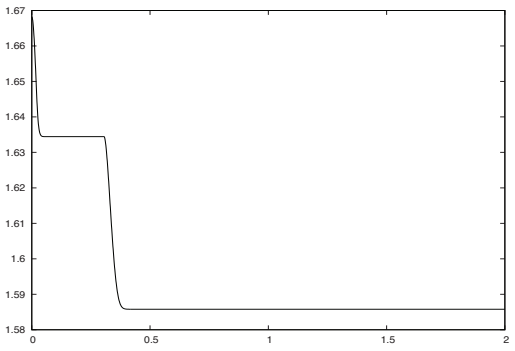
(f) η and u at $t = 0.4$



(g) $\max_x|\eta(x,t)-\eta_0|$ and $\max_x|u(x,t)-u_0|$ vs. time



(h) Magnification of (g)



(i) $\max_x(u-\sqrt{1+\eta})$ vs. time

FIG. 1: Evolution of Gaussian initial profiles, supercritical case.

exit the spatial interval of integration.) The evolution of the numerical solution is depicted in Figure 1(a)-(f). (The approximate η-profiles are on the left and those of u on the right.) The initial Gaussian perturbations of the uniform state $\eta_0 = 1$, $u_0 = 3$ evolve into two unequal pulses for both η and u that travel to the right and exit the computational domain by about $t = 0.4$ without leaving any visible residue or backwards-travelling oscillations as is confirmed by Fig. 1(g) that shows the time history of the quantities $\max_x|\eta(x,t)-\eta_0|$ and $\max_x|u(x,t)-u_0|$. (Here η, u denote the approximate solution.) Due to the presence of the spatial and temporal discretizations, the numerical boundary conditions are not expected to be exactly transparent. However, they are highly absorbing; Figure 1(h) reveals that the residue after the waves exit is of $O(10^{-6})$. The positivity of $\max_x(u-\sqrt{1+\eta})$ for all t checked in Fig.1(i) confirms that the numerical solution has remained supercritical throughout the evolution.

In order to study numerically the stability of the fully discrete scheme (4.1)-(4.2) for this test problem, as there are no exact solutions, we took as a measure of error the residual quantity $\max_x|\eta-\eta_0|$. This is plotted in Figure 1(g) and (h), stabilizes after the waves exit the computational domain, and has the value $9.76E-07$ at $t = 0.45$. We then increased k and observed that this residual was conserved up to about $k/h = 0.3695$ and started increasing afterwards. Since the maximum wave speed c is the speed of the higher rightward-travelling pulse, which is equal to $u+\sqrt{1+\eta} \simeq 4.5$ for the duration of this experiment, we obtain a Courant number restriction of about $ck/h \leqslant 1.67$. (Linear stability theory for this method applied to the model problem $\eta_t + c\eta_x = 0$ with periodic boundary conditions and c constant would give a Courant number restriction $ck/h \leqslant \sqrt{8/3} \simeq 1.633$ which is not far from the experimental result for this small-amplitude nonlinear propagation problem.) It should be noted that the scheme (2.7)-(2.9) discretized in time with the Runge-Kutta method (4.2) yields practically the same numerical results for this test problem.

4.2 *Subcritical case*

We implement the standard Galerkin method for the SW with characteristic boundary conditions in the subcritical case using again piecewise linear continuous functions on a uniform mesh in $[0,1]$. In addition to the notation introduced in the previous subsection for the mesh on $[0,1]$ and the space S_h, we let $S_{h,0}$ consist of the functions in S_h that vanish at $x = 0$ and $x = 1$. We first consider the 'direct'

standard Galerkin semidiscretization of (SW2), i.e. without reducing first the ibvp into the diagonal form (SW2a). We seek accordingly η_h, $u_h : [0,T] \to S_h$, the semidiscrete approximation of the solution of (SW2), satisfying for $0 \leqslant t \leqslant T$

$$(\eta_{ht}, \phi) + (u_{hx}, \phi) + ((\eta_h u_h)_x, \phi) = 0, \quad \forall \phi \in S_h, \tag{4.3}$$

$$(\widetilde{u}_{ht}, \chi) + (\eta_{hx}, \chi) + (\widetilde{u}_h \widetilde{u}_{hx}, \chi) = 0, \quad \forall \chi \in S_{h,0}, \tag{4.4}$$

where $\widetilde{u}_h \in S_{h,0}$, and $u_h(x_i,t) = \widetilde{u}_h(x_i,t)$, $1 \leqslant i \leqslant N-1$,

$$u_h(x_0,t) = -2\sqrt{1+\eta_h(x_0,t)} + u_0 + 2\sqrt{1+\eta_0}, \tag{4.5}$$

$$u_h(x_N,t) = 2\sqrt{1+\eta_h(x_N,t)} + u_0 - 2\sqrt{1+\eta_0}. \tag{4.6}$$

At $t = 0$ we compute $u_h(0)$ and $\eta_h(0)$ as the L^2 projections of the initial data η^0 , u^0 onto S_h. Although the semidiscrete ivp (4.3)-(4.6) has nonlinear boundary conditions, it is easily discretized in time by an explicit scheme such as the 4^{th}-order RK method (4.2), by first advancing from t^n to t^{n+1} the approximations of η_h and of the 'interior' $\widetilde{u}_h$ using the temporal discretizations of (4.3) and (4.4), and then updating the $u_h(x_0,t)$ and $u_h(x_N,t)$ values at $t = t^{n+1}$ by (4.5) and (4.6).

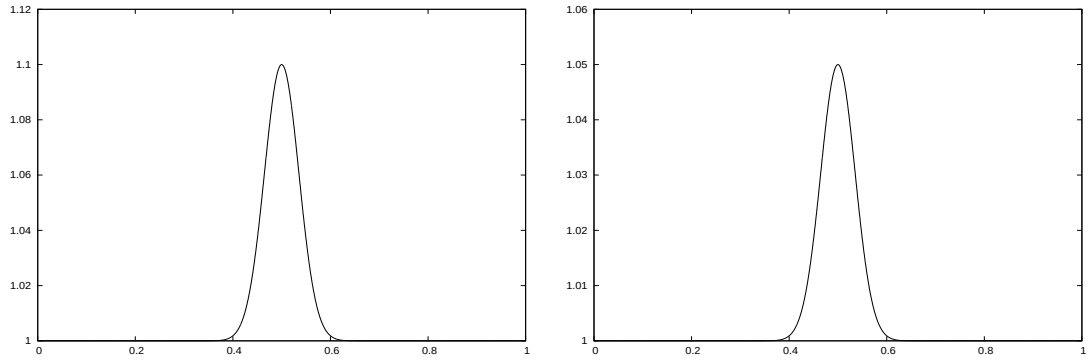
The convergence of this scheme was not analyzed in section 3. The following experiment suggests that in the case of uniform mesh its L^2 errors are of $O(h^2)$. We consider (SW2) with $\eta_0 = 1$, $u_0 = 1$ and its semidiscretization (4.3)-(4.6) with piecewise linear continuous functions. We add appropriate

TABLE 3. *L^2 errors and spatial orders of convergence, subcritical case, semidiscretization (4.3)-(4.6).*

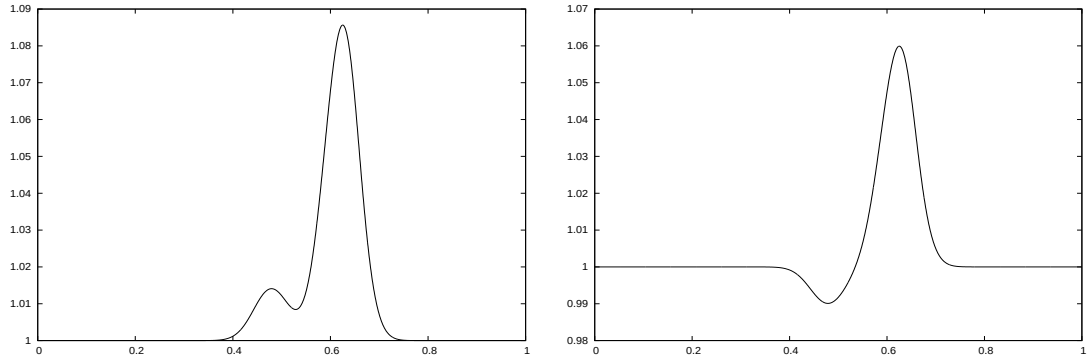
N	η	*order*	u	*order*
40	4.847892(−3)	–	2.932354(−3)	–
80	1.207564(−3)	2.00526	7.414336(−4)	1.98367
160	3.017313(−4)	2.00076	1.860285(−4)	1.99479
320	7.544641(−5)	1.99974	4.657627(−5)	1.99786
480	3.353298(−5)	1.99991	2.071174(−5)	1.99867
520	2.857355(−5)	1.99953	1.764866(−5)	1.99944

right-hand sides to the pde's in (SW2) so that the exact solution of the ibvp is $\eta(x,t) = (x+1)\mathrm{e}^{-xt}$, $u(x,t) = (2x + \cos(\pi x) - 1)\mathrm{e}^t + xA(t) + (1-x)B(t)$, where $A(t) = 2\sqrt{1+\eta(1,t)} + u_0 - 2\sqrt{1+\eta_0}$, $B(t) = -2\sqrt{1+\eta(0,t)} + u_0 + 2\sqrt{1+\eta_0}$. We consider uniform spatial and temporal meshes with $h = 1/N$, $k = h/10$, and discretize the semidiscrete problem in time using again the 4^{th}-order 'classical' RK scheme. (We checked that the temporal error is negligible for the range of N's that we tried.) The resulting L^2 errors and rates of convergence of (essentially) the semidiscrete problem at $T = 1$ are shown in Table 3. The rates are practically equal to 2, i.e. superaccurate, as in the supercritical case. An analogous temporal-order calculation to that appearing in Table 2 was not so robust and gave rates between 3.7 and 3.9; thus some sort of temporal order reduction cannot be ruled out for this scheme. The experimental Courant number restriction was $k/h \leqslant 0.53$.

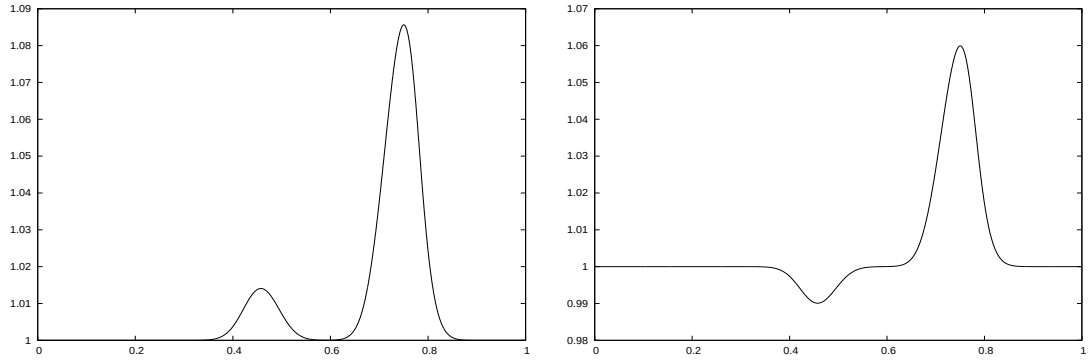
In the following numerical experiment we integrate (SW2) with the same fully discrete scheme, taking $h = 1/N$, $N = 2000$, $k = h/10$, $\eta_0 = u_0 = 1$, and initial conditions $\eta^0(x) = 0.1\exp(-400(x-0.5)^2) + \eta_0$, $u^0(x) = 0.05\exp(-400(x-0.5)^2) + u_0$. The evolution of the numerical solution is shown in Figure 2 (a)-(j).



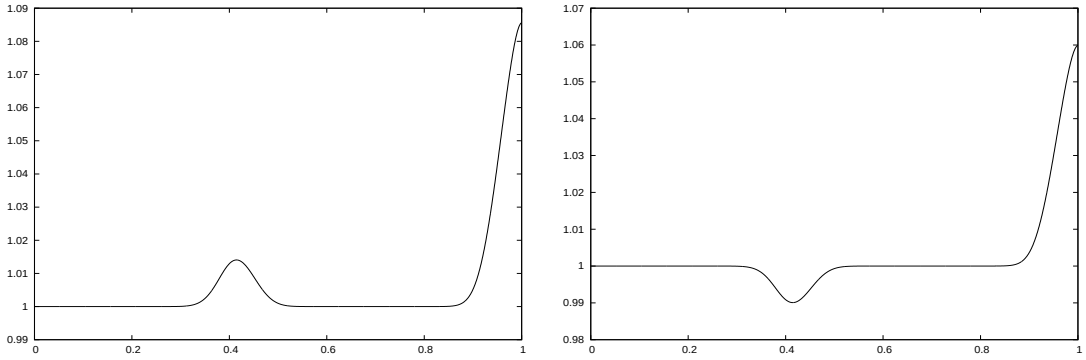
(a) η and u at $t = 0.0$



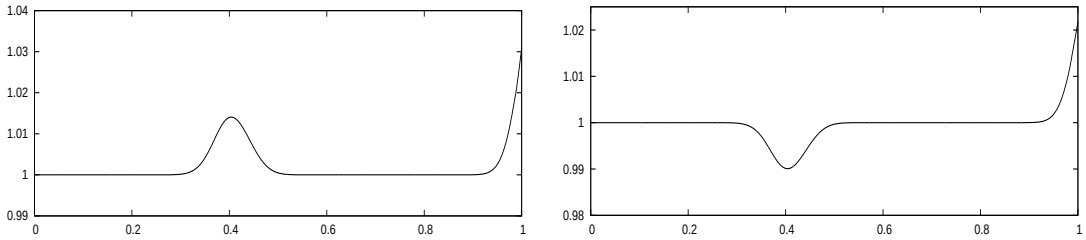
(b) η and u at $t = 0.05$



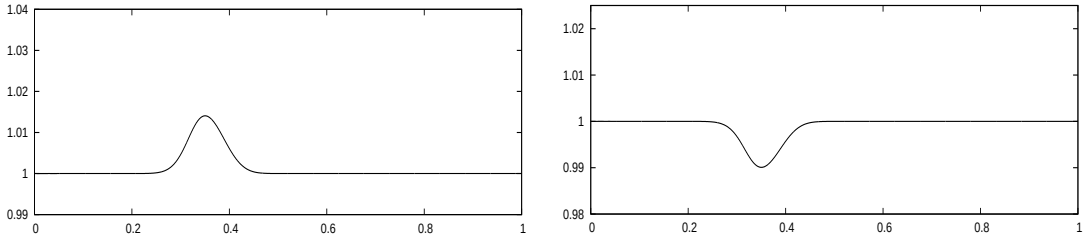
(c) η and u at $t = 0.1$



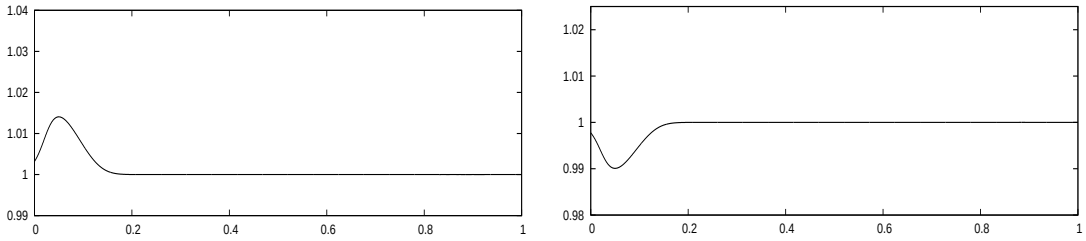
(d) η and u at $t = 0.2$



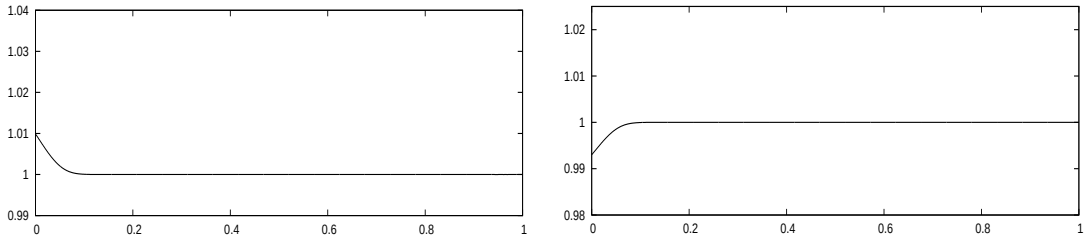
(e) η and u at $t = 0.225$



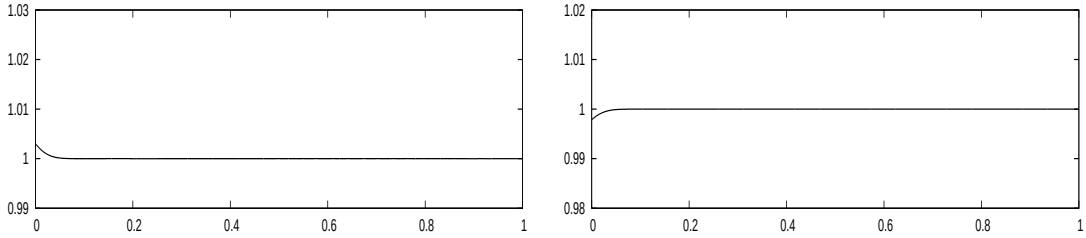
(f) η and u at $t = 0.35$



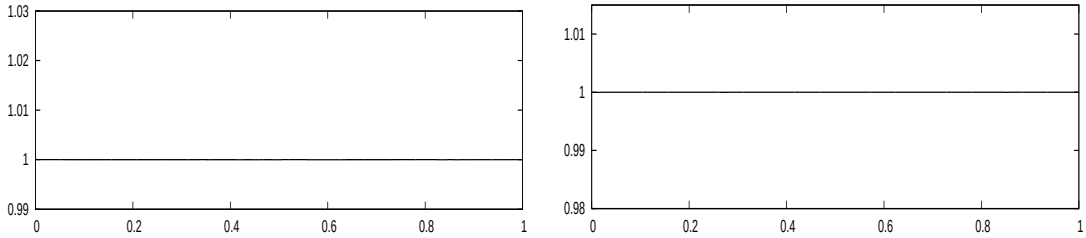
(g) η and u at $t = 1.05$



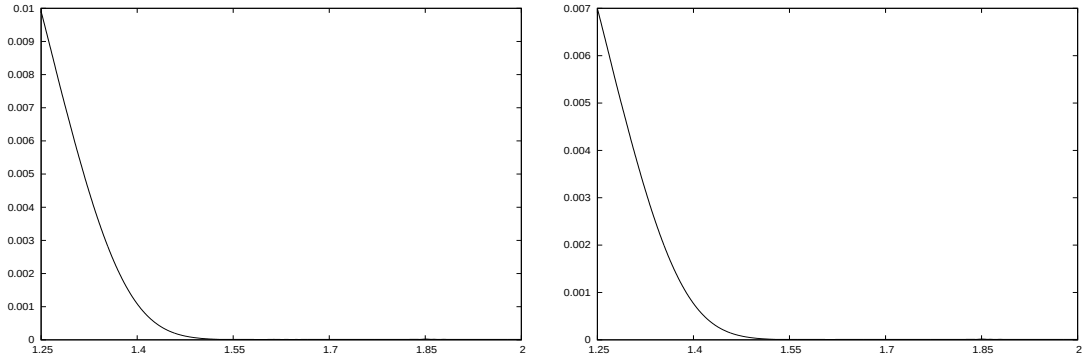
(h) η and u at $t = 1.25$



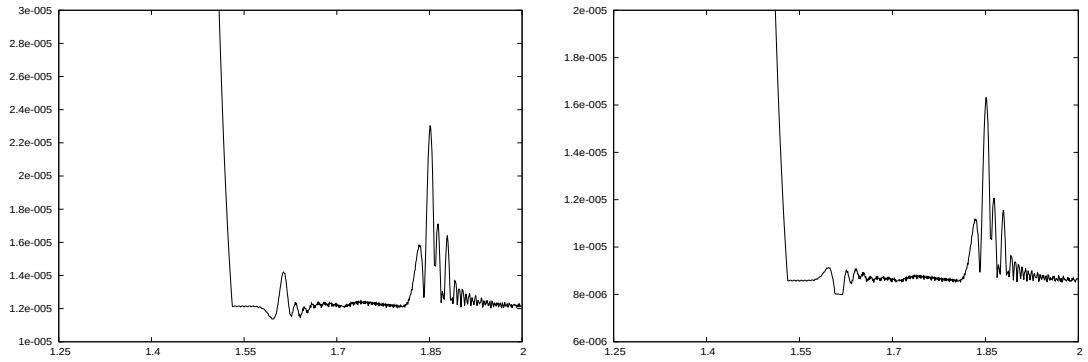
(i) η and u at $t = 1.35$



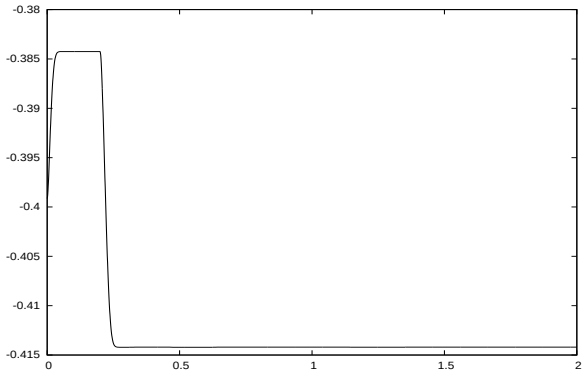
(j) η and u at $t = 1.5$



(k) $\max_x|\eta(x,t)-\eta_0|$ and $\max_x|u(x,t)-u_0|$ vs. time



(l) manification of (k)



(m) $\max_x(u-\sqrt{1+\eta})$ vs. time

FIG. 2: Evolution of Gaussian initial profiles, subcritical case, semidiscretization (4.3)-(4.6).

The initial Gaussian perturbations of the steady state $\eta_0 = 1$, $u_0 = 3$ evolve into two unequal pulses for both components of the solution, which travel to the right and left and exit the computational domain at about $t = 0.25$ and $t = 1.35$, respectively, without leaving any visible residue as confirmed by the temporal history of the maximum deviations of the approximations of η and u from η_0 and u_0 shown in Figures 2(k) and (l). The latter graph shows that the maximum residue after exit is of $O(10^{-5})$, confirming the high degree of absorption of the discrete characteristic boundary conditions. The quantity $\max_x(u-\sqrt{1+\eta})$ remains negative, i.e. the numerical solution is subcritical, throughout the evolution, cf. Fig. 2(m). We investigated the stability of this fully discrete scheme by using again as a measure of error the quantity $\max_x|\eta-\eta_0|$ which stabilizes at $t = 1.55$ to the value $1.21E-05$. The maximum wave speed c for this problem is about 2.5, and the observed Courant number restriction was $ck/h \leqslant 1.67$ in conformity with the analogous value in the supercritical case.

In section 3 we analyzed a different Galerkin semidiscretization of the ibvp under consideration, namely that given by (3.6)-(3.8), a semidiscrete approximation of the diagonal form of the ibvp, i.e.

of (SW2a). As the following experiment suggests, this semidiscretization is also $O(h^2)$ accurate in L^2 if we use piecewise linear continuous functions on a uniform mesh. We consider the inhomogeneous version of (SW2a) with unknowns $2v$ and $2w$ instead of v and w and take as exact solution the functions $v = u - u_0 + 2(\sqrt{1+\eta} - \delta_0)$, $w = u - u_0 - 2(\sqrt{1+\eta} - \delta_0)$, where $\delta_0 = \sqrt{1+\eta_0}$ and $\eta(x,t) = (x+1)\mathrm{e}^{-xt}$, $u(x,t) = (2x+\cos(\pi x)-1)\mathrm{e}^t + xA(t) + (1-x)B(t)$, with $A(t) = 2\sqrt{1+\eta(1,t)} + u_0 - 2\delta_0$, $B(t) = -2\sqrt{1+\eta(0,t)} + u_0 + 2\delta_0$. For $\eta_0 = u_0 = 1$ we approximate this ibvp by the nonhomogeneous analog of the semidiscretization (3.6)-(3.8) (with unknowns $2v_h$ and $2w_h$) using piecewise linear continuous elements with $h = 1/N$, and the 'classical' 4^{th}-order RK scheme for time stepping with $k = h/10$. The

TABLE 4. *L^2 errors and spatial orders of convergence, subcritical case, semidiscretization (3.6)-(3.8).*

N	η	*order*	u	*order*
40	2.470369(−3)	–	9.918820(−4)	–
80	6.172661(−4)	2.00076	2.472869(−4)	2.00398
160	1.543038(−4)	2.00012	6.179903(−5)	2.00053
320	3.857665(−5)	1.99997	1.545737(−5)	1.99929
480	1.714531(−5)	1.99998	6.870865(−6)	1.99967
520	1.460903(−5)	1.99999	5.854663(−6)	1.99958

resulting L^2 errors and rates of convergence at $T = 1$ are given in Table 4. The rates are practically equal to 2 as in the previous cases, due to the uniform mesh. The temporal order of convergence for this scheme was found to be practically equal to 4; the associated temporal-order calculation with $h = 1/50$, $k_{ref} = h/200$ was quite robust.

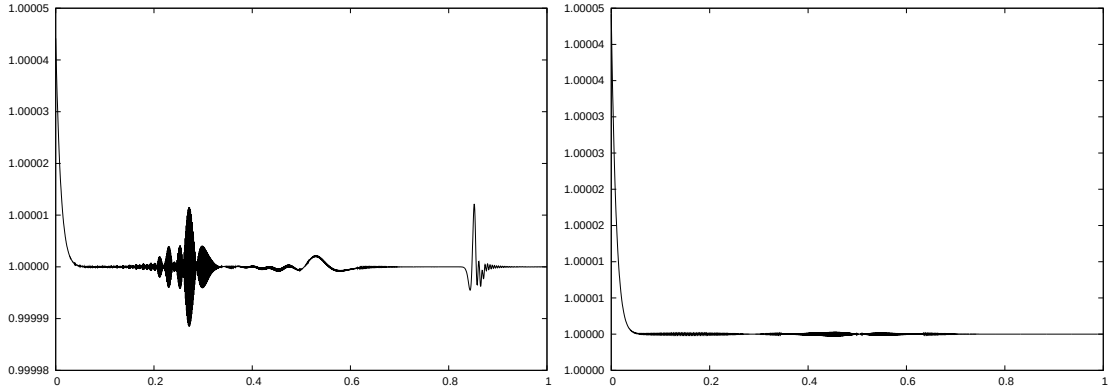
We now compare the solutions of the two semidiscretizations (4.3)-(4.6) and (3.6)-(3.8) by means of a numerical experiment. We consider the ibvp (SW2) with $\eta_0 = u_0 = 1$ and initial values $\eta(x,0) = 0.1\exp(-400(x-0.5)^2) + \eta_0$, $u(x,0) = 0.05\exp(-400(x-0.5)^2) + u_0$. (Recall that the temporal evolution of the numerical solution of this ibvp generated by (4.3)-(4.6) with $N = 2000$, $h = 1/N$, $k = h/10$ is shown in Figure 2.) We will compare the numerical solutions of this ibvp with both discretizations, using the same numerical parameters. Let (η_h, u_h) denote the fully discrete approximations at time t produced by (4.3)-(4.6) as underlying semidiscretization. In addition, let (η_{hD}, u_{hD}) be functions in S_h with point values computed by the formulas (3.10), where v_h, w_h are now the fully discrete approximations at t produced when we use the method (3.6)-(3.8). Define $\varepsilon(t) = \max_{0\leqslant i\leqslant N}|\eta_h(x_i,t) - \eta_{hD}(x_i,t)|$ and $e(t) = \max_{0\leqslant i\leqslant N}|u_h(x_i,t) - u_{hD}(x_i,t)|$. The quantities ε and e for various values of $t = t^n \in [0,1]$

TABLE 5. *Comparison of the discretizations (4.3)-(4.6) and (3.6)-(3.8), subcritical case, experiment of Fig. 2*

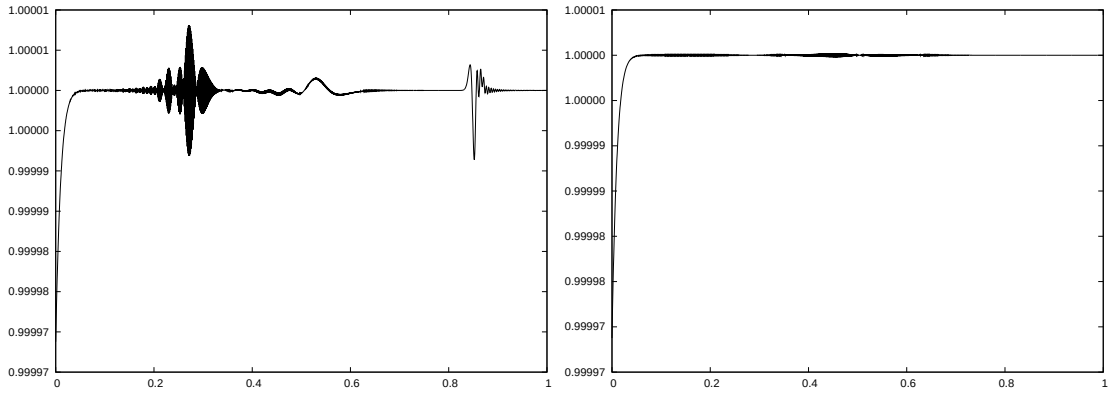
t	0.0	0.05	0.1	0.15	0.2	0.225	0.25
$\varepsilon(t)$	2.0(−8)	2.0(−8)	2.0(−8)	2.0(−8)	4.0(−8)	3.0(−8)	3.0(−8)
$e(t)$	$< 10^{-8}$	$< 10^{-8}$	$< 10^{-8}$	$< 10^{-8}$	2.0(−8)	2.0(−8)	2.0(−8)

t	0.35	0.75	1.05	1.15	1.25	1.35	1.5
$\varepsilon(t)$	1.232(−5)	1.301(−5)	1.237(−5)	1.3(−5)	1.224(−5)	1.23(−5)	1.217(−5)
$e(t)$	8.71(−6)	9.2(−6)	8.75(−6)	9.54(−6)	8.65(−6)	8.75(−6)	8.6(−6)

are given in Table 5. We observe that up to about $t = 0.25$ the errors are of $O(10^{-8})$ and subsequently increase to $O(10^{-5})$. This is due to the fact that about $t = 0.25$ the larger, rightwards-travelling wave completes its exit from the computational domain (see Figure 2(e), and we expect a small residue to be



(a) η_h and η_{hD}, $t = 1.5$



(b) u_h and u_{hD}, $t = 1.5$

FIG. 3: Magnifications of the profiles of the numerical solutions generated by the methods (4.3)-(4.6) (left) and (3.6)-(3.8) (right) at $t = 1.5$. Subcritical case, evolution as in Fig 2.

radiated into $[0,1]$ because the numerical characteristic boundary conditions of (4.3)-(4.6) are not exactly transparent. This residue has a magnitude of $O(10^{-5})$ as evidenced by Fig. 2(l). It is worthwhile to note that the 'diagonal' approximation (3.6)-(3.8) leaves a practically negligible residue. This is suggested by the evidence in Figure 3. In this figure, the two graphs on the left depict what is left in the computational domain of the η-and u-components of the numerical solution generated by (4.3)-(4.6) at $t = 1.5$, after the waves have exited the domain, while those on the right are the analogous η-and u-profiles generated by (3.6)-(3.8). (Thus the graphs on the left are magnifications of the graphs of Fig. 2(j).) We observe that the residues of the usual Galerkin semidiscretizations are of $O(10^{-5})$ while those of the 'diagonal' scheme are much smaller. It is clear, at least for this example, that the 'diagonal' Galerkin method has practically transparent boundary conditions.

4.3 *Remarks*

In section 4 of Antonopoulos & Dougalis (2015) the interested reader may find a description of the *linearized* characteristic boundary conditions for the shallow water equations in the subcritical case and numerical experiments in which the fully discrete Galerkin-finite element method with the nonlinear characteristic boundary conditions is compared to its analog with the linearized conditions. As in Nycander *et al.* (2008) and Shiue *et al.* (2011) our conclusion is that the discretized nonlinear boundary conditions are much more absorbing than their linearized counterparts. Also recorded are the results of numerical experiments with the Galerkin scheme on two test problems of Nycander *et al.* (2008) involving the shallow water equations in their dimensional form.

Acknowledgements

The authors would like to thank Prof. J. Nycander, Dr. A. McC. Hogg, and Dr. L. M. Frankcombe for helpful comments on their numerical experiments in Nycander *et al.* (2008).

REFERENCES

ANTONOPOULOS, D. C. & DOUGALIS V. A. Error estimates for the standard Galerkin-finite element method for the Shallow Water equaions. (To appear in *Math. Comp.*; extended version in arXiv:1403.5699).

ANTONOPOULOS, D. C. & DOUGALIS V. A. (2015) Notes on Galerkin-finite element methods for the Shallow Water equations with characteristic boundary conditions. arXiv:1507.08209.

ANTONOPOULOS, D. C. & DOUGALIS V. A. (2013) Error estimates for Galerkin approximations of the 'classical' Boussinesq system. *Math. Comp.*, **82**, 689–717. (Extended version in arXiv:100.4248).

BONA, J. L., DOUGALIS, V. A., KARAKASHIAN, O. A. & MCKINNEY, W. R. (1995) Conservative, high-order numerical schemes for the generalized Korteweg de-Vries equation. *Phil. Trans. R. Soc. Lond. A* **351**, 107–164.

BOUSQUET, A., PETCU, M., SHIUE, M. - C., TEMAM, R. & TRIBBIA, J. (2013) Boundary conditions for limited area models based on the Shallow Water equations. *Commun. Comput. Phys.*, **14**, 664–702.

DOUGLAS, J. JR., DUPONT, T. & WAHLBIN, L. (1975) Optimal L^∞ error estimates for Galerkin approximations to solutions of two-point boundary value problems. *Math. Comp.*, **29**, 475–483.

DUPONT, T. (1973) Galerkin methods for first order hyperbolics: an example. *SIAM J. Numer. Anal.*, **10**, 890–899.

DUPONT, T. (1974) Galerkin methods for modelling gas pipelines. *Lecture Notes in Math.*, **430**, 112–130. Berlin: Springer-Verlag.

FRANKCOMBE, L. M. & HOGG, A. MCC. (2007) Tidal modulation of two-layer hydraulic exchange flows. *Ocean Sci.*, **3**, 179–188.

HUANG, A., PETCU, M. & TEMAM, R. (2011) The one-dimensional supercritical Shallow Water Equations with topography. *Ann. Univ. Bucharest (Mathematical Series)*, **2 LX**, 63–82.

KURGANOV, A., NOELLE, S. & PETROVA, G. (2001) Semidiscrete central-upwind schemes for conservation laws and Hamilton-Jacobi equations. *SIAM J.Sci.Comput.*, **23**, 707–740.

KURGANOV, A. & PETROVA, G. (2000) Central schemes and contact discontinuities. *ESAIM-M2AN*, **34**, 1259–1275.

MAJDA, A. (1984) *Compressible Fluid Flow and Systems of Conservation Laws in Several Space Variables*. New York: Springer-Verlag.

NYCANDER, J. & DÖÖS, K. (2003) Open boundary conditions for barotropic waves. *J. Geophys. Res.*, **108**, doi:10.1029/2002JC001529.

NYCANDER, J., HOGG, A. MCC., & FRANKCOMBE, L. M. (2008) Open boundary conditions for nonlinear channel flow. *Ocean Modelling*, **24**, 108–121.

PETCU, M. & TEMAM, R. (2011) The one-dimensional shallow water equations with transparent boundary conditions. *Math. Meth. Appl. Sci.*, doi:10.1002/mma.1482.

PETCU, M. & TEMAM, R. (2013) An interface problem : the two-layer shallow water equations. *DCDS-A*, **33**, 5327–5345.

SCHREIBER, R. (1980) Finite element methods of high-order accuracy for singular two-point boundary value problems with non-smooth solutions. *SIAM J. Numer. Anal.*, **17**, 547–566.

SHIUE, M. -C., LAMINIE, J., TEMAM, R. & TRIBBIA, J. (2011) Boundary value problems for the shallow water equations with topography. *J. Geophys. Res.*, **116**, C02015, doi:10.1029/2010JC006315.

WHITHAM, G. B. (1974) *Linear and Nonlinear Waves*. New York: Wiley.