Available online at www.sciencedirect.com
ScienceDirect
Journal of Computational Physics 227 (2008) 1887–1922
JOURNAL OF COMPUTATIONAL PHYSICS
www.elsevier.com/locate/jcp

Discontinuous Galerkin finite element methods for hyperbolic nonconservative partial differential equations

S. Rhebergen *, O. Bokhove, J.J.W. van der Vegt

Department of Applied Mathematics, University of Twente, P.O. Box 217, 7500 AE, Enschede, The Netherlands

Received 18 January 2007; received in revised form 14 August 2007; accepted 1 October 2007
Available online 17 October 2007

Abstract

We present space- and space–time discontinuous Galerkin finite element (DGFEM) formulations for systems containing nonconservative products, such as occur in dispersed multiphase flow equations. The main criterium we pose on the weak formulation is that if the system of nonconservative partial differential equations can be transformed into conservative form, then the formulation must reduce to that for conservative systems. Standard DGFEM formulations cannot be applied to nonconservative systems of partial differential equations. We therefore introduce the theory of weak solutions for nonconservative products into the DGFEM formulation leading to the new question how to define the path connecting left and right states across a discontinuity. The effect of different paths on the numerical solution is investigated and found to be small. We also introduce a new numerical flux that is able to deal with nonconservative products. Our scheme is applied to two different systems of partial differential equations. First, we consider the shallow water equations, where topography leads to nonconservative products, in which the known, possibly discontinuous, topography is formally taken as an unknown in the system. Second, we consider a simplification of a depth-averaged two-phase flow model which contains more intrinsic nonconservative products.

MSC: 35L60; 35L65; 35L67; 65M60; 76M10

PACS: 02.60.Cb; 02.70.Dh; 47.55.−t; 47.85.Dh

Keywords: Nonconservative products; Discontinuous Galerkin finite element methods; Numerical fluxes; Arbitrary Lagrangian Eulerian (ALE) formulation; Two-phase flows

* Corresponding author.
E-mail addresses: s.rhebergen@math.utwente.nl (S. Rhebergen), o.bokhove@math.utwente.nl (O. Bokhove), j.j.w.vandervegt@math.utwente.nl (J.J.W. van der Vegt).

0021-9991/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcp.2007.10.007

1. Introduction

Systems of equations containing nonconservative products cannot be transformed into divergence form, i.e., equations of the form $\partial_t u + \partial_x f(u) + g(u)\partial_x u = 0$ cannot be written as $\partial_t u + \partial_x h(u) = 0$. This causes problems once the solution becomes discontinuous, because the weak solution in the classical sense of distributions then does not exist. Consequently, no classical Rankine–Hugoniot shock conditions can be defined. To overcome these problems we use the theory of Dal Maso, LeFloch and Murat (DLM) [5] for nonconservative products. In this theory a definition is given for nonconservative products $g(u)\partial_x u$, where $g : \mathbb{R}^m \rightarrow \mathbb{R}^m$ is a smooth function, but $u :]a, b[\rightarrow \mathbb{R}^m$ may admit discontinuities. Using this theory, a notion of a weak solution can be given to the Riemann problem for nonconservative hyperbolic partial differential equations. A problem with this theory is, however, the introduction of a path in phase space connecting the left and right state across a discontinuity. It is possible to derive an expression for this path by constructing entropy solutions to the hyperbolic equations (see [14]), but that construction can be a very difficult as well as costly job. In this article we will investigate therefore also the influence of this path in phase space and propose a new discontinuous Galerkin finite element method (DGFEM) suitable for hyperbolic partial differential equations in nonconservative form.

We are particularly interested in solving dispersed two-phase two-fluid models. The use of a DG method for these problems is of interest because it can deal efficiently with unstructured and deforming grids, local mesh refinement (h-adaptation), adjustment of the polynomial order in each element (p-refinement), and parallel computation. These benefits stem from the very compact stencil used in DG methods. Dispersed two-phase two-fluid models contain, however, nonconservative products which are introduced in the governing equations in the modeling procedure [6,7]. This poses serious problems and at present there is no literature available how to genuinely deal with nonconservative products in a DGFEM context, which motivated the research discussed in this article.

Over the years several authors have been developing numerical methods suitable for nonconservative hyperbolic partial differential equations with non-smooth solutions. Toumi [24] introduced a generalized Roe solver based on the DLM theory, which was later applied by Toumi and Kumbaro [25] to shock tube problems and two-fluid problems. The work by Toumi [24] was also used by Parés [16], Castro et al. [2] and Parés and Castro [17] to develop numerical schemes in the finite volume context. An alternative approach is followed by Saurel and Abgrall [19] in which the DLM theory is not used. They apply the criterium in multi-fluid flows, where the phases are separated by well-defined interfaces, that if pressure and velocity are uniform in both fluids, these variables must remain uniform during their temporal evolution (in the absence of surface tension). Using this criterium they construct a Godunov scheme for the conservative part of the system. The nonconservative part is then adjusted to meet the criterium above. They also use this criterium for dispersed two-phase flows, where the interfaces are not well-defined; in this case their approach therefore seems less valid. Recently, Xing and Shu [28] have published work on high order well-balanced finite volume WENO schemes and Runge–Kutta discontinuous Galerkin methods for systems containing nonconservative products. Their schemes are designed such that it maintains properties of the exact preservation of the balance laws for certain steady-state solutions. We use DLM theory to give the nonconservative products a definition even when discontinuities are present.

Here we will use the DLM theory in a DGFEM context. This work differs from the previously mentioned work in that we do not formulate a weak formulation based on generalized Roe solvers. Instead, we present and use a new numerical flux in the context of the DLM theory.

The outline of this article is as follows. We first summarize the main theory of weak solutions for partial differential equations in nonconservative form as proposed by Dal Maso et al. [5] in Section 2, but in space–time. Using this theory we derive the space–time DGFEM formulation in Section 3 and state the space DGFEM formulation as a special case in Appendix A. In DGFEM methods, the numerical flux plays an essential role. In Section 4, we derive therefore the numerical flux for systems with nonconservative products (NCP-flux) which can also be applied to moving grids. In Section 5, we apply DGFEM to two depth-averaged and dispersed multiphase systems and show numerical results using a linear path in phase space. The effect of different paths in phase space on the numerical solution is investigated in Section 6 and conclusions follow in Section 7.

2. Nonconservative hyperbolic partial differential equations

The main topic of this article is the derivation of a formulation for DGFEM suitable for nonlinear hyperbolic partial differential equations in nonconservative form and the numerical investigation of these systems. We use the DLM theory to overcome the absence of a weak solution in the classical sense of distributions for these types of equations. In an article by Dal Maso et al. [5], a definition was given for nonconservative products of the form $g(u)\partial_x u$, where $g : \mathbb{R}^m \to \mathbb{R}^m$ is a smooth function, but $u :]a, b[\to \mathbb{R}^m$ may admit discontinuities. They assumed u to be a function of bounded variation (BV), viz. a Lebesgue integrable function whose first derivative is a bounded Borel measure, and the product $g(u)\partial_x u$ is defined as a Borel measure on $]a, b[$. Such a definition is necessary when g is not the differential of a smooth function q, i.e., there is no q such that $g(u)\partial_x u$ admits a conservative form $\partial_x q$. The following example, given by LeFloch [14], illustrates the DLM theory.

Consider the function $u(x)$ composed of two constant vectors u_L and u_R in $\mathbb{R}^m$ with $u_L \neq u_R$:

$$u(x) = u_L + \mathcal{H}(x - x_d)(u_R - u_L), \quad x \in]a, b[, \tag{1}$$

where $x_d \in]a, b[$ and $\mathcal{H} : \mathbb{R} \to \mathbb{R}$ is the Heaviside function with $\mathcal{H}(x) = 0$ if $x < 0$ and $\mathcal{H}(x) = 1$ if $x > 0$. Consider any smooth function $g : \mathbb{R}^m \to \mathbb{R}^m$. We see immediately that $g(u)\partial_x u$ is not defined at $x = x_d$ since here $|\partial_x u| \to \infty$. Dal Maso et al. [5] introduce therefore a smooth regularization u^ϵ of the discontinuous function u. They show that in this particular case, if the total variation of u^ϵ remains uniformly bounded with respect to ϵ:

$$g(u)\frac{\mathrm{d}u}{\mathrm{d}x} \equiv \lim_{\epsilon \to 0} g(u^\epsilon)\frac{\mathrm{d}u^\epsilon}{\mathrm{d}x}$$

gives a sense to the nonconservative product as a bounded measure. This limit, however, depends on how we choose u^ϵ. Introduce a Lipschitz continuous path $\phi : [0, 1] \to \mathbb{R}^m$, satisfying $\phi(0) = u_L$ and $\phi(1) = u_R$, connecting u_L and u_R in $\mathbb{R}^m$. The following regularization u^ϵ for u then emerges:

$$u^\epsilon(x) = \begin{cases} u_L, & \text{if } x \in]a, x_d - \epsilon[, \\ \phi\left(\frac{x - x_d + \epsilon}{2\epsilon}\right), & \text{if } x \in]x_d - \epsilon, x_d + \epsilon[, \quad \epsilon > 0. \\ u_R, & \text{if } x \in]x_d + \epsilon, b[, \end{cases} \tag{2}$$

Using this regularization, LeFloch [14] states that when ϵ tends to zero, then

$$g(u^\epsilon)\frac{\mathrm{d}u^\epsilon}{\mathrm{d}x} \to C\delta_{x_d} \quad \text{with } C = \int_0^1 g(\phi(\tau))\frac{\mathrm{d}\phi}{\mathrm{d}\tau}(\tau)\mathrm{d}\tau,$$

vaguely in the sense of measures on $]a, b[$, where δ_{x_d} is the Dirac measure at x_d. We see that the limit of $g(u^\epsilon)\partial_x u^\epsilon$ depends on ϕ. There is one exception, namely if an $q : \mathbb{R}^m \to \mathbb{R}$ exists with $g = \partial_u q$. In this case $C = q(u_R) - q(u_L)$. We are, however, interested in the case when such a function q does not exist. We then see that the definition of the nonconservative product $g(u)\partial_x u$ must depend on the path ϕ chosen in the regularization. In Section 6, we will investigate the effect of different paths ϕ on the numerical solution. For now, assume that the path ϕ is given. In Dal Maso et al. [5] it is assumed that the path belongs to a fixed family of paths in $\mathbb{R}^m$. These paths are Lipschitz continuous maps $\phi : [0, 1] \times \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^m$ which satisfy the following properties:

(H1) $\phi(0; u_L, u_R) = u_L, \phi(1; u_L, u_R) = u_R$,
(H2) $\phi(\tau; u_L, u_L) = u_L$,
(H3) $\left|\frac{\partial\phi}{\partial\tau}(\tau; u_L, u_R)\right| \leqslant K|u_L - u_R|$, a.e. in $[0, 1]$.

Dal Maso et al. [5] consider functions $u :]a, b[\to \mathbb{R}^m$ of bounded variation, viz. $u \in \mathrm{BV}(]a, b[, \mathbb{R}^m)$. These are functions of $L^1(]a, b[, \mathbb{R}^m)$ whose first-order derivative is a bounded Borel measure on the interval $]a, b[$. Since u is BV, u admits a countable set of discontinuity points and at each such point x_d, a left trace $u_L = \lim_{\epsilon \downarrow 0} u(x_d - \epsilon)$ and a right trace $u_R = \lim_{\epsilon \downarrow 0} u(x_d + \epsilon)$ exist. For more on Borel measures, BV functions and related topics, see, e.g. [29].

Based on the family of paths satisfying (H1)–(H3), the following theorem is given by Dal Maso et al. [5]:

Theorem 1. *Let $u :]a,b[\to \mathbb{R}^m$ be a function of bounded variation and $g : \mathbb{R}^m \to \mathbb{R}^m$ be a continuous function. Then, there exists a unique real-valued bounded Borel measure μ on $]a,b[$ characterized by the two following properties*:

(1) *If u is continuous on a Borel set $B \subset]a,b[$, then*

$$\mu(B) = \int_B g(u)\frac{\mathrm{d}u}{\mathrm{d}x}\mathrm{d}\lambda,$$

where λ is the Borel measure.

(2) *If u is discontinuous at a point x_d of $]a,b[$, then*

$$\mu(\{x_d\}) = \int_0^1 g(\phi(\tau;u_L,u_R))\frac{\partial\phi}{\partial\tau}(\tau;u_L,u_R)\mathrm{d}\tau.$$

By definition, this measure μ is the nonconservative product of $g(u)$ by $\partial_x u$ and is denoted by $\mu = \left[g(u)\frac{\mathrm{d}u}{\mathrm{d}x}\right]_\phi$.

In this article we will derive a space–time DGFEM weak formulation for nonlinear hyperbolic systems of partial differential equations in nonconservative form in multi-dimensions:

$$U_{i,0} + F_{ik,k} + G_{ikr}U_{r,k} = 0, \quad \bar{x} \in \mathbb{R}^q, \ \ t > 0 \tag{3}$$

with $U \in \mathbb{R}^m$, $F \in \mathbb{R}^m \times \mathbb{R}^q$, $G \in \mathbb{R}^m \times \mathbb{R}^q \times \mathbb{R}^m$; we use the comma notation to denote partial differentiation and the summation convention on repeated indices. Here, $(\cdot)_{,0}$ denotes partial differentiation with respect to time and $(\cdot)_{,k}$ $(k = 1,\ldots,q)$ partial differentiation with respect to the spatial coordinates. In a space–time context, space and time variables are, however, not explicitly distinguished. A point at time $t = x_0$ with position $\bar{x} = (x_1,x_2,\ldots,x_q)$ has Cartesian coordinates $x = (x_0,\bar{x}) \in \mathbb{R}^{q+1}$. We can write (3) then as

$$T_{ikr}U_{r,k} = 0, \quad x \in \mathbb{R}^{q+1}, \ \ x_0 > 0, \ \ k = 0,1,2,\ldots,q \tag{4}$$

with $U \in \mathbb{R}^m$ and $T \in \mathbb{R}^m \times \mathbb{R}^{q+1} \times \mathbb{R}^m$ given by

$$T_{ikr} = \begin{cases} \delta_{ir}, & \text{if } k = 0, \\ D_{ikr}, & \text{otherwise}, \end{cases} \tag{5}$$

where δ represents the Kronecker delta symbol and where $D_{ikr} = \partial F_{ik}/\partial U_r + G_{ikr}$. Dal Maso et al. [5] give a similar theorem to Theorem 1 for the nonconservative term $T_{ikr}U_{r,k}$ in multi-dimensions. As before, assume a given family of Lipschitz continuous paths $\phi : [0,1] \times \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^m$ that satisfy, for some $K > 0$ and for all $U^L, U^R \in \mathbb{R}^m$ and $\tau \in [0,1]$, the properties:

(H1) $\phi_r(0;U^L,U^R) = U^L_r, \phi_r(1;U^L,U^R) = U^R_r$,
(H2) $\phi_r(\tau;U^L,U^L) = U^L_r$,
(H3) $\left|\frac{\partial\phi_r}{\partial\tau}(\tau;U^L,U^R)\right| \leqslant K|U^L_r - U^R_r|$, a.e. in $[0,1]$,
(H4) $\phi_r(\tau;U^L,U^R) = \phi_r(1-\tau;U^R,U^L)$.

Note that property H4 has been added, which does not have to be satisfied in the one-dimensional case. Let $\Omega \subset \mathbb{R}^{q+1}$ with $\Omega = \Omega_u \cup S_u \cup I_u$ where Ω_u is the set of points of approximate continuity, S_u the set of points of approximate jump and I_u contains the irregular points. The DLM theorem then states:

Theorem 2. *Let $U : \Omega \to \mathbb{R}^m$ be a bounded function of bounded variation defined on an open subset Ω of $\mathbb{R}^{q+1}$ and $T : \mathbb{R}^m \to \mathbb{R}^m$ be a locally bounded Borel function. Then there exists a unique family of real-valued bounded Borel measures μ_i on Ω, $i = 1,2,\ldots,m$ such that*

(1) *if B is a Borel subset of Ω_u, then*

$$\mu_i(B) = \int_B T_{ikr}U_{r,k}\,\mathrm{d}\lambda, \tag{6}$$

where λ is the Borel measure;

(2) *if B is a Borel subset of S_u, then*

$$\mu_i(B) = \int_{B\cap S_u} \int_0^1 T_{ikr}(\phi(\tau; U^L, U^R)) \frac{\partial \phi_r}{\partial \tau}(\tau; U^L, U^R) \mathrm{d}\tau \, n_k^L dH^q \tag{7}$$

with U^L and U^R the left and right traces at the discontinuity, where H^q denotes the q-dimensional Hausdorff measure and where we choose n^L the outward normal with respect to the left state;

(3) *if B is a Borel subset of I_u, then $\mu_i(B) = 0$.*

The measure μ_i is the nonconservative product of T_{ikr} by $U_{r,k}$, denoted by:

$$\mu_i = [T_{ikr} U_{r,k}]_\phi. \tag{8}$$

In particular, a piecewise $\mathcal{C}^1$ function U is a weak solution of (4) if and only if the following two conditions are satisfied [2]:

(1) U is a classical solution in the domains where it is $\mathcal{C}^1$.
(2) At a discontinuity U satisfies the generalized Rankine–Hugoniot conditions:

$$-\sigma(U_i^R - U_i^L) + F_{ik}(U^R)\bar{n}_k^L - F_{ik}(U^L)\bar{n}_k^L + \int_0^1 G_{ikr}(\phi(\tau; U^L, U^R)) \frac{\partial \phi_r}{\partial \tau}(\tau; U^L, U^R) \mathrm{d}\tau \, \bar{n}_k^L = 0, \tag{9}$$

where σ is the speed of propagation of the discontinuity, U^L and U^R are the left and right limits of the solution at the discontinuity and $\bar{n}^L$ is the space component of the space–time normal n^L (see e.g. [14]).

When $G(U)$ is the Jacobian of some flux function $Q(U)$, jump conditions (9) are independent of the path and reduce to the Rankine–Hugoniot condition:

$$H_{ik}(U^R)\bar{n}_k^L - H_{ik}(U^L)\bar{n}_k^L = \sigma(U_i^R - U_i^L), \tag{10}$$

where $H = F + Q$.

3. Space–time DGFEM discretization

In this section, we will introduce the formulation for space–time DGFEM for systems of hyperbolic partial differential equations containing nonconservative products. We will start by introducing space–time elements, function spaces, trace operators and basis functions, after which we derive the space–time DG formulation. In Appendix A, we also give the formulation for space DGFEM.

3.1. Space–time elements

In the space–time DGFEM method, the space and time variables are not distinguished. A point at time $t = x_0$ with position vector $\bar{x} = (x_1, x_2, \ldots, x_q)$ has Cartesian coordinates $(x_0, \bar{x})$ in the open domain $\mathcal{E} \subset \mathbb{R}^{q+1}$. At time t, the flow domain $\Omega(t)$ is defined as

$$\Omega(t) := \{\bar{x} \in \mathbb{R}^q : (t, \bar{x}) \in \mathcal{E}\}.$$

By taking t_0 and T as the initial and final time of the evolution of the space–time flow domain, the space–time domain boundary $\partial\mathcal{E}$ consists of the hyper-surfaces:

$$\Omega(t_0) := \{x \in \partial\mathcal{E} : x_0 = t_0\},$$
$$\Omega(T) := \{x \in \partial\mathcal{E} : x_0 = T\},$$
$$\mathcal{Q} := \{x \in \partial\mathcal{E} : t_0 < x_0 < T\}.$$

The time interval $[t_0, T]$ is partitioned using the time levels $t_0 < t_1 < \cdots < T$, where the n-th time interval is defined as $I_n = (t_n, t_{n+1})$ with length $\Delta t_n = t_{n+1} - t_n$. The space–time domain $\mathcal{E}$ is then divided into N_t space–time slabs $\mathcal{E}^n = \mathcal{E} \cap I_n$. Each space–time slab $\mathcal{E}^n$ is bounded by $\Omega(t_n)$, $\Omega(t_{n+1})$ and $\mathcal{Q}^n = \partial\mathcal{E}^n / (\Omega(t_n) \cup \Omega(t_{n+1}))$.

The flow domain $\Omega(t_n)$ is approximated by $\Omega_h(t_n)$, where $\Omega_h(t) \to \Omega(t)$ as $h \to 0$, with h the radius of the smallest sphere completely containing the largest space–time element. The domain $\Omega_h(t_n)$ is divided into N_n non-overlapping spatial elements $K_j(t_n)$. Similarly, $\Omega(t_{n+1})$ is approximated by $\Omega_h(t_{n+1})$. We can relate each element $K_j^n = K_j(t_n)$ to a master element $\widehat{K} \subset \mathbb{R}^q$ through the mapping F_K^n:

$$F_K^n : \widehat{K} \to K_j^n : \bar{\xi} \mapsto \bar{x} = \sum_i x_i(K_j^n)\chi_i(\bar{\xi})$$

with x_i the spatial coordinates of the vertices of the spatial element K_j^n and χ_i the standard Lagrangian shape functions defined on element $\widehat{K}$. The space–time elements $\mathcal{K}_j^n$ are constructed by connecting K_j^n with K_j^{n+1} using linear interpolation in time, resulting in the mapping $G_{\mathcal{K}}^n$ from the master element $\widehat{\mathcal{K}} \subset \mathbb{R}^{q+1}$ to the space–time element $\mathcal{K}^n$:

$$G_{\mathcal{K}}^n : \widehat{\mathcal{K}} \to \mathcal{K}^n : \xi \mapsto (t,\bar{x}) = \left(\frac{1}{2}(t_{n+1}+t_n) + \frac{1}{2}(t_{n+1}-t_n)\xi_0, \frac{1}{2}(1-\xi_0)F_K^n(\bar{\xi}) + \frac{1}{2}(1+\xi_0)F_K^{n+1}(\bar{\xi})\right).$$

The tessellation $\mathcal{T}_h^n$ of the space–time slab $\mathcal{E}_h^n$ consists of all space–time elements $\mathcal{K}_j^n$; thus the tessellation $\mathcal{T}_h$ of the discrete flow domain $\mathcal{E}_h := \cup_{n=0}^{N_t-1}\mathcal{E}_h^n$ then is defined as $\mathcal{T}_h := \cup_{n=0}^{N_t-1}\mathcal{T}_h^n$.

The element boundary $\partial\mathcal{K}_j^n$, which is the union of open faces of $\mathcal{K}_j^n$, consists of three parts: $K_j(t_n^+) = \lim_{\epsilon\downarrow 0} K_j(t_n+\epsilon)$, $K_j(t_{n+1}^-) = \lim_{\epsilon\downarrow 0} K_j(t_{n+1}-\epsilon)$ and $\mathcal{Q}_j^n = \partial\mathcal{K}_j^n/(K_j(t_n^+)\cup K_j(t_{n+1}^-))$. Define the grid velocity $v \in \mathbb{R}^q$ as $v = \Delta\bar{x}/\Delta t$. The outward space–time normal vector at an element boundary point on $\partial\mathcal{K}_j^n$ is given by

$$n = \begin{cases} (1,\bar{0}) & \text{at } K_j(t_{n+1}^-), \\ (-1,\bar{0}) & \text{at } K_j(t_n^+), \\ (-v_k\bar{n}_k,\bar{n}) & \text{at } \mathcal{Q}_j^n, \end{cases} \tag{11}$$

where $\bar{0} \in \mathbb{R}^q$. Note that since the space–time normal vector n has length one, the space component $\bar{n}$ of the space–time normal has a length $|\bar{n}| = 1/\sqrt{1+v\cdot v}$. It can be convenient to split the element boundaries into separate faces. In addition to the faces $K_j(t_n^+)$ and $K_j(t_{n+1}^-)$, we also define therefore interior and boundary faces. An interior face is shared by two neighboring elements $\mathcal{K}_i^n$ and $\mathcal{K}_j^n$, such that $\mathcal{S}_{ij}^n = \mathcal{Q}_i^n \cap \mathcal{Q}_j^n$, and a boundary face is defined as $\mathcal{S}_{Bj}^n = \partial\mathcal{E}^n \cap \mathcal{Q}_j^n$. The set of interior faces in time slab I^n is denoted by $\mathcal{S}_I^n$ and the set of all boundary faces by $\mathcal{S}_B^n$. The total set of faces is denoted by $\mathcal{S}_{I,B}^n = \mathcal{S}_I^n \cup \mathcal{S}_B^n$.

3.2. Function spaces and trace operators

We consider approximations of $U(x,t)$ and functions $V(x,t)$ in the finite element space V_h, which is defined as

$$V_h = \left\{ V \in (L^2(\mathcal{E}_h))^m : V|_{\mathcal{K}} \circ G_{\mathcal{K}} \in (P^p(\widehat{\mathcal{K}}))^m \ \forall \mathcal{K} \in \mathcal{T}_h \right\},$$

where $L^2(\mathcal{E}_h)$ is the space of square integrable functions on $\mathcal{E}_h$ and $P^p(\widehat{\mathcal{K}})$ denotes the space of polynomials of degree at most p on the reference element $\widehat{\mathcal{K}}$. Here m denotes the dimension of U.

We now introduce some operators as defined in Klaij et al. [12]. The trace of a function $f \in V_h$ at the element boundary $\partial\mathcal{K}^L$ is defined as

$$f^L = \lim_{\epsilon\downarrow 0} f(x - \epsilon n^L)$$

with n^L the unit outward space–time normal at $\partial\mathcal{K}^L$. When only the space components of the outward normal vector are considered we will use the notation $\bar{n}^L$. A function $f \in V_h$ has a double valued trace at element boundaries $\partial\mathcal{K}$. The traces of a function f at an internal face $\mathcal{S} = \bar{\mathcal{K}}^L \cap \bar{\mathcal{K}}^R$ are denoted by f^L and f^R. The jump of f at an internal face $\mathcal{S} \in \mathcal{S}_I^n$ in the direction k of a Cartesian coordinate system is defined as

$$[\![f]\!]_k = f^L\bar{n}_k^L + f^R\bar{n}_k^R$$

with $\bar{n}_k^R = -\bar{n}_k^L$. The average of f at $\mathcal{S} \in \mathcal{S}_I^n$ is defined as

$$\{\!\{f\}\!\} = \frac{1}{2}(f^L + f^R).$$

The jump operator satisfies the following product rule at $\mathcal{S} \in \mathcal{S}_I^n$ for $\forall g \in V_h$ and $\forall f \in V_h$, which can be proven by direct verification:

$$[\![g_i f_{ik}]\!]_k = \{\!\{g_i\}\!\}[\![f_{ik}]\!]_k + [\![g_i]\!]_k \{\!\{f_{ik}\}\!\}. \tag{12}$$

Consequently, we can relate element boundary integrals to face integrals:

$$\sum_{\mathcal{K}\in\mathcal{T}_h^n} \int_{\mathcal{Q}} g_i^L f_{ik}^L \bar{n}_k^L \,\mathrm{d}\mathcal{Q} = \sum_{\mathcal{S}\in\mathcal{S}_I^n} \int_{\mathcal{S}} [\![g_i f_{ik}]\!]_k \mathrm{d}\mathcal{S} + \sum_{\mathcal{S}\in\mathcal{S}_B^n} \int_{\mathcal{S}} g_i^L f_{ik}^L \bar{n}_k^L \,\mathrm{d}\mathcal{S}. \tag{13}$$

3.3. Basis functions

Polynomial approximations for the trial function U and the test functions V in each element $\mathcal{K} \in \mathcal{T}_h^n$ are introduced as

$$U(t,\bar{x})|_{\mathcal{K}} = \widehat{U}_m \psi_m(t,\bar{x}) \quad \text{and} \quad V(t,\bar{x})|_{\mathcal{K}} = \widehat{V}_l \psi_l(t,\bar{x}) \tag{14}$$

with ψ_m the basis functions, $\bar{x} \in \mathbb{R}^q$, and expansion coefficients $\widehat{U}_m$ and $\widehat{V}_l$, respectively, for $m, l = 0, 1, 2, \ldots, N$, where N depends on the polynomial degree of the basis functions in V_h and the space dimension q. The basis functions are defined such that the test and trial functions can be split into an element mean at time t_{n+1} and a fluctuating part. The basis functions ψ_m are given by

$$\psi_m = \begin{cases} 1 & \text{for } m = 0, \\ \varphi_m(t,\bar{x}) - \frac{1}{|K_j(t_{n+1}^-)|} \int_{K_j(t_{n+1}^-)} \varphi_m(t,\bar{x}) \mathrm{d}K & \text{for } m = 1, 2, \ldots, N, \end{cases}$$

where the functions $\varphi_m(x)$ in element $\mathcal{K}$ are related to the basis functions $\hat{\varphi}_m(\xi)$, with $\hat{\varphi}_m(\xi) \in P^p(\widehat{\mathcal{K}})$ and ξ the local coordinates in the master element $\widehat{\mathcal{K}}$, through the mapping $G_{\mathcal{K}}$:

$$\varphi_m = \hat{\varphi}_m \circ G_{\mathcal{K}}^{-1}.$$

3.4. Weak formulation

In this section we derive a space–time DGFEM weak formulation for equations containing nonconservative products. Before discussing the space–time DGFEM weak formulation for equations containing nonconservative products, we first introduce as a reference the space–time DGFEM weak formulation for equations in conservative form (see, e.g. [27]).

Consider partial differential equations in conservative form:

$$U_{i,0} + H_{ik,k} = 0, \quad \bar{x} \in \mathbb{R}^q, \ x_0 > 0, \tag{15}$$

where $U \in \mathbb{R}^m$ and $H \in \mathbb{R}^m \times \mathbb{R}^q$. Using the approach discussed in van der Vegt and van der Ven [27], the space–time DG formulation for (15) can be stated as:

Find a $U \in V_h$ such that for all $V \in V_h$:

$$\begin{aligned} 0 = &- \sum_{\mathcal{K}\in\mathcal{T}_h^n} \int_{\mathcal{K}} (V_{i,0} U_i + V_{i,k} H_{ik}) \mathrm{d}\mathcal{K} + \sum_{K\in\mathcal{T}_h^n} \left(\int_{K(t_{n+1}^-)} V_i^L U_i^L \,\mathrm{d}K - \int_{K(t_n^+)} V_i^L U_i^L \,\mathrm{d}K \right) \\ &+ \sum_{\mathcal{S}\in\mathcal{S}_I^n} \int_{\mathcal{S}} (V_i^L - V_i^R) \{\!\{H_{ik} - v_k U_i\}\!\} \bar{n}_k^L \,\mathrm{d}\mathcal{S} + \sum_{\mathcal{S}\in\mathcal{S}_B^n} \int_{\mathcal{S}} V_i^L \left(H_{ik}^L - v_k U_i^L\right) \bar{n}_k^L \,\mathrm{d}\mathcal{S}. \end{aligned} \tag{16}$$

Note that at this point no numerical fluxes have been introduced yet into the DG formulation. We continue now with equations containing nonconservative products. Let $U \in V_h$. We know that the numerical solution is continuous on an element and discontinuous across a face, so, using Theorem 2, U is a weak solution to (4) if

$$0 = \int_{\mathcal{E}_h} V_i \,\mathrm{d}\mu_i, \tag{17}$$

$$\begin{aligned} &= \sum_{\mathcal{K}\in\mathcal{T}_h} \int_{\mathcal{K}} V_i(U_{i,0} + D_{ikr}U_{r,k})\mathrm{d}\mathcal{K} + \sum_{\mathcal{K}\in\mathcal{T}_h} \left(\int_{K(t_{n+1}^-)} \widehat{V}_i \left(\int_0^1 \delta_{ir} \frac{\partial \phi_r}{\partial \tau}(\tau; U_L, U_R)\mathrm{d}\tau\, n_0^L \right) \mathrm{d}K \right. \\ &\quad \left. + \int_{K(t_n^+)} \widehat{V}_i \left(\int_0^1 \delta_{ir} \frac{\partial \phi_r}{\partial \tau}(\tau; U_L, U_R)\mathrm{d}\tau\; n_0^L \right) \mathrm{d}K \right) \\ &\quad + \sum_{\mathcal{S}\in\mathcal{S}_I} \int_{\mathcal{S}} \widehat{V}_i \left(\int_0^1 D_{ikr}(\phi(\tau; U^L, U^R)) \frac{\partial \phi_r}{\partial \tau}(\tau; U^L, U^R)\mathrm{d}\tau\, \bar{n}_k^L + \int_0^1 \frac{\partial \phi_i}{\partial \tau}(\tau; U^L, U^R)\mathrm{d}\tau\, n_0^L \right) \mathrm{d}\mathcal{S}, \end{aligned} \tag{18}$$

$$\begin{aligned} &= \sum_{\mathcal{K}\in\mathcal{T}_h} \int_{\mathcal{K}} V_i(U_{i,0} + D_{ikr}U_{r,k})\mathrm{d}\mathcal{K} \\ &\quad + \sum_{\mathcal{K}\in\mathcal{T}_h} \left(\int_{K(t_{n+1}^-)} \widehat{V}_i(U_i^R - U_i^L) n_0^L \,\mathrm{d}K + \int_{K(t_n^+)} \widehat{V}_i(U_i^R - U_i^L) n_0^L \,\mathrm{d}K \right) \\ &\quad + \sum_{\mathcal{S}\in\mathcal{S}_I} \int_{\mathcal{S}} \widehat{V}_i \left(\int_0^1 D_{ikr}(\phi(\tau; U^L, U^R)) \frac{\partial \phi_r}{\partial \tau}(\tau; U^L, U^R)\mathrm{d}\tau\, \bar{n}_k^L - v_k \delta_{ir} \int_0^1 \frac{\partial \phi_r}{\partial \tau}(\tau; U^L, U^R)\mathrm{d}\tau\, \bar{n}_k^L \right) \mathrm{d}\mathcal{S}, \end{aligned} \tag{19}$$

$$\begin{aligned} &= \sum_{\mathcal{K}\in\mathcal{T}_h} \int_{\mathcal{K}} V_i(U_{i,0} + D_{ikr}U_{r,k})\mathrm{d}\mathcal{K} \\ &\quad + \sum_{\mathcal{K}\in\mathcal{T}_h} \left(\int_{K(t_{n+1}^-)} \widehat{V}_i(U_i^R - U_i^L)\mathrm{d}K - \int_{K(t_n^+)} \widehat{V}_i(U_i^R - U_i^L)\mathrm{d}K \right) \\ &\quad + \sum_{\mathcal{S}\in\mathcal{S}_I} \int_{\mathcal{S}} \widehat{V}_i \left(\int_0^1 D_{ikr}(\phi(\tau; U^L, U^R)) \frac{\partial \phi_r}{\partial \tau}(\tau; U^L, U^R)\mathrm{d}\tau\, \bar{n}_k^L \right) \mathrm{d}\mathcal{S} + \sum_{\mathcal{S}\in\mathcal{S}_I} \int_{\mathcal{S}} \widehat{V}_i [\![v_k U_i]\!]_k \mathrm{d}\mathcal{S}, \end{aligned} \tag{20}$$

where $V \in V_h$ is an arbitrary test function. Furthermore, $\widehat{V}$ is the value (numerical flux) of the test function V on a face $\mathcal{S}$ and δ represents the Kronecker delta symbol. In (20) we used the definition of n_0^L as given in (11). The crucial point in obtaining the DG formulation is the choice of the numerical flux for the test function V. Using $D_{ikr} = \partial F_{ik}/\partial U_r + G_{ikr}$, (20) can be rewritten as

$$\begin{aligned} 0 = &\sum_{\mathcal{K}\in\mathcal{T}_h} \int_{\mathcal{K}} V_i(U_{i,0} + F_{ik,k} + G_{ikr}U_{r,k})\mathrm{d}\mathcal{K} + \sum_{\mathcal{K}\in\mathcal{T}_h} \left(\int_{K(t_{n+1}^-)} \widehat{V}_i(U_i^R - U_i^L)\mathrm{d}K - \int_{K(t_n^+)} \widehat{V}_i(U_i^R - U_i^L)\mathrm{d}K \right) \\ &+ \sum_{\mathcal{S}\in\mathcal{S}_I} \int_{\mathcal{S}} \widehat{V}_i \left(\int_0^1 G_{ikr}(\phi(\tau; U^L, U^R)) \frac{\partial \phi_r}{\partial \tau}(\tau; U^L, U^R)\mathrm{d}\tau\, \bar{n}_k^L \right) \mathrm{d}\mathcal{S} - \sum_{\mathcal{S}\in\mathcal{S}_I} \int_{\mathcal{S}} \widehat{V}_i [\![F_{ik} - v_k U_i]\!]_k \mathrm{d}\mathcal{S}. \end{aligned} \tag{21}$$

We choose the numerical flux for V such that if there exists a Q, with $G_{ikr} = \partial Q_{ik}/\partial U_r$, then the DG formulation for the system containing nonconservative products reduces to the conservative space–time DGFEM weak formulation given by (16) with $H_{ik} = F_{ik} + Q_{ik}$.

Theorem 3. *If the numerical flux $\widehat{V}$ for the test function V in (21) is defined as*

$$\widehat{V} = \begin{cases} \{\!\{V\}\!\} & \text{at } \mathcal{S} \in \mathcal{S}_I, \\ 0 & \text{at } K(t_n) \subset \Omega_h(t_n) \forall n, \end{cases} \tag{22}$$

then the DG formulation (21) will reduce to the conservative space–time DGFEM formulation (16) when there exists a Q such that $G_{ikr} = \partial Q_{ik}/\partial U_r$ so that $H_{ik} = F_{ik} + Q_{ik}$.

Proof. Assume there is a Q, such that $G_{ikr} = \partial Q_{ik}/\partial U_r$. We immediately see that:

$$\int_0^1 G_{ikr}(\phi(\tau; U^L, U^R)) \frac{\partial \phi_r}{\partial \tau}(\tau; U^L, U^R)\mathrm{d}\tau\, \bar{n}_k^L = -[\![Q_{ik}]\!]_k. \tag{23}$$

Integrating by parts the volume integral in (21) and using (23) we obtain

$$0 = -\sum_{\mathcal{K}\in\mathcal{T}_h}\int_{\mathcal{K}}(V_{i,0}U_i + V_{i,k}(F_{ik}+Q_{ik}))\mathrm{d}\mathcal{K} + \sum_{\mathcal{K}\in\mathcal{T}_h}\int_{\partial\mathcal{K}} V_i^L(U_i^L n_0^L + (F_{ik}^L + Q_{ik}^L)\bar{n}_k^L)\mathrm{d}(\partial\mathcal{K})$$
$$+\sum_{\mathcal{K}\in\mathcal{T}_h}\left(\int_{K(t_{n+1}^-)}\widehat{V}_i(U_i^R-U_i^L)\mathrm{d}K - \int_{K(t_n^+)}\widehat{V}_i(U_i^R-U_i^L)\mathrm{d}K\right) - \sum_{\mathcal{S}\in\mathcal{S}_I}\int_{\mathcal{S}}\widehat{V}_i[\![F_{ik}+Q_{ik}-v_kU_i]\!]_k\mathrm{d}\mathcal{S}. \quad (24)$$

We write $H_{ik} = F_{ik} + Q_{ik}$. Using the definition of the normal vector (11), the element boundary integral in (24) becomes

$$\sum_{\mathcal{K}\in\mathcal{T}_h}\int_{\partial\mathcal{K}} V_i^L(U_i^L n_0^L + H_{ik}^L\bar{n}_k^L)\mathrm{d}(\partial\mathcal{K}) = \sum_{\mathcal{K}\in\mathcal{T}_h}\int_{\mathcal{Q}} V_i^L(H_{ik}^L - v_kU_i^L)\bar{n}_k^L\,\mathrm{d}\mathcal{Q}$$
$$+\sum_{\mathcal{K}\in\mathcal{T}_h}\left(\int_{K(t_{n+1}^-)} V_i^LU_i^L\,\mathrm{d}K - \int_{K(t_n^+)} V_i^LU_i^L\,\mathrm{d}K\right). \quad (25)$$

We will now use relations (12) and (13) to write the element boundary integrals as face integrals:

$$\sum_{\mathcal{K}\in\mathcal{T}_h}\int_{\mathcal{Q}} V_i^L(H_{ik}^L - v_kU_i^L)\bar{n}_k^L\,\mathrm{d}\mathcal{Q} = \sum_{\mathcal{S}\in\mathcal{S}_I}\int_{\mathcal{S}}[\![V_i(H_{ik}-v_kU_i)]\!]_k\mathrm{d}\mathcal{S} + \sum_{\mathcal{S}\in\mathcal{S}_B}\int_{\mathcal{S}} V_i^L(H_{ik}^L - v_kU_i^L)\bar{n}_k^L\,\mathrm{d}\mathcal{S}$$
$$= \sum_{\mathcal{S}\in\mathcal{S}_I}\int_{\mathcal{S}}\left(\{\!\{V_i\}\!\}[\![H_{ik}-v_kU_i]\!]_k + (V_i^L - V_i^R)\{\!\{H_{ik}-v_kU_i\}\!\}\bar{n}_k^L\right)\mathrm{d}\mathcal{S}$$
$$+\sum_{\mathcal{S}\in\mathcal{S}_B}\int_{\mathcal{S}} V_i^L(H_{ik}^L - v_kU_i^L)\bar{n}_k^L\,\mathrm{d}\mathcal{S}. \quad (26)$$

Combining (24)–(26) we obtain

$$0 = -\sum_{\mathcal{K}\in\mathcal{T}_h}\int_{\mathcal{K}}(V_{i,0}U_i + V_{i,k}H_{ik})\mathrm{d}\mathcal{K} + \sum_{K\in\mathcal{T}_h}\left(\int_{K(t_{n+1}^-)} V_i^LU_i^L\,\mathrm{d}K - \int_{K(t_n^+)} V_i^LU_i^L\,\mathrm{d}K\right)$$
$$+\sum_{\mathcal{K}\in\mathcal{T}_h}\left(\int_{K(t_{n+1}^-)}\widehat{V}_i(U_i^R-U_i^L)\mathrm{d}K - \int_{K(t_n^+)}\widehat{V}_i(U_i^R-U_i^L)\mathrm{d}K\right)$$
$$+\sum_{\mathcal{S}\in\mathcal{S}_I}\int_{\mathcal{S}}\left(\{\!\{V_i\}\!\}[\![H_{ik}-v_kU_i]\!]_k + (V_i^L - V_i^R)\{\!\{H_{ik}-v_kU_i\}\!\}\bar{n}_k^L\right)\mathrm{d}\mathcal{S}$$
$$-\sum_{\mathcal{S}\in\mathcal{S}_I}\int_{\mathcal{S}}\widehat{V}_i[\![H_{ik}-v_kU_i]\!]_k\mathrm{d}\mathcal{S} + \sum_{\mathcal{S}\in\mathcal{S}_B}\int_{\mathcal{S}} V_i^L(H_{ik}^L - v_kU_i^L)\bar{n}_k^L\,\mathrm{d}\mathcal{S}. \quad (27)$$

The term $\{\!\{V_i\}\!\}[\![H_{ik}-v_kU_i]\!]_k$ is set to zero in the space–time DG formulation for conservative systems by arguing that the formulation must be conservative. For a general nonconservative system we can not use this argument. Instead, we note that by taking $\widehat{V} = \{\!\{V\}\!\}$ on the faces $\mathcal{S}\in\mathcal{S}_I$, the contribution $\int_{\mathcal{S}}\{\!\{V_i\}\!\}[\![H_{ik}-v_kU_i]\!]_k\mathrm{d}\mathcal{S}$ cancels with $-\int_{\mathcal{S}}\widehat{V}_i[\![H_{ik}-v_kU_i]\!]_k\mathrm{d}\mathcal{S}$. Furthermore, taking $\widehat{V}=0$ on the time faces $K(t_n)\subset\Omega_h(t_n)\ \forall n$, we obtain the space–time DGFEM weak formulation for conservative systems given by (16). □

Theorem 3 allows us to finalize the derivation of the DGFEM formulation for hyperbolic nonconservative partial differential equations. First, we start with the volume integral of (21) and integrate by parts, to obtain

$$0 = \sum_{\mathcal{K}\in\mathcal{T}_h}\int_{\mathcal{K}}(-V_{i,0}U_i - V_{i,k}F_{ik} + V_iG_{ikr}U_{r,k})\mathrm{d}\mathcal{K} + \sum_{\mathcal{K}\in\mathcal{T}_h}\left(\int_{K(t_{n+1}^-)} V_i^LU_i^L\,\mathrm{d}K - \int_{K(t_n^+)} V_i^LU_i^L\,\mathrm{d}K\right)$$
$$+\sum_{\mathcal{K}\in\mathcal{T}_h}\left(\int_{K(t_{n+1}^-)}\widehat{V}_i(U_i^R-U_i^L)\mathrm{d}K - \int_{K(t_n^+)}\widehat{V}_i(U_i^R-U_i^L)\mathrm{d}K\right) + \sum_{\mathcal{S}\in\mathcal{S}_I}\int_{\mathcal{S}}\left(\{\!\{V_i\}\!\}[\![F_{ik}-v_kU_i]\!]_k\right.$$
$$\left.+(V_i^L - V_i^R)\{\!\{F_{ik}-v_kU_i\}\!\}\bar{n}_k^L\right)\mathrm{d}\mathcal{S} + \sum_{\mathcal{S}\in\mathcal{S}_B}\int_{\mathcal{S}} V_i^L(F_{ik}^L - v_kU_i^L)\bar{n}_k^L\,\mathrm{d}\mathcal{S}$$
$$+\sum_{\mathcal{S}\in\mathcal{S}_I}\int_{\mathcal{S}}\widehat{V}_i\left(\int_0^1 G_{ikr}(\phi(\tau;U^L,U^R))\frac{\partial\phi_r}{\partial\tau}(\tau;U^L,U^R)\mathrm{d}\tau\,\bar{n}_k^L\right)\mathrm{d}\mathcal{S} - \sum_{\mathcal{S}\in\mathcal{S}_I}\int_{\mathcal{S}}\widehat{V}_i[\![F_{ik}-v_kU_i]\!]_k\mathrm{d}\mathcal{S}, \quad (28)$$

where we used relation (11) for the time component of the space–time normal vector and relations (12) and (13) to write the element boundary integrals as face integrals. For the numerical flux for the test function V in (28) we use (22), and thus obtain

$$\begin{aligned}0 = &\sum_{\mathcal{K}\in\mathcal{T}_h}\int_{\mathcal{K}}(-V_{i,0}U_i - V_{i,k}F_{ik} + V_i G_{ikr}U_{r,k})\mathrm{d}\mathcal{K} + \sum_{\mathcal{K}\in\mathcal{T}_h}\left(\int_{K(t_{n+1}^-)} V_i^L U_i^L\,\mathrm{d}K - \int_{K(t_n^+)} V_i^L U_i^L\,\mathrm{d}K\right)\\ &+\sum_{\mathcal{S}\in\mathcal{S}_I}\int_{\mathcal{S}}(V_i^L - V_i^R)\left\{\!\!\left\{F_{ik} - v_k U_i\right\}\!\!\right\}\bar{n}_k^L\,\mathrm{d}\mathcal{S} + \sum_{\mathcal{S}\in\mathcal{S}_B}\int_{\mathcal{S}} V_i^L(F_{ik}^L - v_k U_i^L)\bar{n}_k^L\,\mathrm{d}\mathcal{S}\\ &+\sum_{\mathcal{S}\in\mathcal{S}_I}\int_{\mathcal{S}}\left\{\!\!\left\{V_i\right\}\!\!\right\}\left(\int_0^1 G_{ikr}(\phi(\tau;U^L,U^R))\frac{\partial\phi_r}{\partial\tau}(\tau;U^L,U^R)\mathrm{d}\tau\,\bar{n}_k^L\right)\mathrm{d}\mathcal{S}.\end{aligned} \tag{29}$$

Theorem 3 states that the weak formulation given by (29) can be reduced to the space–time DGFEM formulation (16), when a Q exists such that $G_{ikr} = \partial Q_{ik}/\partial U_r$. However, this formulation is generally numerically unstable. Problematic in the conservative space–time DGFEM formulation are the interior $(V_i^L - V_i^R)\left\{\!\!\left\{ H_{ik} - v_k U_i\right\}\!\!\right\}\bar{n}_k^L$ and boundary $V_i^L(H_{ik}^L - v_k U_i^L)\bar{n}_k^L$ flux terms, see (16). Generally, a stabilizing term is added to these flux terms, together forming an upwind numerical flux. Furthermore, the following upwind flux is introduced in the conservative space–time DGFEM formulation at the time faces, a formulation naturally ensuring causality in time:

$$\widehat{U} = \begin{cases} U^L & \text{at } K(t_{n+1}^-),\\ U^R & \text{at } K(t_n^+).\end{cases} \tag{30}$$

It replaces the traces of U taken from the interior of $\mathcal{K}\in\mathcal{T}_h^n$. In (29), we also introduce the upwind flux (30) at the time faces. We also need a stabilizing term in (29). To understand how we add our stabilizing term, consider again the conservative space–time formulation. As mentioned above, a stabilizing term is added to $\left\{\!\!\left\{H_{ik} - v_k U_i\right\}\!\!\right\}$. Denote this stabilizing term as H^{stab}, then $(\left\{\!\!\left\{H_{ik} - v_k U_i\right\}\!\!\right\} + H_{ik}^{\mathrm{stab}})\bar{n}_k^L = \widehat{H}_i$, where $\widehat{H}_i$ is the space–time numerical flux. In the nonconservative space–time formulation (29) we add a stabilizing term to the conservative part $\left\{\!\!\left\{F_{ik} - v_k U_i\right\}\!\!\right\}$, but we also need to add a stabilizing part due to the nonconservative product. For the nonconservative product there is no counterpart for $\left\{\!\!\left\{F_{ik} - v_k U_i\right\}\!\!\right\}$. This term is hidden in the volume integral and in the last term of (29). We add the stabilizing term for the nonconservative product P_{ik}^{nc} to the stabilizing term for the conservative product P_{ik}^{c}: $(\left\{\!\!\left\{F_{ik} - v_k U_i\right\}\!\!\right\} + P_{ik}^{c} + P_{ik}^{nc})\bar{n}_k^L = \widehat{P}_i^{nc}$. By introducing a ghost value U^R at the boundary, we can use the same expressions also at a boundary face. An expression for $\widehat{P}_i^{nc}(U^L,U^R,v,\bar{n}^L)$ is derived in Section 4, such that it reduces to the numerical flux in the conservative case, $\widehat{H}_i$. Finally, the space–time DGFEM weak formulation for partial differential equations containing nonconservative products (3) is:

Find a $U\in V_h$ such that for all $V\in V_h$:

$$\begin{aligned}0 = &\sum_{\mathcal{K}\in\mathcal{T}_h^n}\int_{\mathcal{K}}(-V_{i,0}U_i - V_{i,k}F_{ik} + V_i G_{ikr}U_{r,k})\mathrm{d}\mathcal{K} + \sum_{\mathcal{K}\in\mathcal{T}_h^n}\left(\int_{K(t_{n+1}^-)} V_i^L U_i^L\,\mathrm{d}K - \int_{K(t_n^+)} V_i^L U_i^R\,\mathrm{d}K\right)\\ &+\sum_{\mathcal{S}\in\mathcal{S}^n}\int_{\mathcal{S}}(V_i^L - V_i^R)\widehat{P}_i^{nc}\,\mathrm{d}\mathcal{S} + \sum_{\mathcal{S}\in\mathcal{S}^n}\int_{\mathcal{S}}\left\{\!\!\left\{V_i\right\}\!\!\right\}\left(\int_0^1 G_{ikr}(\phi(\tau;U^L,U^R))\frac{\partial\phi_r}{\partial\tau}(\tau;U^L,U^R)\mathrm{d}\tau\,\bar{n}_k^L\right)\mathrm{d}\mathcal{S}.\end{aligned} \tag{31}$$

Note that due to the introduction of the upwind flux at the time faces, each space–time slab only depends on the previous space–time slab so that the summation over all space–time slabs could be dropped.

3.5. Slope limiters

In our space- and space–time DGFEM computations, when the solution may admit discontinuities, we use a slope limiter to deal with overshoots and undershoots. In this article we use a simple minmod function (see e.g. [4]). Let $\overline{U}_k$ represent the mean of U on element $\mathcal{K}_k$ and let $\widehat{U}_k$ represent the slope, then the solution in an element is given by

$$U_k = \overline{U}_k + \psi(x) m(\widehat{U}_k, \overline{U}_{k+1} - \overline{U}_k, \overline{U}_k - \overline{U}_{k-1}),$$

where the minmod function m is defined as

$$m(a_1, a_2, a_3) = \begin{cases} s \min\limits_{1 \leqslant n \leqslant 3} |a_n|, & \text{if } s = \text{sign}(a_1) = \text{sign}(a_2) = \text{sign}(a_3), \\ 0, & \text{otherwise.} \end{cases}$$

3.6. Pseudo-time stepping

By replacing U and V in the weak formulation (31) by their polynomial expansions (14), a system of algebraic equations for the expansion coefficients of U is obtained. For each physical time step, the system can be written as

$$\mathcal{L}(\widehat{U}^n; \widehat{U}^{n-1}) = 0. \tag{32}$$

This system of coupled nonlinear equations is solved by adding a pseudo-time derivative:

$$|K^n| \frac{\partial \widehat{U}^n}{\partial \tau} = -\frac{1}{\Delta t} \mathcal{L}(\widehat{U}^n; \widehat{U}^{n-1}), \tag{33}$$

which is integrated to steady-state in pseudo-time. Following van der Vegt and van der Ven [27] and Klaij et al. [11], we use the explicit Runge–Kutta method for inviscid flow with Melson correction which is given by

Algorithm 1. Five-stage explicit Runge–Kutta scheme:

(1) Initialize $\widehat{V}^0 = \widehat{U}$.
(2) For all stages $s = 1$ to 5:
$$(I + \alpha_s \lambda I)\widehat{V}^s = \widehat{V}^0 + \alpha_s \lambda (\widehat{V}^{s-1} - \mathcal{L}(\widehat{V}^{s-1}, \widehat{U}^{n-1})).$$
(3) Update $\widehat{V} = \widehat{V}^5$.

The coefficient λ is defined as $\lambda = \Delta\tau/\Delta t$, with $\Delta\tau$ the pseudo-time step and Δt the physical time step. The Runge–Kutta coefficients α_s are defined as: $\alpha_1 = 0.0797151$, $\alpha_2 = 0.163551$, $\alpha_3 = 0.283663$, $\alpha_4 = 0.5$ and $\alpha_5 = 1.0$.

4. The NCP numerical flux

In Section 3, we derived a weak formulation for space–time DGFEM for systems of equations containing a nonconservative product. To obtain an expression for the flux $\widehat{P}_i^{nc}(U^L, U^R, v, \bar{n}^L)$ in (31), we first discuss the numerical flux $\widehat{U}$, and then derive the numerical flux for NonConservative Products, or NCP-flux.

Consider the following nonconservative hyperbolic system:

$$\partial_t U + \partial_x F(U) + G(U)\partial_x U = 0, \tag{34}$$

where $U \in \mathbb{R}^m$, with m the number of components of U, similarly $F(U) \in \mathbb{R}^m$, $G(U) \in \mathbb{R}^{m \times m}$ and $x \in \mathbb{R}$ is along the normal of the face. To approximate the Riemann solution of (34) we consider only the fastest left and right moving waves of the system with velocities S_L and S_R and the grid velocity. In the star region (see Fig. 1), which is the domain enclosed by the waves S_L and S_R, the averaged exact solution $\overline{U}^*$ is defined as

$$\overline{U}^* = \frac{1}{T(S_R - S_L)} \int_{TS_L}^{TS_R} U(x, T) \mathrm{d}x. \tag{35}$$

In what follows we obtain a relation for $\overline{U}^*$ from the weak formulation of (34). Using Gauss' theorem we obtain over the control volume $\Omega_1 \cup \Omega_2$ the relation:

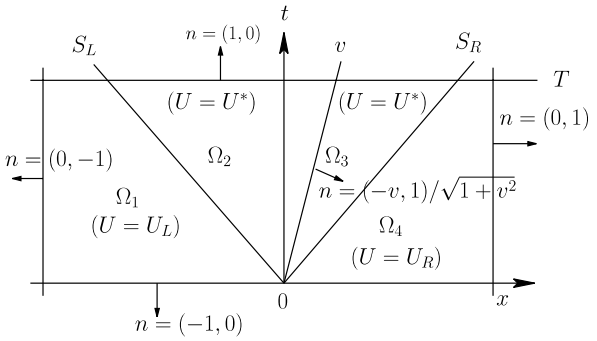


Fig. 1. Wave pattern of the solution for the Riemann problem. Here S_L and S_R are the fastest left and right moving signal velocities and v is the velocity of the element boundary point.

$$\int_{x_L}^{S_L T} U_L \,\mathrm{d}x + \int_{S_L T}^{vT} U(x,T)\,\mathrm{d}x = \int_{x_L}^{0} U_L \,\mathrm{d}x + \int_0^T F_L \,\mathrm{d}t - \int_0^T (F(U(vt,t)) - vU(vt,t))\mathrm{d}t - \int_{\Omega_2} G(U)\partial_x U \,\mathrm{d}x\,\mathrm{d}t - \int_0^T \int_0^1 G(\phi_{LL^*}(\tau; U_L, U_L^*)) \frac{\partial \phi_{LL^*}}{\partial \tau}(\tau; U_L, U_L^*)\mathrm{d}\tau\,\mathrm{d}t, \tag{36}$$

where $F_L = F(U_L)$ and $U_L^* = \lim_{s \downarrow S_L} U^*(st,t)$ is the trace of U^* taken from the interior of Ω_2, which is constant along the wave S_L due to the self similarity of the solution in the star region. Replace the exact integrand in the second integral on the left-hand side of (36) with the approximate solution $\overline{U}^*$. Furthermore, using the self similarity of the solution in the star region [5], we obtain

$$\int_{\Omega_2} G(U)\partial_x U \,\mathrm{d}x\,\mathrm{d}t = \int_{t=0}^{T} \int_{x=S_L t}^{vt} G(U)\partial_x U \,\mathrm{d}x\,\mathrm{d}t = \int_{t=0}^{T} \int_{S_L}^{v} G(U^*)\partial_s U^* \partial_x s |J| \,\mathrm{d}s\,\mathrm{d}t = T \int_{S_L}^{v} G(U^*)\partial_s U^* \,\mathrm{d}s, \tag{37}$$

where we used the coordinate transformation $x = st$, $t = t$, which has a Jacobian $|J| = t$. Introduce the trace of U^* taken from the interior of Ω_2 along the line $x = vt$ as: $U_{Lv}^* = \lim_{s \uparrow v} U^*(st,t)$ and the path $\phi_{L^*v} : [0,1] \times \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^m$ with

$$\phi_{L^*v}(\tau; U_L^*, U_{Lv}^*) = U^*(s), \quad \text{if } S_L < s < v.$$

By connecting these two paths into the path $\phi_{Lv} : [0,1] \times \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^m$, such that $\phi_{Lv}(\tau; U_L, U_{Lv}^*) = \phi_{LL^*} \cup \phi_{L^*v}$, redefining τ and using (37), the integral contributions due to the nonconservative product on the right-hand side of (36) can be combined, resulting in

$$S_L U_L + (v - S_L)\overline{U}^* = F_L - F^v - \int_0^1 G(\phi_{Lv}(\tau; U_L, U_{Lv}^*)) \frac{\partial \phi_{Lv}}{\partial \tau}(\tau; U_L, U_{Lv}^*)\mathrm{d}\tau, \tag{38}$$

where $F^v = F(U(vt,t)) - vU(vt,t)$ which is constant along $x = vt$. Similarly, using Gauss' theorem for the control volume $\Omega_3 \cup \Omega_4$ yields

$$\int_{vT}^{S_R T} U(x,T)\mathrm{d}x + \int_{S_R T}^{x_R} U_R \,\mathrm{d}x = \int_0^{x_R} U_R \,\mathrm{d}x - \int_0^T F_R \,\mathrm{d}t + \int_0^T (F(U(vt,t)) - vU(vt,t))\mathrm{d}t - \int_{\Omega_3} G(U)\partial_x U \,\mathrm{d}x\,\mathrm{d}t - \int_0^T \int_0^1 G(\phi_{R^*R}(\tau; U_R^*, U_R)) \frac{\partial \phi_{R^*R}}{\partial \tau}(\tau; U_R^*, U_R)\mathrm{d}\tau\,\mathrm{d}t, \tag{39}$$

where $F_R = F(U_R)$ and $U_R^* = \lim_{s \uparrow S_R} U^*(st,t)$ is the trace of U^* taken from the interior of Ω_3, which is constant along the wave S_R. Furthermore, denote the trace of U^* taken from the interior of Ω_3 along the line $x = v\,t$ as: $U_{Rv}^* = \lim_{s \downarrow v} U^*(st,t)$. Replace the exact integrand in the first integral on the left-hand side of (39) with the average of the exact solution $\overline{U}^*$. Introduce the path $\phi_{vR^*} : [0,1] \times \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^m$ with

$$\phi_{vR^*}(\tau; U^*_{Rv}, U^*_R) = U^*(s), \quad \text{if } v < s < S_R$$

and the path $\phi_{vR} : [0,1] \times \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^m$ such that $\phi_{vR}(\tau; U^*_{Rv}, U_R) = \phi_{R^*R} \cup \phi_{vR^*}$ after redefining τ. Using the self similarity of the solution in the star region Ω_3, similar to (37), the integral contributions on the right-hand side of (39) can be combined, resulting in

$$(S_R - v)\overline{U}^* - S_R U_R = F^v - F_R - \int_0^1 G(\phi_{vR}(\tau; U^*_{Rv}, U_R)) \frac{\partial \phi_{vR}}{\partial \tau}(\tau; U^*_{Rv}, U_R) \mathrm{d}\tau. \tag{40}$$

Note that $U^*_{Lv} = U^*_{Rv}$ since the solution U is smooth across $\partial\overline{\Omega}_2 \cap \partial\overline{\Omega}_3$, where $\overline{\Omega}_2$ and $\overline{\Omega}_3$ are the closures of Ω_2 and Ω_3. Now, introduce the path $\bar{\phi} : [0,1] \times \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^m$ (see Fig. 2) and redefine τ such that $\bar{\phi}(\tau; U_L, U_R) = \phi_{Lv} \cup \phi_{vR}$ then, by adding (38) and (40) and rearranging terms, we obtain:

$$\overline{U}^* = \frac{S_R U_R - S_L U_L + F_L - F_R}{S_R - S_L} - \frac{1}{S_R - S_L} \int_0^1 G(\bar{\phi}(\tau; U_L, U_R)) \frac{\partial \bar{\phi}}{\partial \tau}(\tau; U_L, U_R) \mathrm{d}\tau. \tag{41}$$

This equation is still exact if we would know the path $\bar{\phi}$. Note from Fig. 1 that outside the star region the solution is still at its initial values at $t = 0$, denoted by U_L and U_R. Within the star region bounded by the slowest and fastest signal speed S_L and S_R, respectively, an averaged star-state solution $\overline{U}^*$ is assumed. We define the numerical flux for U as

$$\widehat{U} = \begin{cases} U_L, & \text{if } v \leqslant S_L, \\ \overline{U}^*, & \text{if } S_L < v < S_R, \\ U_R, & \text{if } v \geqslant S_R, \end{cases}$$

where the averaged star-state solution $\overline{U}^*$ is given by (41) and v is the velocity of the element boundary point. We now continue to derive an expression for $\widehat{P}^{nc}(U_L, U_R, v, \vec{n}^L)$. Define

$$\int_0^\tau G(\bar{\phi}(\tilde{\tau}; U_1, U_2)) \frac{\partial \bar{\phi}}{\partial \tilde{\tau}}(\tilde{\tau}; U_1, U_2) \mathrm{d}\tilde{\tau} \equiv \int_0^\tau \mathrm{d}\mathcal{G}(\bar{\phi}(\tau; U_1, U_2)),$$

so that

$$\int_0^1 G(\bar{\phi}(\tilde{\tau}; U_1, U_2)) \frac{\partial \bar{\phi}}{\partial \tilde{\tau}}(\tilde{\tau}; U_1, U_2) \mathrm{d}\tilde{\tau} = \mathcal{G}(U_2) - \mathcal{G}(U_1)$$

using conditions H1–H4. Denote $\mathcal{G}(U_k) = \mathcal{G}_k$ and introduce $\tilde{\mathcal{G}}_k = \mathcal{G}_k - \{\!\{\mathcal{G}\}\!\}$, for $k = 1,2$ with $\{\!\{\mathcal{G}\}\!\} = (\mathcal{G}_1 + \mathcal{G}_2)/2$. Note that $\mathcal{G}_2 - \mathcal{G}_1 = \tilde{\mathcal{G}}_2 - \tilde{\mathcal{G}}_1$. From (38) and (40), the definition of the paths, conditions H1–H4 and assuming $U^*_{Lv} = U^*_{Rv} = \overline{U}^*$, we then obtain

$$S_L U_L + (v - S_L)\overline{U}^* = F_L - F^v - \widetilde{\mathcal{G}}^* + \widetilde{\mathcal{G}}_L \tag{42}$$

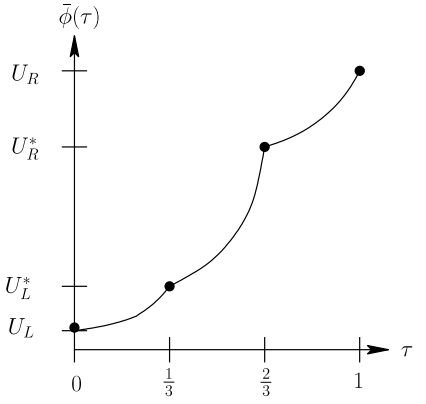


Fig. 2. Combining the paths to form $\bar{\phi}_{LR}(\tau; U_L, U_R) = \phi_{LL^*} \cup \phi_{L^*v} \cup \phi_{vR^*} \cup \phi_{R^*R}$.

and

$$(S_R - v)\overline{U}^* - S_R U_R = F^v - F_R - \widetilde{\mathcal{G}}_R + \widetilde{\mathcal{G}}^*, \tag{43}$$

where $\mathcal{G}_L = \mathcal{G}(U_L)$, $\mathcal{G}_R = \mathcal{G}(U_R)$ and $\mathcal{G}^* = \mathcal{G}(\overline{U}^*)$. Subtracting (43) from (42) and rearranging the terms, we obtain

$$F^v + \widetilde{\mathcal{G}}^* = \{\!\{\widetilde{\mathcal{G}}\}\!\} + \{\!\{F\}\!\} + \frac{1}{2}((S_R - v)\overline{U}^* + (S_L - v)\overline{U}^* - S_L U_L - S_R U_R)$$

with $\{\!\{\widetilde{\mathcal{G}}\}\!\} \equiv (\widetilde{\mathcal{G}}_L + \widetilde{\mathcal{G}}_R)/2 = 0$. Similarly, by adding (42) and (43) together and rearranging terms, we obtain

$$F_L + \widetilde{\mathcal{G}}_L = F_L - \frac{1}{2}\int_0^1 G(\bar{\phi}(\tau; U_L, U_R))\frac{\partial \phi}{\partial \tau}(\tau; U_L, U_R)\mathrm{d}\tau$$

and

$$F_R + \widetilde{\mathcal{G}}_R = F_R + \frac{1}{2}\int_0^1 G(\bar{\phi}(\tau; U_L, U_R))\frac{\partial \phi}{\partial \tau}(\tau; U_L, U_R)\mathrm{d}\tau.$$

The NCP numerical flux $\widehat{P}^{nc}(U_L, U_R, v, \bar{n}^L)$ is defined in Ω_1 as $F_L + \widetilde{\mathcal{G}}_L$, in $\Omega_2 \cup \Omega_3$ as $F^v + \widetilde{\mathcal{G}}^*$ and in Ω_4 as $F_R + \widetilde{\mathcal{G}}_R$ (see also (31)). The NCP-flux is thus given by

$$\widehat{P}_i^{nc}(U_L, U_R, v, \bar{n}^L) = \begin{cases} F_{ik}^L \bar{n}_k^L - \frac{1}{2}\int_0^1 G_{ikr}(\bar{\phi}(\tau; U_L, U_R))\frac{\partial \bar{\phi}_r}{\partial \tau}(\tau; U_L, U_R)\mathrm{d}\tau \bar{n}_k^L, & \text{if } S_L > v, \\ \{\!\{F_{ik}\}\!\}\bar{n}_k^L + \frac{1}{2}((S_R - v)\overline{U}_i^* + (S_L - v)\overline{U}_i^* - S_L U_i^L - S_R U_i^R), & \text{if } S_L < v < S_R, \\ F_{ik}^R \bar{n}_k^L + \frac{1}{2}\int_0^1 G_{ikr}(\bar{\phi}(\tau; U_L, U_R))\frac{\partial \bar{\phi}_r}{\partial \tau}(\tau; U_L, U_R)\mathrm{d}\tau \bar{n}_k^L, & \text{if } S_R < v \end{cases} \tag{44}$$

with $\overline{U}^*$ given by (41). Note that if G is the Jacobian of some flux function Q, then $\widehat{P}^{nc}(U_L, U_R, v, \bar{n}^L)$ is exactly the HLL flux derived for moving grids in van der Vegt and van der Ven [27].

5. Test cases

5.1. The one-dimensional shallow water equations with topography

We consider a non-dimensional form of the shallow water system with topography. The system reads

$$U_{i,0} + F_{i,1} + G_{ij}U_{j,1} = 0 \quad \text{for } i, j = 1, 2, 3 \tag{45}$$

with

$$U = \begin{bmatrix} b \\ h \\ hu \end{bmatrix}, \quad F = \begin{bmatrix} 0 \\ hu \\ hu^2 + \frac{1}{2}\mathsf{F}^{-2}h^2 \end{bmatrix}, \quad G(U) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \mathsf{F}^{-2}h & 0 & 0 \end{bmatrix}. \tag{46}$$

Here b is the topography, h the water depth, u the flow velocity and F the Froude number defined as $\mathsf{F} = u_0^*/\sqrt{g^* h_0^*}$, where the starred values denote reference values. The eigenvalues of $\partial F/\partial U + G(U)$ are given by

$$\lambda_1 = u - \sqrt{\mathsf{F}^{-2}h}, \quad \lambda_2 = 0, \quad \lambda_3 = u + \sqrt{\mathsf{F}^{-2}h}. \tag{47}$$

When taking $\phi = U_L + \tau(U_R - U_L)$, the NCP-flux for (45) on a fixed grid becomes:

$$\widehat{P}^{nc} = \begin{cases} F^L - \frac{1}{2}V^{nc}, & \text{if } S_L > 0, \\ F^{hll} - (S_R + S_L)V^{nc}/(2(S_R - S_L)), & \text{if } S_L < 0 < S_R, \\ F^R + \frac{1}{2}V^{nc}, & \text{if } S_R < 0, \end{cases}$$

in which F^{hll} is the HLL-flux [23]:

$$F^{hll} = \frac{S_R F_L - S_L F_R + S_L S_R (U_R - U_L)}{S_R - S_L}$$

and V^{nc} appears in the extra term due to the nonconservative product:

$$V^{nc} = [0, 0, -\mathsf{F}^{-2}\{\!\{h\}\!\}[\![b]\!]]^{\mathrm{T}}.$$

In the numerical flux, as derived in Section 4, we take

$$S_L = \min(u_L - \sqrt{\mathsf{F}^{-2}h_L}, u_R - \sqrt{\mathsf{F}^{-2}h_R}) \quad \text{and} \quad S_R = \max(u_L + \sqrt{\mathsf{F}^{-2}h_L}, u_R + \sqrt{\mathsf{F}^{-2}h_R}).$$

5.1.1. Test cases 1 and 2: Rest flow

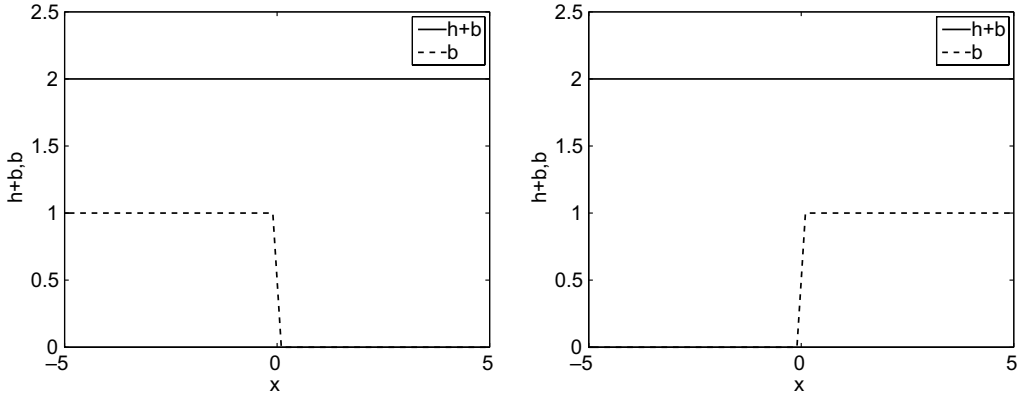
For test cases 1 and 2 we only consider the solution determined with space–time DGFEM calculations using linear basis functions and the linear path $\phi = U_L + \tau(U_R - U_L)$. Consider flow at rest over a discontinuous topography with initial and boundary conditions:

- *Test case 1*. Initial conditions: $b(x,0) = 1$ if $x < 0$ and $b(x,0) = 0$ if $x > 0$, $h(x,0) + b(x,0) = 2$, $hu(x,0) = 0$. Boundary conditions: $b(-5,t) = 1$, $h(-5,t) = 1$, $u(-5,t) = 0$, $b(5,t) = 0$, $h(5,t) = 2$, $u(5,t) = 0$.
- *Test case 2*. Initial conditions: $b(x,0) = 0$ if $x < 0$ and $b(x,0) = 1$ if $x > 0$, $h(x,0) + b(x,0) = 2$, $hu(x,0) = 0$. Boundary conditions: $b(-5,t) = 0$, $h(-5,t) = 2$, $u(-5,t) = 0$, $b(5,t) = 1$, $h(5,t) = 1$, $u(5,t) = 0$.

In Fig. 3, we show the steady-state solution, calculated using a time step of $\Delta t = 10^{21}$ on a grid with 100 cells and a Froude number of $\mathsf{F} = 0.2$. We solve the system of nonlinear equations using a pseudo-time stepping integration method (see [27]). As stopping criterium in the pseudo-time stepping calculation we take that the maximum residual must be smaller than 10^{-13}. A pseudo-time stepping CFL number of $\mathrm{CFL}^{\mathrm{pseudo}} = 0.8$ is used.

For the space DGFEM weak formulation we prove theoretically, that when using linear basis functions and taking the path $\phi = U_L + \tau(U_R - U_L)$, rest flow remains at rest. Consider the one-dimensional version of the space DGFEM weak formulation (A.11) for the shallow water equations:

$$0 = \sum_k \int_{K_k} (V_i U_{i,0} - V_{i,1} F_i + V_i G_{ij} U_{j,1}) \mathrm{d}K + \sum_{\mathcal{S}\in\mathcal{S}_I} \int_{\mathcal{S}} \{\!\{V_i\}\!\} \left(\int_0^1 G_{ij}(\phi(\tau; U_L, U_R)) \frac{\partial \phi_j}{\partial \tau}(\tau; U_L, U_R) \mathrm{d}\tau \right) \bar{n}^L \mathrm{d}\mathcal{S}$$
$$+ \sum_{\mathcal{S}\in\mathcal{S}_I} \int_{\mathcal{S}} (V_i^L - V_i^R) \widehat{P}_i^{nc} \mathrm{d}\mathcal{S}.$$



(a) Test case 1. (b) Test case 2.

Fig. 3. Flow at rest over a discontinuous topography. $\mathsf{F} = 0.2$, 100 cells, $\Delta t = 10^{21}$.

We only consider cell K_k where the contributions satisfy

$$0 = \int_{K_k} (V_i U_{i,0} - V_{i,1} F_i + V_i G_{ij} U_{j,1}) \mathrm{d}K + \int_{\mathcal{S}_{k+1}} \frac{1}{2} V_i^L \left(\int_0^1 G_{ij}(\phi(\tau; U_L, U_R)) \frac{\partial \phi_j}{\partial \tau}(\tau; U_L, U_R) \mathrm{d}\tau \right) \bar{n}^L + V_i^L \widehat{P}_i^{nc} \, \mathrm{d}\mathcal{S} + \int_{\mathcal{S}_k} \frac{1}{2} V_i^R \left(\int_0^1 G_{ij}(\phi(\tau; U_L, U_R)) \frac{\partial \phi_j}{\partial \tau}(\tau; U_L, U_R) \mathrm{d}\tau \right) \bar{n}^L - V_i^R \widehat{P}_i^{nc} \, \mathrm{d}\mathcal{S}. \tag{48}$$

For the numerical flux we take the star-state solution given by (41). For rest flow, using $\phi = U_L + \tau(U_R - U_L)$ and $h_L + b_L = h_R + b_R$ the star-state solution is given by

$$\overline{U}^* = \frac{1}{S_R - S_L} [\, S_R b_R - S_L b_L, \quad S_R h_R - S_L h_L, \quad 0 \,]^{\mathrm{T}}, \tag{49}$$

so that the numerical flux $\widehat{P}^{nc} = \{\!\{F\}\!\} + \frac{1}{2}(S_L(\overline{U}^* - U_L) + S_R(\overline{U}^* - U_R))$ is given by

$$\widehat{P}^{nc} = \left[\frac{S_L S_R (b_R - b_L)}{S_R - S_L}, \quad \frac{S_L S_R (h_R - h_L)}{S_R - S_L}, \quad \frac{1}{4} \mathsf{F}^{-2} (h_L^2 + h_R^2) \right]^{\mathrm{T}}. \tag{50}$$

Also, using $\phi = U_L + \tau(U_R - U_L)$ and $h_L + b_L = h_R + b_R$ we can show that

$$\int_0^1 G_{ij}(\phi(\tau; U_L, U_R)) \frac{\partial \phi_j}{\partial \tau}(\tau; U_L, U_R) \mathrm{d}\tau = \left[0, 0, -\mathsf{F}^{-2} [\![b]\!] \{\!\{ h \}\!\} \right]^{\mathrm{T}}.$$

We can write (48) now as

$$0 = \int_{K_k} (V_i U_{i,0} - V_{i,1} F_i + G_{ij} U_{j,1}) \mathrm{d}K + \int_{\mathcal{S}_{k+1}} V_i^L \mathcal{P}_i^p \, \mathrm{d}\mathcal{S} - \int_{\mathcal{S}_k} V_i^R \mathcal{P}_i^m \, \mathrm{d}\mathcal{S}, \tag{51}$$

where $\mathcal{P}^p$ and $\mathcal{P}^m$ are given by

$$\begin{aligned}
\mathcal{P}^p &= \frac{1}{2} \int_0^1 G_{ij}(\phi(\tau; U_L, U_R)) \frac{\partial \phi_j}{\partial \tau}(\tau; U_L, U_R) \mathrm{d}\tau + \widehat{P}^{nc}, \\
&= \left[\frac{S_L S_R (b_R - b_L)}{S_R - S_L}, \quad \frac{S_L S_R (h_R - h_L)}{S_R - S_L}, \quad \frac{1}{2} \mathsf{F}^{-2} h_L^2 \right]^{\mathrm{T}} \\
\mathcal{P}^m &= \frac{1}{2} \int_0^1 G_{ij}(\phi(\tau; U_L, U_R)) \frac{\partial \phi_j}{\partial \tau}(\tau; U_L, U_R) \mathrm{d}\tau - \widehat{P}^{nc} \\
&= \left[-\frac{S_L S_R (b_R - b_L)}{S_R - S_L}, \quad -\frac{S_L S_R (h_R - h_L)}{S_R - S_L}, \quad \frac{1}{2} \mathsf{F}^{-2} h_R^2 \right]^{\mathrm{T}}.
\end{aligned}$$

Using linear basis functions we can evaluate the integrals as follows:

$$\int_{K_k} V_i U_{i,0} \, \mathrm{d}K = \Delta x \overline{V}_i|_{K_k} \partial_t \overline{U}_i|_{K_k} + \frac{\Delta x}{3} \widehat{V}_i|_{K_k} \partial_t \widehat{U}_i|_{K_k}, \tag{52a}$$

$$\begin{aligned}
-\int_{K_k} V_{i,1} F_i \, \mathrm{d}K &= -\int_{-1}^1 \widehat{V}_i|_{K_k} F(\overline{U}_i|_{K_k} + \widehat{U}_i|_{K_k} \xi) \mathrm{d}\xi \\
&= -\widehat{V}_i|_{K_k} \left[0, \quad 0, \quad \frac{1}{3} \mathsf{F}^{-2} \hat{h}_k^2 + \mathsf{F}^{-2} \bar{h}_k^2 \right]^{\mathrm{T}},
\end{aligned} \tag{52b}$$

$$\begin{aligned}
\int_{K_k} V_i G_{ij} U_{j,1} \, \mathrm{d}K &= \int_{-1}^1 (\overline{V}_i|_{K_k} + \widehat{V}_i|_{K_k} \xi) G (\overline{U}_i|_{K_k} + \widehat{U}_i|_{K_k} \xi) \widehat{U}_i|_{K_k} \, \mathrm{d}\xi \\
&= \overline{V}_i|_{K_k} \left[0 \quad 0, \quad 2\mathsf{F}^{-2} \overline{h}_k \hat{b}_k \right]^{\mathrm{T}} + \widehat{V}_i|_{K_k} \left[0, \quad 0, \quad \frac{2}{3} \mathsf{F}^{-2} \hat{h}_k \hat{b}_k \right]^{\mathrm{T}}
\end{aligned} \tag{52c}$$

$$\int_{\mathcal{S}_{k+1}} V_i^L \mathcal{P}_i^p \, \mathrm{d}\mathcal{S} = (\overline{V}|_{K_k} + \widehat{V}|_{K_k}) \begin{bmatrix} \frac{S_{k+1}^L S_{k+1}^R (b_{k+1}^R - b_{k+1}^L)}{S_{k+1}^R - S_{k+1}^L} \\ \frac{S_{k+1}^L S_{k+1}^R (h_{k+1}^R - h_{k+1}^L)}{S_{k+1}^R - S_{k+1}^L} \\ \frac{1}{2} \mathsf{F}^{-2} (\bar{h}_k + \hat{h}_k)^2 \end{bmatrix}, \tag{52d}$$

$$-\int_{\mathcal{S}_k} V_i^R \mathcal{P}_i^m \mathrm{d}\mathcal{S} = -(\overline{V}|_{K_k} - \widehat{V}|_{K_k}) \begin{bmatrix} \frac{S_k^L S_k^R (b_k^R - b_k^L)}{S_k^R - S_k^L} \\ \frac{S_k^L S_k^R (h_k^R - h_k^L)}{S_k^R - S_k^L} \\ \frac{1}{2}\mathsf{F}^{-2}(\bar{h}_k - \hat{h}_k)^2 \end{bmatrix}, \tag{52e}$$

where $\overline{(\cdot)}$ and $\widehat{(\cdot)}$ are the means and slopes, respectively, of the approximation for U and V. Adding the vectors (52b)–(52e), we note that the third element of this sum is zero using $h_L + b_L = h_R + b_R$ and the fact that the slope of $h + b = 0$ (so $\widehat{U}|_{K_k} = (-\hat{h}_k, \hat{h}_k, 0)$). Note that in (52d) and (52e) we have $b_{k+1}^R - b_{k+1}^L + h_{k+1}^R - h_{k+1}^L = 0$ and $b_k^R - b_k^L + h_k^R - h_k^L = 0$, respectively so that

$$\partial_t(\bar{h}_k + \bar{b}_k) = 0, \quad \partial_t(\hat{h}_k + \hat{b}_k) = 0, \quad \partial_t \overline{hu}_k = 0, \quad \partial_t \widehat{hu}_k = 0$$

meaning that for rest flow $h + b$ remains constant.

5.1.2. Test case 3: Subcritical flow over a bump

We now consider subcritical flow with a Froude number of $\mathsf{F} = 0.2$ over a bump. The topography reads

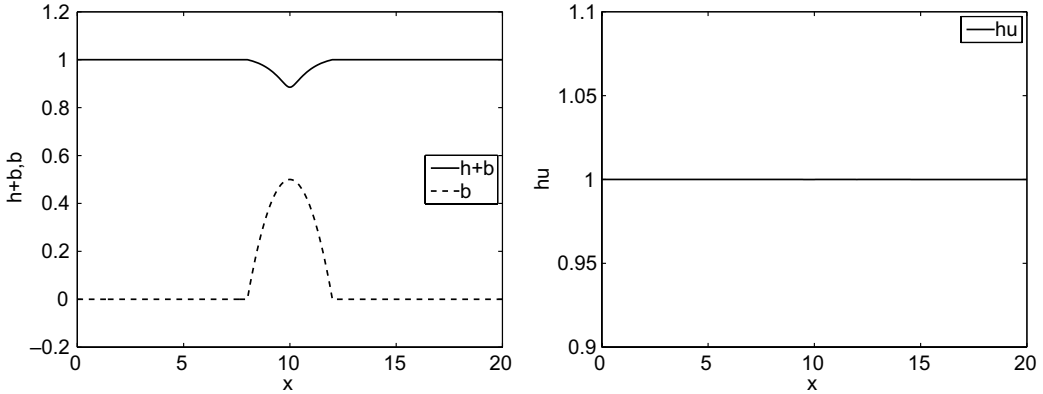
$$b(x) = \begin{cases} a(b - (x - x_p))(b + (x - x_p))b^{-2} & \text{for } |x - x_p| \leqslant b, \\ 0 & \text{otherwise.} \end{cases} \tag{53}$$

We use $x_p = 10$, $a = 0.5$ and $b = 2$ as in [20]. The exact steady-state solution for this test case is found by solving the following third-order equation in u [9,20]:

$$\mathsf{F}^2 u^3/2 + (b - \mathsf{F}^2/2 - 1)u + 1 = 0 \quad \text{with } hu = 1. \tag{54}$$

The domain $x \in [0, 20]$ is divided into 40, 80, 160 and 320 cells. We consider DGFEM and STDGFEM calculations using the linear path $\phi = U_L + \tau(U_R - U_L)$. For space DGFEM calculations, a CFL number of $\mathrm{CFL} = 0.8$ is taken and when the residuals are smaller than 10^{-11} the calculation is stopped. For STDGFEM calculations we consider the solution after one physical time step of $\Delta t = 10^{21}$. We can do this because we want to consider the steady-state solution. As stopping criterium in the pseudo-time stepping calculation we take that the maximum residual must be smaller than 10^{-11}. A pseudo-time stepping CFL number of $\mathrm{CFL}^{\mathrm{pseudo}} = 0.8$ is used.

The initial condition is $h + b = 1$ and $hu = 1$ and the boundary conditions are: $b(0,t) = 0$, $h(0,t) = 1$, $u(0,t) = 1$, $b(1,t) = 0$, $h(1,t) = 1$ and $u(1,t) = 1$. The steady-state solution is given in Fig. 4. The order of convergence is determined by looking at the L^2 and the $L^{\max}$ norm of the numerical error in $z = h + b$ and hu with respect to the exact solution:



(a) The water level $h(x) + b(x)$. (b) The mass flow $hu(x)$.

Fig. 4. Test case 3: steady-state solution calculated using space DGFEM, $\mathsf{F} = 0.2$, 320 cells.

$$\|z_{\text{num}} - z_{\text{exact}}\|_2 = \left(\sum_{k=1}^{N_{\text{cells}}} \int_{K_k} (z_{K_k}^{\text{num}} - z_{K_k}^{\text{exact}})^2 \right)^{1/2} \tag{55}$$

and

$$\|z_{\text{num}} - z_{\text{exact}}\|_{\max} = \max\{|z_{\text{num}}^i - z_{\text{exact}}^i| : 1 \leqslant i \leqslant N_{\text{cells}}\}. \tag{56}$$

The order of convergence using DGFEM and STDGFEM is given in Table 1 using linear basis functions and in Table 2 using quadratic basis functions. Using linear basis functions we obtain second-order convergence and using quadratic basis functions we obtain third-order convergence for both space-DGFEM and space–time DGFEM calculations.

5.1.3. Test case 4: Supercritical flow over a bump

Next, we consider supercritical flow with a Froude number of $\mathsf{F} = 1.9$ over a bump. We use the same topography (53) and the exact solution can be found by solving (54). The domain $x \in [0, 20]$ is again divided into 40, 80, 160 and 320 cells and we consider DGFEM and STDGFEM calculations using the linear path $\phi = U_L + \tau(U_R - U_L)$. For space DGFEM calculations, time steps of $\Delta t = 0.01$ are made. Using linear basis functions, a CFL number of $\text{CFL} = 0.3$ is taken and when the residuals are smaller than 10^{-11} the calculation is stopped. For the STDGFEM calculation we consider again the solution after one physical time step of $\Delta t = 10^{21}$. The same stopping criteria as in the subcritical flow case are used. Using linear basis functions,

Table 1
L^2 and $L^{\max}$ error for $h + b$ and hu using DGFEM and STDGFEM for test case 3

N_{cells}	$h + b$				hu			
	L^2 error	p	$L^{\max}$ error	p	L^2 error	p	$L^{\max}$ error	p
DGFEM								
40	0.1133×10^{-2}	–	0.6513×10^{-2}	–	0.1265×10^{-2}	–	0.3302×10^{-2}	–
80	0.3193×10^{-3}	1.8	0.2387×10^{-2}	1.4	0.1944×10^{-3}	2.7	0.8030×10^{-3}	2.0
160	0.8364×10^{-4}	1.9	0.6989×10^{-3}	1.8	0.2764×10^{-4}	2.8	0.1369×10^{-3}	2.6
320	0.2119×10^{-4}	2.0	0.1847×10^{-3}	1.9	0.3798×10^{-5}	2.9	0.2931×10^{-4}	2.2
STDGFEM								
40	0.1141×10^{-2}	–	0.6559×10^{-2}	–	0.1262×10^{-2}	–	0.3285×10^{-2}	–
80	0.3194×10^{-3}	1.8	0.2387×10^{-2}	1.5	0.1943×10^{-3}	2.7	0.8029×10^{-3}	2.0
160	0.8365×10^{-4}	1.9	0.6989×10^{-3}	1.8	0.2763×10^{-4}	2.8	0.1369×10^{-3}	2.6
320	0.2119×10^{-4}	2.0	0.1847×10^{-3}	1.9	0.3797×10^{-5}	2.9	0.2929×10^{-4}	2.2

Second-order convergence rates are shown for $\mathsf{F} = 0.2$.

Table 2
L^2 and $L^{\max}$ error for $h + b$ and hu using DGFEM and STDGFEM for test case 3

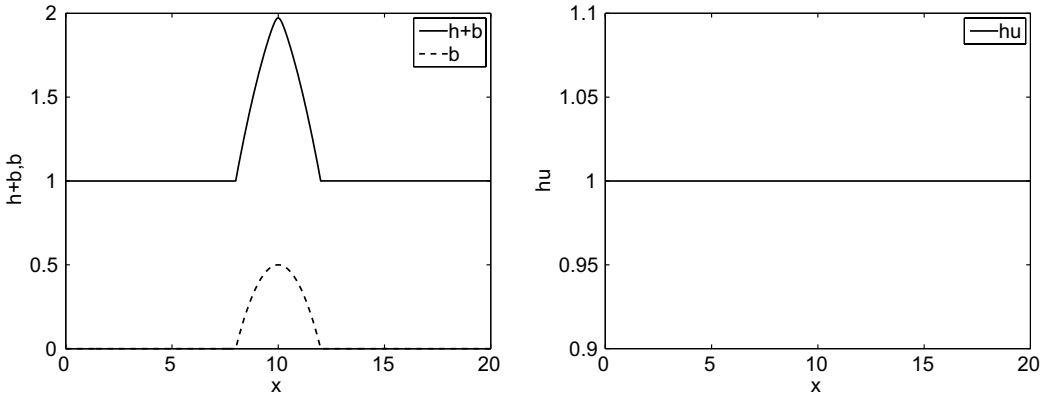
N_{cells}	$h + b$				hu			
	L^2 error	p	$L^{\max}$ error	p	L^2 error	p	$L^{\max}$ error	p
DGFEM								
40	0.3210×10^{-3}	–	0.1466×10^{-2}	–	0.8352×10^{-3}	–	0.3124×10^{-2}	–
80	0.4622×10^{-4}	2.8	0.2670×10^{-3}	2.5	0.1269×10^{-3}	2.7	0.5562×10^{-3}	2.5
160	0.6303×10^{-5}	2.9	0.3567×10^{-4}	2.9	0.1689×10^{-4}	2.9	0.7186×10^{-4}	3.0
320	0.7931×10^{-6}	3.0	0.4459×10^{-5}	3.0	0.2144×10^{-5}	3.0	0.8860×10^{-5}	3.0
STDGFEM								
40	0.3278×10^{-3}	–	0.1836×10^{-2}	–	0.2339×10^{-3}	–	0.1170×10^{-2}	–
80	0.4433×10^{-4}	2.9	0.3195×10^{-3}	2.5	0.3721×10^{-4}	2.7	0.2401×10^{-3}	2.3
160	0.4556×10^{-5}	3.3	0.3142×10^{-4}	3.3	0.5513×10^{-5}	2.8	0.3596×10^{-4}	2.7
320	0.5522×10^{-6}	3.0	0.4407×10^{-5}	2.8	0.7489×10^{-6}	2.9	0.5218×10^{-5}	2.8

Third-order convergence rates are shown for $\mathsf{F} = 0.2$.

we use a pseudo-time stepping CFL number of $\mathrm{CFL}^{\mathrm{pseudo}} = 0.8$. For quadratic basis functions, on the grids with 40 and 160 cells, a pseudo-time stepping CFL number of $\mathrm{CFL}^{\mathrm{pseudo}} = 0.4$ is employed and on the grids with 80 and 320 cells a pseudo-time stepping CFL number of $\mathrm{CFL}^{\mathrm{pseudo}} = 0.8$.

The initial condition is $h + b = 1$ and $hu = 1$ and transmissive boundary conditions are given at $x = 0$ and at $x = 20$, i.e., $U^b = U^L$, where U^b is the vector of the boundary data and U^L is the vector with the data calculated at the boundary from inside the domain. The steady-state solution is shown in Fig. 5. The order of convergence is again determined by computing the L^2 and the $L^{\max}$ norm of the numerical error in $h + b$ and hu with respect to the exact solution as defined in (55) and (56). The order of convergence using DGFEM and STDGFEM is given in Table 3 using linear basis functions and in Table 4 using quadratic basis functions.

We see that the space- and space–time DGFEM calculations results in second-order convergence for $h + b$ using linear basis functions and in third-order convergence for $h + b$ using quadratic basis functions. We do not show the order of convergence for hu because the error for hu is of the order of machine precision on all



(a) The water level $h(x) + b(x)$. (b) The mass flow $hu(x)$.

Fig. 5. Test case 4: steady-state solution calculated using space DGFEM, $\mathsf{F} = 1.9$, 320 cells.

Table 3
L^2 and $L^{\max}$ error for $h + b$ using DGFEM and STDGFEM for test case 4

N_{cells}	DGFEM $h + b$				STDGFEM $h + b$			
	L^2 error	p	$L^{\max}$ error	p	L^2 error	p	$L^{\max}$ error	p
40	0.7543×10^{-2}	–	0.4619×10^{-1}	–	0.7543×10^{-2}	–	0.4619×10^{-1}	–
80	0.1281×10^{-2}	2.6	0.9406×10^{-2}	2.3	0.1281×10^{-2}	2.6	0.9406×10^{-2}	2.3
160	0.3188×10^{-3}	2.0	0.2615×10^{-2}	1.8	0.3188×10^{-3}	2.0	0.2615×10^{-2}	1.8
320	0.7914×10^{-4}	2.0	0.6883×10^{-3}	1.9	0.7914×10^{-4}	2.0	0.6883×10^{-3}	1.9

Second-order convergence rates are shown for $\mathsf{F} = 1.9$.

Table 4
L^2 and $L^{\max}$ error for $h + b$ using DGFEM and STDGFEM for test case 4

N_{cells}	DGFEM $h + b$				STDGFEM $h + b$			
	L^2 error	p	$L^{\max}$ error	p	L^2 error	p	$L^{\max}$ error	p
40	0.1293×10^{-2}	–	0.5034×10^{-2}	–	0.9181×10^{-3}	–	0.4946×10^{-2}	–
80	0.1944×10^{-3}	2.7	0.9383×10^{-3}	2.4	0.1624×10^{-3}	2.5	0.1127×10^{-2}	2.1
160	0.2892×10^{-4}	2.7	0.1545×10^{-3}	2.6	0.1830×10^{-4}	3.1	0.1382×10^{-3}	3.0
320	0.3724×10^{-5}	3.0	0.2111×10^{-4}	2.9	0.2253×10^{-5}	3.0	0.2002×10^{-4}	2.8

Third-order convergence rates are shown for $\mathsf{F} = 1.9$.

meshes for the space DGFEM calculations and stabilizes around 10^{-8} for the space–time DGFEM calculations.

5.1.4. Test case 5: Transcritical flow over a bump

For this test case we consider the steady-state solution of transcritical flow with a shock over a bump. The topography is given by

$$b(x) = \begin{cases} 0.2 - 0.05(x-10)^2 & \text{if } 8 \leqslant x \leqslant 12, \\ 0 & \text{otherwise,} \end{cases}$$

which is the same as that used by Xing and Shu [28]. The initial condition is $h + b = 0.5$ and $hu = 0$ and the boundary conditions are: $b(0,t) = 0$, $hu(0,t) = 0.18$, $b(25,t) = 0$, $h(25,t) = 0.33$, $hu(25,t) = 0.18$. The remaining boundary data are set equal to the data calculated at the boundary from inside the domain. In our computations, we take $\mathsf{F}^{-2} = 9.812$. Simulations concern space–time DGFEM. We consider the solution after one physical time step of $\Delta t = 10^{21}$ on a grid with 200 cells using a pseudo-time stepping CFL number of $\mathrm{CFL}^{\mathrm{pseudo}} = 0.8$. To deal with the shock, we used the slope limiter as discussed in Section 3.5. The solution is given in Fig. 6 and compares well with results in [9].

5.1.5. Test case 6: Perturbation of a steady-state solution

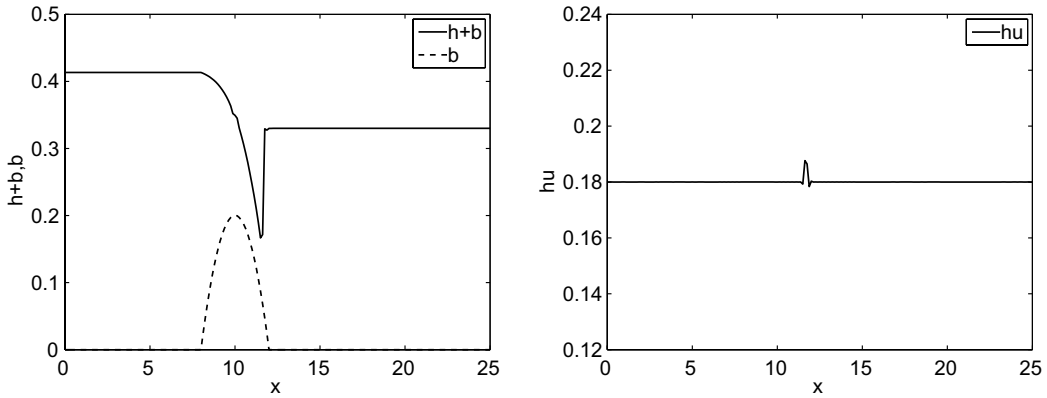
We repeat a test case as was formulated in Xing and Shu [28] which was originally proposed by LeVeque [15]. Consider a topography given by

$$b(x) = \begin{cases} 0.25(\cos(10\pi(x-1.5)) + 1), & \text{if } 1.4 \leqslant x \leqslant 1.6, \\ 0, & \text{otherwise.} \end{cases}$$

The initial conditions are given by

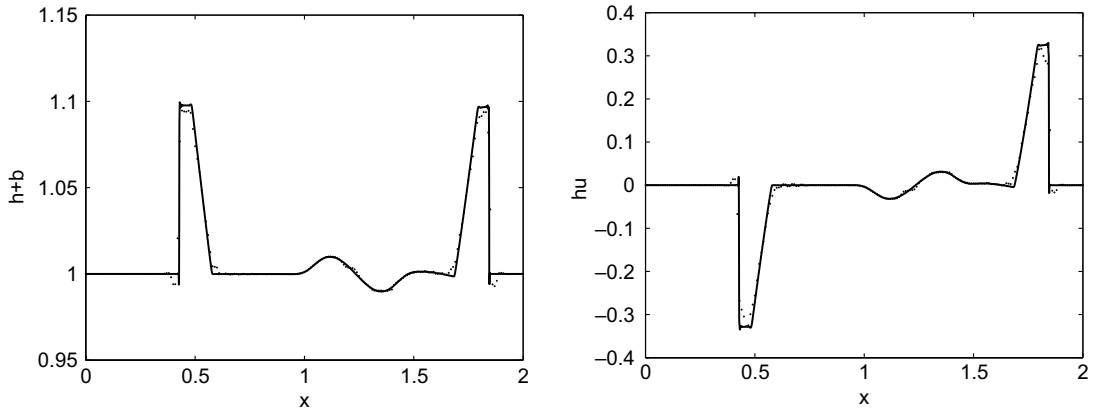
$$hu(x,0) = 0, \quad h(x,0) = \begin{cases} 1 - b(x) + \epsilon, & \text{if } 1.1 \leqslant x \leqslant 1.2, \\ 1 - b(x), & \text{otherwise.} \end{cases}$$

At the boundaries, we use transmissive boundary conditions. We take $\mathsf{F}^{-2} = 9.812$. The same two cases as in Xing and Shu [28] were run: $\epsilon = 0.2$ (big pulse) and $\epsilon = 0.001$ (small pulse). We used space–time DGFEM to compute the solution on a uniform grid with 200 cells and 3000 cells. On the grid with 200 cells, a physical time step of $\Delta t = 0.0002$ was used. On the grid with 3000 cells, we used a physical time step of $\Delta t = 0.00002$. A pseudo-time stepping CFL number of $\mathrm{CFL}^{\mathrm{pseudo}} = 0.4$ was used. In Figs. 7 and 8, we show the fine and coarse mesh solution, as in [28], for the water level $h(x) + b(x)$ and mass flow $hu(x)$ at time $t = 0.2$ for the big pulse test case and the small pulse test case, respectively.



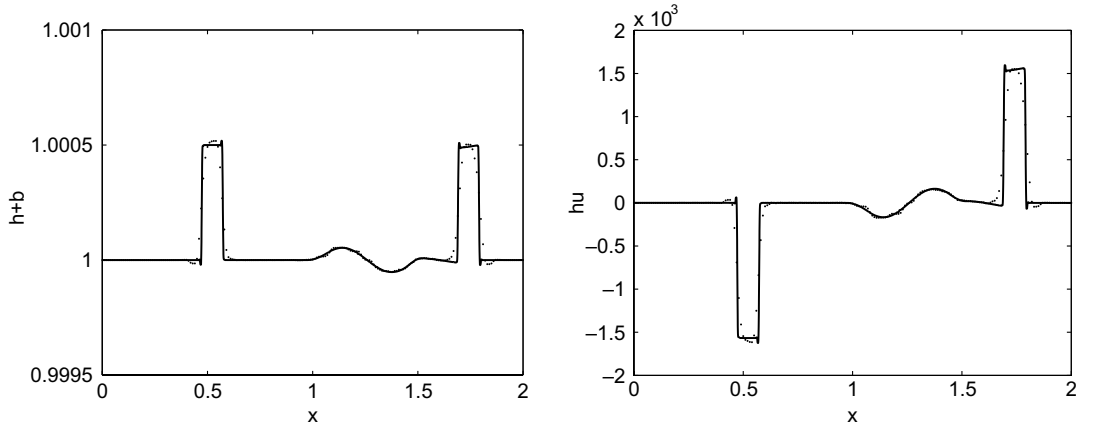
(a) The water level $h(x) + b(x)$. (b) The mass flow $hu(x)$.

Fig. 6. Test case 5: steady-state transcritical flow with a shock, $\Delta t = 10^{21}$, $N_{\mathrm{cells}} = 200$, $\mathrm{CFL}^{\tau} = 0.8$, $\mathsf{F}^{-2} = 9.812$.



(a) The water level $h(x) + b(x)$. (b) The mass flow $hu(x)$.

Fig. 7. Test case 6: perturbation of a steady-state solution with a big pulse at time $t = 0.2$, $\epsilon = 0.2$. Line: $N_{\text{cells}} = 3000$. Dots: $N_{\text{cells}} = 200$.



(a) The water level $h(x) + b(x)$. (b) The mass flow $hu(x)$.

Fig. 8. Test case 6: perturbation of a steady-state solution with a small pulse at time $t = 0.2$, $\epsilon = 0.001$. Line: $N_{\text{cells}} = 3000$. Dots: $N_{\text{cells}} = 200$.

5.1.6. Test case 7: Dam break problem over a rectangular bump

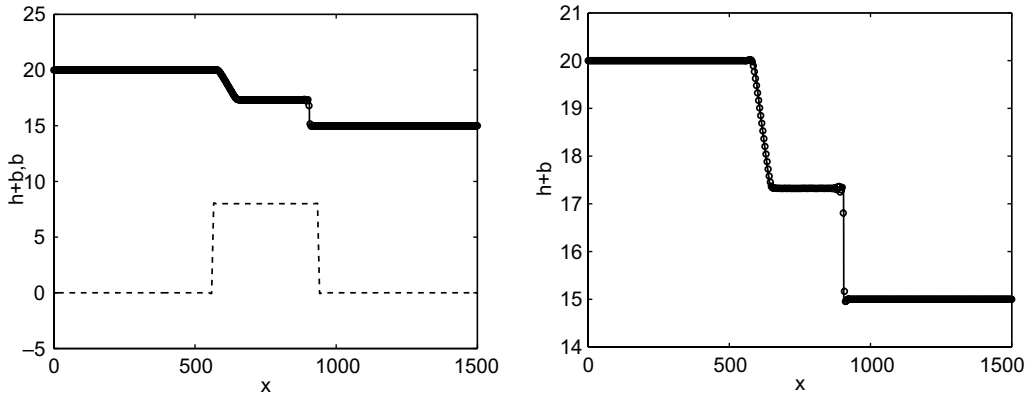
A dam break problem is simulated over a rectangular hump, as in [28]. The topography is given by

$$b(x) = \begin{cases} 8, & \text{if } |x - 750| \leqslant 1500/8, \\ 0, & \text{otherwise}, \end{cases}$$

for $x \in [0, 1500]$. The initial conditions are given by

$$hu(x,0) = 0, \quad h(x,0) = \begin{cases} 20 - b(x) & \text{if } x \leqslant 750, \\ 15 - b(x) & \text{otherwise} \end{cases}$$

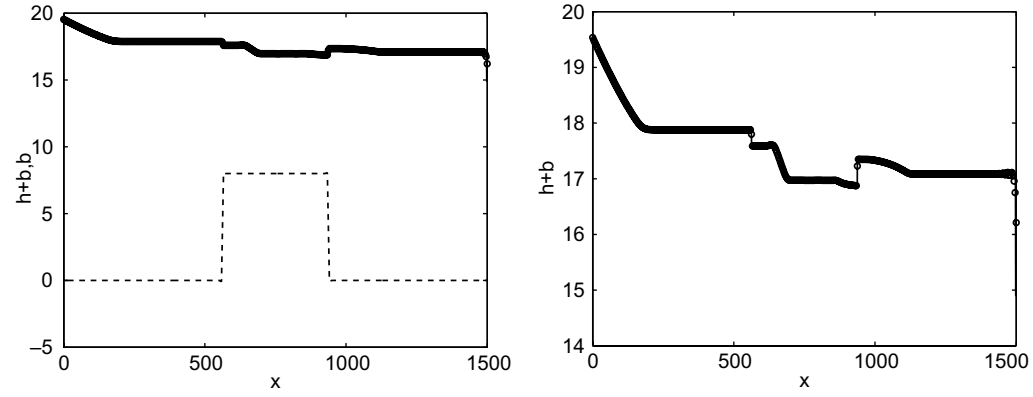
and as boundary conditions we take: $b(0,t) = 0$, $h(0,t) = 20$, $hu(0,t) = 0$, $b(1500,t) = 0$, $h(1500,t) = 15$ and $hu(1500,t) = 0$. We take $\mathsf{F}^{-2} = 9.812$. With space–time DGFEM the solution was computed on a uniform grid with 400 cells and 4000 cells. On the grid with 400 cells, a physical time step of $\Delta t = 0.02$ was used and on the grid with 4000 cells, the physical time step was $\Delta t = 0.002$. The pseudo-time stepping CFL number was $\text{CFL}^{\text{pseudo}} = 0.8$. In Figs. 9 and 10, we show the solution for the water level $h(x) + b(x)$ at time $t = 15$ and at time $t = 60$, respectively.



(a) The numerical solution of the water level and the topography.

(b) The numerical solution of the water level.

Fig. 9. Test case 7: the dam breaking problem at time $t = 15$. Line: 4000 cells. Dots: 400 cells.



(a) The numerical solution of the water level and the topography.

(b) The numerical solution of the water level.

Fig. 10. Test case 7: the dam breaking problem at time $t = 60$. Line: 4000 cells. Dots: 400 cells.

5.1.7. Conclusions

For the shallow water equations with topography we showed numerical results of seven test cases calculated using the space- and/or space–time DGFEM discretizations we developed for nonconservative hyperbolic partial differential equations. For all test cases we obtained good results. For test cases 1 and 2 we showed that rest flow remained unchanged despite having discontinuities in the topography. In test cases 3 and 4 we solved subcritical and supercritical flow over a bump demonstrating that the scheme is second-order accurate for linear basis functions and third-order accurate for quadratic basis functions. In test cases 5, 6 and 7 we showed that we resolved also more complex test cases with discontinuous solutions.

5.2. Two-dimensional shallow water and morphological flow

5.2.1. Test case 8: Hydraulic and sediment transport through a contraction

Consider the non-dimensional form of the shallow water equations and the bed evolution equation (for details see [21,22]):

$$A_{ir}U_{r,0} + F_{ik,k} + G_{ikr}U_{r,k} = 0, \tag{57}$$

where $U = [h, hu_1, hu_2, b]^{\mathrm{T}}$ and

$$A = \begin{bmatrix} \epsilon & 0 & 0 & 0 \\ 0 & \epsilon & 0 & 0 \\ 0 & 0 & \epsilon & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad F = \begin{bmatrix} hu_1 & hu_2 \\ hu_1^2 + \mathsf{F}^{-2}h^2/2 & hu_1u_2 \\ hu_1u_2 & hu_2^2 + \mathsf{F}^{-2}h^2/2 \\ |u|^{\beta-1}u_1 & |u|^{\beta-1}u_2 \end{bmatrix}, G_{k=1} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathsf{F}^{-2}h \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

$$G_{k=2} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathsf{F}^{-2}h \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

where ϵ is the ratio between the sediment and hydrodynamic discharge and β is a constant. In most rivers far less sediment than water is transported so that $\epsilon \ll 1$. In our calculations we take $\epsilon = 0$, $\beta = 3$ and $\mathsf{F} = 0.1$.

An extra complication in this test case is matrix A in (57) since it is a singular matrix when $\epsilon = 0$. This is a problem when deriving the numerical flux and the wave speeds S_L and S_R. However, since we solve the system of algebraic equations in pseudo-time, we need the numerical flux on the space faces only in the space–time normal direction. To obtain the numerical flux on a fixed grid, note that the normal in the time direction is 0, so that, after augmenting with a pseudo-time derivative, (57) is changed to

$$\partial_\tau U_r + F_{ik,k} + G_{ikr}U_{r,k} = 0. \tag{58}$$

The numerical flux is then determined in the space normal direction to a face (see [21,22]). For one-dimensional numerical examples solving (57) including convergence rates with space and space–time DGFEM we refer to Tassi et al. [21].

In this test case we consider hydraulic and sediment transport through a contraction. The mesh considered is given in Fig. 11. In Tassi et al. [21] we show results of this test case using space DGFEM and here we use space–time DGFEM. The physical time step is $\Delta t = 0.0001$. For the pseudo-time stepping, the pseudo-time CFL number is $\mathrm{CFL}^{\mathrm{pseudo}} = 0.8$. Furthermore, if residuals converged to a tolerance of 10^{-6} in the pseudo-time integration, we considered the system to be solved. In Figs. 12–14 we show the mass

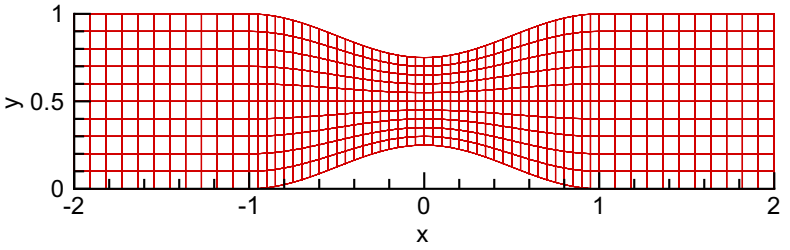


Fig. 11. Test case 8: the mesh.

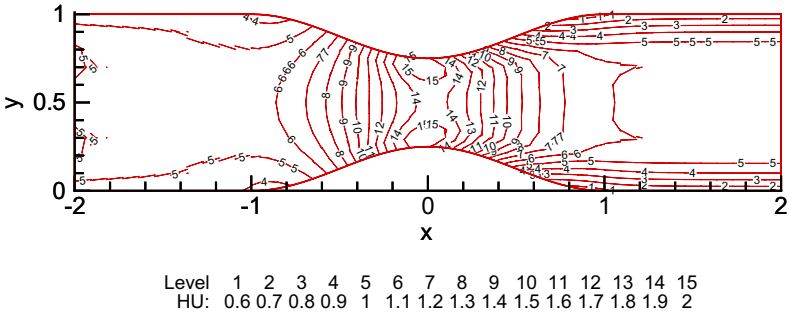


Fig. 12. Test case 8: flow and sediment transport in a contraction channel: mass flow $hu(x)$ at time $t = 0.005$.

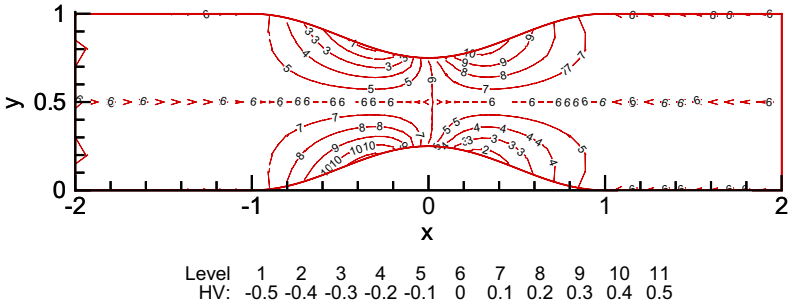


Fig. 13. Test case 8: flow and sediment transport in a contraction channel: mass flow $hv(x)$ at time $t = 0.005$.

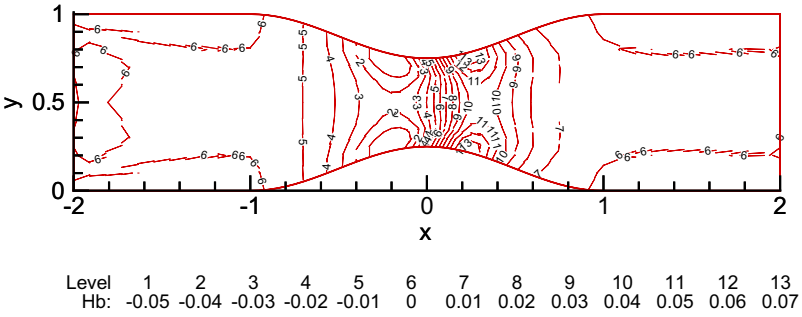


Fig. 14. Test case 8: flow and sediment transport in a contraction channel: bottom profile $b(x)$ at time $t = 0.005$.

flow hu, hv and the bed elevation b at time $t = 0.005$ which in physical time corresponds to a few months. As in Kubatko et al. [13], we observe that the bed experiences erosion in the converging part of the channel due to an increase in the flow velocity and the development of a mound in the diverging part of the channel. The results compare qualitatively well with those presented [13] and are the same as we obtained using space DGFEM in Tassi et al. [21].

5.3. The depth-averaged two-fluid model

In this section we consider two-fluid models (also known as Eulerian models) in which the particle phase is treated as a continuum by averaging over individual particles. Two frequently used models for two-fluid equations, are those derived by Anderson and Jackson [1], and Drew and Lahey [6] and Enwald et al. [7]. Apart from their derivation, the difference between these systems of equations is how the fluid-phase shear stress (if included) is multiplied by the solid volume fraction in the momentum equations (see also [26]). In the limiting case that pressure is the only fluid stress, both formulations are equal.

We will consider a simplification of these equations, namely the depth-averaged two-fluid model derived by Pitman and Le [18]. They start with the system of Anderson and Jackson [1] and use the shallow flow assumption, $H/L \ll 1$, where H is the characteristic length of the flow in the z-direction and L the characteristic length of the flow in the y-direction. The derivation is similar to the way the shallow water equations are derived from the Navier–Stokes equations. Since the pressure is the only fluid stress, the same depth-averaged two-fluid model also follows from the system derived by Drew and Lahey [6] and Enwald et al. [7].

The dimensionless depth-averaged two-fluid model of Pitman and Le [18], ignoring source terms for simplicity, can be written as

$$U_{i,0} + F_{i,1} + G_{ij}U_{j,1} = 0 \quad \text{for } i,j = 1,2,3,4, \tag{59}$$

where

$$U = \begin{bmatrix} h(1-\alpha) \\ h\alpha \\ h\alpha v \\ hu(1-\alpha) \\ b \end{bmatrix}, \quad F = \begin{bmatrix} h(1-\alpha)u \\ h\alpha v \\ h\alpha v^2 + \frac{1}{2}\varepsilon(1-\rho)\alpha_{xx}gh^2\alpha \\ hu^2 + \frac{1}{2}\varepsilon gh^2 \\ 0 \end{bmatrix},$$
$$G(U) = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \varepsilon\rho\alpha gh & \varepsilon\rho\alpha gh & 0 & 0 & \varepsilon(1-\rho)\alpha_{xx}gh\alpha + \varepsilon\rho\alpha gh \\ \frac{2u^2\alpha}{1-\alpha} - \alpha u^2 - \varepsilon gh\alpha & -\varepsilon gh\alpha - \alpha u^2 & u(\alpha-1) & u\alpha - \frac{2u\alpha}{1-\alpha} & (1-\alpha)\varepsilon gh \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{60}$$

Again we have taken the topography b as unknown. The meaning of the different symbols are: $h(x,t)$ is the depth of the flow, $v(x,t)$ the velocity of the solid phase, $u(x,t)$ the velocity of the fluid phase, $\alpha(x,t)$ the volume fraction of the solid phase, $b(x)$ the topography term, $\varepsilon = H/L$, ρ is the ratio between the fluid density and the solid density, $\alpha_{xx} = k_{ap}$, where k_{ap} is the Earth pressure coefficient and g is the z-component of the scaled gravity. Note that in the limit $\alpha \to 0$, this model reduces to the shallow water equations with εg akin to F^{-2}:

$$\begin{aligned} &\partial_t h + \partial_x(hu) = 0, \\ &\partial_t(hu) + \partial_x\left(hu^2 + \frac{1}{2}\varepsilon gh^2\right) = -\varepsilon gh\partial_x b. \end{aligned} \tag{61}$$

In the limit $\alpha \to 1$, the depth-averaged two-fluid model reduces to

$$\begin{aligned} &\partial_t h + \partial_x(hv) = 0, \\ &\partial_t(hv) + \partial_x\left(hv^2 + \frac{1}{2}\varepsilon k_{ap} gh^2\right) = -\varepsilon k_{ap} gh\partial_x b, \end{aligned} \tag{62}$$

which is the Savage–Hutter model without source terms, a model that simulates avalanches of dry granular matter [10].

In our simulations, we set the Earth pressure coefficient to be $\alpha_{xx} = 1$ and take $\varepsilon = 1$. To compute the eigenvalues of $\partial F/\partial U + G(U)$, we use the LAPACK package. The biggest eigenvalue is used for S_R and the smallest eigenvalue is used for S_L in the NCP numerical flux.

5.3.1. Test case 9: Two-phase subcritical flow

As in the case of the shallow water equations with topography, also for the two-phase flow model we consider the steady-state solution for subcritical flow over a bump. We consider the same topography (53). The reference solution is found by solving

$$\partial_x U = A^{-1}S, \tag{63}$$

where U, A and S are given by

$$U = [h(1-\alpha), \quad h\alpha]^{\mathrm{T}}, \quad S = \begin{bmatrix} -(1-\alpha)hg\partial_x b \\ -gh\alpha\partial_x b \end{bmatrix}$$
$$A = \begin{bmatrix} u^2(1-\alpha) - 2u^2 + gh(1-\alpha) & u^2(1-\alpha) + gh(1-\alpha) \\ \frac{1}{2}(1+\rho)gh\alpha & \frac{1}{2}(1-\rho)g(1+\alpha)h + g\rho h\alpha - v^2 \end{bmatrix} \tag{64}$$

with the topography derivative a known function and steady-state discharges:

$$hu(1-\alpha) = q_1, \quad hv\alpha = q_2 \tag{65}$$

with q_1 and q_2 integration constants. Here we take $q_1 = 0.2$, $q_2 = 0.1$, $g = 1$ and $\rho = 0.5$ and as initial condition $h(1-\alpha) = 1$, $h\alpha = 0.6$, $hu(1-\alpha) = 0.2$ and $hv\alpha = 0.1$. We use the STDGFEM formulation to calculate the solution. We consider one physical time step of $\Delta t = 10^{21}$ and use a pseudo-time stepping integration method to solve the system of nonlinear equations. We determine the solution on a domain $x \in [0, 20]$ divided into 40, 80, 160 and 320 cells. As stopping criterium in the pseudo-time stepping method we take that the maximum residual must be smaller that 10^{-8}. The pseudo-time stepping CFL number is $\mathrm{CFL}^{\mathrm{pseudo}} = 0.1$. At the boundaries, we define the exterior trace to be the same as the initial condition. The numerical flux decides then what to do with this information. The steady-state solution is given in Fig. 15. The order of convergence is determined by computing the L^2 and $L^{\max}$ norm of the error, similar as to what is done in (55) and (56). The order of convergence is given in Table 5. Using linear basis functions, we obtain second-order convergence as expected.

5.3.2. Test case 10: Two-phase supercritical flow

We will now consider the steady-state solution of two-phase supercritical flow over a bump with (53) as topography. The exact solution is found by solving (63)–(65), now with $q_1 = 4$ and $q_2 = 2$. Other constants

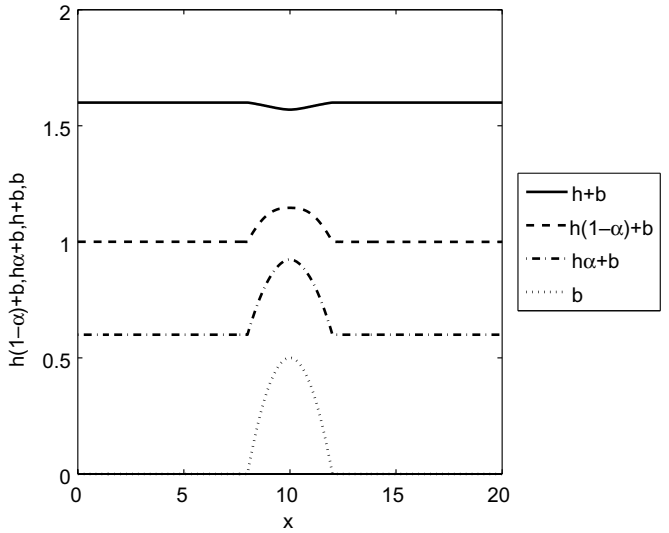


Fig. 15. Test case 9: steady-state solution for a subcritical two-phase flow calculated with STDGFEM using 320 cells. Shown are the total flow height $h + b$, the flow height due to the fluid phase $h(1-\alpha)$, the flow height due to the solids phase $h\alpha$ and the topography b.

Table 5
L^2 and $L^{\max}$ error for $h(1-\alpha)+b$, $h\alpha+b$, $hu(1-\alpha)$ and $hv\alpha$ using STDGFEM for test case 9

N_{cells}	$h(1-\alpha)+b$				$h\alpha+b$			
	L^2 error	p	$L^{\max}$ error	p	L^2 error	p	$L^{\max}$ error	p
STDGFEM								
40	0.8171×10^{-3}	–	0.2308×10^{-2}	–	0.1404×10^{-2}	–	0.4194×10^{-2}	–
80	0.2025×10^{-3}	2.0	0.5584×10^{-3}	2.0	0.3537×10^{-3}	2.0	0.9903×10^{-3}	2.1
160	0.4871×10^{-4}	2.1	0.1322×10^{-3}	2.1	0.8511×10^{-4}	2.1	0.2306×10^{-3}	2.1
320	0.9789×10^{-5}	2.3	0.2651×10^{-4}	2.3	0.1712×10^{-4}	2.3	0.4597×10^{-4}	2.3
	$hu(1-\alpha)$				$hv(\alpha)$			
40	0.3672×10^{-4}	–	0.1442×10^{-3}	–	0.1212×10^{-4}	–	0.3409×10^{-4}	–
80	0.5911×10^{-5}	2.6	0.3448×10^{-4}	2.1	0.1791×10^{-5}	2.8	0.8054×10^{-5}	2.1
160	0.1049×10^{-5}	2.5	0.8471×10^{-5}	2.0	0.3807×10^{-6}	2.2	0.2048×10^{-5}	2.0
320	0.1723×10^{-6}	2.6	0.2078×10^{-5}	2.0	0.5115×10^{-7}	2.9	0.4861×10^{-6}	2.1

remain as in test case 9 and we use the same solution strategy. The steady-state solution is given in Fig. 16 and the order of convergence is given in Table 6. Again, using linear basis functions, we obtain second-order convergence for the variables $h(1-\alpha)+b$ and $h\alpha+b$. We do not see second-order convergence for the variables $hu(1-\alpha)$ and $hv\alpha$ because the error for these solutions stabilizes around 10^{-8}, the value of the maximum residual.

5.3.3. Test case 11: A two-phase dam break problem

For the depth-averaged two-phase flow model we consider a dam break type test case. Consider two mixtures separated by a membrane. The left mixture has a solid volume fraction of $\alpha = 0.4$ and the right mixture has a solid volume fraction of $\alpha = 0.6$. At time $t = 0$ we remove the membrane. We want to know how the mixtures behave. We consider the solution on the domain $[0,1]$. As initial condition we take $U(x,0) = U_L$ if $x < 0.5$ and $U(x,0) = U_R$ if $x > 0.5$, where $U_L = [1.8,\ 1.2,\ 0,\ 0,\ 0]^{\mathrm{T}}$ and $U_R = [1.2,\ 1.8,\ 0,\ 0,\ 0]^{\mathrm{T}}$. The constants in the computation are taken as $g = 1$ and $\rho = 0.5$. We compute the solution on a domain with 16, 32, 64, 128, 256, 512 or 1024 elements. We consider DGFEM calculations using the linear path $\phi = U_L + \tau(U_R - U_L)$. The solution is determined at $t = 0.175$ using a time step of $\Delta t = 0.0001$. The solutions of $h(1-\alpha)$, $h\alpha$, b and h are depicted in Fig. 17(a), the solutions of $hu(1-\alpha)$ and $hv\alpha$ are depicted in Fig. 17(b) and the solution of α is depicted in Fig. 17(c) in which we compare the solutions on a grid with 128 elements to the solutions computed on a grid with 10,000 elements. Apart from some small spurious oscillations obtained on the grid with 128 elements, the solutions compare very well with the solutions obtained on

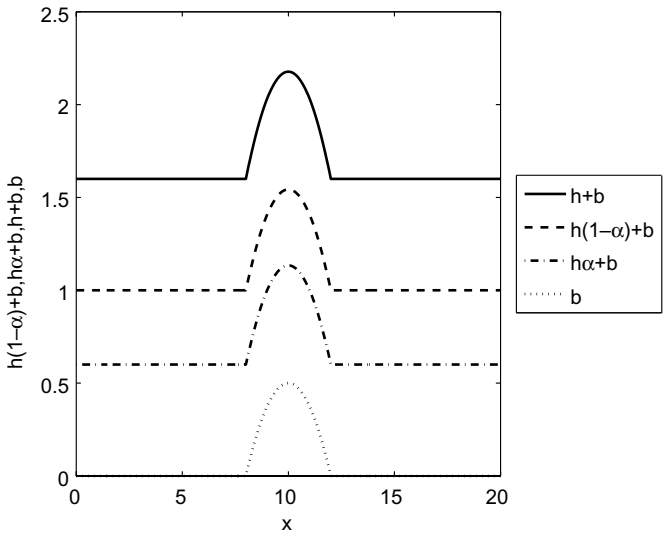
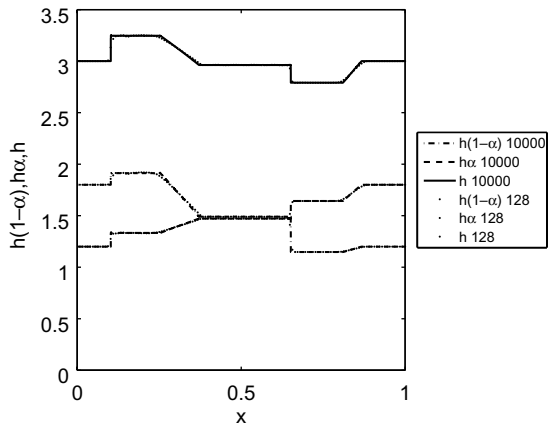


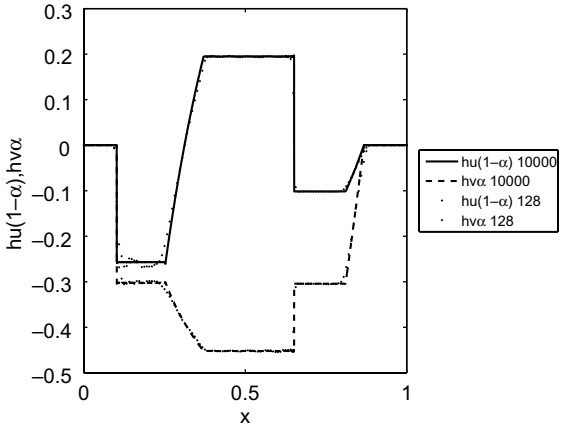
Fig. 16. Test case 10: steady-state solution for a supercritical two-phase flow calculated using STDGFEM using 320 cells. Shown are the total flow height $h + b$, the flow height due to the fluid phase $h(1-\alpha)$, the flow height due to the solids phase $h\alpha$ and the topography b.

Table 6
L^2 and $L^{\max}$ error for $h(1-\alpha)+b$, $h\alpha+b$, $hu(1-\alpha)$ and $hv\alpha$ using STDGFEM for test case 10

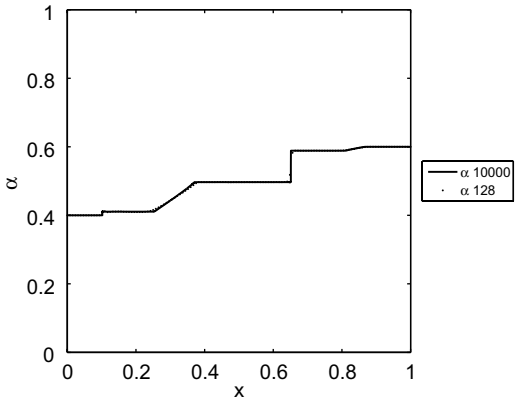
N_{cells}	$h(1-\alpha)+b$				$h\alpha+b$			
	L^2 error	p	$L^{\max}$ error	p	L^2 error	p	$L^{\max}$ error	p
STDGFEM								
40	0.2400×10^{-2}	–	0.5674×10^{-2}	–	0.2359×10^{-2}	–	0.5575×10^{-2}	–
80	0.6060×10^{-3}	2.0	0.1402×10^{-2}	2.0	0.5958×10^{-3}	2.0	0.1378×10^{-2}	2.0
160	0.1459×10^{-3}	2.1	0.3339×10^{-3}	2.1	0.1434×10^{-3}	2.1	0.3280×10^{-3}	2.1
320	0.2933×10^{-4}	2.3	0.6678×10^{-4}	2.3	0.2884×10^{-4}	2.3	0.6561×10^{-4}	2.3



(a) Solution of $h(1-\alpha)$, $h\alpha$, b and h.



(b) Solution of $hu(1-\alpha)$ and $hv\alpha$.



(c) Solution of α.

Fig. 17. Test case 11. The solution computed on a mesh with 128 elements compared to the solution computed on a mesh with 10000 elements at time $t = 0.175$ using space DGFEM.

the grid with 10000 elements. Since we do not have an exact solution, we compute the order behavior using the following approach:

$$\frac{\|U_N - U_{2N}\|_2}{\|U_{2N} - U_{4N}\|_2} = 2^p, \tag{66}$$

where p is the order of convergence, U_N the solution on a mesh consisting of N cells, and $\|\cdot\|_2$ is the L^2 norm. The order behavior is shown in Table 7. Due to the presence of shocks we cannot obtain second-order accuracy. Instead we obtain a convergence rate of approximately $\mathcal{O}(h^{1/2})$.

Table 7
L^2 error and convergence rate for $h(1-\alpha)$, $h\alpha$, $hu(1-\alpha)$ and $hv\alpha$ using DGFEM for test case 11

N_{cells}	L^2 of $h(1-\alpha)$	p	L^2 of $h\alpha$	p	L^2 of $hu(1-\alpha)$	p	L^2 of $hv\alpha$	p
DGFEM								
32	0.1238×10^{-1}	–	0.7030×10^{-2}	–	0.1263×10^{-1}	–	0.1384×10^{-1}	–
64	0.1125×10^{-1}	0.1	0.5780×10^{-2}	0.3	0.1155×10^{-1}	0.1	0.8164×10^{-2}	0.8
128	0.6231×10^{-2}	0.9	0.3391×10^{-2}	0.8	0.7114×10^{-2}	0.7	0.4465×10^{-2}	0.9
256	0.4379×10^{-2}	0.5	0.2751×10^{-2}	0.3	0.4494×10^{-2}	0.7	0.3828×10^{-2}	0.2
512	0.3085×10^{-2}	0.5	0.1875×10^{-2}	0.6	0.3536×10^{-2}	0.3	0.3275×10^{-2}	0.2

The convergence rates are shown for the solution at $t = 0.175$. With L^2 of U we mean $\|U_N - U_{2N}\|_2$.

6. Effect of the path in phase space on the numerical solution

6.1. Polynomial paths

In the numerical test cases discussed in the previous section a linear path was taken: $\phi = U^L + \tau(U^R - U^L)$. In this section, we will investigate the effect of different paths on our numerical results. To determine this effect we again consider test case 11 in Section 5.3 for which we expect to find the biggest effect of the path due to the shock waves in the solution. We use the following paths and note that in one dimension property (H4) can be neglected:

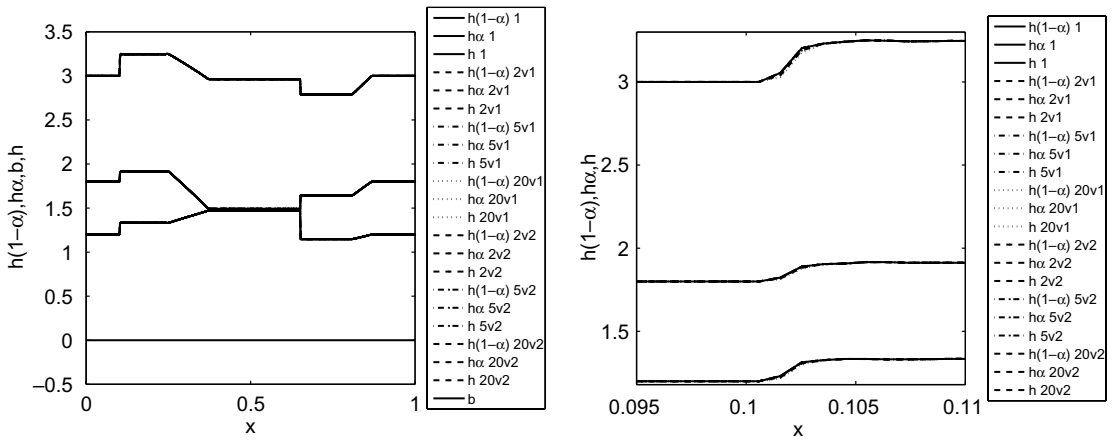
$$\begin{aligned}
&\phi_{2v1} = U^L + \tau^2(U^R - U^L), \quad \phi_{2v2} = U^R + (1-\tau)^2(U^L - U^R),\\
&\phi_{5v1} = U^L + \tau^5(U^R - U^L), \quad \phi_{5v2} = U^R + (1-\tau)^5(U^L - U^R),\\
&\phi_{20v1} = U^L + \tau^{20}(U^R - U^L), \quad \phi_{20v2} = U^R + (1-\tau)^{20}(U^L - U^R).
\end{aligned} \tag{67}$$

In Fig. 18, $h(1-\alpha)$, $h\alpha$, b and h are shown on the whole domain and also a zoom-in on the left shock wave. The deviations shown in these figures are approximately also seen in the mass flow variables and the void fraction.

In these computations it is important to have a good numerical integration scheme to approximate the path integral. Incorrectly approximating the path integral results in solutions having incorrect faster or slower shock speeds. A two-point Gauss integration scheme is sufficient when taking ϕ linear or when using ϕ_{2v1} and ϕ_{2v2}. For the other paths we split the domain $[0,1]$ into eight nonintersecting uniform intervals and within each interval we evaluate the integral in the two Gauss points corresponding to that particular interval. To conclude for this test case, when properly integrated any choice of paths in (67) leads to the same numerical solution with only minor differences.

6.2. Toumi paths

In this section we will consider paths similar to those chosen in Toumi [24]. These paths are different from those of the previous section in that these paths are C^0. We will compare the solutions determined with the following five paths with the solution determined with a linear path:



(a) The solution on the whole domain.

(b) The solution zoomed in on the left shock wave.

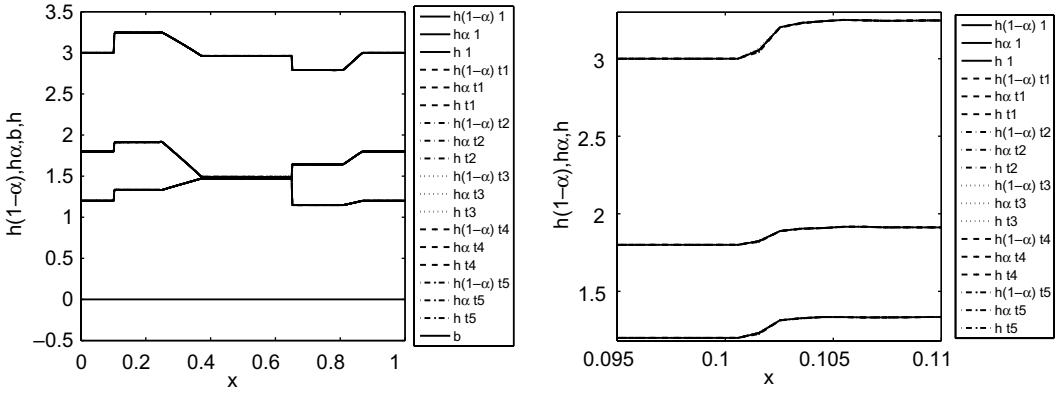
Fig. 18. Solution of $h(1-\alpha)$, $h\alpha$, b and h calculated on a mesh with 1024 elements at time $t = 0.175$ using the paths defined in (67).

$$
\begin{aligned}
\phi_{T1}(\tau;U^L,U^R) &= \begin{cases} (U_1^L + 2\tau(U_1^R - U_1^L), U_2^L, U_3^L, U_4^L, U_5^L), & \text{for } \tau \in \left[0,\tfrac{1}{2}\right], \\ (U_1^R, U_2^L + (2\tau-1)(U_2^R - U_2^L), U_3^L + (2\tau-1)(U_3^R - U_3^L) & \\ U_4^L + (2\tau-1)(U_4^R - U_4^L), U_5^L + (2\tau-1)(U_5^R - U_5^L)) & \text{for } \tau \in \left[\tfrac{1}{2},1\right], \end{cases} \\
\phi_{T2}(\tau;U^L,U^R) &= \begin{cases} (U_1^L, U_2^L + 2\tau(U_2^R - U_2^L), U_3^L, U_4^L, U_5^L) & \text{for } \tau \in \left[0,\tfrac{1}{2}\right], \\ (U_1^L + (2\tau-1)(U_1^R - U_1^L), U_2^R, U_3^L + (2\tau-1)(U_3^R - U_3^L), & \\ U_4^L + (2\tau-1)(U_4^R - U_4^L), U_5^L + (2\tau-1)(U_5^R - U_5^L)) & \text{for } \tau \in \left[\tfrac{1}{2},1\right], \end{cases} \\
\phi_{T3}(\tau;U^L,U^R) &= \begin{cases} (U_1^L, U_2^L, U_3^L + 2\tau(U_3^R - U_3^L), U_4^L, U_5^L) & \text{for } \tau \in \left[0,\tfrac{1}{2}\right], \\ (U_1^L + (2\tau-1)(U_1^R - U_1^L), U_2^L + (2\tau-1)(U_2^R - U_2^L), U_3^R, & \\ U_4^L + (2\tau-1)(U_4^R - U_4^L), U_5^L + (2\tau-1)(U_5^R - U_5^L)) & \text{for } \tau \in \left[\tfrac{1}{2},1\right], \end{cases} \\
\phi_{T4}(\tau;U^L,U^R) &= \begin{cases} (U_1^L, U_2^L, U_3^L, U_4^L + 2\tau(U_4^R - U_4^L), U_5^L) & \text{for } \tau \in \left[0,\tfrac{1}{2}\right], \\ (U_1^L + (2\tau-1)(U_1^R - U_1^L), U_2^L + (2\tau-1)(U_2^R - U_2^L), & \\ U_3^L + (2\tau-1)(U_3^R - U_3^L), U_4^R, U_5^L + (2\tau-1)(U_5^R - U_5^L)) & \text{for } \tau \in \left[\tfrac{1}{2},1\right], \end{cases} \\
\phi_{T5}(\tau;U^L,U^R) &= \begin{cases} (U_1^L, U_2^L, U_3^L, U_4^L, U_5^L + 2\tau(U_5^R - U_5^L)) & \text{for } \tau \in \left[0,\tfrac{1}{2}\right], \\ (U_1^L + (2\tau-1)(U_1^R - U_1^L), U_2^L + (2\tau-1)(U_2^R - U_2^L), & \\ U_3^L + (2\tau-1)(U_3^R - U_3^L), U_4^L + (2\tau-1)(U_4^R - U_4^L), U_5^R) & \text{for } \tau \in \left[\tfrac{1}{2},1\right]. \end{cases}
\end{aligned} \tag{68}
$$

In the implementation the integrals are computed using a two-point Gauss integration rule. In Fig. 19, $h(1-\alpha)$, $h\alpha$, b and h are shown on the whole domain and also zoomed in on the left shock wave. The deviations shown in these figures are approximately also seen in the mass flow variables and the void fraction. We see that the final solution determined with the paths given in (68) are all very similar. The choice of one of these paths does not have a big effect on the final solution compared to the linear path.

6.3. Refining the mesh

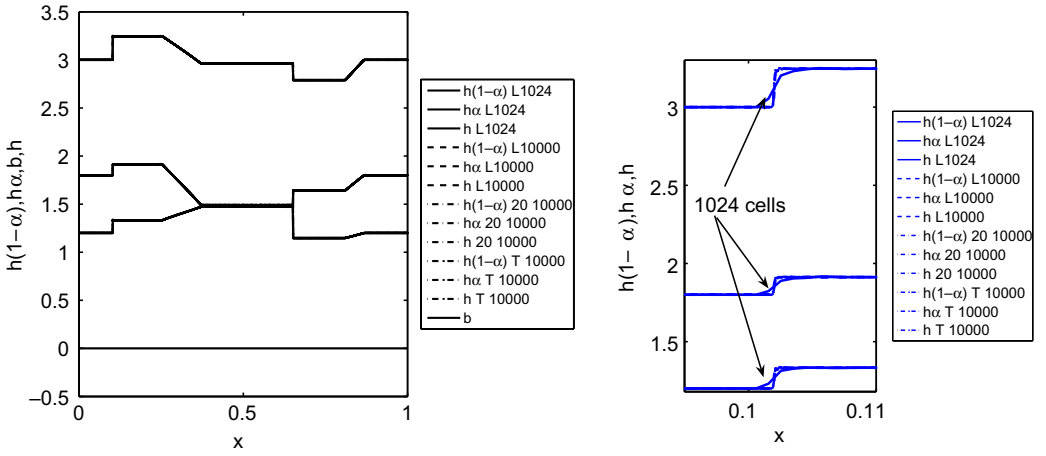
As a final check we further refine our mesh. We will calculate the solution on a mesh with 10,000 elements. We only do this for the linear path, ϕ_{20v1} (see (67)) and ϕ_{T1} (see (68)) and compare these solutions with the numerical solution determined with the linear path on a mesh with 1024 elements. In Fig. 20, $h(1-\alpha)$, $h\alpha$, b



(a) The solution on the whole domain.

(b) The solution zoomed in on the left shock wave.

Fig. 19. Solution of $h(1-\alpha)$, $h\alpha$, b and h calculated on a mesh with 1024 elements at time $t = 0.175$ using the paths defined in (68).



(a) The solution on the whole domain.

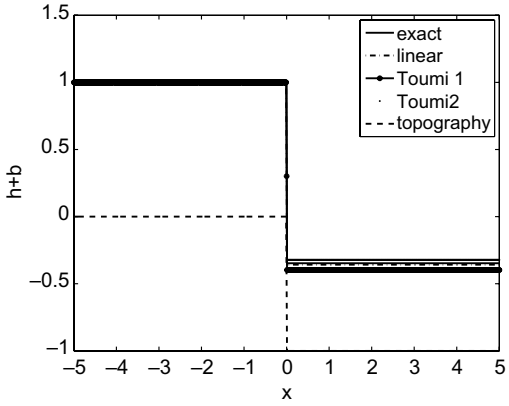
(b) The solution zoomed in on the left shock wave.

Fig. 20. Solution of $h(1-\alpha)$, $h\alpha$, b and h calculated on a mesh with 10,000 elements at time $t = 0.175$ using the linear path, ϕ_{20v1} and ϕ_{T1}.

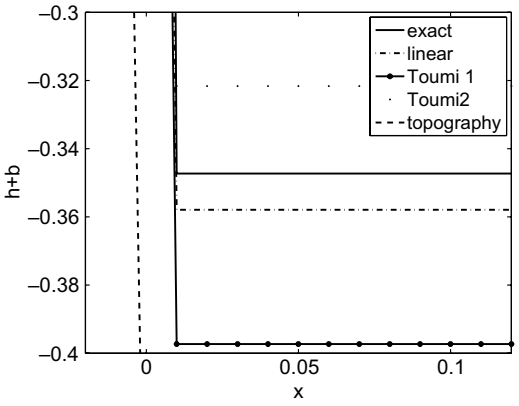
and h are shown on the whole domain and also zoomed in on the left shock wave. The deviations shown in these figures are approximately also seen in the mass flow variables and the void fraction. To obtain these figures, the integral of the nonconservative product for each path was evaluated differently. For the linear path a two-point Gauss integration scheme was used for the whole domain $[0,1]$. For the path ϕ_{20v1} we divided the domain $[0,1]$ into 16 nonintersecting uniform domains and within each domain we used again a two-point Gauss integration scheme. For the path ϕ_{T1} the domain $[0,1]$ was divided into eight nonintersecting uniform domains and within each domain we used a two-point Gauss integration scheme. As we see in these figures, the differences in the numerical solution for all the paths are minimal. The slight differences in the shock speed are more likely to be caused by the numerical integration scheme than the difference in the path. If we were to determine the numerical solution using the path ϕ_{20v1} by dividing the domain $[0,1]$ into eight nonintersecting uniform domains instead of 16, the differences in shock speed in comparison to the other paths will increase, so it is important to have a good approximation for the integral of the nonconservative product. We conclude that it is important to have a good numerical integration scheme to approximate the path integral. Using a linear path, a two-points Gauss integration scheme, without refinement, suffices. We saw that it does not matter which path is chosen, but choosing the linear path, due to the simple integration scheme, is by far the cheapest and easiest choice.

6.4. Contact waves

In Parés and Castro [17] a test case is presented for the shallow water equations in which they state that the selection of the path is critical in order to satisfactorily capture stationary contact discontinuities related to bottom discontinuities (see [17]). We repeat this test case. Consider the shallow water equations given by (45) and (46). Following Parés and Castro [17], the initial condition is given by $U = U^L$ if $x \leqslant 0$ and $U = U^R$ if $x > 0$, where $U^L = [0,\ 1,\ \sqrt{2g}]^{\mathrm{T}}$ and $U^R = [-1,\ 0.6527036446614,\ \sqrt{2g}]^{\mathrm{T}}$ if $x > 0$, where $g = 9.81$. These initial conditions are such that the states U^L and U^R are connected by an entropic contact discontinuity (see [17]). The boundary conditions are given by: $b(-5,t) = 0$, $h(-5,t) = 1$, $hu(-5,t) = \sqrt{2g}$, $b(5,t) = -1$. The remaining boundary data are set equal to the data calculated at the boundary from inside the domain. The steady-state solution is calculated using a physical time step of $\Delta t = 10^{21}$ and a pseudo-CFL number of $\mathrm{CFL}^{\mathrm{pseudo}} = 0.8$. We consider the solution on a mesh with 1000 elements on which the contact wave falls exactly on a face, and on a mesh with 999 elements so that the contact wave falls exactly in the middle of an element. The effect of three paths are considered, namely the linear path $\phi(\tau; U^L, U^R) = U^L + \tau(U^R - U^L)$ and two Toumi like paths:



(a) Comparison of the computed water level at the stationary state with the exact solution.



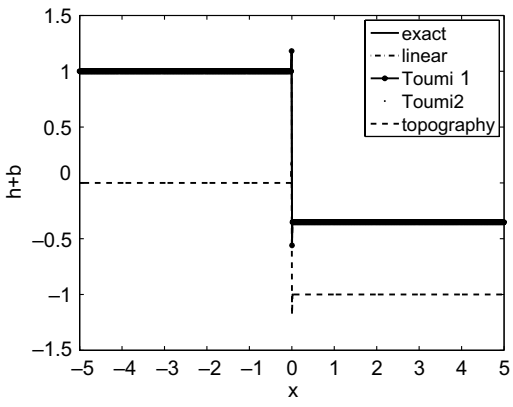
(b) Zoom of the water level right of the discontinuity in the topography.

Fig. 21. A comparison of the computed solution of a contact discontinuity related to the discontinuous topography with the exact solution. The solution was computed on a mesh with 1000 elements using a physical time step of $\Delta t = 10^{21}$ and a pseudo-CFL number $\mathrm{CFL}^{\mathrm{pseudo}} = 0.8$.

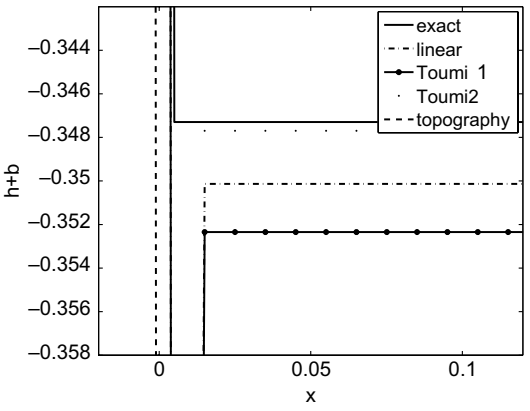
$$\phi_{T1}(\tau; U^L, U^R) = \begin{cases} (U_1^L + 2\tau(U_1^R - U_1^L), U_2^L) & \text{for } \tau \in [0, \frac{1}{2}], \\ (U_1^R, U_2^L + (2\tau - 1)(U_2^R - U_2^L)) & \text{for } \tau \in [\frac{1}{2}, 1], \end{cases}$$

$$\phi_{T2}(\tau; U^L, U^R) = \begin{cases} (U_1^L, U_2^L + 2\tau(U_2^R - U_2^L)) & \text{for } \tau \in [0, \frac{1}{2}], \\ (U_1^L + (2\tau - 1)(U_1^R - U_1^L), U_2^R) & \text{for } \tau \in [\frac{1}{2}, 1]. \end{cases}$$

Note that the path for $U_3 = hu$ is irrelevant since the nonconservative product for the shallow water equations only involve b and h. The solution on the mesh with 1000 elements is shown in Fig. 21 and the solution on the mesh with 999 elements is shown in Fig. 22. We see that the solution of a steady contact discontinuity experiences a similar dependence on the path as observed by Parés and Castro [17], also after refining the mesh to 10,000 and 9999 elements, respectively. The numerical dissipation introduced when the contact discontinuity is not exactly at an element face has, however, a strong regularizing effect (compare Figs. 21 and 22) and significantly reduces the dependence of the solution on the path. This effect will even be stronger in multi-dimen-



(a) Comparison of the computed water level at the stationary state with the exact solution.



(b) Zoom of the water level right of the discontinuity in the topography.

Fig. 22. A comparison of the computed solution of a contact discontinuity related to the discontinuous topography with the exact solution. The solution was computed on a mesh with 999 elements using a physical time step of $\Delta t = 10^{21}$ and a pseudo-CFL number $\mathrm{CFL}^{\mathrm{pseudo}} = 0.8$.

sional problems since the discontinuities then rarely coincide with mesh lines, but it is always important to check the dependence of the solution on the chosen path.

7. Conclusions

In this article we have derived weak formulations for space- and space–time DGFEM for nonconservative hyperbolic partial differential equations. We also introduced a numerical flux for systems with nonconservative products (NCP-flux) suitable for DGFEM.

As test cases we considered the shallow water equations with and without dynamic topography (1D and 2D) and a simplified depth-averaged two-phase flow model. For the shallow water equations we considered rest flow over discontinuous topography and showed, both numerically and theoretically, that rest flow is preserved. We also considered subcritical and supercritical flow over a bump. For these test cases we obtained second- and third-order accuracy for suitable basis functions. We also considered more complex test cases: steady-state transcritical flow with a shock, a perturbation of a steady-state solution over a discontinuous topography and a dam breaking problem over a rectangular bump. For the two-dimensional shallow water equations with dynamic topography, we considered hydraulic and morphological transport through a contraction.

For the simplified depth-averaged two-phase flow model we also considered subcritical and supercritical flow over a bump and again obtained second-order accuracy using linear basis functions. A dam break test case was further used to investigate the effect of the path on the numerical solution. The effect of the path was very small in the numerical solutions. Taking different paths did not lead to relevant changes in the final solution. We did see, however, that for certain paths it is not sufficient to simply use a two-point Gauss integration scheme over the whole domain of integration for the path integral, but higher order integration rules were required. It resulted in significantly larger computational cost which is undesirable.

Finally, we examined the effect of the path across a contact wave and saw that we could not capture the stationary contact discontinuity. By making the mesh such that the contact wave falls within an element we did see that the numerical error made is a full-order smaller than if the contact wave falls exactly on a face. The numerical dissipation has a regularizing effect decreasing the effect of the path, but at the moment it is still unclear how to choose the path in case of a contact discontinuity and this is a topic of further research. The regularizing effect due to numerical dissipation across shock waves is much larger explaining why we did not experience any significant effect of the path in test cases containing shock waves.

Acknowledgements

We would like to thank C.M. Klaij (Twente) who contributed to the development of the numerical flux for systems with nonconservative products. O.B. acknowledges a fellowship from The Royal Netherlands Academy of Arts and Sciences (KNAW), and S.R. support from the Institute of Mechanics, Processes and Control Twente. For J.V. this research was partly funded by the ICT project BRICKS (http://www.bsik-bricks.nl), theme MSV1.

Appendix A. Derivation of the weak formulation for space DGFEM

In this section we derive a space DGFEM weak formulation for hyperbolic nonconservative partial differential equations (see also e.g. [3] for more on the Runge–Kutta discontinuous Galerkin method for conservative hyperbolic systems). As opposed to the derivation of the weak formulation for space–time DGFEM, we now only consider fixed grids. We first introduce the function spaces and basis functions after which we derive the weak formulation.

Let $\Omega \subset \mathbb{R}^q$ be the bounded flow domain approximated by Ω_h such that $\Omega_h \to \Omega$ as $h \to 0$, with h the radius of the smallest sphere completely containing the largest element K_j. Consider approximations of $U(x,t)$ and the test function $V(x,t)$ in the finite element space defined as

$$W_h = \left\{ V \in (L^2(\Omega_h))^m : V|_{K_j} \circ F_K \in (P^p(\widehat{K}))^m \right\}, \tag{A.1}$$

where m denotes the dimension of U. Polynomial approximations for the trial function U and the test function V in each element K_j are introduced:

$$U(t,\bar{x})|_{K_j} = \widehat{U}_m \psi_m(\bar{x}) \quad \text{and} \quad V(t,\bar{x})|_{K_j} = \widehat{V}_l \psi_l(\bar{x})$$

for $m,l = 0,1,2,\ldots,M$, where M depends on the order of accuracy and the space dimension, and where the basis functions ψ are given by

$$\psi_m = \begin{cases} 1 & \text{for } m = 0 \\ \varphi_m(\bar{x}) - \frac{1}{|K_j|}\int_{K_j} \varphi_m(\bar{x})\mathrm{d}K & \text{for } m = 1,2,\ldots,M, \end{cases}$$

where the functions φ in element K_j are related to the basis functions $\hat{\varphi}$ on the master element $\widehat{K}$ through the mapping F:

$$\varphi_m = \hat{\varphi}_m \circ F_K^{-1}$$

with $\hat{\varphi}_m(\xi) \in P^p(\widehat{K})$ and ξ the local coordinates in the master element $\widehat{K}$.

The weak formulation for space DGFEM can be derived in a similar manner as that for space–time DGFEM, except that now we consider fixed grids. Before discussing the space DGFEM weak formulation for equations containing nonconservative products, we first introduce as a reference the space DGFEM weak formulation for equations in conservative form (see e.g. [20]).

Consider partial differential equations in conservative form:

$$U_{i,0} + H_{ik,k} = 0, \quad \bar{x} \in \mathbb{R}^q, \quad t > 0, \tag{A.2}$$

where $U \in \mathbb{R}^m$ and $H \in \mathbb{R}^m \times \mathbb{R}^q$. Using the approach discussed in Tassi et al. [20], the space DG formulation for (A.2) can be stated as:

Find a $U \in W_h$ such that for all $V \in W_h$:

$$0 = \sum_j \int_{K_j} (V_i U_{i,0} - V_{i,k} H_{ik})\mathrm{d}K + \sum_{\mathcal{S}\in\mathcal{S}_I} \int_{\mathcal{S}} [\![V_i]\!]_k \{\!\{H_{ik}\}\!\}\mathrm{d}\mathcal{S} + \sum_{\mathcal{S}\in\mathcal{S}_B} \int_{\mathcal{S}} V_i^L H_{ik}^L \bar{n}_k^L \,\mathrm{d}\mathcal{S}. \tag{A.3}$$

Note that at this point no numerical fluxes have been introduced yet into the DG formulation. We now continue with equations containing nonconservative products. Let $U \in W_h$ (see (A.1)). We know that the numerical solution is continuous on an element and discontinuous across a face, so, using Theorem 2, U is a weak solution to (3) if

$$0 = \int_{\Omega_h} V_i U_{i,0}\,\mathrm{d}K + \int_{\Omega_h} V_i\,\mathrm{d}\bar{\mu}_i \tag{A.4}$$

$$= \sum_j \int_{K_j} V_i (U_{i,0} + D_{ikr} U_{r,k})\mathrm{d}\mathcal{K} + \sum_{\mathcal{S}\in\mathcal{S}_I} \int_{\mathcal{S}} \widehat{V}_i \left(\int_0^1 D_{ikr}(\phi(\tau;U^L,U^R)) \frac{\partial \phi_r}{\partial \tau}(\tau;U^L,U^R)\mathrm{d}\tau\, \bar{n}_k^L \right) \mathrm{d}\mathcal{S}, \tag{A.5}$$

where $V \in W_h$ is an arbitrary test function. Furthermore, $\widehat{V}$ is the value (numerical flux) of the test function V on a face $\mathcal{S}$. Note that Theorem 2 is applied to nonconservative products in space–time where space and time variables are not explicitly distinguished. In space DGFEM this is the case and we only need the space part of the measure in Theorem 2. This measure is denoted in (A.4) as $\bar{\mu}_i$. The crucial point in obtaining the DG formulation is the choice of the numerical flux for the test function V. Using $D_{ikr} = \partial F_{ik}/\partial\, U_r + G_{ikr}$, (A.5) can be rewritten as

$$0 = \sum_j \int_{K_j} V_i (U_{i,0} + F_{ik,k} + D_{ikr} U_{r,k})\mathrm{d}\mathcal{K} - \sum_{\mathcal{S}\in\mathcal{S}_I} \int_{\mathcal{S}} \widehat{V}_i [\![F_{ik}]\!]_k \mathrm{d}\mathcal{S}$$
$$+ \sum_{\mathcal{S}\in\mathcal{S}_I} \int_{\mathcal{S}} \widehat{V}_i \left(\int_0^1 D_{ikr}(\phi(\tau;U^L,U^R)) \frac{\partial \phi_r}{\partial \tau}(\tau;U^L,U^R)\mathrm{d}\tau\, \bar{n}_k^L \right) \mathrm{d}\mathcal{S}. \tag{A.6}$$

We choose the numerical flux for V such that if there exists a Q such that $G_{ikr} = \partial Q_{ik}/\partial U_r$, then the DG formulation for the system containing nonconservative products reduces to the conservative space DGFEM weak formulation given by (A.3) with $H_{ik} = F_{ik} + Q_{ik}$.

Theorem 4. *If the numerical flux* $\widehat{V}$ *for the test function* V *in (A.6) is defined as* $\widehat{V} = \{\!\{V\}\!\}$, *then the weak formulation (A.6) will reduce to the conservative space DGFEM formulation (A.3) when there exists a* Q *such that* $G_{ikr} = \partial Q_{ik}/\partial U_r$ *so that* $H_{ik} = F_{ik} + Q_{ik}$.

Proof. Assume there is a Q such that $G_{ikr} = \partial Q_{ik}/\partial U_r$. We immediately see

$$\int_0^1 G_{ikr}(\phi(\tau; U^L, U^R))\frac{\partial \phi_r}{\partial \tau}(\tau; U^L, U^R)\mathrm{d}\tau \bar{n}_k^L = -[\![Q_{ik}]\!]_k. \tag{A.7}$$

Integrating by parts the volume integral in (A.6) we obtain

$$0 = \sum_k \int_{K_k} (V_i U_{i,0} - V_{i,k}(F_{ik} + Q_{ik}))\mathrm{d}K + \sum_k \int_{\partial K_k} V_i^L (F_{ik}^L + Q_{ik}^L)\bar{n}_k^L \,\mathrm{d}(\partial K) - \sum_{\mathcal{S}\in\mathcal{S}_I} \int_{\mathcal{S}} \widehat{V}_i [\![F_{ik} + Q_{ik}]\!]_k \mathrm{d}\mathcal{S}. \tag{A.8}$$

We write $H_{ik} = F_{ik} + Q_{ik}$. Use relations (12) and (13) to write the element boundary integrals as face integrals:

$$\begin{aligned}\sum_j \int_{\partial K_j} V_i^L H_{ik}^L \bar{n}_k^L \,\mathrm{d}(\partial K) &= \sum_{\mathcal{S}\in\mathcal{S}_I} \int_{\mathcal{S}} [\![V_i H_{ik}]\!]_k \mathrm{d}\mathcal{S} + \sum_{\mathcal{S}\in\mathcal{S}_B} \int_{\mathcal{S}} V_i^L H_{ik}^L \bar{n}_k^L \,\mathrm{d}\mathcal{S} \\ &= \sum_{\mathcal{S}\in\mathcal{S}_I} \int_{\mathcal{S}} (\{\!\{V_i\}\!\}[\![H_{ik}]\!]_k + (V_i^L - V_i^R)\{\!\{H_{ik}\}\!\}\bar{n}_k^L)\mathrm{d}\mathcal{S} + \sum_{\mathcal{S}\in\mathcal{S}_B} \int_{\mathcal{S}} V_i^L H_{ik}^L \bar{n}_k^L \,\mathrm{d}\mathcal{S}.\end{aligned} \tag{A.9}$$

Combining (A.8) and (A.9) we obtain

$$\begin{aligned}0 = &\sum_j \int_{K_j} V_i U_{i,0} - V_{i,k} H_{ik} \mathrm{d}K + \sum_{\mathcal{S}\in\mathcal{S}_I} \int_{\mathcal{S}} (\{\!\{V_i\}\!\}[\![H_{ik}]\!]_k + (V_i^L - V_i^R)\{\!\{H_{ik}\}\!\}\bar{n}_k^L)\mathrm{d}\mathcal{S} \\ &+ \sum_{\mathcal{S}\in\mathcal{S}_B} \int_{\mathcal{S}} V_i^L H_{ik}^L \bar{n}_k^L \mathrm{d}\mathcal{S} - \sum_{\mathcal{S}\in\mathcal{S}_I} \int_{\mathcal{S}} \widehat{V}_i [\![H_{ik}]\!]_k \mathrm{d}\mathcal{S}.\end{aligned} \tag{A.10}$$

The term $\{\!\{V_i\}\!\}[\![H_{ik}]\!]_k$ is set to zero in the space DG formulation for conservative systems arguing that the formulation must be conservative. For a general nonconservative system we can not use this argument. Instead, we note that by taking $\widehat{V} = \{\!\{V\}\!\}$ on the faces $\mathcal{S}$, the contribution $\int_{\mathcal{S}} \{\!\{V_i\}\!\}[\![H_{ik}]\!]_k \mathrm{d}\mathcal{S}$ cancels with $-\int_{\mathcal{S}} \widehat{V}_i [\![H_{ik}]\!]_k \mathrm{d}\mathcal{S}$. We now obtain the weak formulation given by (A.3). □

Theorem 4 allows us to finalize the derivation of the DGFEM weak formulation, similar to the space–time DG formulation, to:

Find a $U \in W_h$ such that for all $V \in W_h$:

$$\begin{aligned}0 = &\sum_j \int_{K_j} (V_i U_{i,0} - V_{i,k} F_{ik} + V_i G_{ikr} U_{r,k})\mathrm{d}K + \sum_{\mathcal{S}} \int_{\mathcal{S}} (V_i^L - V_i^R)\widehat{P}_i^{nc} \,\mathrm{d}\mathcal{S} \\ &+ \sum_{\mathcal{S}} \int_{\mathcal{S}} \{\!\{V_i\}\!\}\left(\int_0^1 G_{ikr}(\phi(\tau; U^L, U^R))\frac{\partial \phi_r}{\partial \tau}(\tau; U^L, U^R)\mathrm{d}\tau \bar{n}_k^L\right)\mathrm{d}\mathcal{S}.\end{aligned} \tag{A.11}$$

Note that we combined the fluxes at interior and boundary faces by using a ghost value U^R at the boundary. In this article, to integrate in time, we use an explicit TVD third-order Runge–Kutta method (see e.g. [8]).

References

[1] T.B. Anderson, R. Jackson, A fluid mechanical description of fluidized beds, Ind. Eng. Chem. Fundam. 6 (1967) 527.
[2] M. Castro, J.M. Gallardo, C. Parés, High order finite volume schemes based on reconstruction of states for solving hyperbolic systems with nonconservative products. Application to shallow-water systems, Math. Comput. 75 (2006) 1103.
[3] B. Cockburn, C.W. Shu, The Runge–Kutta discontinuous Galerkin method for conservation laws V, J. Comput. Phys. 141 (1998) 199.
[4] B. Cockburn, C.W. Shu, Runge–Kutta discontinuous Galerkin methods for convection-dominated problems, J. Sci. Comput. 16 (2001) 173.

[5] G. Dal Maso, P.G. LeFloch, F. Murat, Definition and weak stability of nonconservative products, J. Math. Pures Appl. 74 (1995) 483.
[6] D.A. Drew, R.T. Lahey, Analytical modeling of multiphase flow, in: M.C. Roco (Ed.), Particulate Two-Phase Flows, Butterworth-Heinemann, Boston, 1993.
[7] H. Enwald, E. Peirano, A.E. Almstedt, Eulerian two-phase flow theory applied to fluidization, Int. J. Multiphase Flow 22 (1996) 21.
[8] S. Gottlieb, C.W. Shu, Total variation diminishing Runge–Kutta schemes, Math. Comput. 67 (1998) 73.
[9] D.D. Houghton, A. Kasahara, Nonlinear shallow fluid flow over an isolated ridge, Commun. Pure Appl. Math. 21 (1968) 1.
[10] K. Hutter, Y. Wang, S.P. Pudasaini, The Savage–Hutter avalanche model: how far can it be pushed? Philos. Trans. Roy. Soc. A 363 (2005) 1507.
[11] C.M. Klaij, J.J.W. van der Vegt, H. van der Ven, Pseudo-time stepping methods for space–time discontinuous Galerkin discretizations of the compressible Navier–Stokes equations, J. Comput. Phys. 219 (2006) 622.
[12] C.M. Klaij, J.J.W. van der Vegt, H. van der Ven, Space–time discontinuous Galerkin method for the compressible Navier–Stokes equations, J. Comput. Phys. 217 (2006) 589.
[13] E.J. Kubatko, J.J. Westerink, C. Dawson, An unstructured grid morphodynamic model with a discontinuous Galerkin method for bed evolution, Ocean Model. 15 (2006) 71.
[14] P.G. LeFloch, Shock waves for nonlinear hyperbolic systems in nonconservative form, Report 593, Institute for Mathematics and its Applications, Minneapolis, MN, 1989.
[15] R.J. LeVeque, Balancing source terms and flux gradients in high-resolution Godunov methods: the quasi-steady wave-propagation algorithm, J. Comput. Phys. 146 (1998) 346.
[16] C. Parés, Numerical methods for nonconservative hyperbolic systems: a theoretical framework, SIAM J. Numer. Anal. 44 (2006) 300.
[17] C. Parés, M. Castro, On the well-balance property of Roe's method for nonconservative hyperbolic systems. Applications to shallow-water systems, ESIAM: Math. Model. Numer. Anal. 38 (2004) 821.
[18] E.B. Pitman, L. Le, A two-fluid model for avalanche and debris flows, Philos. Trans. Roy. Soc. A 363 (2005) 1573.
[19] R. Saurel, R. Abgrall, A multiphase Godunov method for compressible multifluid and multiphase flows, J. Comput. Phys. 150 (1999) 425.
[20] P.A. Tassi, O. Bokhove, C.A. Vionnet, Space discontinuous Galerkin method for shallow water flows – kinetic and HLLC flux, and potential vorticity generation, Adv. Water Resour. 30 (2007) 998.
[21] P.A. Tassi, S. Rhebergen, C.A. Vionnet, O. Bokhove, A discontinuous Galerkin finite element model for river bed evolution under shallow flows, Comput. Methods Appl. Mech. Eng., submitted for publication.
[22] P.A. Tassi, S. Rhebergen, C.A. Vionnet, O. Bokhove, A discontinuous Galerkin finite element model for river bed evolution under shallow flows. Additional appendices, 2007. <http://eprints.eemcs.utwente.nl/9962/>.
[23] E.F. Toro, Riemann Solvers and Numerical Methods for Fluid Dynamics, Springer-Verlag, 1997.
[24] I. Toumi, A weak formulation of Roe's approximate Riemann solver, J. Comput. Phys. 102 (1992) 360.
[25] I. Toumi, A. Kumbaro, An approximate linearized Riemann solver for a two-fluid model, J. Comput. Phys. 124 (1996) 286.
[26] B.G.M. van Wachem, J.C. Schouten, C.M. van den Bleek, Comparative analysis of CFD models of dense gas–solid systems, AIChE J. 47 (2001) 1035.
[27] J.J.W. van der Vegt, H. van der Ven, Space–time discontinuous galerkin finite element method with dynamic grid motion for inviscid compressible flows: I. General formulation, J. Comput. Phys. 182 (2002) 546.
[28] Y. Xing, C.W. Shu, High order well-balanced finite volume WENO schemes and discontinuous Galerkin methods for a class of hyperbolic systems with source terms, J. Comput. Phys. 214 (2006) 567.
[29] W.P. Ziemer, Weakly Differentiable Functions: Sobolev Spaces and Functions of Bounded Variation, Springer-Verlag, New York Inc, 1989.