Available online at www.sciencedirect.com
ScienceDirect
Ocean Modelling 15 (2006) 101–123
Ocean Modelling
www.elsevier.com/locate/ocemod

A one-dimensional benchmark for the propagation of Poincaré waves

Laurent White [a,b,*], Vincent Legat [a], Eric Deleersnijder [a,b], Daniel Le Roux [c]

[a] Centre for Systems Engineering and Applied Mechanics (CESAME), Université Catholique de Louvain, 4, Avenue Georges Lemaître, B-1348 Louvain-la-Neuve, Belgium
[b] G. Lemaître Institute of Astronomy and Geophysics (ASTR), Université Catholique de Louvain, 2, Chemin du Cyclotron, B-1348 Louvain-la-Neuve, Belgium
[c] Département de Mathématiques et de Statistique, Université Laval, Québec, Que., Canada G1K7P4

Received 23 December 2004; received in revised form 10 November 2005; accepted 30 November 2005
Available online 3 January 2006

Abstract

Several numerical methods are employed to solve the linear shallow-water equations describing the propagation of Poincaré waves within a one-dimensional finite domain. An analytical solution to the problem, set off by a discontinuous steplike elevation, is known and allows for assessing the accuracy and robustness of each method and in particular their ability to capture the traveling discontinuities without generating spurious oscillations. The following methods are implemented: the method of characteristics, the Galerkin finite-element method (FEM) and the discontinuous Galerkin FEM with two different ways of computing the numerical fluxes.

Keywords: Poincaré waves; Method of characteristics; Discontinuous finite elements; Riemann solver

1. Introduction

Motion in the ocean spans a very wide range of timescales. While the large-scale circulation is characterized by velocities on the order of up to one meter per second and timescales that can be as large as hundreds of years, the fast-propagating inertia–gravity waves exhibit phase velocities on the order of hundreds of meters per second and much smaller timescales. Internal gravity waves propagate with velocities on the order of one meter per second or less. The vast disparity of ocean processes timescales poses a challenge in numerical ocean modeling. If an explicit time step is used, it is limited by the so-called

* Corresponding author. Address: Centre for Systems Engineering and Applied Mechanics (CESAME), Université Catholique de Louvain, 4, Avenue Georges Lemaître, B-1348 Louvain-la-Neuve, Belgium. Tel.: +32 10472357.
E-mail address: lwhite@mema.ucl.ac.be (L. White).

1463-5003/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ocemod.2005.11.001

Courant–Friedrichs–Lewy (CFL) condition, which states that the time step should not be larger than the travel time of the fastest physical process over the smallest space increment. In free surface ocean models that allow for the existence of external inertia–gravity (Poincaré) waves, the upper bound on the time step is far smaller than more practical time steps that would permit time integration over thousands of years on today's computers. The first attempt at circumventing this problem by replacing the free surface by a rigid lid—thereby eliminating external inertia–gravity waves—has been widely dismissed. Among the rationales for such a design are that a rigid lid distorts the properties of large-scale barotropic Rossby waves, does not permit tidal modeling and complicates inclusion of fresh water flux surface boundary condition (Killworth et al., 1991; Dukowicz and Smith, 1994; Deleersnijder and Campin, 1995; Hallberg, 1997; Higdon and de Szoeke, 1997).

A common alternative no longer relies on the rigid-lid approximation. The ocean surface is free and remains a prognostic variable but the governing equations are split into subsystems that model the fast and slow motions separately. These subsystems are generally referred to as the barotropic and baroclinic systems, respectively, or the external and internal modes, respectively. Fast motions are approximately independent of the vertical coordinate z so that the external mode is two-dimensional and is well represented by the shallow-water equations that model the motion of fluid layers of constant density. Slow motions are fully three-dimensional, however, but the restriction on the time step is dictated by the internal dynamics, of which timescales are several orders of magnitude larger than that of the external mode. The latter can be solved explicitely with small time steps or implicitely with larger time steps. Choosing an implicit treatment eliminates the constraint imposed by the CFL condition but leads to large systems to be solved at each time step. This choice can be made for tidal and tsunami calculations provided that a reduced time step be used. If an explicit approach is considered for the barotropic mode, the number of small barotropic time steps for each large baroclinic time step is roughly the ratio of barotropic inertia–gravity wave speed to baroclinic internal gravity wave speed (Killworth et al., 1991). Details on mode splitting implementations can be found in Blumberg and Mellor (1987), Hallberg (1997), Higdon and de Szoeke (1997) and Higdon (2002).

Large-scale oceanic motions roughly obey the geostrophic equilibrium. When imbalances occur, the geostrophic balance is restored by means of Poincaré waves. In strongly stratified seas, internal inertia–gravity waves are generated when displacement of density surfaces occurs. Those waves respond to the same physical mechanism as external Poincaré waves (Gill, 1982). In models allowing for the existence of inertia–gravity waves, it is of paramount importance to represent those waves accurately. In that respect, the coupled issues of time and space discretization ought to be focused on. Time stepping is not the subject of this paper (see e.g., Beckers and Deleersnijder, 1993) as we concentrate on spatial discretization. A one-dimensional benchmark for the propagation of Poincaré waves is proposed. This problem bears many similarities with the classical geostrophic adjustment initially studied by Rossby and further investigated by Gill (1976) for the linear part and Kuo and Polvani (1996) for its nonlinear counterpart. In this paper, the linearized shallow-water equations, in which homogeneity is assumed in the y-direction, are solved in a domain of finite length with an initial discontinuous elevation field. The design difference with adjustment problems lies in the finiteness of the domain in the x-direction. Whereas in adjustment problems, an infinite domain in the x-direction is considered, we study the case of Poincaré waves propagation in a finite domain. In so doing, no end state is ever reached and, in the absence of friction, wave propagation goes on forever within the domain. The persistence of the discontinuities is the prominent feature of the time-dependent solution presented by Gill (1976). It also appears in the solution to our benchmark, thereby posing a challenge for classical numerical methods to solve the problem. A numerical method will be appraised based upon its ability to capture the traveling discontinuity without generating spurious oscillations. The following methods are considered in this paper: the method of characteristics, the Galerkin finite-element method (FEM) and the discontinuous Galerkin FEM with two different ways of computing the numerical fluxes.

2. A one-dimensional benchmark

The linearized governing equations for a single, inviscid, homogeneous shallow layer of fluid on an f-plane are the shallow-water equations, given by

$$\begin{aligned}
&\frac{\partial u}{\partial t} - fv = -g\frac{\partial \eta}{\partial x},\\
&\frac{\partial v}{\partial t} + fu = -g\frac{\partial \eta}{\partial y},\\
&\frac{\partial \eta}{\partial t} + h\frac{\partial u}{\partial x} + h\frac{\partial v}{\partial y} = 0,
\end{aligned} \tag{1}$$

where u and v are the vertically averaged horizontal velocity components in the x- and y-directions, respectively. The reference layer thickness is constant and denoted by h while η represents the free surface elevation. The Coriolis parameter f is taken to be constant under the f-plane approximation. Finally, g is the gravitational acceleration.

Linearization implies getting rid of advective terms and assuming that the free surface elevation be much smaller than the constant reference depth (i.e., $\eta \ll h$). The disposal of advective terms is legitimate as long as the Rossby number is much smaller than 1, in which case inertial terms are not dominant. We decide to focus on a set of linear equations, mainly for the sake of simplicity and because we will be able to interpret the results in the best way.

Within the frame of this work, we will further assume homogeneity in the y-direction so that all derivatives with respect to y vanish. The domain is thus infinite in the y-direction, which reduces the problem to a one-dimensional case. The domain remains finite in the x-direction. It should be noted that the problem we propose to solve does not consist of an adjustment problem as in Gill (1976) in which the domain is infinite—or large enough so that it can be deemed so numerically, as explained in Kuo and Polvani (1996). In that respect, we do not focus on the final state, which does not exist for finite domains. Instead, we study the wave propagation phenomenon. Reducing the system (1) to the unique x-direction yields

$$\begin{aligned}
&\frac{\partial u}{\partial t} - fv = -g\frac{\partial \eta}{\partial x},\\
&\frac{\partial v}{\partial t} + fu = 0,\\
&\frac{\partial \eta}{\partial t} + h\frac{\partial u}{\partial x} = 0,
\end{aligned} \tag{2}$$

where $x \in [-L/2, L/2]$ and $t \geqslant 0$. The boundary conditions are $u(x = \pm L/2, t) = 0$, which merely consists of boundary impermeability. We study the time evolution of an initially motionless fluid layer with a discontinuity in the elevation field. Thus, at $t = 0$

$$u(x,0) = v(x,0) = 0,$$
$$\eta(x,0) = \eta_0 \mathrm{sign}\,(x) = \begin{cases} -\eta_0 & \text{if } -L/2 \leqslant x < 0,\\ \eta_0 & \text{if } 0 < x \leqslant L/2. \end{cases}$$

Nondimensionalization of (2) is obtained by introducing the following characteristic scales: f^{-1}, L, η_0, $Lh^{-1}f\eta_0$, for the time, the space, the elevation and the velocities, respectively. Using the same symbols, the nondimensional equations become

$$\frac{\partial u}{\partial t} - v = -\alpha^2\frac{\partial \eta}{\partial x}, \tag{3}$$

$$\frac{\partial v}{\partial t} + u = 0, \tag{4}$$

$$\frac{\partial \eta}{\partial t} + \frac{\partial u}{\partial x} = 0. \tag{5}$$

We have defined $\alpha = \frac{\sqrt{gh}}{fL}$, which is the ratio of the Rossby radius of deformation to the length scale, or a nondimensional Rossby radius of deformation. Note that (3)–(5) is now defined for $t \geqslant 0$ and $x \in [-1/2, 1/2]$. Boundary and initial conditions are adapted accordingly.

2.1. Analytical solution

As a first step, we present the analytical solution to (3)–(5). Differentiation of (3) and (5) with respect to t and x, respectively, gives rise to

$$\frac{\partial^2 u}{\partial t^2} - \frac{\partial v}{\partial t} = -\alpha^2 \frac{\partial^2 \eta}{\partial t \partial x},$$
$$\frac{\partial^2 \eta}{\partial x \partial t} + \frac{\partial^2 u}{\partial x^2} = 0.$$

Elimination of the mixed derivative and substitution of $-\frac{\partial v}{\partial t}$ by u from (4) leads to a single equation for the zonal velocity u:

$$\frac{\partial^2 u}{\partial t^2} + u = \alpha^2 \frac{\partial^2 u}{\partial x^2}. \tag{6}$$

Eq. (6) can be analytically solved using the separation of variables method. This is shown in details in Appendix A. Solution to (3)–(5) is

$$\begin{aligned} u(x,t) &= \sum_{n=1}^{\infty} H_n (-1)^{n+1} \frac{\alpha^2 k_n}{\omega_n} \sin(\omega_n t)\cos(k_n x), \\ v(x,t) &= \sum_{n=1}^{\infty} H_n (-1)^{n+1} \frac{\alpha^2 k_n}{\omega_n^2} [\cos(\omega_n t) - 1]\cos(k_n x), \\ \eta(x,t) &= \sum_{n=1}^{\infty} H_n (-1)^{n} \sin(k_n x)\left\{1 - \frac{\alpha^2 k_n^2}{\omega_n^2}[1 - \cos(\omega_n t)]\right\}, \end{aligned} \tag{7}$$

where coefficients H_n amount to $H_n = \frac{4(-1)^n}{k_n}$. In Fig. 1, we show the solution (7) for the elevation at different times and compare it with Gill's analytical solution to the adjustment problem (Gill, 1976). Solutions were computed with $\alpha = \sqrt{10}/10$. Left panels of Fig. 1 show the solution within the left part of the *finite* domain ($x < 0$). Right panels show the solution within the right part of the *infinite* domain ($x > 0$). Thus, the panels separation is the axis $x = 0$. In both situations, the front moves at a speed equal to α, to the left and to the right, for the left and right panels, respectively. As long as the front does not hit the boundary of the finite domain, both solutions are the same (although antisymmetric). After reflection at the boundary, Poincaré waves evolve within the finite domain. For the adjustment problem, the front keeps moving to the right, trailing a wake of Poincaré waves behind it.

2.2. A hyperbolic problem

Because (3)–(5) is a system of first-order hyperbolic equations, there exist three real characteristics. We can write the system in compact form:

$$\mathbf{A}\frac{\partial \mathbf{u}}{\partial t} + \mathbf{B}\frac{\partial \mathbf{u}}{\partial x} = \mathbf{d},$$

where $\mathbf{A}$, $\mathbf{B}$, $\mathbf{u}$ and $\mathbf{d}$ are defined to obtain the following expression:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \frac{\partial}{\partial t} \begin{bmatrix} \eta \\ u \\ v \end{bmatrix} + \begin{bmatrix} 0 & 1 & 0 \\ \alpha^2 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \frac{\partial}{\partial x} \begin{bmatrix} \eta \\ u \\ v \end{bmatrix} = \begin{bmatrix} 0 \\ v \\ -u \end{bmatrix}.$$

In order to reduce (3)–(5) to a system of three ordinary differential equations (ODEs), we now compute the eigenvalues and eigenvectors of the generalized problem:

$$\mathbf{Z}_i^{\mathrm{T}} \cdot (\mathbf{B} - \lambda_i \mathbf{A}) = 0,$$
$$\det(\mathbf{B} - \lambda_i \mathbf{A}) = 0$$

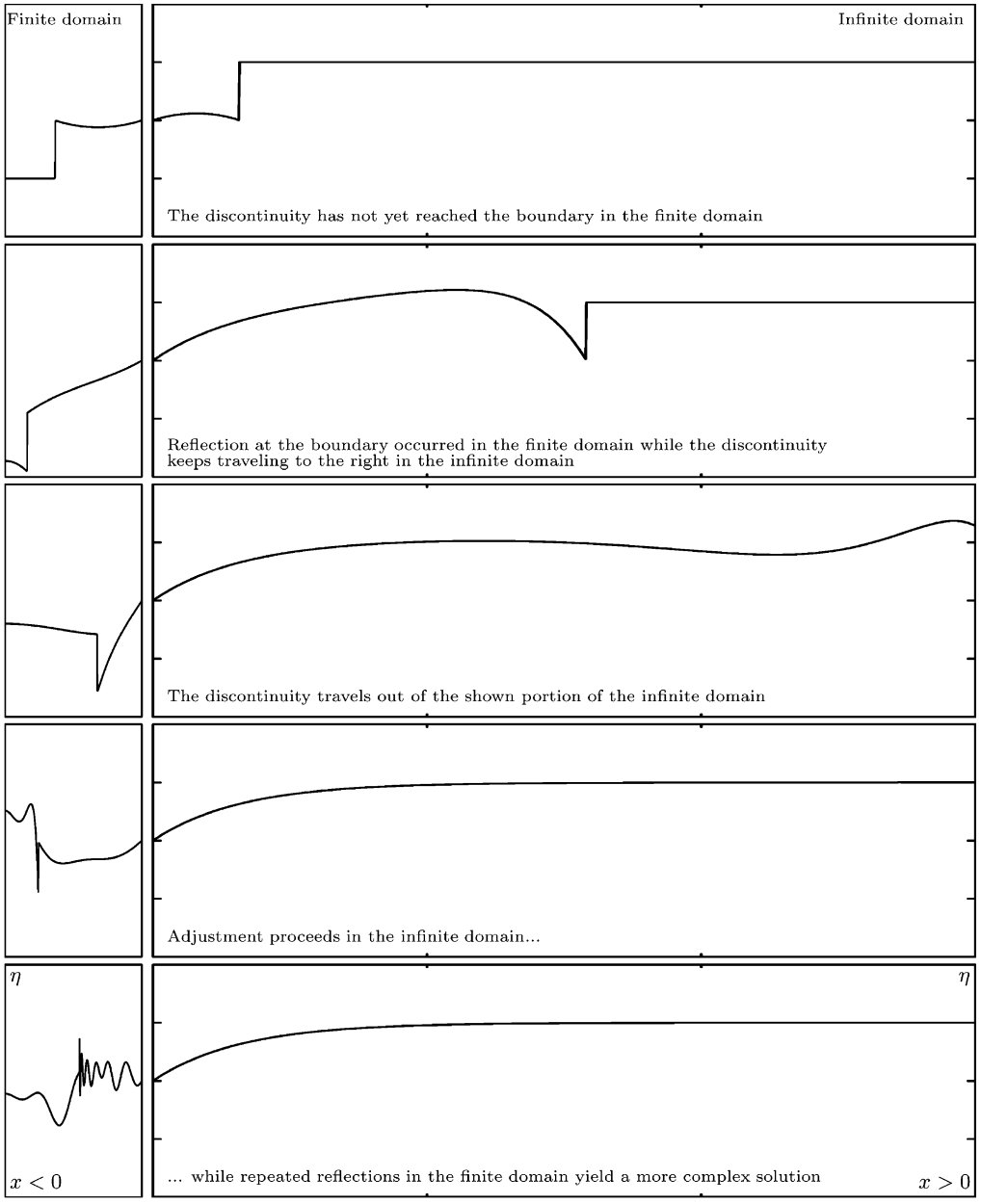

Fig. 1. Exact solution for the elevation η. Left panels show solutions for the finite domain ($x < 0$) and right panels show solutions for the adjustment problem ($x > 0$), as provided by Gill (1976). The axis $x = 0$ separates left and right panels. Left panels are 0.5-unit long and right panels are 3-unit long. The ticks on the y-axis are one unit of elevation apart, the middle one being 0. From top to bottom, solutions are shown at $t = 1$, $t = 5$, $t = 10$, $t = 100$ and $t = 1000$. The parameter α is $\sqrt{10}/10$.

for which we have

$$\lambda_1 = 0, \quad \mathbf{z}_1 = [0\ 0\ 1]^{\mathrm{T}},$$
$$\lambda_2 = \alpha, \quad \mathbf{z}_2 = [\alpha\ 1\ 0]^{\mathrm{T}},$$
$$\lambda_3 = -\alpha, \quad \mathbf{z}_3 = [\alpha\ -1\ 0]^{\mathrm{T}}.$$

For each eigenvector $\mathbf{z}_i$, an ODE is obtained by computing the following expression:

$$\mathbf{z}_i^{\mathrm{T}} \cdot \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{u} = \mathbf{z}_i^{\mathrm{T}} \cdot \mathbf{d}.$$

The system of ODEs then is

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t} v = -u, & \text{on } \dfrac{\mathrm{d}x}{\mathrm{d}t} = 0, \\ \dfrac{\mathrm{d}}{\mathrm{d}t}(\alpha\eta + u) = v, & \text{on } \dfrac{\mathrm{d}x}{\mathrm{d}t} = \alpha, \\ \dfrac{\mathrm{d}}{\mathrm{d}t}(\alpha\eta - u) = -v, & \text{on } \dfrac{\mathrm{d}x}{\mathrm{d}t} = -\alpha. \end{cases} \tag{8}$$

The foregoing procedure has allowed for transforming the system of partial differential Eqs. (3)–(5) into the system of ODEs (8) in the characteristic variables v, $\alpha\eta + u$ and $\alpha\eta - u$. Each ordinary differential equation is written on a characteristic curve $(x(t), t)$ defined by $\frac{\mathrm{d}x}{\mathrm{d}t} = \lambda_i$, where $\lambda_1 = 0$, $\lambda_2 = \alpha$ and $\lambda = -\alpha$, for the first, second and third ODE. Because the position is dependent on time, only time integration needs be performed to compute the characteristic variables, as long as we remain located on the associated characteristic curve.

3. Analysis of some numerical methods

From our standpoint, the main interest of this problem lies in its ability to be a benchmark for numerical methods. Therefore, we may compare the accuracy and robustness between several numerical techniques to solve (3)–(5). The difficulty in solving these equations lies in the presence of the discontinuity. Any numerical scheme ought to be assessed based upon its ability to capture this discontinuity without generating spurious oscillations. In this section, we present the following methods: the method of characteristics, the Galerkin finite-element method (FEM), the discontinuous Galerkin FEM and the discontinuous Riemann–Galerkin FEM. All numerical experiments were conducted with $f = 10^{-4}\,\mathrm{s}^{-1}$, $g = 10\,\mathrm{m\,s}^{-2}$, $h = 100\,\mathrm{m}$, $L = 10^6\,\mathrm{m}$, $\eta_0 = 1\,\mathrm{m}$, leading to $\alpha = \sqrt{10}/10$.

3.1. Method of characteristics

Classical finite difference schemes may now be employed to solve (8), for which we are constrained to use a time step and a spatial increment satisfying $\frac{\Delta x}{\Delta t} = \alpha$, as suggested in Fig. 2. For the sake of clarity, let us define the characteristic variables $w \doteq \alpha\eta + u$ and $q \doteq \alpha\eta - u$. A forward Euler stencil applied to (8) yields

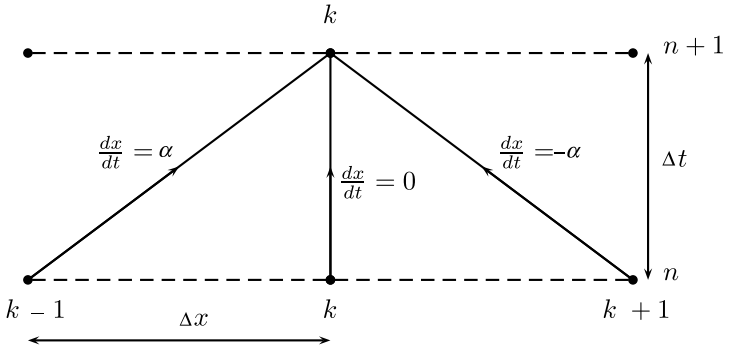

Fig. 2. Time integration must be performed along characteristics. Indices k and n identify space and time discretization points, respectively.

$$\begin{cases} \dfrac{v_k^{n+1} - v_k^n}{\Delta t} = u_k^n, \\ \dfrac{w_k^{n+1} - w_{k-1}^n}{\Delta t} = v_{k-1}^n, \\ \dfrac{q_k^{n+1} - q_{k+1}^n}{\Delta t} = -v_{k+1}^n, \end{cases} \tag{9}$$

where all information at time step n has been taken along appropriate characteristics.

The essence of the method of characteristics resides in its ability to carry the information along characteristics, which allows to focus solely on time integration. Therefore, we expect the method to be able to capture the traveling discontinuity at any time step provided that the time integration be sufficiently accurate. This issue is illustrated in Fig. 3, where the forward Euler and the second-order Runge–Kutta stencils have been used with $\Delta t = 0.01$. The solution for the elevation η is compared with the exact solution at dimensionless time $t = 200$. Notice how the approximate solution obtained with the first-order Euler scheme captures the discontinuity at the right location but is highly inaccurate overall. The second-order Runge–Kutta method performs much better, with an L^2-norm that is more than 20 times smaller. To assess the extra computational cost incurred by the use of the second-order Runge–Kutta method, a run with 400,000 time steps ($\Delta t = 0.001$) has been carried out with both methods. The forward Euler integration yields the solution after 54 s while the second-order Runge–Kutta integration does so after 83 s. Hence, there is roughly a 50% extra computational cost in using the latter method. It should be borne in mind that, however efficient the method of characteristics may be for this benchmark, a major drawback lies in the fact that such an approach cannot be straightforwardly extended to two-dimensional computations.

3.2. Continuous Galerkin

The continuous Galerkin method is the simplest of the considered methods to implement in two and three dimensions. A variational formulation can be derived by first time-discretizing (3)–(5). Each resulting equation

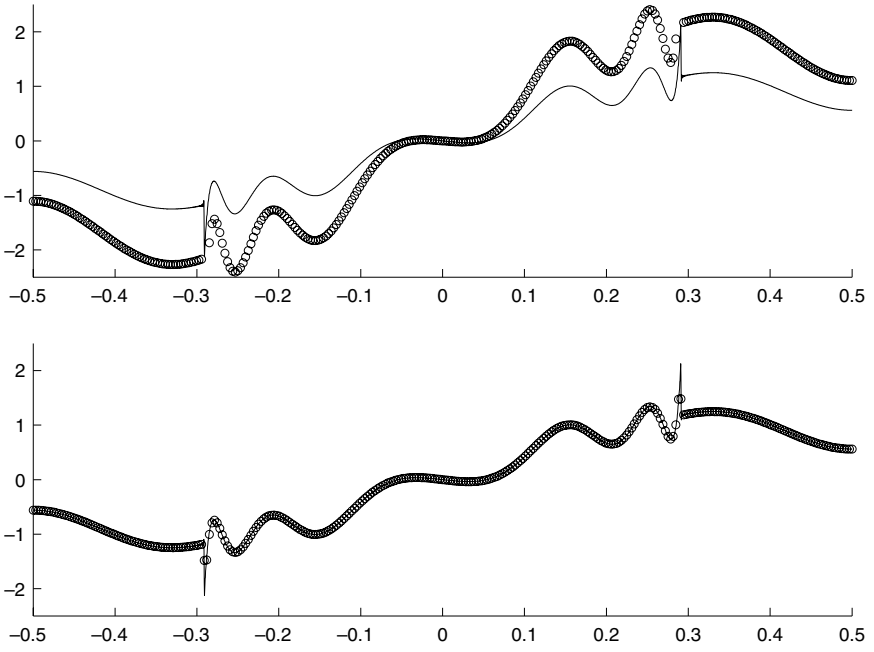

Fig. 3. Approximate and exact solutions for η at dimensionless time $t = 200$ for the first-order forward Euler method (top) and the second-order Runge–Kutta method (bottom) with a time step of $\Delta t = 0.01$. The solid line represents the exact solution. The circles represent the approximate solution at grid points.

is then mutliplied by a test function (symbolized by a hat) and integrated over the entire domain $\Omega = [-1/2, 1/2]$. If a so-called θ-scheme is employed for time discretization, the variational formulation consists in finding $\mathbf{u}^{n+1} = (u^{n+1}, v^{n+1}, \eta^{n+1}) \in \mathscr{U} = (\mathscr{U}, \mathscr{V}, \mathscr{E})$ such that

$$\begin{aligned}
&\int_\Omega \left(\frac{u^{n+1} - u^n}{\Delta t} \hat{u} - v^{n+\theta} \hat{u} + \alpha^2 \frac{\partial \eta^{n+\theta}}{\partial x} \hat{u} \right) \mathrm{d}x = 0 \quad \forall \hat{u} \in \widehat{\mathscr{U}}, \\
&\int_\Omega \left(\frac{v^{n+1} - v^n}{\Delta t} \hat{v} + u^{n+\theta} \hat{v} \right) \mathrm{d}x = 0 \quad \forall \hat{v} \in \widehat{\mathscr{V}}, \\
&\int_\Omega \left(\frac{\eta^{n+1} - \eta^n}{\Delta t} \hat{\eta} + \frac{\partial u^{n+\theta}}{\partial x} \hat{\eta} \right) \mathrm{d}x = 0 \quad \forall \hat{\eta} \in \widehat{\mathscr{E}},
\end{aligned} \tag{10}$$

where $a^{n+\theta} = \theta a^{n+1} + (1-\theta) a^n$ and θ is an adjustable parameter that allows for choosing between time schemes. The so-called Crank–Nicolson scheme is obtained with $\theta = 0.5$. Note that u^n, v^n and η^n denote the functions evaluated at the previous time step and live in the same functional spaces as the unknowns. That is to say, a finite-element problem is solved at each time step. We may also consider using the following alternative scheme that likens the classical forward–backward scheme, in which case a variational formulation consists in finding $\mathbf{u}^{n+1} \in \mathscr{U}$ such that

$$\begin{aligned}
&\int_\Omega \left(\frac{u^{n+1} - u^n}{\Delta t} \hat{u} - \frac{1}{2}(v^{n+1} + v^n) \hat{u} + \alpha^2 \frac{\partial \eta^{n+1}}{\partial x} \hat{u} \right) \mathrm{d}x = 0 \quad \forall \hat{u} \in \widehat{\mathscr{U}}, \\
&\int_\Omega \left(\frac{v^{n+1} - v^n}{\Delta t} \hat{v} + \frac{1}{2}(u^{n+1} + u^n) \hat{v} \right) \mathrm{d}x = 0 \quad \forall \hat{v} \in \widehat{\mathscr{V}}, \\
&\int_\Omega \left(\frac{\eta^{n+1} - \eta^n}{\Delta t} \hat{\eta} + \frac{\partial u^n}{\partial x} \hat{\eta} \right) \mathrm{d}x = 0 \quad \forall \hat{\eta} \in \widehat{\mathscr{E}},
\end{aligned} \tag{11}$$

where η^{n+1} is first computed from the continuity equation and used in the subsequent calculation of (u^{n+1}, v^{n+1}). The Coriolis term is treated semi-implicitely in both formulations so as to not artificially generate nor dissipate energy, which complies with the fact that no work is done by the Coriolis force. In formulations (10) and (11), $\mathbf{u}^{n+1}$ and $\hat{\mathbf{u}} = (\hat{u}, \hat{v}, \hat{\eta})$ belong to suitable infinite-dimensional function spaces. Each variable a^{n+1} is approximated as follows

$$a^{n+1} \simeq a_h^{n+1} = \sum_{j=1}^{N} A_j^{n+1} \phi_j(x),$$

where A_j^{n+1} are the nodal values and ϕ_j are the polynomial basis functions. The approximation $\mathbf{u}_h^{n+1} = (u_h^{n+1}, v_h^{n+1}, \eta_h^{n+1}) \in \mathscr{U}_h^{\mathrm{c}} = (\mathscr{U}_h^{\mathrm{c}}, \mathscr{V}_h^{\mathrm{c}}, \mathscr{E}_h^{\mathrm{c}})$, which are finite-dimensional subspaces of $(\mathscr{U}, \mathscr{V}, \mathscr{E})$. Note that the superscript c stands for *continuous*. Following the notation by Hughes et al. (2000), the test functions $\hat{\mathbf{u}}$ are similarly approximated by $\hat{\mathbf{u}}_h = (\hat{u}_h, \hat{v}_h, \hat{\eta}_h) \in \widehat{\mathscr{U}}_h^{\mathrm{c}} = (\widehat{\mathscr{U}}_h^{\mathrm{c}}, \widehat{\mathscr{V}}_h^{\mathrm{c}}, \widehat{\mathscr{E}}_h^{\mathrm{c}})$, which are finite-dimensional subspaces of $\widehat{\mathscr{U}} = (\widehat{\mathscr{U}}, \widehat{\mathscr{V}}, \widehat{\mathscr{E}})$. Linear approximations are used for the test functions and for all variables for the sake of simplicity and for an easier interpretation. Hence, $\mathbf{u}_h^{n+1}$ and $\hat{\mathbf{u}}_h$ are continuous across Ω, and piecewise linear over each element Ω_e. We bear in mind, however, that pressure modes may appear in two and three dimensions when the same interpolant order is used for the velocity and the elevation. Experiments with quadratic elements for the velocity and linear elements for the elevation, as well as linear elements for the velocity and constant elements for the elevation, have been conducted. The conclusions are the same as those presented hereafter.

In Fig. 4, we show the elevation field obtained at time $t = 2$ using the forward–backward scheme. Spurious oscillations pollute the 100-element and the 400-element approximations. Experiments with finer meshes have been carried out and no improvement is brought about by the use of smaller element sizes. Nevertheless, if we set off the time integration with a smoother initial condition, the use of smaller elements eliminates spurious oscillations. In that respect, a hyperbolic tangent profile has been chosen for the initial elevation field, that is,

$$\eta(x, 0) = \tanh(Rx), \tag{12}$$

where R, the steepness parameter, controls how steep the transition is between -1 and 1. The larger R, the closer this initial condition will be to the sign function. The foregoing experiments have been repeated with

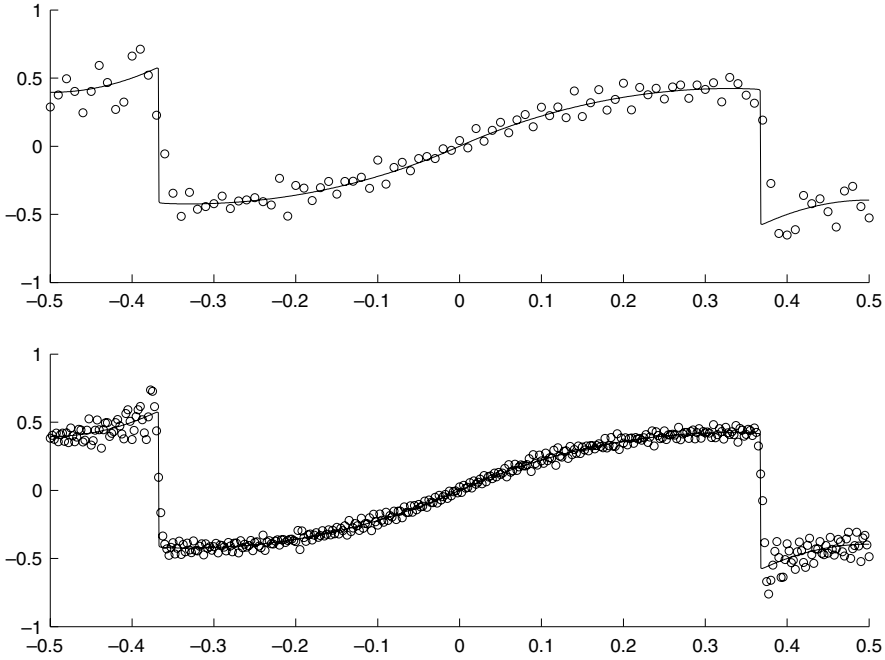

Fig. 4. The Galerkin finite-element approximations at dimensionless time $t = 2$ with 100 elements (top) and 400 elements (bottom) when the steplike initial elevation field is used. The time step is 0.001. The solid line is the exact solution.

the hyperbolic tangent initial condition (12), with a steepness parameter $R = 100$, and results are shown in Fig. 5. Note that in the case of a hyperbolic tangent initial elevation field, coefficients H_n that appear in the exact solution (7) must be numerically evaluated.

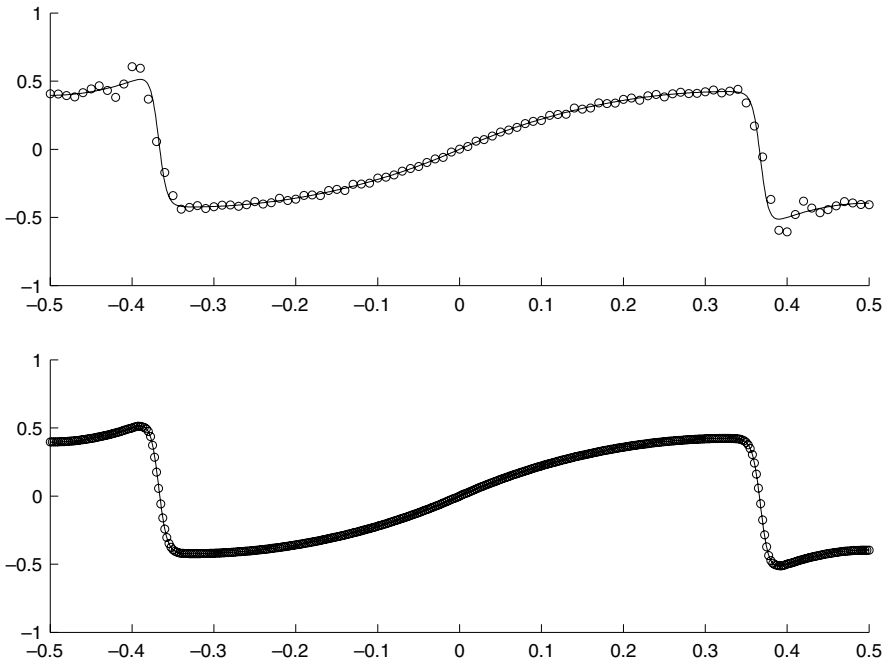

Fig. 5. The Galerkin finite-element approximations at dimensionless time $t = 2$ with 100 elements (top) and 400 elements (bottom) when a hyperbolic tangent profile is used for the initial elevation field ($R = 100$). The time step is 0.001. The solid line is the exact solution.

The assessment of the finite-element scheme is not trivial because it includes both time and space discretizations. We do not wish to go into details regarding time discretization techniques in this paper and for the convergence analysis only the forward–backward (FB) scheme has been explored. A comparison between approximate and exact solutions at dimensionless time $t = 1$ was performed on gradually refined uniform meshes. It is reported in Section 3.5.

3.3. Discontinuous Galerkin

The discontinuous Galerkin method (DGM) provides an appealing approach to address problems having discontinuities. Another advantage of the DGM is that it is inherently locally conservative while continuous Galerkin methods are locally conservative provided that subsequent postprocessing be carried out (Hughes et al., 2000). A broad review may be found in Cockburn et al. (2000). In the DGM, the solution is a piecewise-continuous function relative to a mesh (Flaherty et al., 2002). As such, it is not required that the sought solution assume the same value at each physical mesh node because two computational nodes belong to the same physical node (in a one-dimensional mesh—see Fig. 6). This property provides more flexibility in representing steep gradients and discontinuities. A steplike initial condition for the elevation field will be exactly represented, which is not the case with continuous methods.

In continuous finite-element methods, two neighboring elements share a common computational node. This common node allows information to be conveyed from one element to its neighbor. In discontinuous methods, all the nodes lie in their respective element so that, a priori, there is no transfer of information between neighboring elements. One has to keep that in mind when deriving the weak formulation. In that respect, the weak formulation (11) will be altered in such a way that neighboring elements are able to exchange information between them. As for the continuous case, a variational formulation is obtained from the time-discretized equations. For the forward–backward scheme, the problem consists in finding $\mathbf{u}^{n+1}$ in $\mathscr{U}$ such that

$$
\begin{aligned}
&\sum_{e=1}^{N_e}\int_{\Omega_e}\left(\frac{u^{n+1}-u^n}{\Delta t}\hat{u}-\frac{1}{2}(v^{n+1}+v^n)\hat{u}+\alpha^2\frac{\partial\eta^{n+1}}{\partial x}\hat{u}\right)\mathrm{d}x+\underbrace{\sum_{e=1}^{N_e}|a(\hat{u})[\alpha^2\eta^{n+1}]|_{\partial\Omega_e}}_{\mathscr{S}_1}=0\quad\forall\hat{u}\in\hat{\mathscr{U}},\\
&\sum_{e=1}^{N_e}\int_{\Omega_e}\left(\frac{v^{n+1}-v^n}{\Delta t}\hat{v}+\frac{1}{2}(u^{n+1}+u^n)\hat{v}\right)\mathrm{d}x=0\quad\forall\hat{u}\in\widehat{\mathscr{V}},\\
&\sum_{e=1}^{N_e}\int_{\Omega_e}\left(\frac{\eta^{n+1}-\eta^n}{\Delta t}\hat{\eta}+\frac{\partial u^n}{\partial x}\hat{\eta}\right)\mathrm{d}x+\underbrace{\sum_{e=1}^{N_e}|a(\hat{\eta})[u^n]|_{\partial\Omega_e}}_{\mathscr{S}_2}=0\quad\forall\hat{\eta}\in\widehat{\mathscr{E}}.
\end{aligned}
\tag{13}
$$

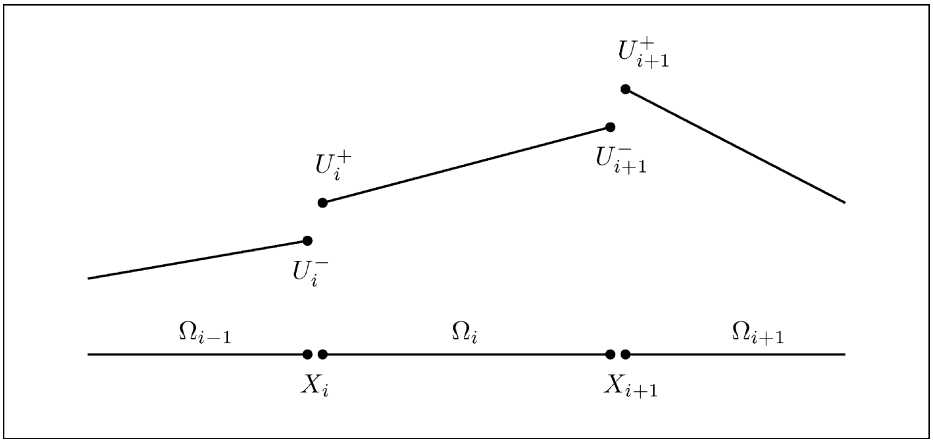

Fig. 6. One-dimensional mesh for the discontinuous Galerkin method: there are two computational nodes (i.e., two nodal values, U_i^- and U_i^+) at each physical node, X_i.

where N_e is the number of elements. An approximation $\mathbf{u}_h^{n+1} = (u_h^{n+1}, v_h^{n+1}, \eta_h^{n+1})$ is sought within $\mathscr{U}_h^{\mathrm{d}} = (\mathscr{U}_h^{\mathrm{d}}, \mathscr{V}_h^{\mathrm{d}}, \mathscr{E}_h^{\mathrm{d}})$, which are finite-dimensional subspaces of $\mathscr{U}$. The d superscript stands for *discontinuous*. Similarly, the test functions $\hat{\mathbf{u}}$ are approximated by $\hat{\mathbf{u}}_h = (\hat{u}_h, \hat{v}_h, \hat{\eta}_h) \in \widehat{\mathscr{U}}_h^{\mathrm{d}} = (\widehat{\mathscr{U}}_h^{\mathrm{d}}, \widehat{\mathscr{V}}_h^{\mathrm{d}}, \widehat{\mathscr{E}}_h^{\mathrm{d}})$, which are finite-dimensional subspaces of $\widehat{\mathscr{U}}$. As for the Galerkin method, a linear approximation is used for the test functions and all variables. However, because the discontinuous Galerkin method is employed here, the finite-dimensional subspaces $\mathscr{U}_h^{\mathrm{d}}$ and $\widehat{\mathscr{U}}_h^{\mathrm{d}}$ allow discontinuities across elements:

$$\mathscr{U}_h^{\mathrm{d}} = \widehat{\mathscr{U}}_h^{\mathrm{d}} = \{v \in L^2(\Omega) \mid v_{|\Omega_e} \in P^1(\Omega_e)\}^3,$$

where $P^1(\Omega_e)$ is the set of linear polynomials on element Ω_e. Note that the following relationships hold for finite-dimensional subspaces of the Galerkin and discontinuous Galerkin methods: $\mathscr{U}_h^{\mathrm{c}} \subset \mathscr{U}_h^{\mathrm{d}} \subset \mathscr{U}$ and $\widehat{\mathscr{U}}_h^{\mathrm{c}} \subset \widehat{\mathscr{U}}_h^{\mathrm{d}} \subset \widehat{\mathscr{U}}$. The role of $\mathscr{S}_1$ and $\mathscr{S}_2$ in the first and third equations is to weakly enforce continuity of η^{n+1} and u^{n+1}, respectively. The vertical bars indicate that expressions must be evaluated along the boundary of element Ω_e, that is at the extremities of element Ω_e for one-dimensional problems. The function $a(\hat{u})$ is defined as

$$a(\hat{u}) \doteq \left(\lambda - \frac{1}{2}\mathrm{sign}(\hat{n})\right)\hat{u},$$

where $\hat{n}$ is the outward-pointing normal at each element boundary $\partial\Omega$. The interelement jump in the nodal values at a given physical node is defined as $[u^n(X_i)] = U_i^- - U_i^+$. The parameter $\lambda \in [-1/2,1/2]$ is tunable in the sense that it allows for the interelement jump to be weighted. For example, the jump $[u^n]$ evaluated at the physical node X_i in Fig. 6 is weighted by $(\lambda - 1/2)$ on computational node i^- and by $(\lambda + 1/2)$ on computational node i^+, given that the signs of the normal $\hat{n}$ at nodes i^- and i^+, are $+1$ and -1, respectively. A centered scheme is obtained by choosing $\lambda = 0$, in which case no preference is given to any of the nodes i^- or i^+. For transport problems, it is common to give more weight to node i^+ (or node i^-) if the advective flux is known to travel from left to right (respectively from right to left). As in Hanert et al. (2004), an alternative formulation can be derived by integrating the spatial derivatives by parts. In so doing, (13) expands to

$$\begin{aligned}
&\sum_{e=1}^{N_e}\int_{\Omega_e}\left(\frac{u^{n+1}-u^n}{\Delta t}\hat{u} - \frac{1}{2}(v^{n+1}+v^n)\hat{u} - \alpha^2\eta^{n+1}\frac{\partial\hat{u}}{\partial x}\right)\mathrm{d}x \\
&\quad + \alpha^2\sum_{i=1}^{N_v}\{\langle\eta^{n+1}(X_i)\rangle[\hat{u}(X_i)] + [\eta^{n+1}(X_i)]\langle\hat{u}(X_i)\rangle\} + \alpha^2\sum_{i=1}^{N_v}[a(\hat{u}(X_i))][\eta^{n+1}(X_i)] = 0, \\
&\sum_{e=1}^{N_e}\int_{\Omega_e}\left(\frac{v^{n+1}-v^n}{\Delta t}\hat{v} + \frac{1}{2}(u^{n+1}+u^n)\hat{v}\right)\mathrm{d}x = 0, \\
&\sum_{e=1}^{N_e}\int_{\Omega_e}\left(\frac{\eta^{n+1}-\eta^n}{\Delta t}\hat{\eta} - u^n\frac{\partial\hat{\eta}}{\partial x}\right)\mathrm{d}x \\
&\quad + \sum_{i=1}^{N_v}\{\langle u^n(X_i)\rangle[\hat{\eta}(X_i)] + [u^n(X_i)]\langle\hat{\eta}(X_i)\rangle\} + \sum_{i=1}^{N_v}[a(\hat{\eta}(X_i))][u^n(X_i)] = 0,
\end{aligned} \tag{14}$$

where N_v is the number of physical nodes and $\langle f(X_i)\rangle$ denotes the average of f at X_i, that is

$$\langle f(X_i)\rangle = \frac{1}{2}(f(X_i^-) + f(X_i^+)).$$

By combining all the terms involved in the summations, the foregoing formulation reduces to

$$\begin{aligned}
&\sum_{e=1}^{N_e}\int_{\Omega_e}\left(\frac{u^{n+1}-u^n}{\Delta t}\hat{u} - \frac{1}{2}(v^{n+1}+v^n)\hat{u} - \alpha^2\eta^{n+1}\frac{\partial\hat{u}}{\partial x}\right)\mathrm{d}x + \alpha^2\sum_{i=1}^{N_v}\langle\eta^{n+1}(X_i)\rangle_\lambda[\hat{u}(X_i)] = 0, \\
&\sum_{e=1}^{N_e}\int_{\Omega_e}\left(\frac{v^{n+1}-v^n}{\Delta t}\hat{v} + \frac{1}{2}(u^{n+1}+u^n)\hat{v}\right)\mathrm{d}x = 0, \\
&\sum_{e=1}^{N_e}\int_{\Omega_e}\left(\frac{\eta^{n+1}-\eta^n}{\Delta t}\hat{\eta} - u^n\frac{\partial\hat{\eta}}{\partial x}\right)\mathrm{d}x + \sum_{i=1}^{N_v}\langle u^n(X_i)\rangle_\lambda[\hat{\eta}(X_i)] = 0,
\end{aligned} \tag{15}$$

where $\langle f(X_i)\rangle_\lambda$ is the weighted average of f at X_i, defined as

$$\langle f(X_i)\rangle_\lambda = \left(\frac{1}{2}+\lambda\right) f(X_i^-) + \left(\frac{1}{2}-\lambda\right) f(X_i^+).$$

In Appendix B, we show how formulations (14) and (15) are derived.

The discontinuous finite-element formulation (13) has been used to solve our benchmark problem with 100 and 400 elements. Results are shown in Fig. 7 where approximate and exact solutions are compared at $t = 2$. A centered scheme is employed here ($\lambda = 0$). Severe oscillations pollute the solutions. The classical forward–backward time scheme is employed for better stability properties when boundary terms $\mathscr{S}_1$ and $\mathscr{S}_2$ are involved. In Fig. 8, the top panel reproduces the 400-element solution with $\lambda = 0$ while the bottom panel shows the solution obtained with $\lambda = 0.001$. Hence, Fig. 8 permits to compare a centered and a slightly off-centered scheme. The aim of these numerical experiments is twofold. Firstly, we wish to verify whether weakly enforcing continuity on u^h and η^h ensures stability of the formulation (13). Secondly, we would like to lower the level of arbitrariness associated with the weak enforcement of continuity by appraising the sensitivity of the parameter λ. Looking at Fig. 8, we see that both choices for λ—the centered and the slightly off-centered schemes—do no prevent spurious oscillations. Moreover, the off-centered scheme makes it even worse, suggesting the importance of symmetry in the problem. Other experiments have been performed to test higher values (as well as negative values) of λ, only to further conclude that $\lambda = 0.0$ gives rise to the least severe oscillations. In Fig. 9, we show how the solution behaves when the hyperbolic tangent (12) is used as initial condition (with $R = 100$). The same experiment as with the continuous Galerkin method has been conducted here. Fig. 9 is to be compared with Fig. 5 showing the solution obtained with the continuous Galerkin method. The latter clearly outperforms the DGM. The presence of spurious oscillations for all values of λ suggests that the wrong field is upwinded. The following question thus arises: What variables should we weakly enforce the continuity of?

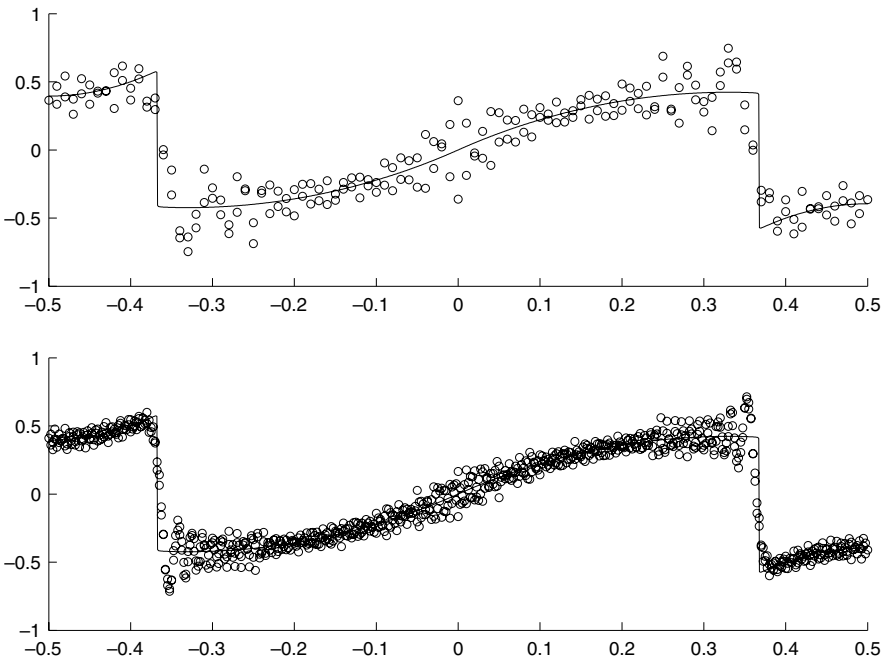

Fig. 7. Discontinuous Galerkin finite-element approximation with 100 elements (top) and 400 elements (bottom) at dimensionless time $t = 2$ with a steplike initial condition. The time step is 0.001. Continuity is weakly enforced using $\lambda = 0.0$.

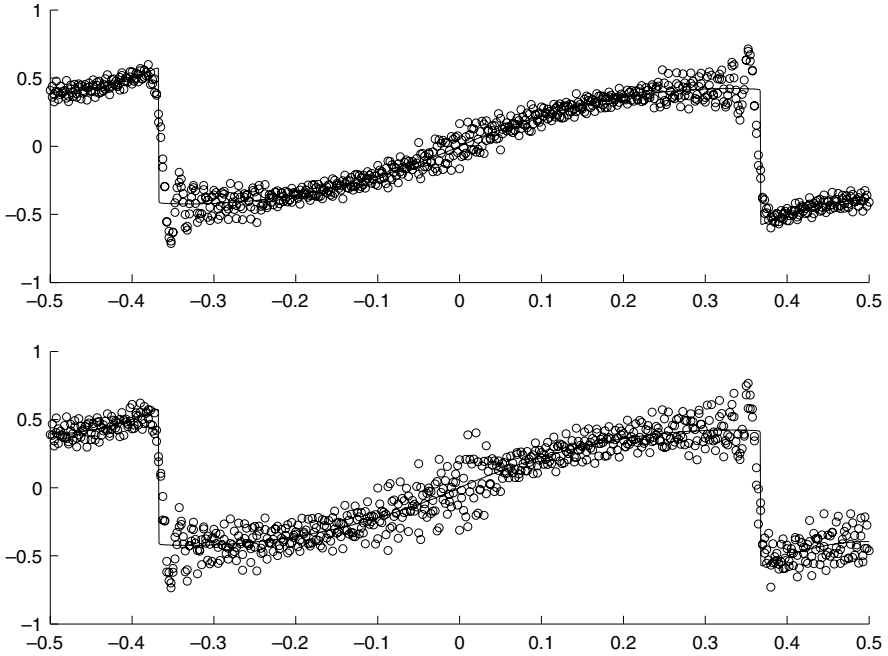

Fig. 8. Discontinuous Galerkin finite-element approximation with 400 elements at dimensionless time $t = 2$ with a steplike initial condition. The time step is 0.001. Continuity is weakly enforced using $\lambda = 0.0$ (top) and $\lambda = 0.001$ (bottom).

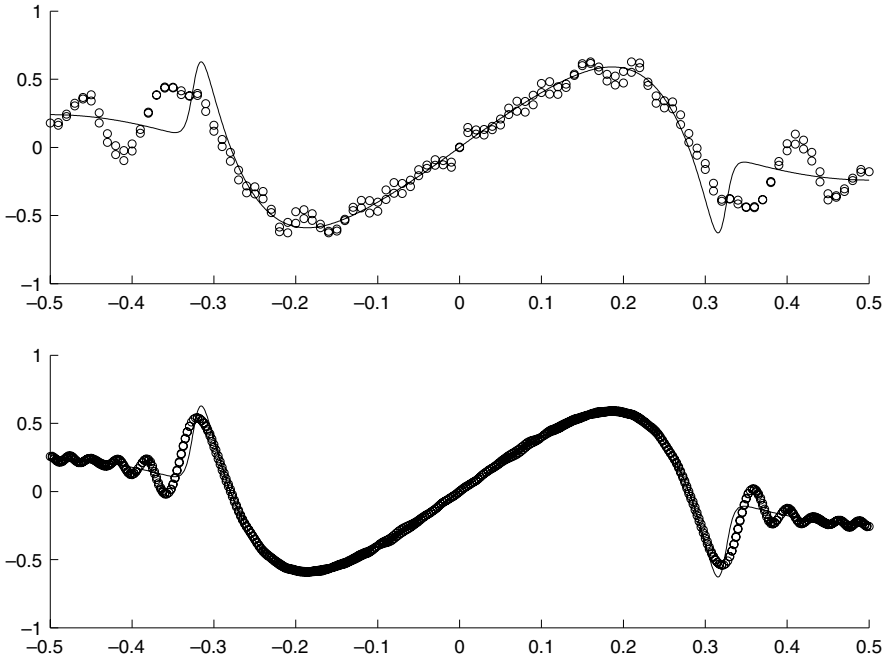

Fig. 9. The discontinuous Galerkin finite-element approximations at dimensionless time $t = 20$ with 100 elements (top) and 400 elements (bottom) when a hyperbolic tangent profile is used for the initial elevation field ($R = 100$). The time step is 0.001 and continuity is weakly enforced with $\lambda = 0$. The solid line is the exact solution.

3.4. Discontinuous Riemann–Galerkin

To answer the previous question, a closer look at the way information is propagating is advisable. Since information is carried along characteristic curves by characteristic variables, a better approach would be to enforce continuity of those very variables that transport information. In addition, we know the direction of propagation of those variables so that weighting can adequately be adapted. This approach is commonly referred to as a Riemann solver (Roe, 1981; Schwanenberger and Kongeter, 2000; Cockburn and Shu, 2001; Remacle et al., submitted for publication). A variational formulation similar to (13) may be derived. The difference will lie in the way continuity is enforced. The problem consists in finding $\mathbf{u}^h$ in $\mathscr{U}$ such that

$$
\begin{aligned}
&\sum_{e=1}^{N_e}\int_{\Omega_e}\left(\frac{u^{n+1}-u^n}{\Delta t}\hat{u}-\frac{1}{2}(v^{n+1}+v^n)\hat{u}+\alpha^2\frac{\partial\eta^{n+1}}{\partial x}\hat{u}\right)\mathrm{d}x\\
&\quad+\sum_{e=1}^{N_e}|a(\hat{u})[\alpha u^n+\alpha^2\eta^{n+1}]|_{\partial\Omega_e}+\sum_{e=1}^{N_e}|b(\hat{u})[\alpha u^n-\alpha^2\eta^{n+1}]|_{\partial\Omega_e}=0\quad\forall\hat{u}\in\widehat{\mathscr{U}},\\
&\sum_{e=1}^{N_e}\int_{\Omega_e}\left(\frac{v^{n+1}-v^n}{\Delta t}\hat{v}-\frac{1}{2}(u^{n+1}+u^n)\hat{v}\right)\mathrm{d}x=0\quad\forall\hat{v}\in\widehat{\mathscr{V}},\\
&\sum_{e=1}^{N_e}\int_{\Omega_e}\left(\frac{\eta^{n+1}-\eta^n}{\Delta t}\hat{\eta}+\frac{\partial u^n}{\partial x}\hat{\eta}\right)\mathrm{d}x\\
&\quad+\sum_{e=1}^{N_e}|a(\hat{\eta})[\alpha\eta^n+u^n]|_{\partial\Omega_e}+\sum_{e=1}^{N_e}|b(\hat{\eta})[\alpha\eta^n-u^n]|_{\partial\Omega_e}=0\quad\forall\hat{\eta}\in\widehat{\mathscr{E}},
\end{aligned}
\tag{16}
$$

where functions $a(\hat{u})$ and $b(\hat{u})$ are defined as follows:

$$
\begin{aligned}
a(\hat{u})&\doteq\frac{1}{2}\left(\frac{1}{2}-\lambda\,\mathrm{sign}\,(\hat{n})\right)\hat{u},\\
b(\hat{u})&\doteq\frac{1}{2}\left(\frac{1}{2}+\lambda\,\mathrm{sign}\,(\hat{n})\right)\hat{u},
\end{aligned}
$$

where we usually take $\lambda=1/2$. Again, an alternative formulation can be obtained by integrating the spatial derivatives by parts and combining the sums, as we have achieved for the previous DG formulation. It can be shown that (16) is equivalent to

$$
\begin{aligned}
&\sum_{e=1}^{N_e}\int_{\Omega_e}\left(\frac{u^{n+1}-u^n}{\Delta t}\hat{u}-\frac{1}{2}(v^{n+1}+v^n)\hat{u}-\alpha^2\eta^{n+1}\frac{\partial\hat{u}}{\partial x}\right)\mathrm{d}x\\
&\quad+\frac{1}{2}\alpha\sum_{i=1}^{N_v}[\hat{u}(X_i)]\{(\alpha\eta^{n+1}(X_i^-)+u^n(X_i^-))+(\alpha\eta^{n+1}(X_i^+)-u^n(X_i^+))\}\\
&\quad+(1-2\lambda)\sum_{i=1}^{N_v}[\alpha^2\eta^{n+1}(X_i)]\langle\hat{u}(X_i)\rangle=0,\\
&\sum_{e=1}^{N_e}\int_{\Omega_e}\left(\frac{v^{n+1}-v^n}{\Delta t}\hat{v}+\frac{1}{2}(u^{n+1}+u^n)\hat{v}\right)\mathrm{d}x=0,\\
&\sum_{e=1}^{N_e}\int_{\Omega_e}\left(\frac{\eta^{n+1}-\eta^n}{\Delta t}\hat{\eta}-u^n\frac{\partial\hat{\eta}}{\partial x}\right)\mathrm{d}x\\
&\quad+\frac{1}{2}\sum_{i=1}^{N_v}[\hat{\eta}(X_i)]\{(\alpha\eta^n(X_i^-)+u^n(X_i^-))-(\alpha\eta^n(X_i^+)-u^n(X_i^+))\}\\
&\quad+(1-2\lambda)\sum_{i=1}^{N_v}[u^n(X_i)]\langle\hat{\eta}(X_i)\rangle=0.
\end{aligned}
\tag{17}
$$

Setting $\lambda = 1/2$ further reduces the foregoing formulation and we obtain

$$\begin{aligned}
&\sum_{e=1}^{N_e} nt_{\Omega_e}\left(\frac{u^{n+1}-u^n}{\Delta t}\hat{u} - \frac{1}{2}(v^{n+1}+v^n)\hat{u} - \alpha^2\eta^{n+1}\frac{\partial \hat{u}}{\partial x}\right)\mathrm{d}x \\
&\quad + \frac{1}{2}\alpha\sum_{i=1}^{N_v}[\hat{u}(X_i)]\{(\alpha\eta^{n+1}(X_i^-) + u^n(X_i^-)) + (\alpha\eta^{n+1}(X_i^+) - u^n(X_i^+))\} = 0, \\
&\sum_{e=1}^{N_e}\int_{\Omega_e}\left(\frac{v^{n+1}-v^n}{\Delta t}\hat{v} + \frac{1}{2}(u^{n+1}+u^n)\hat{v}\right)\mathrm{d}x = 0, \\
&\sum_{e=1}^{N_e}\int_{\Omega_e}\left(\frac{\eta^{n+1}-\eta^n}{\Delta t}\hat{\eta} - u^n\frac{\partial \hat{\eta}}{\partial x}\right)\mathrm{d}x \\
&\quad + \frac{1}{2}\sum_{i=1}^{N_v}[\hat{\eta}(X_i)]\{(\alpha\eta^n(X_i^-) + u^n(X_i^-)) - (\alpha\eta^n(X_i^+) - u^n(X_i^+))\} = 0.
\end{aligned} \tag{18}$$

Formulation (18) is elegant. In the first equation, the summation involves an average of characteristic variables at each physical node X_i. In particular, the average is computed by taking the characteristic variables $\alpha\eta + u$ and $\alpha\eta - u$ at nodes X_i^- and X_i^+, which merely reflects the way information propagates. A similar comment can be made on the third equation where jumps of characteristic variables make up the summation.

Now, to understand the seemingly complicated formulation (16), let us evaluate the expressions that weakly enforce continuity of the characteristic variables. We focus on the first equation and assume $\hat{u} = \phi_i^-$, that is the shape function associated with computational node i^-. We further assume that the shape function is evaluated at node X_i^-. The outward-pointing normal is $+1$ so that the functions a and b take on the following expressions

$$\begin{aligned}
a(\phi_i^-) &= \frac{1}{2}(1/2 - \lambda), \\
b(\phi_i^-) &= \frac{1}{2}(1/2 + \lambda)
\end{aligned}$$

and the expression associated with node i^- is

$$\frac{1}{2}(1/2-\lambda)[\alpha u^n + \alpha^2\eta^{n+1}] + \frac{1}{2}(1/2+\lambda)[\alpha u^n - \alpha^2\eta^{n+1}].$$

If we take $\lambda = 1/2$, the latter expression simply becomes $\frac{1}{2}[\alpha u^n - \alpha^2\eta^{n+1}]$. Concretely, this is what has to be added to row i^- of the linear system. The same reasoning applied to node i^+ (i.e., shape function ϕ_i^+) gives rise to $\frac{1}{2}[\alpha u^n + \alpha^2\eta^{n+1}]$. One can see that in both expressions, a linear combination of one of the characteristic variables is involved. The jump of $\alpha(u - \alpha\eta)$ is associated with node i^- while the jump of $\alpha(u + \alpha\eta)$ is associated with node i^+. This pattern consistently translates the way information is conveyed. So as to compare with the previous discontinuous method, the same experiment has been performed (a 400-element mesh and a solution analyzed at $t = 2$) with the Riemann–Galerkin formulation (16). Results are shown in Fig. 10, where the superiority of the Riemann–Galerkin formulation is manifest when compared with Fig. 7. Let us emphasize that the quality of the approximate solution suffers from numerical dissipation when long time integration is performed, a trend already observed by Kuo and Polvani (1996) with their shock-capturing numerical methods. This effect is illustrated in Fig. 11 where the approximate solution is unable to capture higher-frequency features that make up the exact solution. Higher-order time discretization schemes should be able to tackle this problem, though, and it is indispensable to investigate the effect of such techniques on the accuracy.

3.5. Comparison between methods

Before comparing methods, it is of interest to assess the convergence rate of each of them by computing the L^2-norm of the error on gradually refined meshes. The time step used in the following experiments is very small in order for the time discretization error to be negligible in contrast to the space discretization error.

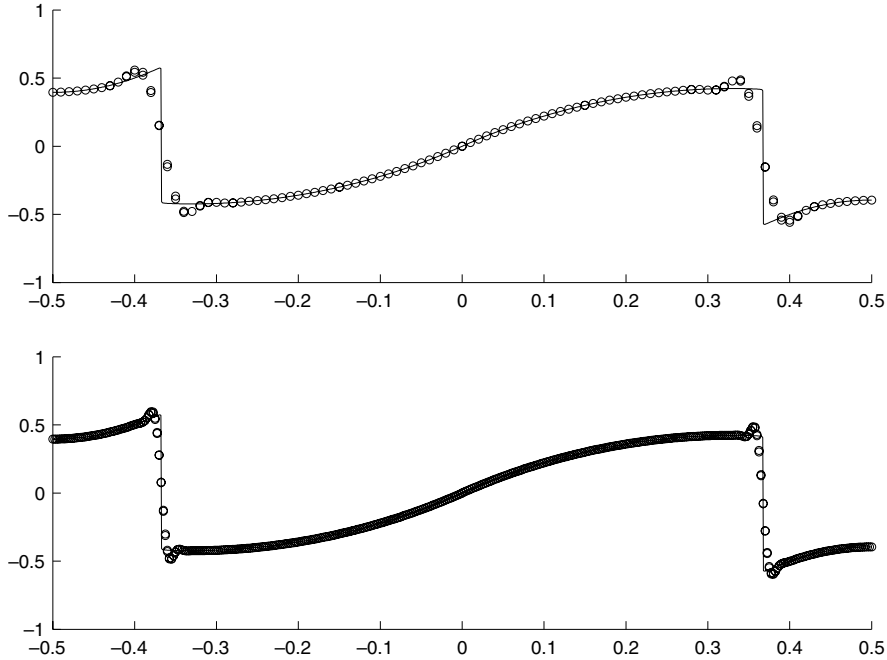

Fig. 10. Discontinuous Riemann–Galerkin finite-element approximation with 100 elements (top) and 400 elements (bottom) at dimensionless time $t = 2$ with a steplike initial condition. The time step is 0.001.

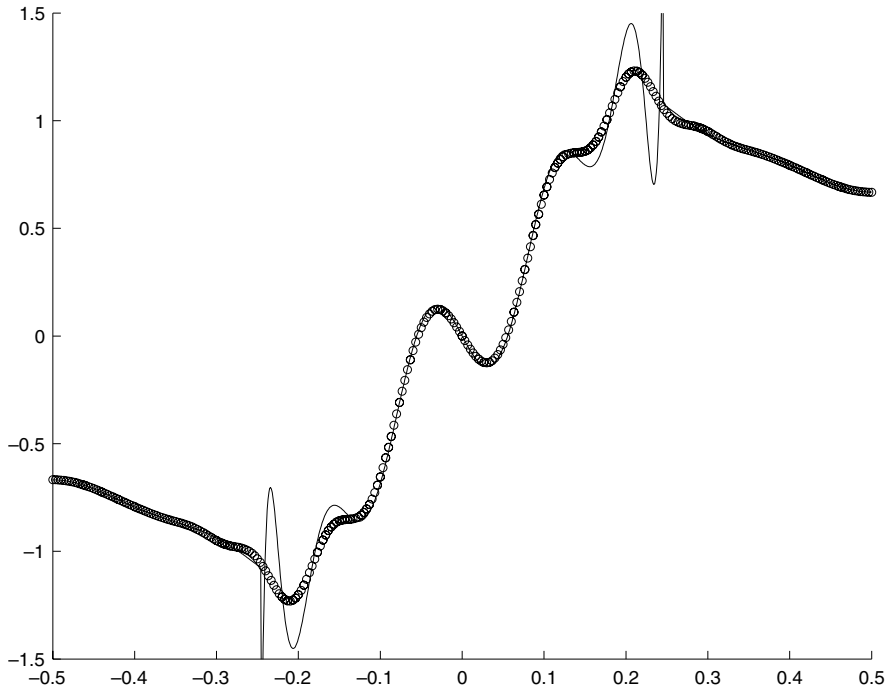

Fig. 11. Discontinuous Riemann–Galerkin finite-element approximation with 300 elements at dimensionless time $t = 200$ with a steplike initial condition. The time step is 0.002.

A time step of $\Delta t = 10^{-5}$ is used and the error at time $t = 1$ is computed. Meshes containing 25, 50, 100, 200 and 400 elements are used. The results of the convergence analysis are reported in Fig. 12. The hyperbolic

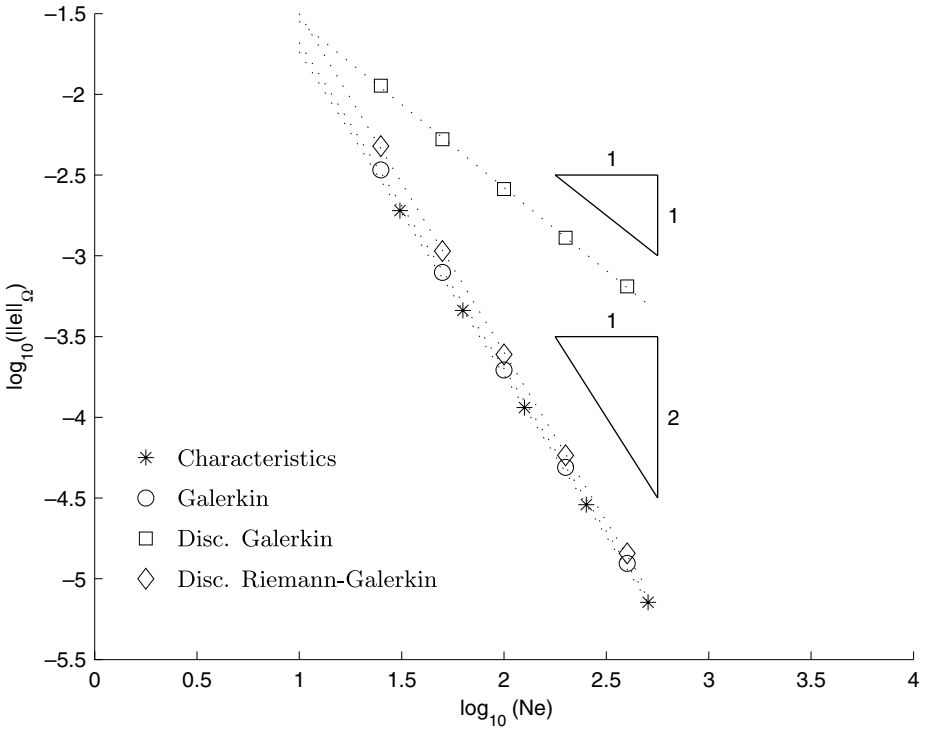

Fig. 12. L^2-norm ($\|e\|_\Omega$) of the error in the elevation η on gradually refined meshes for the three FEM with the hyperbolic tangent initial condition ($R = 10$) at $t = 1$. Notice the second-order rate of convergence obtained with the Galerkin and discontinuous Riemann–Galerkin methods while the discontinuous Galerkin method yields a first-order rate. The dotted lines represent least-square approximations to experimental errors. The error is plotted versus the number of elements.

tangent initial condition may be used to compare the three methods for different values of the steepness parameter. Results are shown on the top graph of Fig. 13 where we can observe that for smooth initial conditions, the Galerkin method performs the best while for sharp initial conditions, the discontinuous Riemann–Galerkin method yields the best approximation. It should be pointed out, though, that the errors remain close to one another and that none of the methods could be immediately ruled out based upon this quantitative analysis. Moreover, the gap between the errors obtained for sharp initial conditions does not increase when using higher-resolution meshes. The bottom graph of Fig. 13 shows the L^2-norm of the error computed on the restricted domain $\Omega_r = [-0.25, 0.25]$ that does not contain any of the discontinuities, as can be seen in Fig. 4. In so doing, the error for the discontinuous Riemann–Galerkin method remains very close to 10^{-4} while it increases up to 10^{-2} for the two other methods. This behavior is caused by the spreading of spurious oscillations toward the inner part of the domain, where the solution should remain smooth. These oscillations do not exist for the discontinuous Riemann–Galerkin method, thereby leading to an error that is two orders of magnitude smaller for sharp initial conditions. A last comment may be made regarding the use of the L^2-norm. The latter may be misleading in the sense that, by examining the top graph of Fig. 13, we are tempted to conclude that all methods are equivalent for sharp initial conditions. This is untrue and the problem is that the error is closely concentrated around the discontinuities for the discontinuous Riemann–Galerkin method (and reaches about 10^{-2}) while it remains as low as 10^{-4} away from the discontinuities. By contrast, as we can observe on the bottom graph of Fig. 13, the error reaches 10^{-2} away from the discontinuities for the Galerkin and discontinuous Galerkin methods.

In Fig. 14, the Galerkin and the discontinuous Riemann–Galerkin FEM are compared when solving the same problem with different mesh resolutions, starting at 0.1 and increasing it to 0.02 and 0.005. For the discontinuous Riemann–Galerkin method, using a coarse mesh does not produce spurious oscillations, even though high-frequency features are filtered out due to numerical dissipation. The same experiment has been

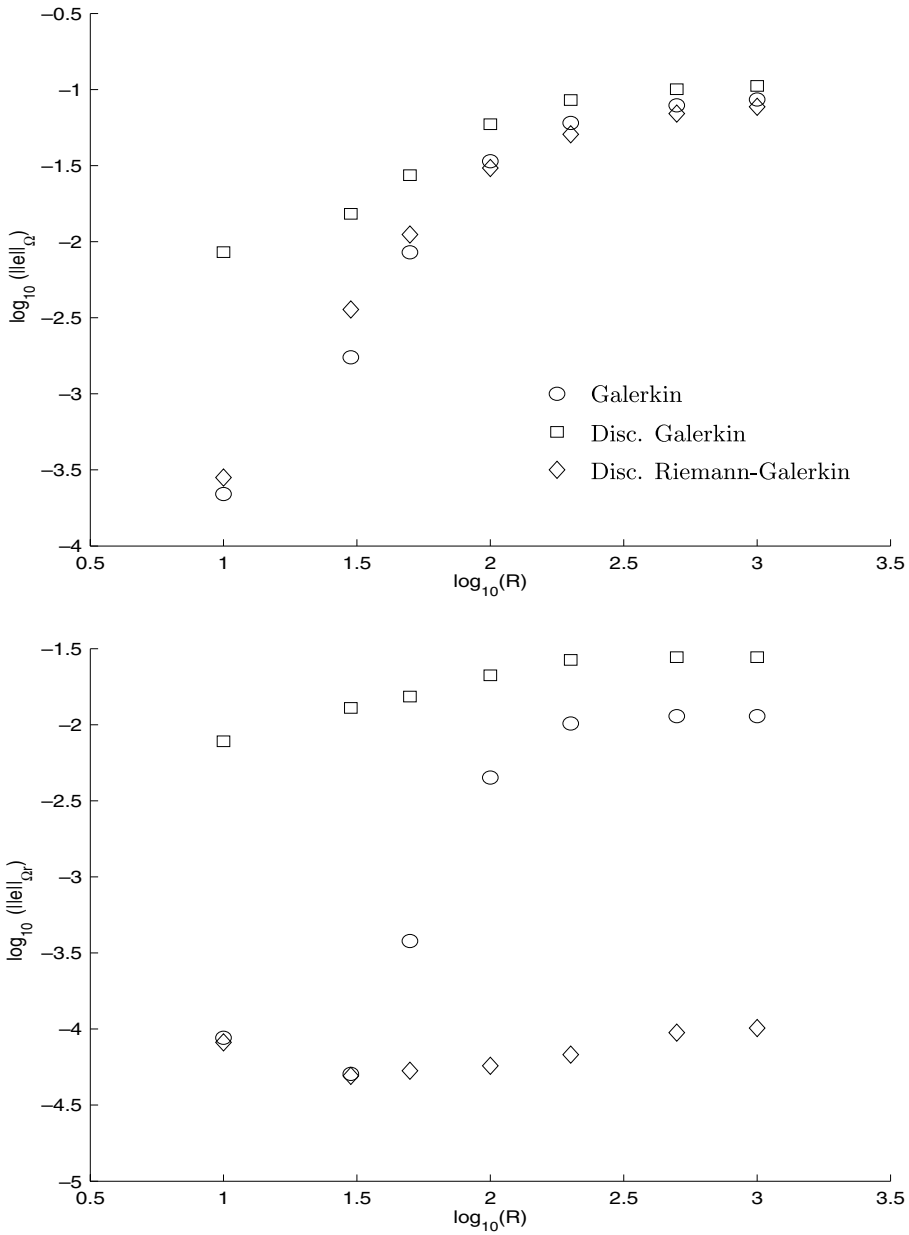

Fig. 13. The top graph shows the L^2-norm of the error in the elevation η on a mesh containing 100 uniform elements for increasing steepness parameter R with the hyperbolic tangent initial condition at $t = 2$. The bottom graph differs from the top graph in the calculation of the error: the error is computed on the restricted domain $\Omega_r = [-0.25, 0.25]$ that does not contain any of the discontinuities. The same symbols are used for both graphs. The bottom graph shows that for the Galerkin and discontinuous Galerkin methods, oscillations spread out to reach the inner region while the latter remains devoid of spurious oscillations for the discontinuous Riemann–Galerkin method.

carried out with the continuous Galerkin FEM, only to conclude that oscillations that characterize the method amplify when the resolution decreases. They do, however, remain finite. Note that no stabilization whatsoever has been used for the continuous Galerkin method so that care must be taken when comparing the latter with the Riemann–Galerkin method where characteristic variables are upwinded. As a final note, it must be stressed that such high resolutions as those previously employed are never used in large-scale ocean models. This is why the last experiment, carried out on low-resolution meshes, was presented. Namely to highlight the usability of the discontinuous Riemann–Galerkin method on low-resolution meshes. Nevertheless, it must be

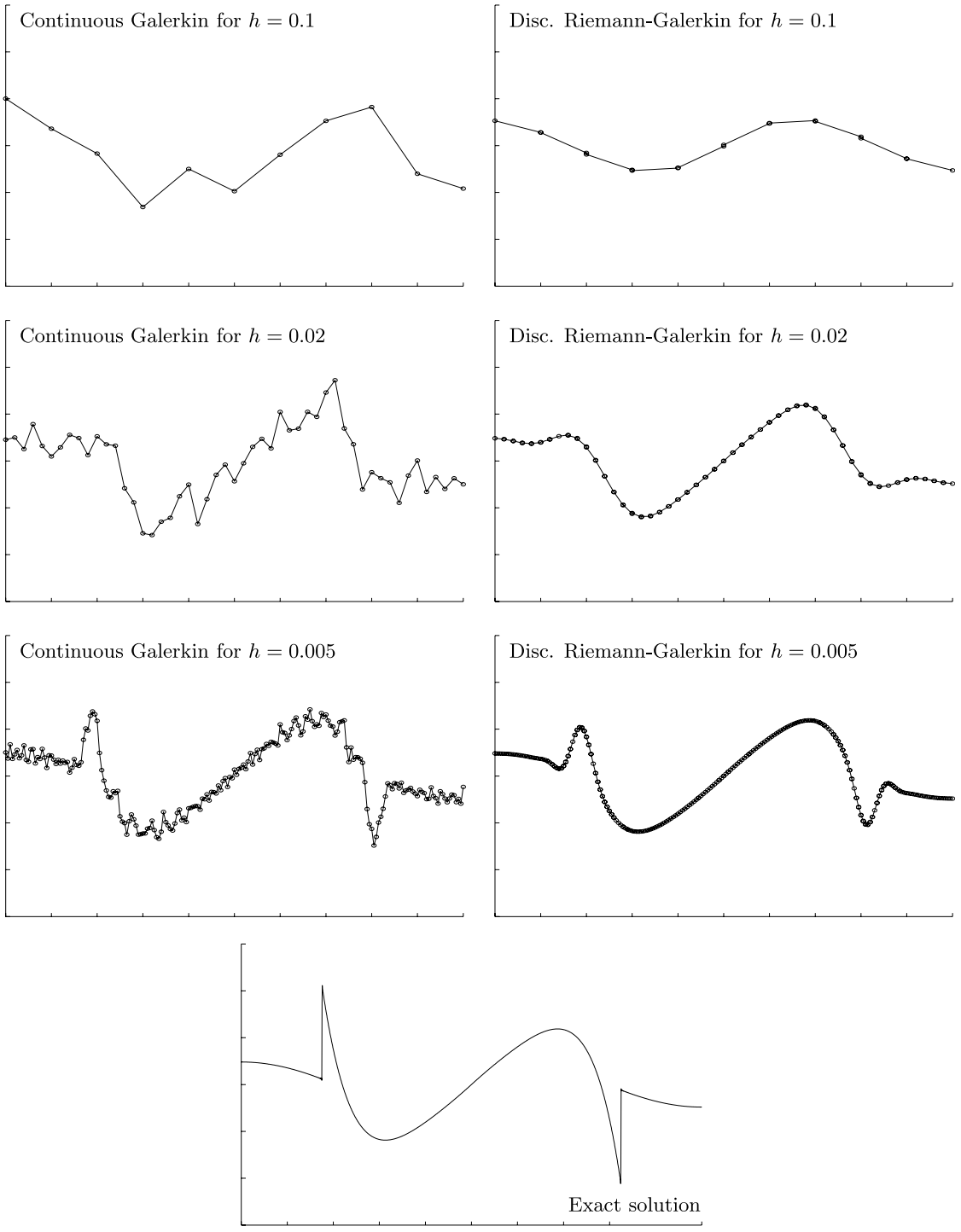

Fig. 14. Comparison of the Galerkin and the discontinuous Riemann–Galerkin FEM at time $t = 20$ for a time step of 0.001. Left and right panels are the solutions for the Galerkin and the discontinuous Riemann–Galerkin method, respectively. The first, second and third rows show results for meshes containing 10 ($h = 0.1$), 50 ($h = 0.02$) and 200 ($h = 0.005$) elements. The bottom graph is the exact solution.

stressed that the use of discontinuous methods implies increasing the number of unknowns compared with continuous methods on meshes having the same resolution.

Finally, another way of comparing the three finite-element methods is to determine the CFL condition for each of them. A von Neumann stability analysis allows to find—after quite tedious and lengthy computations—the maximum Courant number $C = \alpha\Delta t/\Delta x$ that guarantees numerical stability. For the continuous Galerkin method, we have $C \leqslant 2\sqrt{3}/3 \simeq 1.15$. For the discontinuous Galerkin method—that involves the determinant of a 6-by-6 matrix—we have $C \leqslant 0.5$. Finally, the discontinuous Riemann–Galerkin method yields the following condition: $C \leqslant 0.2564$. The latter was determined numerically while the first two were determined analytically.

4. Conclusions

A benchmark for the propagation of Poincaré waves within a one-dimensional finite domain has been proposed and a comparison between four numerical methods to resolve it has been accomplished. The use of a steplike—and thus discontinuous—initial elevation field makes it challenging for numerical techniques to capture the traveling discontinuity without spawning spurious oscillations. Because the equations describing the physics of the problem are hyperbolic, the method of characteristics is a suitable way of solving for the wave propagation. If a sufficiently accurate time scheme is employed, this technique is able to solve the benchmark very satisfyingly.

More commonly used numerical methods were then presented. In the considerations that follow, we bear in mind that the issue of time discretization must be thoroughly investigated as well. As we already said it, this was not the subject of this work. The classical continuous Galerkin FEM has difficulties capturing steep gradients, let alone discontinuities. This was revealed by the experiment carried out with the hyperbolic tangent initial elevation field. Increasing the number of elements is not really a solution by itself, for an infinite number is necessary to resolve the discontinuity. In that respect, the discontinuous Galerkin (DG) method is appealing for its ability to exactly represent discontinuities. However, this may constitute an asset as much as a drawback in the sense that one has to carefully choose the variable of which continuity is weakly enforced. That statement is illustrated by comparing the classical DG method and the so-called discontinuous Riemann–Galerkin (DRG) method. In the former, we enforce continuity of the variables whose spatial derivatives appear in the formulation. Usual DG schemes where upwind weighting is naively applied to the primitive variables (velocity and elevation) appear to poorly perform for all values of λ. It is then mandatory to impose the continuity of suitable combinations of the primitive variables. It is well known that enforcing the weak continuity of the so-called Riemann variables would perform quite better. Such an approach is known as the DG method with a Riemann solver and its numerical performances have been well documented in the literature (Roe, 1981; Schwanenberger and Kongeter, 2000; Cockburn and Shu, 2001; Flaherty et al., 2002; Remacle et al., submitted for publication). In the one-dimensional framework, this established method is presented as the DG formulation expressed in terms of Riemann variables. The main contribution of this benchmark is to show that the one-dimensional counterpart of the DGM with a Riemann solver is the optimal technique. However, as it is quite impossible to extend the method of characteristics to 2D and 3D cases, the definition of Riemann variables in higher dimensions is not obvious and the classical approach consists in considering a simplified version of the one-dimensional Riemann problem along the normal direction of each segment. Therefore, our conclusion is that the natural 2D or 3D extension of the DRG technique already exists and this paper only shows that its 1D counterpart leads to a very intuitive and physical scheme.

The continuous Galerkin and the discontinuous Galerkin methods can be both easily extended in higher dimensions without too much effort and the extensions of our results can be immediately derived. This benchmark appears to be very illustrative of the numerical behavior of wave propagation problems that model the barotropic systems of ocean models.

Acknowledgements

Laurent White and Eric Deleersnijder are Research fellow and Research associate, respectively, with the Belgian National Fund for Scientific Research (FNRS). The present study was carried out within the scope

of the project "A second-generation model of the ocean system", which is funded by the Communauté Française de Belgique, as Actions de Recherche Concertées, under contract ARC 04/09-316. This work is a contribution to the construction of SLIM, the Second-Generation Louvain-la-Neuve Ice-ocean Model (http://www.astr.ucl.ac.be/SLIM). We gratefully thank Benoît Cushman-Roisin for his insights regarding the physics of the problem.

Appendix A. Analytical solution

The solution to (6) on [0, 1], subject to an arbitrary initial condition on the elevation, say $\eta_0(x)$, is developed herein. Using the method of separation of variables, we define $u(x,t)$ to be

$$u(x,t) = F(x)T(t),$$

so that replacing u by that product into (6) yields

$$T''F + TF = \alpha^2 TF''$$

or

$$\frac{T''}{T} = \alpha^2 \frac{F''}{F} - 1 = C,$$

where C is a constant expressing the fact that both sides of the first equality must not depend upon neither x nor t. The solution to the time-dependent part, $T(t)$, must be of the form

$$T(t) = A\sin(\omega t)$$

to account for the initial condition on u. Note that the constant C is deemed negative to avoid growing exponential-type solutions in time. By twice differentiating T, the constant C is found to be: $C = -\omega^2$. The space-dependent part, $F(x)$, obeys

$$F'' = \frac{\omega^2 - 1}{\alpha^2} F,$$

where it is required that $\omega^2 > 1$ to avoid an exponential dependence on x, which could not satisfy the boundary conditions. For the same reason, solutions involving cosine cannot exist. Thus, we have

$$F(x) = B\sin(kx),$$

where $k^2 = \frac{\omega^2 - 1}{\alpha^2}$. Now, to satisfy both boundary conditions, we must have $k = k_n = (2n-1)\pi$, which constrains ω to $\omega = \omega_n = \sqrt{1 + \alpha^2 k_n^2}$. Combining the time and space dependences, the velocity $u(x,t)$ is given by an infinite sum of those harmonics:

$$u(x,t) = \sum_{n=1}^{\infty} D_n \sin(\omega_n t)\sin(k_n x), \tag{19}$$

where the constant D_n is to be determined. To do so, we may write Eq. (3) at $t = 0$:

$$\alpha^2 \frac{\partial \eta}{\partial x} = -\frac{\partial u}{\partial t} = -\sum_{n=1}^{\infty} D_n \omega_n \sin(k_n x).$$

This equality is satisfied provided that the initial elevation field $\eta_0(x)$ take the following form

$$\eta_0(x) = \sum_{n=1}^{\infty} H_n \cos(k_n x),$$

where the coefficients H_n are given by

$$H_n = 2\int_0^1 \eta_0(x)\cos(k_n x)\mathrm{d}x. \tag{20}$$

Thus, for each n, we have

$$D_n = \frac{\alpha^2 k_n}{\omega_n} H_n$$

and the final expression for $u(x,t)$ is

$$u(x,t) = \sum_{n=1}^{\infty} H_n \frac{\alpha^2 k_n}{\omega_n} \sin(\omega_n t) \sin(k_n x). \tag{21}$$

Now that $u(x,t)$ is known, we may seek the expression for $v(x,t)$ by using Eq. (4) and the initial condition $v(x,0) = 0$, which yields

$$v(x,t) = \sum_{n=1}^{\infty} H_n \frac{\alpha^2 k_n}{\omega_n} [\cos(\omega_n t) - 1] \sin(k_n t). \tag{22}$$

Finally, the elevation field $\eta(x,t)$ is easily inferred from Eq. (3). A few algebraic manipulations lead to

$$\eta(x,t) = \sum_{n=1}^{\infty} H_n \cos(k_n x) \left\{ 1 - \frac{\alpha^2 k_n^2}{\omega_n^2} [1 - \cos(\omega_n t)] \right\}. \tag{23}$$

Depending on the initial condition, an analytical expression can be found for H_n. For the sign function, coefficients H_n amount to

$$H_n = \frac{4(-1)^n}{k_n}.$$

Appendix B. Derivation of the variational formulation for the DGM

We focus on the continuity equation to show how formulations (14) and (15) are derived. Integration by parts of the term involving the spatial derivative generates an extra term, as shown hereafter:

$$\sum_{e=1}^{N_e} \int_{\Omega_e} \frac{\partial u^n}{\partial x} \hat{\eta} \, \mathrm{d}x = -\sum_{e=1}^{N_e} \int_{\Omega_e} u^n \frac{\partial \hat{\eta}}{\partial x} \, \mathrm{d}x + \sum_{e=1}^{N_e} |u^n \hat{\eta}|_{\partial\Omega_e}. \tag{24}$$

The last sum of (24) may be expanded so that the index now runs on physical nodes:

$$\sum_{e=1}^{N_e} = |u^n \hat{\eta}|_{\partial\Omega_e} \sum_{i=1}^{N_v} \{u^n(X_i^-)\hat{\eta}(X_i^-) - u^n(X_i^+)\hat{\eta}(X_i^+)\} = \sum_{i=1}^{N_v} \{\langle u^n(X_i)\rangle [\hat{\eta}(X_i)] + [u^n(X_i)]\langle \hat{\eta}(X_i)\rangle\}, \tag{25}$$

where $\langle (fX_i)\rangle$ and $[f(X_i)]$ are the average and jump of f at physical node X_i, defined as

$$\langle f(x_i)\rangle = \frac{1}{2}(f(X_i^-) + f(X_i^+)),$$
$$[f(X_i)] = f(X_i^-) - f(X_i^+).$$

The last sum of (25) is obtained from the following equality:

$$ac - bd = \frac{1}{2}(a+b)(c-d) + \frac{1}{2}(a-b)(c+d).$$

Next, the sum $\mathscr{S}_2$ in (13) may be rewritten so as to run on physical node indices as well. We have

$$\sum_{e=1}^{N_e} |a(\hat{\eta})[u^n]|_{\partial\Omega_e} = \sum_{e=1}^{N_e} a(\hat{\eta}(X_{e+1}^-))[u^n(X_{e+1})] - a(\hat{\eta}(X_e^+))[u^n(X_e)] = \sum_{i=1}^{N_v} [a(\hat{\eta}(X_i))][u^n(X_i)]. \tag{26}$$

Combining (24)–(26) yields formulation (14). Finally, we arrive at formulation (15) by putting together both sums. That is, we can write

$$\sum_{i=1}^{N_v}\langle u^n(X_i)\rangle[\hat{\eta}(X_i)] + [u^n(X_i)]\langle\hat{\eta}(X_i)\rangle + [a(\hat{\eta}(X_i))][u^n(X_i)]$$
$$= \sum_{i=1}^{N_v}\langle u^n(X_i)\rangle[\hat{\eta}(X_i)] + [u^n(X_i)]\left(\langle\hat{\eta}(X_i)\rangle + \left(\lambda - \frac{1}{2}\right)\hat{\eta}(X_i^-) - \left(\lambda + \frac{1}{2}\right)\hat{\eta}(X_i^+)\right)$$
$$= \sum_{i=1}^{N_v}\langle u^n(X_i)\rangle[\hat{\eta}(X_i)] + [u^n(X_i)]\lambda[\hat{\eta}(X_i)] = \sum_{i=1}^{N_v}[\hat{\eta}(X_i)]\langle u^n(X_i)\rangle_\lambda, \tag{27}$$

where $\langle f(X_i)\rangle_\lambda$ is a weighted average:

$$\langle f(X_i)\rangle_\lambda = \left(\frac{1}{2} + \lambda\right)f(X_i^-) + \left(\frac{1}{2} - \lambda\right)f(X_i^+).$$

References

Beckers, J.-M., Deleersnijder, E., 1993. Stability of a FBTCS scheme applied to the propagation of shallow-water inertia–gravity waves on various space grids. Journal of Computational Physics 108 (1), 95–104.

Blumberg, A.F., Mellor, G.L., 1987. A description of a three-dimensional coastal ocean circulation model. In: Heaps, N.S. (Ed.), Three-dimensional Coastal Ocean Models. American Geophysical Union, Washington, DC, pp. 1–16.

Cockburn, B., Karniadakis, G.E., Shu, C.W., 2000. Discontinuous Galerkin Methods. Theory, Computation and Applications. In: Lectures Notes in Computational Science and Engineering. Springer.

Cockburn, B., Shu, C.-W., 2001. Runge–Kutta discontinuous Galerkin methods for convection-dominated problems. Journal of Scientific Computing 16 (3), 173–261.

Deleersnijder, E., Campin, J.-M., 1995. On the computation of the barotropic mode of a free-surface world ocean model. Annales Geophysicae 13, 675–688.

Dukowicz, J.K., Smith, R.D., 1994. Implicit free-surface method for the Bryan–Cox–Semtner ocean model. Journal of Geophysical Research 99 (4), 7991–8014.

Flaherty, J.E., Krivodonova, L., Remacle, J.-F., Shephard, M.S., 2002. Aspects of discontinuous Galerkin methods for hyperbolic conservation laws. Finite Element Analysis and Design 38, 889–908.

Gill, A.E., 1976. Adjustment under gravity in a rotating channel. Journal of Fluid Mechanics 77, 603–621.

Gill, A.E., 1982. Atmosphere–Ocean Dynamics. Academic Press.

Hallberg, R., 1997. Stable split time stepping schemes for large-scale ocean modeling. Journal of Computational Physics 135, 54–65.

Hanert, E., Le Roux, D.Y., Legat, V., Deleersnijder, E., 2004. Advection schemes for unstructured grid ocean modelling. Ocean Modelling 7, 39–58.

Higdon, R.L., de Szoeke, R.A., 1997. Barotropic–baroclinic time splitting for ocean circulation modeling. Journal of Computational Physics 135, 30–53.

Higdon, R.L., 2002. A two-level time-stepping method for layered ocean circulation models. Journal of Computational Physics 177, 59–94.

Hughes, T.J.R., Engel, L., Mazzei, L., Larson, M.G., 2000. The continuous Galerkin method is locally conservative. Journal of Computational Physics 163, 467–488.

Killworth, P.D., Stainforth, D., Webb, D.J., Paterson, S.M., 1991. The development of a free-surface Bryan–Cox–Semtner ocean model. Journal of Physical Oceanography 21, 1333–1348.

Kuo, A.C., Polvani, L.M., 1996. Time-dependent fully nonlinear geostrophic adjustment. Journal of Physical Oceanography 27, 1614–1634.

Remacle, J.-F., Hillewaert, K., Chevaugeon, N., submitted for publication. Optimal numerical parameterization of discontinuous Galerkin method applied to wave propagation problems. Journal of Computational Physics.

Roe, P.L., 1981. Approximate Riemann solvers, parameter vectors and difference schemes. Journal of Computational Physics 43, 357–372.

Schwanenberger, D., Kongeter, J., 2000. A discontinuous Galerkin method for the shallow water equations with source terms. Discontinuous Galerkin Methods. Lectures Notes in Computational Science and Engineering, vol. 11, Berlin, pp. 419–424.